ING. GERARDO JAVIER KURI MONGE

MODELING AND PREDICTING POLLUTANT PARTICLE EXCEEDANCES THROUGH SWARM INTELLIGENCE TECHNIQUES

2022


**UNIVERSIDAD AUTÓNOMA DE QUERÉTARO**
**FACULTAD DE INGENIERÍA**

# "MODELING AND PREDICTING POLLUTANT PARTICLE EXCEEDANCES THROUGH SWARM INTELLIGENCE TECHNIQUES"

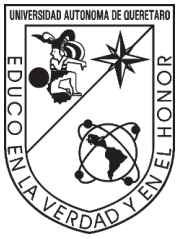
## TESIS

QUE COMO PARTE DE LOS REQUISITOS PARA OBTENER EL DIPLOMA DE LA

MAESTRÍA EN CIENCIAS EN INTELIGENCIA ARTIFICIAL

PRESENTA:
ING. GERARDO JAVIER KURI MONGE

DIRIGIDO POR:
DR. MARCO ANTONIO ACEVES FERNÁNDEZ




Universidad Autónoma de Querétaro

Facultad de Ingeniería

Maestría en Ciencias en Inteligencia Artificial

# Modelado y predicción de excedencias de partículas contaminantes mediante técnicas de inteligencia de enjambre

# Modeling and predicting pollutant particle exceedances through swarm intelligence techniques

Tesis

Que como parte de los requisitos para obtener el Grado de
Maestro en Inteligencia Artificial

Presenta
Gerardo Javier Kuri Monge

Dirigido por
Dr. Marco Antonio Aceves Fernández

Dr. Marco Antonio Aceves Fernández
Presidente (Director)

Dra. Karen Esquivel Escalante
Secretario

Dr. Saúl Tovar Arriaga
Vocal

Dr. Jesús Pedraza Ortega
Suplente

Dr. Efrén Gorrostieta Hurtado
Suplente

# Dedicatorias

A mi padre y a mi madre, por ser ese ejemplo y esa inspiración de vida para siempre proponerme nuevos retos y “hacerle frente” hasta lograrlos. Por ser esas personas que siempre han sido mi referencia de lo correcto e incorrecto y han dedicado su vida para que mi hermana Daniela y yo seamos personas humildes de bien. Por aquellas enseñanzas brindadas y regaños para rectificarme, este logro es suyo, pues fueron ustedes quienes hicieron todo esto posible desde el principio, por eso y más los amo como no tienen idea, gracias por todo.

A mi hermana, porque, aunque esté más chiquita me ha enseñado como dedicarle tiempo y trabajo a las cosas que realmente valen la pena. Por ser ese ejemplo de responsabilidad y dedicación en la familia, gracias por todo y sé que vas a tener éxito en tu vida.

A mi abuela, por ser ese ejemplo de fortaleza emocional al cual siempre he aspirado llegar. Por brindarme todos esos consejos sabios y llenos de amor siempre que la busqué cuando tenía alguna duda o necesitaba ayuda.

A todas esas personas que forman parte de mi familia y de mis amigos que me han brindado su apoyo, su tiempo y su amor, gracias por todo.

# Table of Contents

# List of figures

List of tables

## ABSTRACT

The lack of air quality affects population´s health exposed to it. This makes it a topic of current interest. There are different pollutants that contribute to this problem, such as particulate matter (PM) generated mainly by industrial development and traffic flow. Since the Metropolitan Zone of the Valley of Mexico's geographic characteristics do not allow proper ventilation and due to its population's density a significant quantity of poor air quality events are registered. The World Health Organization (WHO) stipulates air quality guidelines globally based in their risk assessment which allows certain airborne pollution. This thesis proposes a methodology to improve the prediction of exceedances and modelling of $PM_{10}$ and $PM_{2.5}$ made by a recurrent long-term/short term memory (LSTM) network using the Ant Colony Optimization (ACO) algorithm. The results show an improvement on the classification of exceedances of an averaged 2.57% in accuracy, 1.88% in precision, 3.58% in recall and 3.63% in F1-score, and reducing the error by around 13.00% in RMSE and 14.82% in MAE using as reference the results obtained with the LSTM deep neural network. Overall, the current study proposes a methodology to be studied in the future more profoundly to improve different modeling and classification techniques in real life applications where there's no short-time prediction condition.

# RESUMEN

La mala calidad del aire afecta la salud de la población expuesta a esta. Este factor lo convierte en un tema de interés actual. Existen diferentes contaminantes que contribuyen a este problema, como la materia particulada (PM por sus siglas en inglés) generada principalmente por el desarrollo industrial y el flujo de tráfico. Dado que las características geográficas de la Zona Metropolitana del Valle de México no permiten una ventilación adecuada y debido a su alta densidad poblacional se registra una cantidad significativa de eventos de mala calidad del aire. La Organización Mundial de la Salud (OMS) estipula pautas de calidad del aire a nivel mundial basadas en su evaluación de riesgos que permiten cierta contaminación atmosférica. Esta tesis propone una metodología para mejorar la predicción de excedencias y modelado de $PM_{10}$ y $PM_{2.5}$ realizada por una red recurrente de memoria a largo/corto plazo (LSTM por sus siglas en inglés) utilizando el algoritmo de Optimización por Colonia de Hormigas (ACO por sus siglas en inglés). Obteniendo mejoras en la clasificación de excedencias que promedian 2.57% de mejor en la tasa de clasificación, 1.88% en precisión, 3.58% en sensibilidad y 3.63% en la métrica F1-score. Al igual que, reduciendo el error aproximadamente en 13.00% en RMSE y 14.82% en MAE usando como referencia los resultados obtenidos con una red neuronal profunda LSTM. En resumen, la tesis actualpropone una metodología que se debería estudiar en el futuro con mayor profundidad para mejorar diferentes técnicas de modelado y clasificación en aplicaciones de la vida real donde no exista una condición de predicción de corto plazo.

# I. INTRODUCTION

## A. Airborne Pollution

By seeking our comfort and development in society as human beings we have developed technological advances that have facilitated transportation, daily habits, and the manufacture of various products. Just as these technological advances increase, so does environmental deterioration which significantly threatens our health and current development [1]. Among this environmental deterioration, one of the most challenging issues is atmospheric pollution. This pollution can affect the population's health exposed to it even in low concentrations of these pollutants [2]. By atmospheric pollution, we can refer to the presence in the air of substances or compounds in an amount that involves discomfort or risk to the health of the population exposed to it. The main routes for this pollution to enter the population's organism are through the respiratory route and ingestion [3]. The global mortality related to air pollution was found to be 1 million premature deaths in rural and urban areas in 2000, it was increased to 3.1 million in 2012, and increased to 4.2 million in 2016, deaths from lung cancer, respiratory infections, strokes, ischemic heart disease, and obstructive pulmonary diseases [4, 5]. According to Babatola and North, *et al*., global air pollution is the leading environmental cause of death estimated to contribute to approximately 14 million deaths annually, while leading to over 3.2% of global disease in 2019 [6, 7]. In the Metropolitan Zone of the Valley of Mexico (ZMVM for its acronym in Spanish) airborne pollution, especially particulate matter, has been a growing concern in the field of health and environment due to its evident trend to growing motorization and industrialization in the area [8].

Particulate matter has become a relevant subject of research between these pollutants due to $PM_{10}$ (particulate matter having an effective aerodynamic diameter smaller than 10 μm) and its high correlation to the increase in hospital admissions due to various health outcomes such as respiratory diseases, cardiovascular diseases, and pregnancy outcomes [9, 10]. In most of the studies, $PM_{10}$ is introduced as the primary pollution in cities with similar characteristics as the ones in ZMVM

[11]; For example, in Pecan (China), which is a city with a similar population to the ZMVM in terms of industries, transportation, and population, $PM_{10}$ is responsible for approximately 80% of annual air pollution [12]. $PM_{2.5}$ (particulate matter having an effective aerodynamic diameter smaller than 2.5 µm) impacts more negatively the population´s health exposed to it than $PM_{10}$ since it penetrates more deeply in the respiratory system due to its smaller size [3]. The World Health Organization (WHO) has stated "There is a strong evidence to conclude that fine particles ($PM_{2.5}$) are more hazardous than coarse particles ($PM_{10}$) in terms of mortality and cardiovascular and respiratory endpoints in panel studies" [13]. Figure 1 compares the size between coarse particles ($PM_{10}$), fine particles ($PM_{2.5}$), and an average thickness of a human hair (50-70 µm).

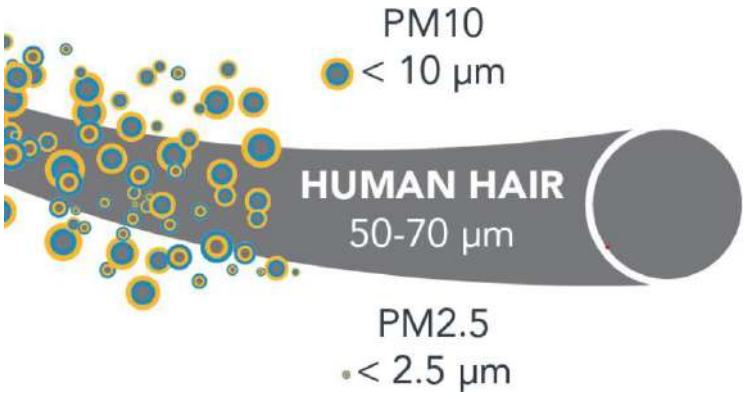



*Figure 1. $PM_{10}$ and $PM_{2.5}$ size comparison [14].*

Exceedance of a particle is determined when such particle exceeds some defined standard. The air quality guidelines for particulate matter stipulated by the World Health Organization or WHO defines a daily average of 50 µg/$m^3$ and 20 µg/$m^3$ annually for $PM_{10}$ and a daily average of 25 µg/m3 and 10 µg/$m^3$ annually for $PM_{2.5}$ as the permitted value of each, any value above the averages mentioned is considered an exceedance [13]. These standards define the maximum amount of pollutants that can be present in outdoor air without harming human health [14]. Figure 2 shows how deeply fine particles ($PM_{2.5}$) and coarse particles ($PM_{10}$) penetrate the respiratory system of human beings exposed to them.

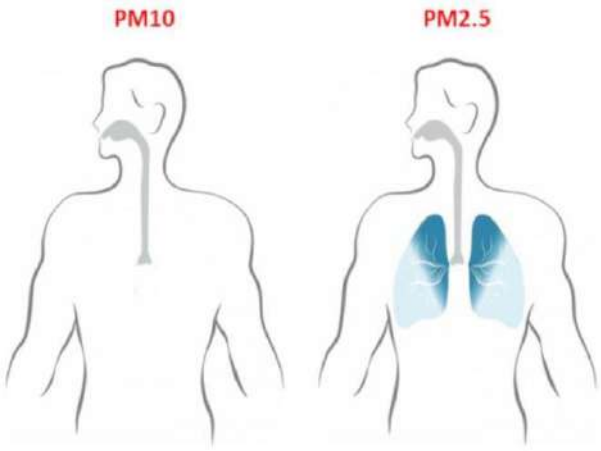



*Figure 2. Pulmonary penetration of PM10 and PM2.5 [15].*

## B. Artificial Intelligence

The concept of using computers to simulate intelligent behavior and critical thinking was first described by Alan Turing in 1950 [16]. In the book "Computers and Intelligence" he described a simple test (which later became known as the "Turing test") to determine whether computers are capable of human intelligence [17]. Six years later, John McCarthy described the term artificial intelligence (AI) as "the science and engineering of making intelligent machines" [18]. AI began as a simple series of "if-then rules" and has advanced over several decades to include more-complex algorithms that perform similar to the human brain. AI is among today's most actively debated developments in information technology with the potential for tremendous impact on individuals, organizations, and societies over the next decades [19]. Based on early evidence, our average simulation shows around 70 percent of companies at least one of these types of AI technologies by 2030, and less than a half of large companies may be using the full range of AI technologies across their organizations. By 2030, AI could potentially deliver an additional economic output of around $13 trillion of USD [20]. AI can perform complex tasks that were previously thought possible only for humans to perform. Machine learning (ML) is an area of AI-related to both cybernetics and computer science, attracting recently an overwhelming interest both of professionals and of the general public. In the last few years, thanks to the successes of computer science (the emergence of

GPUs, leading to significant improvements in the performance of computers and the development of special software, allowing to work with big data) machine learning is often attributed to computer science [21].

The origin of machine learning in its modern sense is usually associated with the name of the psychologist Frank Rosenblatt from Cornell University, who, based on ideas about the work of the human nervous system, created a group that built a machine recognizing the letters of the alphabet Rosenblatt [22]. The machine called the “perceptron” by its creator used both analog and discrete signals and included a threshold element that converted analog signals into discrete ones. It became the prototype of modern artificial neural networks (ANN), and the model of its learning was close to the models of animal and human learning developed in psychology [23]. ANN, specifically the recent branch termed deep learning, has gained an unprecedented revived interest and attention in the past decade from both academia and industry. They have achieved the state-of-the-art in many fields, ranging from computer vision, natural language processing to niche applications, such as self-driving cars [24].

For these reasons and possible consequences, it is highly important an accurate and precise model of airborne pollution for its analysis and to foresee events and trends to take due precautions. Airborne pollution can be classified as a time-dependent practical problem which can be analyzed and predicted as a time series [25]. These time series problems are more complicated than other statistical data due to the long-term trends, cyclical variations, seasonal variations, and irregular movements. Predicting such highly fluctuating and irregular data is usually subject to large errors [26, 27]. Learning these long-range dependencies that are embedded in time series is often an obstacle for most algorithms.

For datasets with the characteristics mentioned above, with a sequential nature, Recurrent Neural Networks (RNN) have been applied successfully due to their capacity to model highly non-linear data. RNN’s are a powerful model for processing sequential data such as airborne pollution [28]. The RNN’s refer to neural networks

that take as input their previous state, this means that the neural network will have two inputs, the new information entered in the network and its previous state, shown in Figure 3.

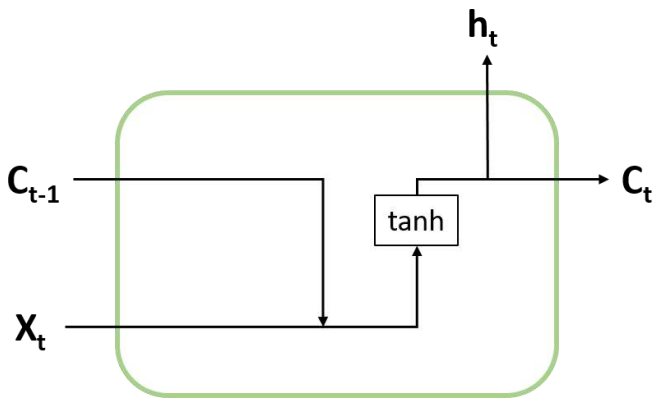



*Figure 3. RNN structure. Adapted figure from [29].*

Where X refers to the input of the model, t refers to the time unit being evaluated, C refers to the current state in the unit and h refers to the output of the model.

Despite traditional recurrent neural networks' inability to handle long sequences of data, a new class of network architectures with learnable gates has been shown to effectively alleviate this problem. The most popular of this variant is the Long Short Term Memory (LSTM) network architecture [24]. In 1997, Hochreiter and Schmidhuder proposed the LSTM model [30], unlike a simple recurrent neural network that has a long-term memory in the form of weights, which are modified during the training of the network and short-term memory defined as activation functions between the communication of the nodes of neurons [31]. The LSTM model introduces a block of internal memory, composed of simple blocks connectedin a specific way [25], as shown in Figure 4.

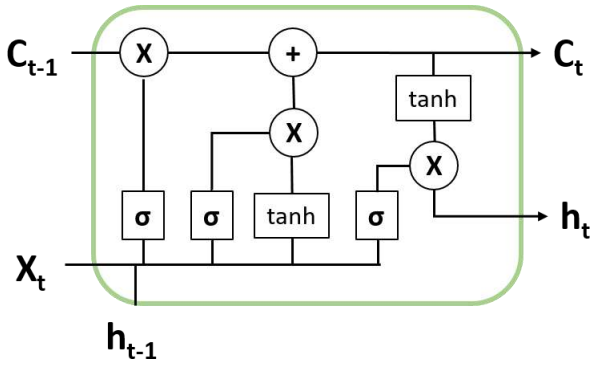



*Figure 4. LSTM structure. Adapted figure from [18].*

A typical LSTM unit have the building blocks shown in Figure 5.

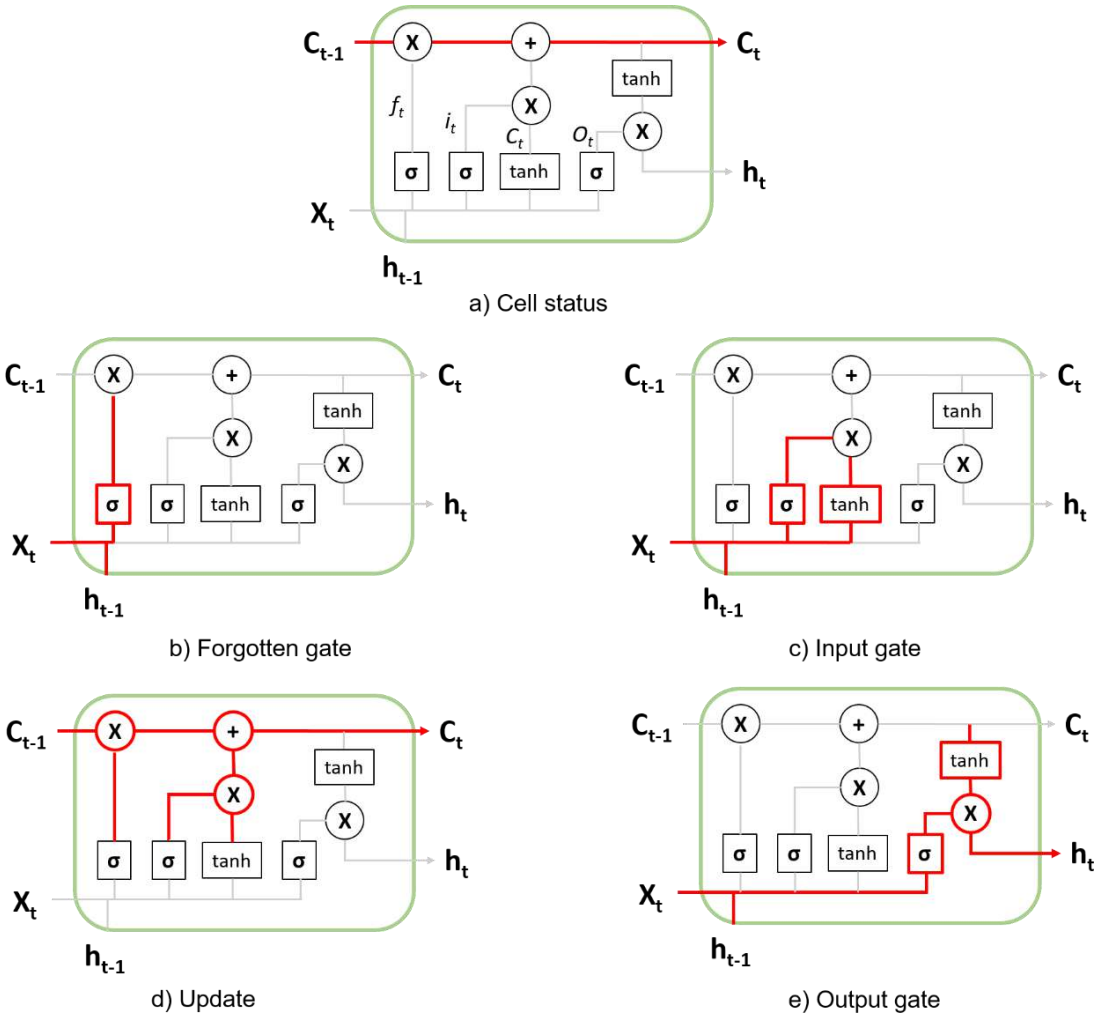



*Figure 5. Components of LSTM. Adapted figure from [29].*

a) Cell status: A conveyor of information with only a few linear interactions.
b) Forgotten gate: This gate is used to control the state of the cell where the previous state and the input of the LSTM unit pass through a sigmoid function producing a value between 0 (do not allow any data to pass through) and 1 (allow every data to pass through).
c) Input gate: This gate determines the added information to the state cell. It contains two parts: a sigmoid input (input signal control) and a tanh function (the input content).
d) Update: Updates the cell status.
e) Output gate: This gate determines the output of the unit. It contains two parts: a sigmoid function (output signal control) and a tanh function (the output content).

### C. Swarm Intelligence

As computer-related technologies are more widely used, many nondeterministic polynomial problems or NP-hard problems have emerged in areas like big data, spam detection, image processing, and environmental concerns [32, 33, 34, 35]. However, when traditional gradient descent and other deterministic methods are used to solve these complicated issues, the degree of finding the global optima, or an acceptable approach, is extremely unsatisfactory in these scenarios [36]. Recently, Swarm Intelligence (SI), or Bio-inspired computation, has gained a lot of attention approaching this kind of problem [37]. The term swarm intelligence was first introduced in 1989 by Gerardo Beni and Jing Wang to describe the dynamics of cellular robots that could be framed as a form of intelligent collective behavior [38]. This marks the point at which swarm behaviors were started to be studied outside natural sciences, although animal behavior has always continued to be a major source of inspiration for swarm intelligence development [39]. There are various reasons responsible for the growing popularity of such swarm intelligence-based algorithms, most importantly being the flexibility and versatility offered by these algorithms [40].

Among the concept of swarm intelligence, there are algorithms based on animal behaviors that demonstrate social intelligence but not particular intelligence as an individual. Within these algorithms, we can highlight the algorithm of ant colony optimization or ACO [41]. The ACO algorithm was proposed by Dorigo in 1992 [42]. It is a well-explored metaheuristic evolutionary algorithm based on population, which is inspired by the research results of the collective behavior of real ants in nature. The ACO algorithm relies on the activities of many individualities and feedback of information. Although the activity of an ant is very simple, the activity of a whole ant colony is characterized by its intelligent behavior [43]. This algorithm is based on the pheromone paths that real ants deposit and follow, this behavior is simulated through simple units that process information (artificial ants), which interact with each other through artificial pheromone modifying the computational environment, which characterizes this interaction as indirect and local, only artificial ants in the modified area perceive this modification [44]. In Figure 6, a graphic description of ants' behavior is shown.

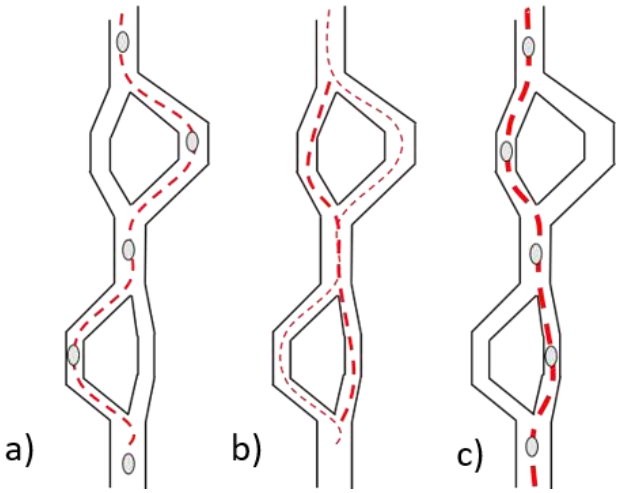



*Figure 6. Ants' behavior in ACO algorithm and their pheromone traces (own authorship).*

In Figure 6 it can be seen how the first iteration (a) the artificial ant chooses a route or solution randomly, but in such a way that the algorithm iterates (b) the weight of the pheromone traces, represented by the thickness of the dotted red line, the optimal route changes, and gains more and more weight. In (c) it is shown how the

pheromone trail of less optimal paths disappears completely. This algorithm belongs to metaheuristic algorithms, which refer to algorithms that are designed to solve combinatorial problems.

This algorithm works as follows: artificial ants use pheromone traces in a search space to construct routes, choosing different positions or nodes in the search space, through the possible solutions in the optimization problem combinatorial. Each of the artificial ants chooses its next move (to the next position) according to the pheromone concentration value, which is normally initialized as 0 or a constant. This means that the first solutions constructed by the artificial ant colony are random. This trace of artificial pheromones' weight or concentration, which is updated after each solution is built, is inversely related to the solution's cost which is the variable that it is being reduced, therefore, optimizing the problem's solution. The increase in the pheromones' concentration will have as a consequence a tendency of the artificial ants to take that route or solution, the traces not taken by the ants will have a tendency of losing their pheromones through the process of evaporation through time, or in this case, iterations [45]. This process is repeated until the stop condition is met. A simple version of the ACO algorithm is presented in Figure 7 as a flowchart.

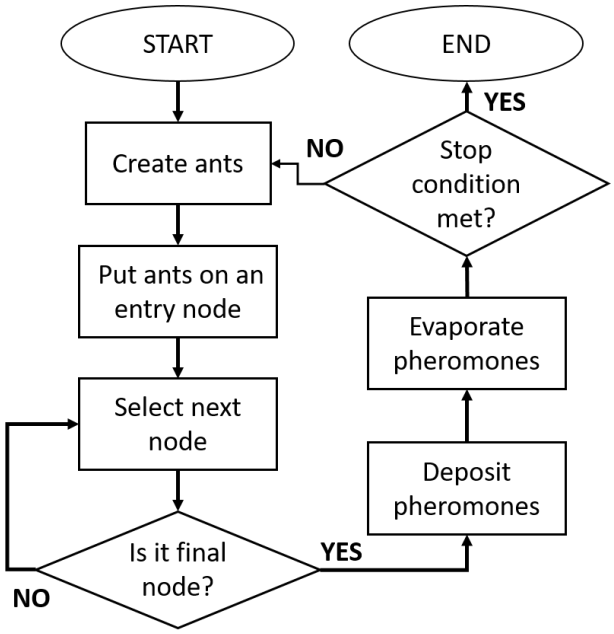



*Figure 7. Ant Colony Optimization algorithms (own authorship).*

In Figure 7 can be seen how a simple version of the ACO algorithm works, where ants are created and introduced into an entry state in a search space composed of multiple nodes. An artificial ant has built a solution (passing through all nodes) pheromones in the solution are deposited. This pheromone concentration deposited is directly related to how efficient the solution is. At the end of each iteration pheromones in the search space will evaporate at some rate. The evaporation takes place to allow less efficient solutions to get lost through iterations. Finally, if convergence is reached, a stop condition is met where a solution can be identified because it possesses the highest pheromone concentration deposited by the artificial ants [46], [47].

# II. BACKGROUND

As seen in the section above, $PM_{2.5}$ and $PM_{10}$ have high impacts on the population's health exposed to them. The development of a robust model for $PM_{2.5}$ and $PM_{10}$ concentrations and classifier of exceedances provides invaluable information for local authorities to take precautionary measures and implement significant actions to improve air pollution status. Due to the current impact of this topic, multiple authors are exploring different modeling techniques to predict $PM_{10}$ and $PM_{2.5}$ concentrations and classification techniques to classify them as exceedance or not exceedance. These different modeling techniques will be described in this section and results will be compared with the ones obtained in this work in results and discussion.

## A. Air pollution modeling approaches

A $PM_{10}$ and $PM_{2.5}$ modeling methodology was proposed by Shtein, *et. al.* based in a spatio-temporal hybrid, based in spatial (predictors based on land-use, population, density, normalized difference vegetation index, elevation, roads density) and temporal (predictors based on air temperature, relative humidity, wind speed, rainfall and Nitrogen Oxides concentrations) modeling approach. This modeling technique consists of three stages which are calibration (using concentrations obtained from stations), estimation (modeling for locations with satellite information) and modeling (modeling for locations without satellite information) [48].

A different approach for modeling $PM_{10}$ was proposed by Abdullah, *et. al.* based on linear (Multiple Linear Regression, MLR) and nonlinear (Multilayer Perceptron) models forecasting capability in the industrial area of Pasir Gudang, Johor. This study was conducted based on 8 years of data covered from the year 2007-2014. The data was divided into two parts; 70% for model training and 30% for model testing. $PM_{10}$, relative humidity, wind speed, ambient temperature, and gaseous pollutants (CO, $NO_2$, and $SO_2$) were known as input parameters (7 inputs) and the output parameter is the next day $PM_{10}$ concentration [49]. A similar approach was

proposed by Feng, *et al.* using an MLP is used to analyze and predict ambient $PM_{2.5}$ in eight regional core cities in China to resolve clashes. This model had as an input temperature, relative humidity, atmospheric pressure, wind speed, and wind direction having as output the concentration of $PM_{2.5}$ of the next day [50].

A model called Prophet Forecasting Model (PFM) developed by Facebook®, which forecasts the desired variable with respect to time. PFM's novel ability to forecast accurately without plenty of complex parameters (meteorology) drastically increases the versatility and applications of PFM. The PFM was used by Shen, *et. al.* to predict both short-term and long-term air pollution in South Korea, Seoul, which has experienced high levels of air pollution. This model was implemented using input parameters like changepoints, seasonality, and holidays [51].

A land-use regression (LUR) model, which has been proving to be an effective method for predicting the spatial distribution of pollutants, was proposed by Han, *et. al.* to analyze and model $PM_{10}$ and $PM_{2.5}$. These LUR models consist in working using pollutant concentration data collected at a limited number of monitoringstations in conjunction with characteristic variables such as land use information to evaluate pollutant concentrations in areas that lack monitoring stations. To constructthe $PM_{10}$ and $PM_{2.5}$ models a total of 87 independent variables were extracted, which are significantly correlated with each other [52]. A similar approach was proposed byMiri, *et. al.* who used LUR models to predict $PM_{2.5}$ and $PM_{10}$ in Sabzevar, Iran. In this work, 104 predictive variables were used as input parameters (ranked by strength of correlation) [53].

In the study implemented by Pak, *et. al.* a spatiotemporal convolutional neural network (CNN) and LSTM (CNN-LSTM) model was proposed and used to predict the next day's daily average $PM_{2.5}$ concentrations in Beijing City. The CNN was designed to efficiently extract the inherent feature essential for the prediction of $PM_{2.5}$from input data; the LSTM was designed to fully represent the long-term historic process of input time series data. The input parameters used in this study were air quality and meteorological data from 384 monitoring stations which represent the

whole area of China with Beijing City [54]. A similar approach was proposed by Zhang, *et. al.* which used CNN-BiLSTM to predict the $PM_{2.5}$ concentrations in Shunyi District, Beijing. A BiLSTM-based structure also allows the training of the prediction model to use both the future features and the past features for a specific time range efficiently, which improves the prediction accuracy to a certain extent [55].

To approach this airborne pollution issue in Liverpool, England, Collazo, *et. al*. compares the Fuzzy C-Means Clustering (FCM) methodology, which consists of an iterative optimization algorithm that minimizes a cost function using clustering techniques that have membership values between 0 and 1, and a Fuzzy Clustering Subtractive (FCS), which assumes each data point is a potential cluster center and calculates a measure of the likelihood that each data point would define the cluster center. The proposed methodology uses the FCM algorithm to partition each point to belong to several clusters with membership values between 0 and 1, and the FCS algorithm to determine the number of clusters of the data being proposed, to model $PM_{10}$ in Liverpool [56].

A number of studies have proposed different methodologies to model airborne pollution in Mexico City, as in the present work. A combination of the Support Vector Machine (SVM) algorithm using the Gaussian, Polynomial, and Spline kernel functions to model different airborne pollutants, $PM_{10}$ between them. This methodology was proposed by Sotomayor, *et. al*. [57]. A Gated Recurrent Unit (GRU) deep neural networks, that was proposed by Cho, *et. al.* in 2014 [58], was used to approach this issue in Mexico City was proposed by Becerra, *et. al*. which isan RNN that is considered a variation of the LSTM algorithm, in the sense that usesthe same functions, but they are organized in a different way. Similar to this work anRNN was used to model $PM_{10}$ in Mexico City [59]. To approach the $PM_{10}$ modeling issue in Mexico City a methodology was proposed by Aceves, *et. al*. in which Deep Convolutional Neural Networks (CNNs), which its main characteristic is the convolutional layer was used as an application of a filter to the input data. To forecast these concentrations PM10, temperature, wind direction, wind speed, relative

humidity, solar ultraviolet radiation type A and solar ultraviolet radiation type B were used as input parameters [60]. Table 1 shows an overview of the techniques used for modeling, the location in which the methodology was applied, the particle modeled, and the year in which was proposed, these works are presented in chronological order. A model optimization approach was proposed by Cabrera, *et. al.* which modeled $PM_{10}$ in Mexico City using an Adaptive Neuro Fuzzy Inference System (ANFIS) which is later optimized using a swarm intelligence technique, named Bacteria Foraging Optimization Algorithm (BFOA) [61]. A similar approach was proposed by Ordóñez, *et. al.* in which $PM_{10}$ and $PM_{2.5}$ were modeled through ANFIS and were later optimized by another swarm intelligence algorithm called Particle Swarm Optimization (PSO) [62].

*Table 1. Modeling approaches for PM10 and PM2.5.*

| Authors | Year | Region | Techniques | Particle |
|---|---|---|---|---|
| Collazo, *et. al.* | 2010 | Liverpool, England | Fuzzy clustering | $PM_{10}$ |
| Sotomayor, *et. al.* | 2013 | Mexico City, Mexico | SVM/Kernel functions | $PM_{10}$ |
| Shtein, *et. al.* | 2018 | Israel | Spatial-temporal | $PM_{10}$ and $PM_{2.5}$ |
| Abdullah, *et. al*. | 2018 | Pasir Gudang, Malaysia | Multilayer Perceptron | $PM_{10}$ |
| Miri, *et. al.* | 2019 | Sabzevar, Iran | Land use regression | $PM_{10}$ and $PM_{2.5}$ |
| Pak, *et. al.* | 2019 | Beijing, China | CNN | $PM_{2.5}$ |
| Cabrera, *et. al.* | 2019 | Mexico City, Mexico | ANFIS-BFOA | $PM_{10}$ |
| Ordónez, *et. al.* | 2019 | Mexico City, Mexico | ANFIS-PSO | $PM_{10}$ and $PM_{2.5}$ |
| Shen, *et. al.* | 2020 | Seoul, South Korea | Prophet Forecast Model | $PM_{10}$ and $PM_{2.5}$ |
| Han, *et. al.* | 2020 | Guanzhong basin, China | Land use regression | $PM_{10}$ and $PM_{2.5}$ |
| Feng, *et. al.* | 2020 | China | Multilayer Perceptron | $PM_{2.5}$ |
| Becerra, *et. al.* | 2020 | Mexico City, Mexico | GRU | $PM_{10}$ |
| Aceves, *et. al.* | 2020 | Mexico City, Mexico | CNN | $PM_{10}$ |
| Zahng, *et. al.* | 2021 | Beijing, China | CNN-BiLSTM | $PM_{2.5}$ |

## B. Exceedance classification approaches

As seen in the section above, exceedances are defined when a particle exceeds some defined standard stipulated. These standards define the maximum amount of pollutants that can be present in outdoor air without harming human health [14]. Due to this definition involving a standard defined by the WHO and its direct correlation with the impact in the population's health exposed to it, a number of authors are constantly proposing new techniques to classify and predict exceedances as accurately as possible.

A method based on Decision Trees called CART was proposed by Snezhana, *et. al*. which is used for solving classification or regression predictive modeling problems in the class of machine learning data mining methods. This method was used to classify exceedances of $PM_{10}$, averaging daily air data for the city of Pleven, Bulgaria for a period of 5 years. This model uses seven meteorological variables, time variables, as well as lagged $PM_{10}$ variables and some lagged meteorological variables, delayed by 1 or 2 days with respect to the initial time series, respectively [63].

A combination of ordered multiple computational intelligence techniques was proposed by Dotse, *et. al.* which consists in using the random forest algorithm and genetic algorithms to initially determine the optimal set of inputs from the initial data sets of largely available meteorological, persistency of high pollution levels, short and long term variations of emissions rates parameters. This optimal set of inputs is used to classify and predict $PM_{10}$ exceedances in Brunei Darussalam through back propagation neural networks model [64].

An MLR model was proposed by Biancofiore, *et. al.* which is one of the statistical techniques used in several research applications, it can be applied to analyze the relationship among various variables and predict the outcome of a response variable. This model was used to classify $PM_{10}$ exceedances in central Italy using as inputs temperature, pressure, humidity, wind speed, direction, and $PM_{10}$ concentrations [65].

This work will be directly compared with the proposal made by Ramírez, *et. al*. which consists in using the LSTM algorithm to predict and classify exceedances of airborne pollution in Mexico City where the input parameter was the PM10 concentration and the output parameters were obtained the exceedance predictions for the next 24, 48 and 72 hours [66]. Table 2 shows an overview of the techniques used to classify exceedances, the location in which the methodology was applied, the particle modeled, and the year in which was proposed.

*Table 2. Exceedance classification approaches for PM10 and PM2.5.*

| Authors | Year | Region | Techniques | Particle |
|---|---|---|---|---|
| Biancofiore, *et. al.* | 2017 | Pescara, Italy | MLR | $PM_{10}$ and $PM_{2.5}$ |
| Dotse, *et. al.* | 2017 | Brunei Darussalam | Multiple computational techniques | $PM_{10}$ |
| Snezhana, *et. al*. | 2018 | Pleven, Bulgaria | CART | $PM_{10}$ |
| Ramírez-Motañez | 2019 | Mexico City, Mexico | LSTM | $PM_{10}$ |

# III. HYPOTHESIS

The application of swarm intelligence techniques in $PM_{10}$ and $PM_{2.5}$ models will improve the accuracy of the exceedance predictions as well as the performance of this modeling.

# IV. OBJECTIVES

## General objective

To Develop an algorithm capable of modeling the pollutant densities, classify them as exceedances according to air quality guidelines stipulated by the World Health Organization, in order to optimize it using swarm intelligence techniques, which will improve modeling and classification results.

## Specific objectives

- ✓ To use the database obtained by the Automatic Atmospheric Monitoring Network to be able to carry out a predictive algorithm based on real data.
- ✓ To identify erroneous data and correct it through multiple imputation techniques in order to be able to use it in the algorithm.
- ✓ To make $PM_{10}$ and $PM_{2.5}$ density predictions based on the model obtained by deep learning techniques.
- ✓ To classify these predictions as pollutants exceedances and obtain the precision results of this classification according to the air quality guidelines stipulated by the World Health Organization.
- ✓ To optimize this algorithm using swarm intelligence techniques to improve prediction results.

# V. MATERIAL, METHODS AND EVALUATION

## A. Materials

The metropolitan area of the valley of Mexico has a continuous atmospheric monitoring network called Automatic Atmospheric Monitoring Network (or RAMA for its acronym in Spanish). There are 24 stations that belong to this network. Each of these stations registers the concentrations of different pollutants, including $PM_{10}$ and $PM_{2.5}$, among others [67]. This database is maintained and updated by RAMA. The stations used in this work were chosen to take into account two considerations:

1. The availability of $PM_{10}$ and $PM_{2.5}$ data from 2012 to 2019.
2. The available data must have a maximum of 30% missing data during the whole evaluated period. Otherwise, the modeling and prediction may be biased.

Based on these considerations 6 stations were chosen. These stations are San Agustín (SAG), Tlanepantla (TLA), Merced (MER), Xalostoc (XAL), Camarones (CAM) and Hospital General de México (HGM). The stations chosen differ in traffic flow density by a wide range calculated by Tellez using the number of vehicles registered per 1000 habitats [68]. These categories of traffic flow density are: low density (XAL, TLA and SAG), intermediate density (CAM and MER) and high density (HGM). The locations of these stations are shown in Figure 8 as yellow circles.

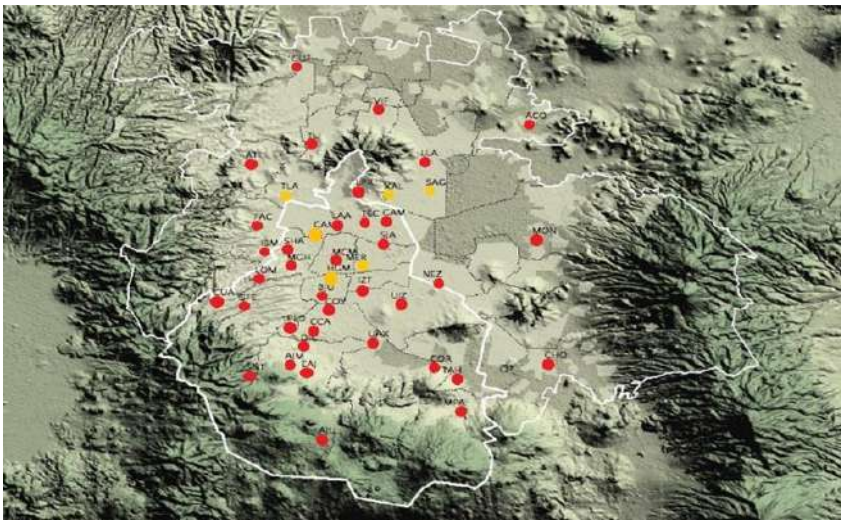


*Figure 8. All stations available in RAMA (red), stations used in this work (yellow) [67].*

The databases for these stations available in RAMA give the concentration in $\mu g/m^3$ which is captured every hour. The period chosen for this work is from 2012 to 2019 given that the stations chosen meet the requirements mentioned above only during this period. The ideal size of each of these databases should be 70,080. Nevertheless, all of the stations have a certain percentage of missing data which are in the range of 13.29% to 26.35%. Since it is necessary for the LSTM algorithm to be trained by a database without missing data, an imputation algorithm needed to be implemented.

## B. Methodology

To solve the missing data problem, the Multivariate Imputation by Chained Equations or MICE algorithm was applied to the database with the shape of $x_1$, $x_2$, ..., $x_n$, where n refers to the length of the dataset and a subset of missing values is present [69]. To implement the MICE algorithm, the first step is to fill the missing data with random values. The first missing value of the set, for instance $x_1$, is then regressed on the other variables $x_2$, ..., $x_n$. The next missing value, $x_2$, is regressed on all the other variables $x_1$, $x_3$, ..., $x_n$. The process is repeated for all other missing values in turn, this is called a cycle Royston and White (2011). Most authors report that a total of 5 cycles is enough to reach convergence [70-73]. The full methodology for the MICE algorithm is shown in Figure 9.

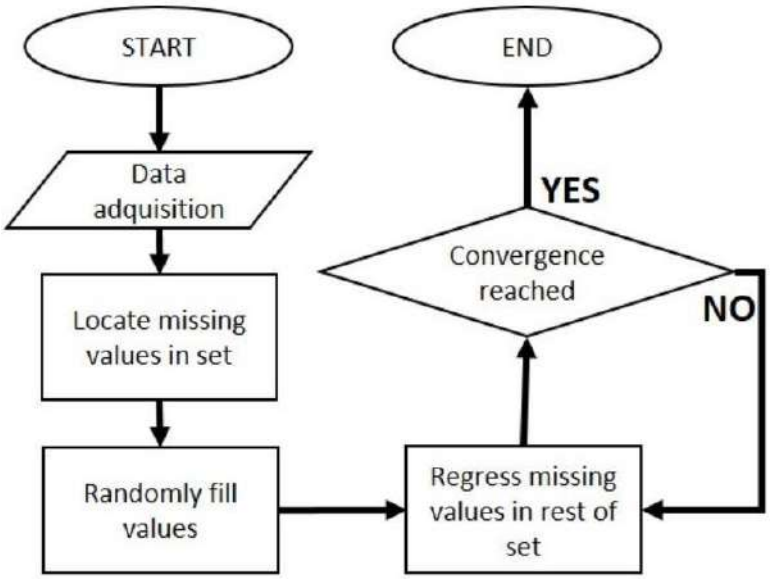



*Figure 9. Multiple Imputation by Chained Equations (MICE) (own authorship).*

Once the data is completed with both raw and imputed data, the training stage using LSTM is performed as in [66]. The first step is to normalize the data into a scale from -1 to 1, this normalization intends to facilitate training in the LSTM modelby decreasing the non-linearity of the data [74].

The network used for modeling consists in:

- ✓ Layer 1 consists of 50 LSTM neurons, which will take the first 50 data, the window being evaluated, and expand them to feed the second layer.
- ✓ Layer 2 is the hidden layer that consists of 256 LSTM neurons, this layer is in charge of controlling memory in the network.
- ✓ Layer 3 is a simple neuron, which based on the previously recorded data will generate a new value, successively.

The predictive neural network of the model consists in:

- ✓ Layer 1 each neuron receives a value from the input data vector which generates an output response, this vector's length depends the prior number of hours being evaluated, in this case, 24 hours.

- ✓ Layer 2 receives the results of the first network and generates a classification, given that initially the days with exceedances are known. The network uses a continuous regression to adjust its weights and obtain the expected result.

The complete model of the LSTM predictive network is shown in Figure 10 where the first three layers correspond to the modeling network and layers 4 and 5 correspond to the classification network.

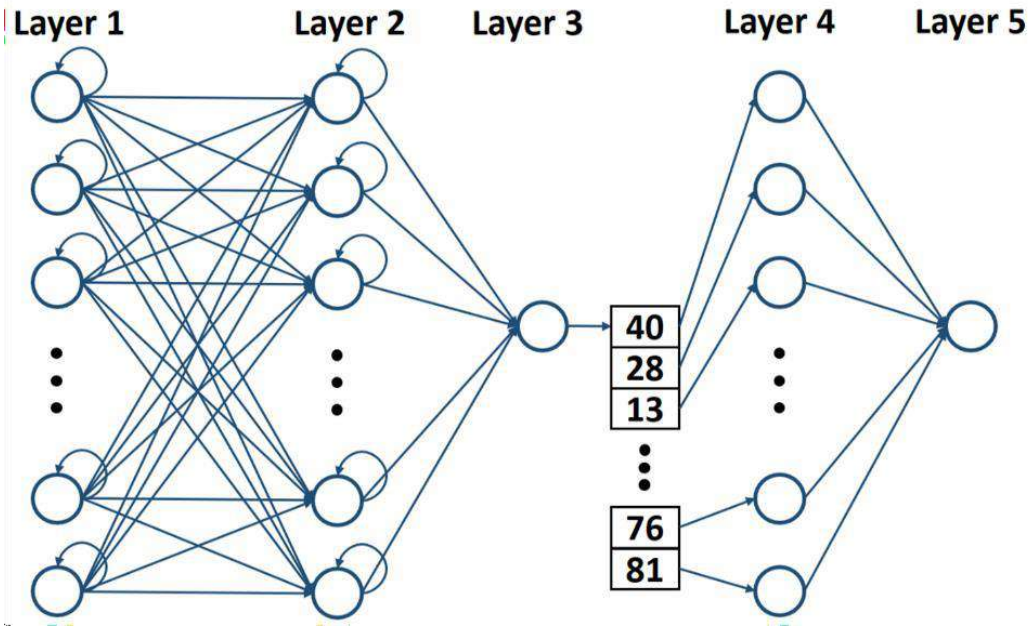



*Figure 10. Complete predictive model [66].*

The construction of the ACO algorithm needs to consider that the search space in which the ant colony will look for the optimized model is a matrix of n by m, n being the quantity of LSTM algorithms trained and m being the vector size that describes the days' concentration prediction from the model. The ACO algorithm has to select one of these daily concentrations given by one of the multiple LSTM models to have as a result an optimized LSTM-ACO model.

To accomplish this, the cost matrix must be initialized. The cost matrix refers to a matrix that describes how "distantly" distributed the nodes in the problem are. This "distance" refers to the variable trying to be minimized or maximized [75]. For this

work, the variable being optimized is the difference between the evaluated value with the centroid of the n LSTM-given values. This means that ants will tend to go where the concentration of these values given by the LSTM algorithm is higher, the cost matrix was generated using Equations (1) and (2):

$$y_{ij} = \left|\left(\frac{1}{n} * \sum_{j=1}^{n} x_{ij}\right) - x_{ij}\right| \quad (1)$$

$$cost\ matrix = \begin{bmatrix} y_{11} & \cdots & y_{1k} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nk} \end{bmatrix} \quad (2)$$

Where $y_{ij}$ refers to the cost value of the *i*-th row and the *j*-th column. In this work, the columns refer to the day being evaluated and rows refer to the prediction of the LSTM model being evaluated. n refers to the quantity of LSTM models included in the algorithm and *k* refers to the quantity of days that compose the prediction.

The next stage consists of the pheromone matrix being updated. This matrix describes how pheromones are distributed in each of the nodes. The pheromone matrix is initialized with a single value chosen for the whole matrix or with custom values chosen by the user. This method is used if the user wants to give ants a preference of a solution. For this work the pheromone matrix is going to be initialized with 1 single value tau calculated with Equation (3) [76]:

$$tau = 10 * \left(\frac{1}{n} * \sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij}\right) \quad (3)$$

Once the matrices have been initialized, the following stage is the ant colony parameter initialization, in which each ant is going to refer to a solution of the problem, where the ant will be constructing the solution by deciding which node it is going to choose next. This decision is made using the probability described in Equation (4):

$$P_{ij} = \frac{(r_{ij})^{\alpha}(\eta_{ij})^{\beta}}{\sum(r_{ij})^{\alpha}(\eta_{ij})^{\beta}} \quad (4)$$

Where $P_{ij}$ refers to the probability of traveling the path between node *i* and node*j* and $\tau_{ij}$ refers to the concentration of pheromones between node *i* and node *j*, this pheromone concentration is obtained from pheromone matrix and is updated in each iteration. Where $\eta_{ij}$ refers to the feasibility between node *i* and node *j*. This feasibility is obtained through the reciprocal value of the cost matrix describing the cost between node *i* and node *j*. Also α refers to the weight that pheromone concentration will be given and β refers to the weight that feasibility will be given [77].

When the ants have the solutions fully constructed the pheromone matrix is updated through Equation (5), and Equation (6):

$$\Delta r^{f}_{ij} = \frac{1}{L_f} \quad (5)$$

Where $L_f$ refers to the summation of all costs in the trajectory that are part of the solution of ant *f*.

$$r^{f}_{ij} = (1-\rho) r_{ij} + \sum_{f=1}^{g} \Delta r^{f}_{ij} \quad (6)$$

Where ρ refers to the evaporation rate of the pheromones and g to the quantity of ants.

A flowchart of the entire methodology is shown in Figure 11. The preprocessing and prediction stage are covered in greater detail in [66].

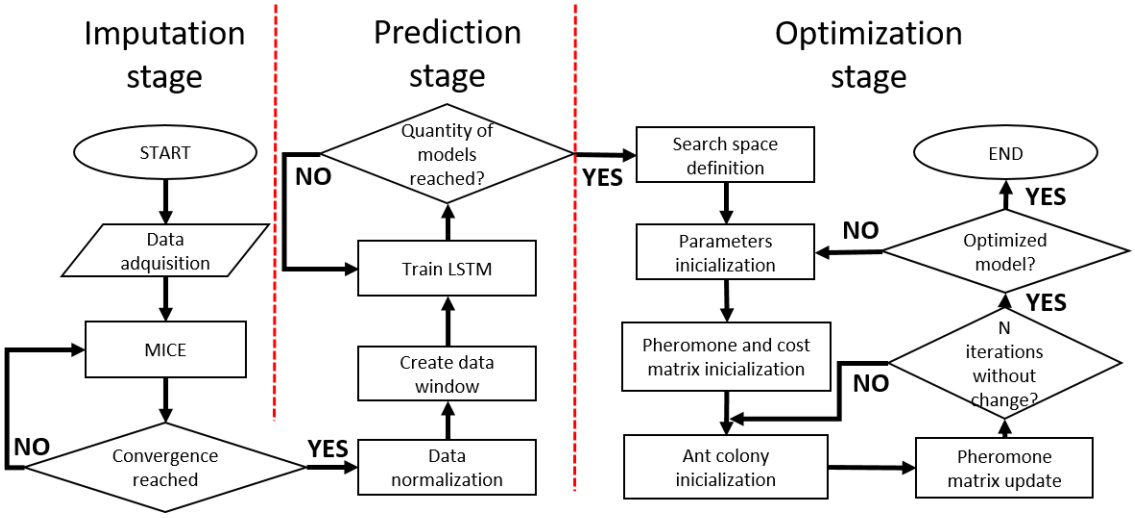



*Figure 11. Proposed methodology (own authorship).*

## C. Evaluation

To evaluate the model, a number of metrics will be used throughout the methodology. To evaluate the MICE algorithm, the standard deviation was calculated for the imputed and raw database using Equation (7):

$$Std.Dev. = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{7}$$

Where $x_i$ refers to the *i*-th element of the database, $x$ refers to the mean concentration for the whole data set and $n$ refers to the number of the data points in the data set.

To evaluate the LSTM-ACO model and successfully compare it with the previous LSTM model, the root mean squared error (RMSE), which represents a more appropriate error to represent model performance when the error distribution is expected to be Gaussian, the correlation coefficient (r), the coefficient of determination ($r^2$), which is a well-defined for linear regression models, and is popularly used in practice as a measure of goodness-of-fit of the underlying models, and mean absolute error (MAE), which gives the same weight to every error of the model, were implemented. The comparison was made between the resulting LSTM-

ACO model and the mean of the measurements between the n LSTM models trained [78,79,80]. According to Zhou, et. al. “a combination of metrics is often required to accurately evaluate model performance” [81]. The RMSE represents the standard deviation between actual and predicted values as Equation (8) describes:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(YM_i - YR_i)^2} \quad (8)$$

Where $YM_i$ refers to the *i*-th element of the prediction model and $YR_i$ refers to the *i*-th element of the real data.

RMSE is a representation of how close distance between the prediction and real values are. This value will be used to decide the optimal number of LSTM models to use as the search space in the ACO algorithm [80].

In order to evaluate the precision of the model obtained, the MAE was implemented using Equation (9):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|YM_i - YR_i| \quad (9)$$

Where $YM_i$ refers to the *i*-th element of the prediction model and $YR_i$ refers to the *i*-th element of the real data.

To evaluate how strong is the relationship between the predicted model and the real data the correlation coefficient was implemented [79]. The correlation coefficient’s values may vary between -1, a perfect negative correlation between data sets, and 1, a perfect correlation between data sets. The correlation coefficient (r) was implemented using Equation (10):

$$r = \frac{n\left(\sum_{i=1}^{n}(YM_i)(YR_i)\right) - \left(\sum_{i=1}^{n} YM_i\right)\left(\sum_{i=1}^{n} YR_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} YM_i^2 - \left(\sum_{i=1}^{n} YM_i\right)^2\right] - \left[n\sum_{i=1}^{n} YR_i^2 - \left(\sum_{i=1}^{n} YR_i\right)^2\right]}} \quad (10)$$

$i=1 \quad i \qquad i=1 \quad i \qquad i=1 \quad i \qquad i=1 \quad i$

Where $YM_i$ refers to the *i*-th element of the prediction model and $YR_i$ refers to the *i*-th element of the real data.

In order to explain how much variability of one factor can be caused by its relationship to another factor the coefficient of determination was implemented ($r^2$) which is calculated by squaring the correlation coefficient [79].

A classification evaluation was implemented in which a confusion matrix is calculated with 4 different values that are described below:

- ✓ True positive or TP: Which is interpreted as predicted positive and it's true.
- ✓ False positive or FP: Which is interpreted as predicted negative and it's true.
- ✓ True negative or TN: Which is interpreted as predicted positive and it's false.
- ✓ False negative or FN: Which is interpreted as predicted negative and it's false.

These measurements fit right with this classification problem because it may have 1 of 2 possible classifications, exceedance or not exceedance, 50 μg/m$^3$ daily permissible for $PM_{10}$ and 25 μg/m$^3$ daily permissible for $PM_{2.5}$ [82]. Each of the following equations describe the measurements implemented for this classification problem:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$Accuracy = \frac{TP + TN}{Total} \tag{13}$$

$$F1 - score = \frac{2 * TP}{(2 * TP + FP + FN)} \tag{14}$$

# VI. RESULTS AND DISCUSSION

As mentioned in the materials section above, the only stations considered for this work were the ones with less than a 30% of data missing in both particles evaluated, $PM_{2.5}$ and $PM_{10}$. Initial missing data is shown in Table 3 per particle and station.

*Table 3. Initial missing data per particle and station.*

| Station | Initial data | Missing |
|---|---|---|
| | $PM_{2.5}$ | $PM_{10}$ |
| SAG | 26.35% | 26.35% |
| TLA | 19.27% | 19.24% |
| MER | 13.64% | 13.29% |
| XAL | 18.28% | 18.21% |
| HGM | 16.49% | 16.49% |
| CAM | 24.02% | 24.02% |

To evaluate the effect, the MICE imputation has upon the preprocessed data, the mean and the standard deviation was calculated before and after the imputation.

The relationship between the percentage of the initial missing data in the database and the difference between the mean of the imputed database and the mean of the initial database is shown in Figure 9 a) and the relationship between the percentage of the initial missing data in the database and the absolute difference between the standard deviation of the imputed database and the standard deviation of the initial database is shown in Figure 9 b).

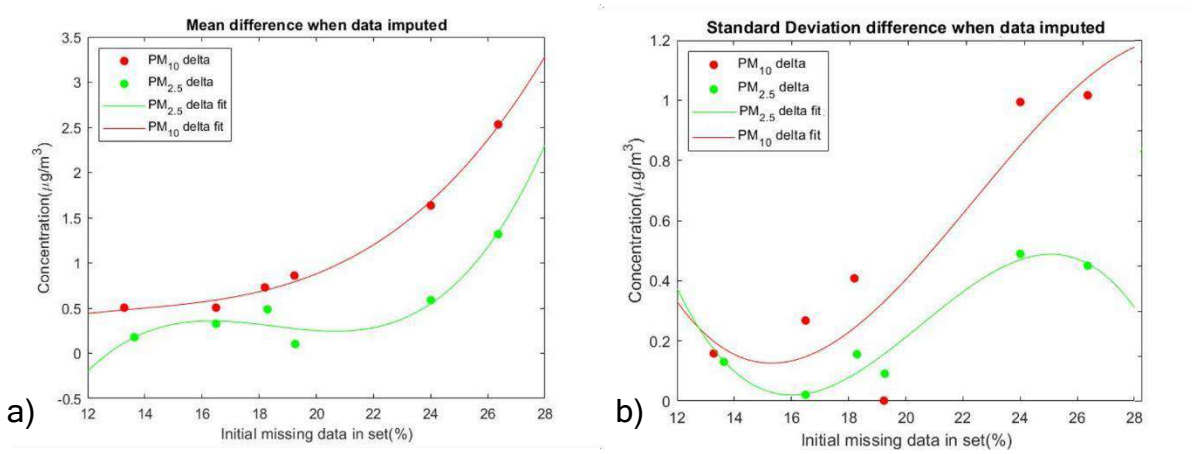



*Figure 12. Relationship between initial missing data in set and the absolute difference of mean after imputation was implemented (own authorship).*

As shown in Figure 9 a), the trend in the relationship between initial missing data in set and absolute difference of mean after imputation is exponential according to the results obtained. This seems to indicate that the greater percentage of missing data in databases is considered, less reliable the model will be related to the real $PM_x$ concentration in the area evaluated. Conversely, the highest difference presented in this work being the case of $PM_{10}$ in SAG station only presented a decrease of 4.91% in its mean. As shown in Figure 12 b) the trend in the relationship between initial missing data in set and absolute difference of standard deviation after imputation is similar to the trend obtained in Figure 12 a). The same case of $PM_{10}$ for SAG station is shown as the highest difference in the standard deviation after imputed but, only presented a rise of 3.09% in its standard deviation.

An example of how the MICE algorithm imputed the database is shown in Figure 13, showing the case of the Merced station between January and February of 2018.

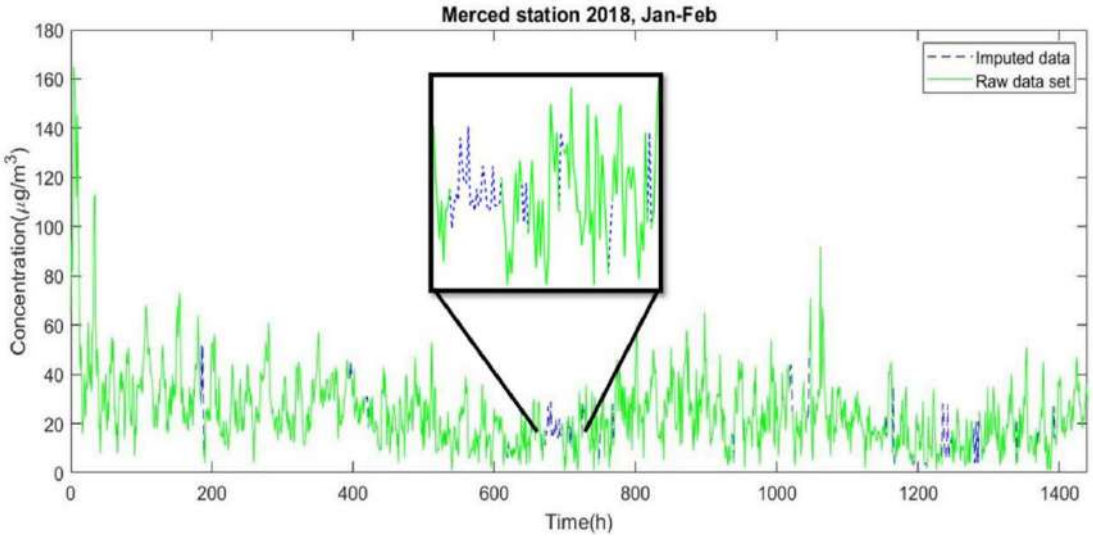



*Figure 13. Example of how MICE algorithm imputes the data base (own authorship).*

To decide the optimal number of LSTM models that make up the ACO search space, an experiment was implemented. This experiment consists in executing the ACO algorithm using a search space of 3 LSTM models following by a comparison between the RMSE error of the LSTM and the LSTM optimized model. Subsequently, 1 LSTM will be concatenated to the search space and the process will be repeated. The objective of this experimentation is to find a correlation between the quantity of LSTM that compose the search space with the RMSE. These results are shown in Figure 14.

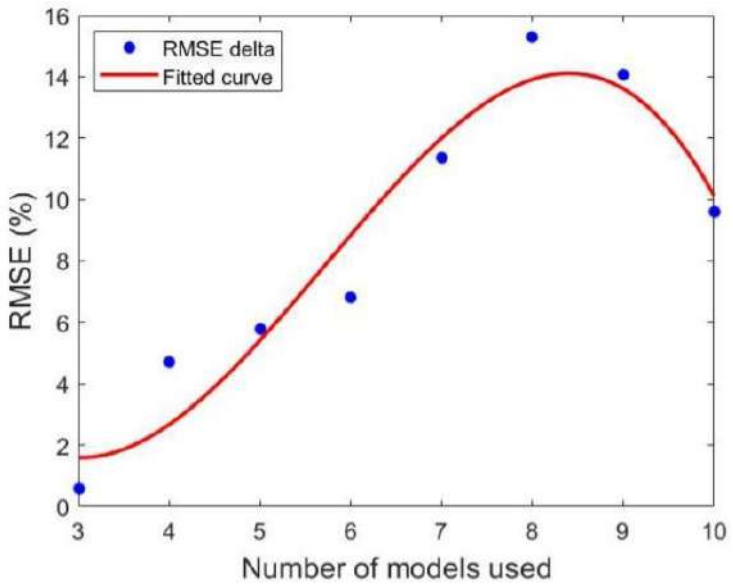



*Figure 14. RMSE improvement percentage vs number of LSTM models that compose the search space (own authorship).*

In Figure 14 the data shown represents the difference between the LSTM-ACO model RMSE and the mean RMSE of the LSTM models being evaluated. Also, Figure 14 shows an improvement in RMSE can be seen with a quantity of models equal or greater than 3. These experiments indicate that the best results are given with 8 LSTM models. Also, this experiment was implemented using the $PM_{10}$ data from the Merced station, since this is the station with less missing $PM_x$ data, the missing $PM_{10}$ data represented 13.29% from the chosen range from 2012 to 2019.

Once the quantity of LSTM models was defined, the next stage consists of an experiment to determine the optimum parameter combination for the ACO algorithm. The definition of efficiency considered in the experiment was to reach the smallest RMSE difference between the obtained model and the imputed data using the least amount of time. This experiment was repeated 25 times to observe repeatability through the standard deviation of the RMSE results per combination of parameters. Table 4 shows the results for time, root mean square error (RMSE) and standard deviation of such experiments using different hyperparameter combinations. This test consisted in combining values for the number of ants (g in Equation (6)) in the

algorithm and for the evaporation rate (ρ in Equation (6)), the weight of the pheromone concentration (α in Equation (4)) and the weight of feasibility (β in Equation (4)) were defined to 1 in order to avoid any bias for the parameter initialization. The results considering a quantity of ants of 1, 3, 5,10, 25 and 50, and the evaporation rate of 0.1%, 1%,5%, 10% and 20% are shown in Table 4, this testing was developed using $PM_{10}$ data of station in Xalostoc. The termination criteriafor the algorithm considered the ants to reach a 25 iterations without finding a bettersolution than the local best.

*Table 4. Evaporation rates and ant quantities tested with respective scale of colors.*

| | ρ=0,1% | | | ρ=1% | | | ρ=5% | | | ρ=10% | | | ρ=20% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **No.** | t(s) | rmse | std | t(s) | rmse | std | t(s) | rmse | std | t(s) | rmse | std | t(s) | rmse | std |
| **1** | 2,68 | 5,53 | 0,06 | 2,46 | 5,51 | 0,04 | 2,02 | 5,48 | 0,04 | 1,27 | 5,49 | 0,03 | 1,05 | 5,5 | 0,04 |
| **3** | 5,1 | 5,51 | 0,03 | 4,29 | 5,49 | 0,04 | 3,31 | 5,47 | 0,05 | 2,28 | 5,49 | 0,05 | 2,28 | 5,48 | 0,05 |
| **5** | 7,48 | 5,49 | 0,05 | 5,24 | 5,48 | 0,05 | 5,5 | 5,48 | 0,03 | 3,41 | 5,46 | 0,03 | 2,97 | 5,48 | 0,04 |
| **10** | 9,54 | 5,49 | 0,04 | 8,79 | 5,48 | 0,02 | 6,64 | 5,46 | 0,03 | 6,96 | 5,47 | 0,05 | 5,17 | 5,47 | 0,05 |
| **25** | 23,12 | 5,49 | 0,03 | 19,73 | 5,46 | 0,04 | 17,52 | 5,47 | 0,03 | 14,06 | 5,47 | 0,05 | 17,12 | 5,48 | 0,05 |
| **50** | 35,6 | 5,49 | 0,03 | 30,42 | 5,46 | 0,05 | 26,73 | 5,48 | 0,05 | 24,41 | 5,46 | 0,04 | 21,8 | 5,48 | 0,06 |

| | | | | | | |
|---|---|---|---|---|---|---|
| t(s) | >17,51 | 14,01-17,5 | 10,51-14 | 7,01-10,5 | 3,51-7 | <3,5 |
| rmse | >5,513 | 5,5-5,512 | 5,487-5,499 | 5,474-5,486 | 5,461-5,473 | <5,46 |
| std | >0,07 | 0,051-0,06 | 0,041-0,05 | 0,031-0,04 | 0,021-0,03 | <0,02 |

Results in Table 4 presented ρ= 10% and ants =5 as the most efficient combination of parameters show the lowest RMSE, 5.46, having a low amount of time, 3.41 seconds. The repeatability of this combination of parameters is considered as high as the standard deviation of the 25 RMSE's obtained is 0.03. This experimentation demonstrates that a better result, in this case, not necessarily corresponds to a greater number of ants but it seems that a greater evaporation rate is directly related to execution time. The challenge here was to find a combination of parameters that got the best results with high repeatability in the least amount of time.

## A. Concentration modelling results

To evaluate the optimized LSTM-ACO model the RMSE, MAE, and the coefficient of determination was implemented, it is noteworthy that when evaluatingthe RMSE and MAE the lower the value the better [78] and, when evaluating the coefficient of determination, the higher the value the better [79]. The purpose of these repetitions is to compare them with the averaged result of the 8 LSTM modelsobtained for each station for both $PM_{2.5}$ and $PM_{10}$, and to compare the results with abaseline using different methodologies including a MultiLayer Perceptron (MLP) model, a hybrid model using spacial and temporal variables, land use regression models, a forecasting model developed by Facebook called Prophet, a Convolutional Neural Network (CNN), and Convolutional Neural Network-LSTM model for the particle evaluated obtained from a number of different authors [48–55,60]. These methodologies will be compared by its corresponding evaluation metric and the particle modeled above from Figure 15 to Figure 20.

The results of the predicted $PM_{10}$ RMSE for the 25 LSTM-ACO models, the averaged RMSE of the 8 LSTM models and the RMSE obtained by [49,52,48] is shown in Figure 15. The RMSE presented in [49] obtained through the MLP model equals to 11.389 $\mu g/m^3$, which seems to indicate that the results obtained in this contribution are 61.717% better than a baseline MLP model. The RMSE presented in [52] obtained through the land use regression model equals to 14.842 $\mu g/m^3$, which seems to indicate that the results obtained in this contribution are 70.624% better than a baseline land use regression model. The RMSE presented in [60] obtained through CNN equals to 19.44 $\mu g/m^3$, which seems to indicate that the results obtained in this contribution are 77.572% better than a baseline CNN model. The RMSE presented in [48] obtained through the spatio-temporal hybrid model equals to 19.94 $\mu g/m^3$, which seems to indicate that the results obtained in this contribution are 78.134% better than a baseline spatio-temporal hybrid model and

9.892% better than the LSTM recurrent network without the optimized model for $PM_{10}$.

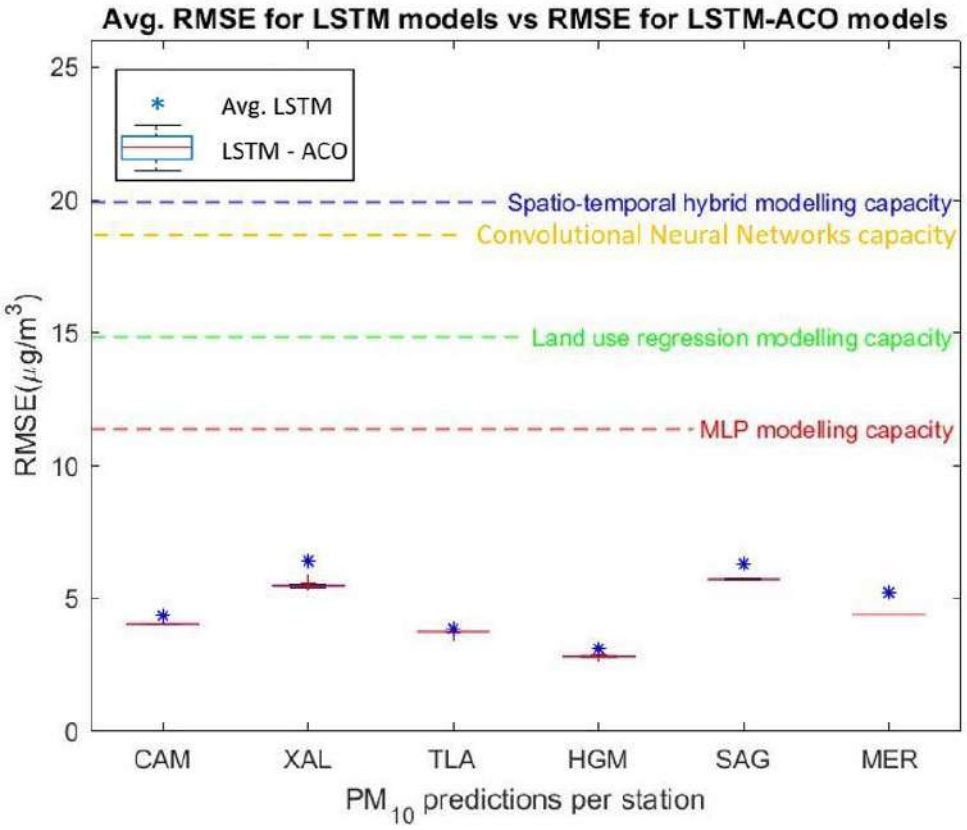



*Figure 15. PM10 RMSE of LSTM models, LSTM-ACO models and MLP implemented by [49], land use regression model implemented by [52] and spatio-temporal hybrid model implemented by [48].*

The results of the predicted $PM_{2.5}$ RMSE for the 25 LSTM-ACO models, the averaged RMSE of the 8 LSTM models and the RMSE obtained by [50,52,48] is shown in Figure 16. The averaged RMSE for 8 cases in China presented in [50] equals to 19.2125 μg/m$^3$, which seems to indicate that the results obtained in this contribution are 84.786% better than a baseline MLP model. The RMSE presented in [52] obtained through the land use regression model equals to 8.355 μg/m$^3$, which seems to indicate that the results obtained in this contribution are 65.937% better than a baseline land use regression model. The RMSE presented in [48] obtained through the spatio-temporal hybrid model equals to 6.16 μg/m$^3$, which seems to indicate that the results obtained in this contribution are 53.799% better than a

baseline spatio-temporal hybrid model and 16.113% better than the LSTM recurrent network without the optimized model for $PM_{2.5}$.

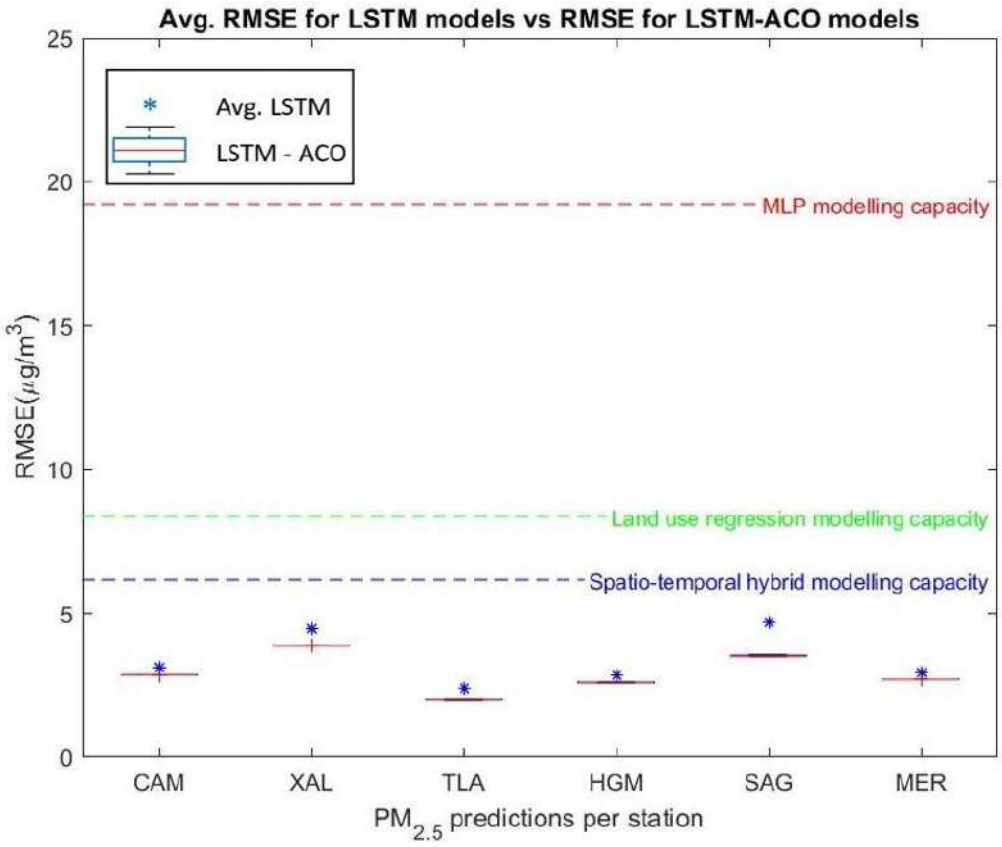



*Figure 16. PM2.5 RMSE of LSTM models, LSTM-ACO models and MLP implemented by [50], land use regression model implemented by [52] and spatio-temporal hybrid model implemented by [48].*

The results of the predicted $PM_{10}$ MAE for the 25 LSTM-ACO models, the averaged MAE of the 8 LSTM models and MAE obtained by [49,51] is shown in Figure 17. The MAE presented in [49] equals to 8.519 $\mu g/m^3$, which seems to indicate that the results obtained in this contribution are 61.04% better than a baseline MLP model. The MAE presented in [51] obtained through the Prophet model equals to 7.6 $\mu g/m^3$, which seems to indicate that the results obtained in this

contribution are 55.526% better than a baseline Prophet model and 10.933% better than the LSTM recurrent network without the optimized model for $PM_{10}$.

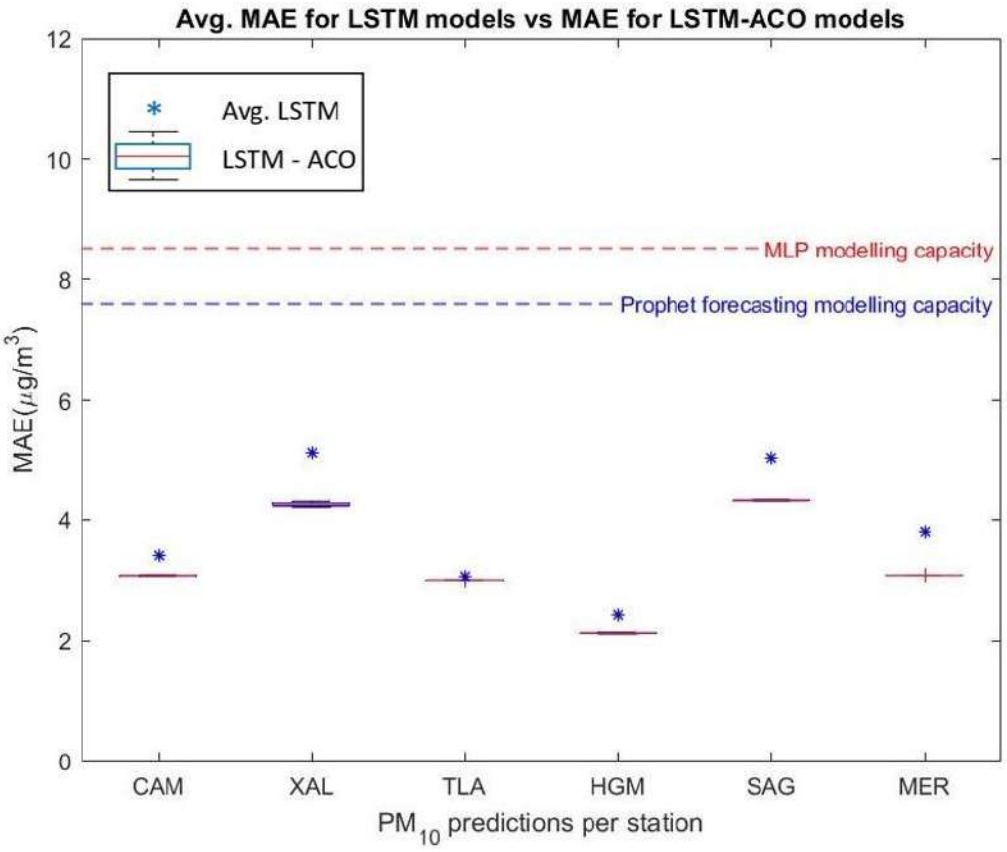



*Figure 17. PM10 MAE of LSTM models, LSTM-ACO models and MLP implemented by [49] and Prophet forecasting implemented by [51].*

The results of the predicted $PM_{2.5}$ MAE for the 25 LSTM-ACO models, the averaged MAE of the 8 LSTM models and MAE obtained by [54,51,55] is shown in Figure 18. The MAE presented in [54] equals to 28.881 $\mu g/m^3$, which seems to indicate that the results obtained in this contribution are 92.31% better than a baseline MLP model. The MAE presented in [51] obtained through the Prophet model equals to 12.6 $\mu g/m^3$, which seems to indicate that the results obtained in this contribution are 85.159% better than a baseline Prophet model. The MAE presented in [55] obtained through the land use regression model equals to 7.032 $\mu g/m^3$, which seems to indicate that the results obtained in this contribution are 73.41% better than

a baseline Prophet model and 18.71% better than the LSTM recurrent network without the optimized model for $PM_{2.5}$.

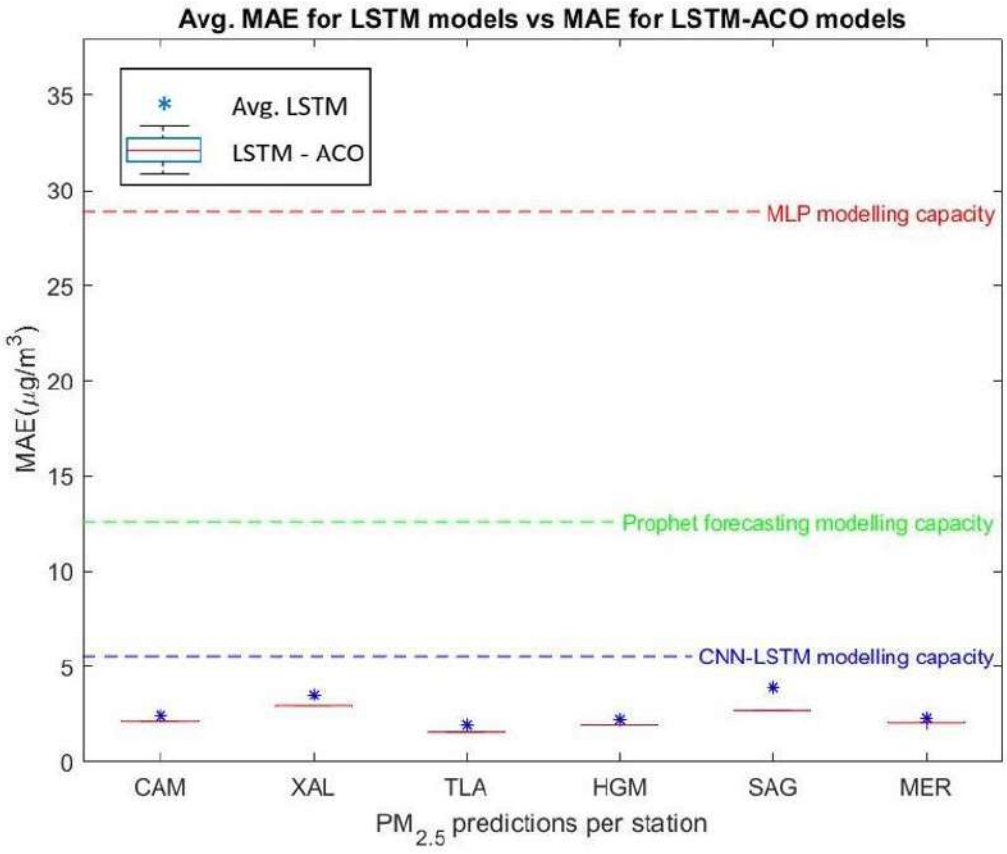



*Figure 18. PM2.5 MAE of LSTM models, LSTM-ACO models and MLP implemented by [54], Prophet forecasting implemented by [51] and CNN-LSTM implemented by [55].*

The results of the predicted $PM_{10}$ coefficient of determination for the 25 LSTM-ACO models, the averaged coefficient of determination of the 8 LSTM models and coefficient of determination obtained by [49,48] is shown in Figure 19. The coefficient of determination presented in [49] equals to 0.839, which seems to indicate that the results obtained in this contribution are 16.508% better than a baseline MLP model. The coefficient of determination presented in [48] obtained through the spatio-temporal hybrid model equals to 0.92, which seems to are 6.52% better than a

baseline spatio-temporal hybrid model and 0.167% better than the LSTM recurrent network without the optimized model for $PM_{10}$.

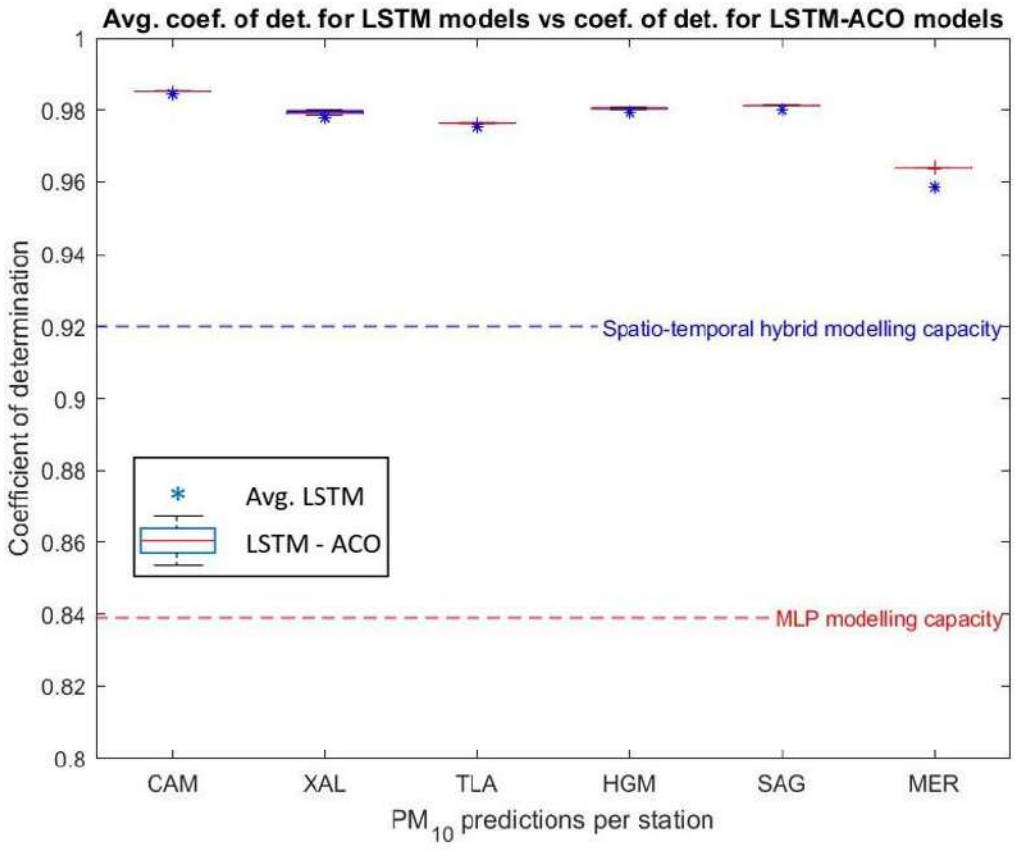



*Figure 19. PM10 coefficient of determination of LSTM models, LSTM-ACO models and MLP implemented by [49] and spatio-temporal hybrid model implemented by [48].*

The results of the predicted $PM_{2.5}$ coefficient of determination for the 25 LSTM-ACO models, the averaged coefficient of determination of the 8 LSTM models and the coefficient of determination obtained by [50,48,55] is shown in Figure 20. The averaged coefficient of determination for 8 cases in China presented in [50] equals to 0.79625, which seems to indicate that the results obtained in this contribution are 22.441% better than a baseline MLP model. The coefficient of determination presented in [48] obtained through the spatio-temporal hybrid model equals to 0.87, which seems to indicate that the results obtained in this contribution are 11.49% better than a baseline spatio-temporal hybrid model. The coefficient of determination

presented in [55] obtained through the CNN-LSTM model equals to 0.906, which seems to indicate that the results obtained in this contribution are 7.06% better than a baseline CNN-LSTM model and 0.168% better than the LSTM recurrent network without the optimized model for $PM_{2.5}$.

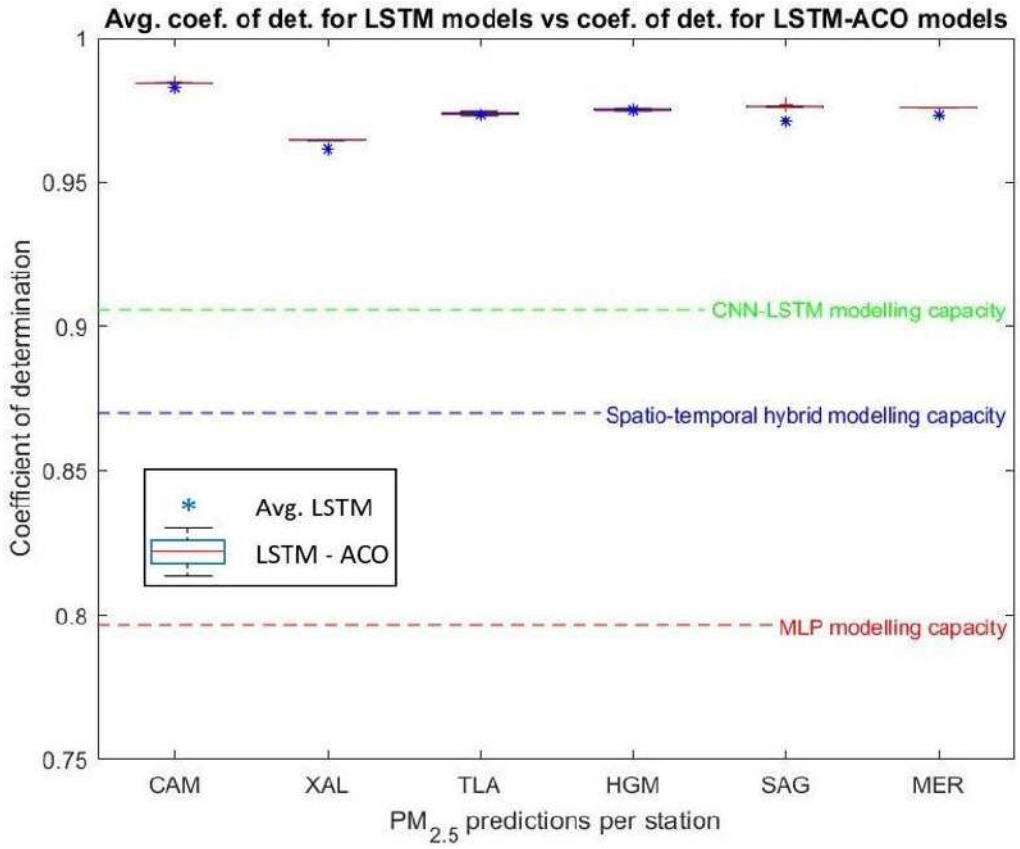



*Figure 20. PM2.5 coefficient of determination of LSTM models, LSTM-ACO models and MLP implemented by [50], spatio-temporal hybrid model implemented by [48] and CNN-LSTM implemented by [55].*

The synthesized results for the LSTM-ACO models obtained for $PM_{10}$ and $PM_{2.5}$ are shown in Table 5 and Table 6 respectively. These results include mean RMSE and its standard deviation, mean MAE and its standard deviation and mean coefficient of determination and its standard deviation for the 6 stations evaluated.

*Table 5. Evaluation metrics of LSTM-ACO models predicting PM10 concentrations.*

| Stations | RMSE (µg/m$^3$) | | MAE (µg/m$^3$) | | Coef. of det. | |
|---|---|---|---|---|---|---|
| | Mean | Std. Dev. | Mean | Std. Dev. | Mean | Std. Dev. |
| SAG | 5.726 | 0.013 | 4.341 | 0.008 | 0.981 | 0.000 |
| TLA | 3.733 | 0.002 | 3.011 | 0.003 | 0.976 | 0.000 |
| MER | 4.386 | 0.001 | 3.088 | 0.002 | 0.964 | 0.000 |
| XAL | 5.474 | 0.045 | 4.268 | 0.028 | 0.979 | 0.000 |
| HGM | 2.829 | 0.015 | 2.124 | 0.008 | 0.98 | 0.000 |
| CAM | 4.039 | 0.006 | 3.084 | 0.006 | 0.985 | 0.000 |

*Table 6. Evaluation metrics of LSTM-ACO models predicting PM2.5 concentrations.*

| Stations | RMSE (µg/m$^3$) | | MAE (µg/m$^3$) | | Coef. of det. | |
|---|---|---|---|---|---|---|
| | Mean | Std. Dev. | Mean | Std. Dev. | Mean | Std. Dev. |
| SAG | 3.523 | 0.019 | 2.683 | 0.006 | 0.976 | 0.000 |
| TLA | 1.985 | 0.012 | 1.589 | 0.006 | 0.974 | 0.000 |
| MER | 2.698 | 0.001 | 2.053 | 0.001 | 0.976 | 0.000 |
| XAL | 3.87 | 0.006 | 2.932 | 0.002 | 0.965 | 0.000 |
| HGM | 2.597 | 0.008 | 1.939 | 0.004 | 0.975 | 0.000 |
| CAM | 2.863 | 0.003 | 2.131 | 0.005 | 0.984 | 0.000 |

It is worth highlighting that the standard deviation results shown in Table 5 and Table 6 where in all of the cases the standard deviation represents less than 1% of the evaluation metric being evaluated. Having no outliers for both tables, the convergence of the LSTM-ACO models can be assured. These results may be translated to robust models with high replicability of both $PM_{2.5}$ and $PM_{10}$.

The LSTM-ACO model was built by the ACO algorithm with the characteristics mentioned above using the 8 LSTM models as the ant's search space and minimizing the cost in Equation (1). This is why it is necessary to compare the results for the LSTM models and the results for the LSTM-ACO models. The results shown are taken by Equation (15):

$$delta = \left(\frac{optimized\ value - Initial\ value}{Initial\ value}\right) * 100 \qquad (15)$$

Where the optimized value refers to the LSTM-ACO models mean result and the initial value refers to the LSTM models mean result.

These percentage differences are shown in Table 7 ($PM_{2.5}$) and Table 8 ($PM_{10}$).

*Table 7. Difference between LSTM-ACO model and LSTM model predicting PM2.5 concentrations.*

| Stations | RMSE | MAE | Coef. of det |
|---|---|---|---|
| SAG | -25.15% | -30.17% | +0.13% |
| TLA | -18.26% | -18.46% | +0.11% |
| MER | -22.24% | -24.90% | +0.15% |
| XAL | -14.51% | -16.71% | +0.30% |
| HGM | -8.48% | -11.21% | +0.16% |
| CAM | -8.04% | -10.81% | +0.16% |

*Table 8. Difference between LSTM-ACO model and LSTM model predicting PM10 concentrations.*

| Stations | RMSE | MAE | Coef. of det |
|---|---|---|---|
| SAG | -9.01% | -7.57% | +0.00% |
| TLA | -4.63% | -1.77% | +0.11% |
| MER | -15.61% | -18.59% | +0.57% |
| XAL | -14.31% | -15.85% | +0.07% |
| HGM | -9.00% | -11.89% | +0.11% |
| CAM | -6.79% | -9.93% | +0.04% |

Table 7 and Table 8 show the impact in which the ACO algorithm improve an LSTM modeling capacity, for the case evaluated in this thesis, by a significant delta that varies between 4.63% - 25.15% for RMSE, and 1.77% - 30.17% for MAE. A possible explanation for the wide range of improvements shown is, LSTM-ACO models "merge" LSTM predictions and when these LSTM predictions differ a lot from each other, the LSTM-ACO algorithm makes a better "merging" between them. Summarizing, LSTM-ACO models describe closer predictions to real concentrations than LSTM models.

### B. Exceedance classification results

Once the 8 LSTM models per particle and station were trained they were evaluated using precision (Equation 11), recall (Equation 12), accuracy (Equation

13) and f1-score (Equation 14). The mean results of the 8 LSTM models are shown in Table 9 ($PM_{2.5}$) and Table 10 ($PM_{10}$).

*Table 9. Evaluation metrics of LSTM models predicting PM2.5 exceedances.*

| Stations | Precision | Accuracy | Recall | F1-Score |
|---|---|---|---|---|
| SAG | 83.77% | 84.95% | 75.41% | 73.39% |
| TLA | 89.87% | 91.87% | 89.42% | 88.63% |
| MER | 88.74% | 91.38% | 93.47% | 90.47% |
| XAL | 84.27% | 86.20% | 91.68% | 86.43% |
| HGM | 93.12% | 93.52% | 88.08% | 89.81% |
| CAM | 94.13% | 93.08% | 91.50% | 92.41% |

*Table 10. Evaluation metrics of LSTM models predicting PM10 exceedances.*

| Stations | Precision | Accuracy | Recall | F1-Score |
|---|---|---|---|---|
| SAG | 93.97% | 95.46% | 94.75% | 94.22% |
| TLA | 91.81% | 94.47% | 93.61% | 92.47% |
| MER | 94.07% | 93.48% | 93.02% | 93.35% |
| XAL | 91.93% | 91.41% | 96.31% | 93.81% |
| HGM | 95.06% | 96.85% | 89.23% | 91.71% |
| CAM | 93.90% | 95.32% | 95.72% | 94.58% |

The LSTM-ACO model was built by the ACO algorithm with the characteristics mentioned above using the 8 LSTM models as the ant's search space and minimizing the cost in Equation (1). The improvement in the results of the evaluation metrics applied to the LSTM-ACO model are shown in Table 11 ($PM_{2.5}$) and Table 12 ($PM_{10}$). The results shown are taken by a simple difference between the LSTM-ACO result and the LSTM result.

*Table 11. Difference between LSTM-ACO model and LSTM model predicting PM2.5 exceedances.*

| Stations | Precision | Accuracy | Recall | F1-Score |
|---|---|---|---|---|
| SAG | +6.39% | +9.28% | +14.75% | +16.77% |
| TLA | +1.57% | +3.67% | +7.25% | +5.35% |
| MER | +1.21% | +1.65% | +0.97% | +1.67% |
| XAL | +4.12% | +6.60% | +6.18% | +6.45% |
| HGM | +2.26% | +0.63% | +3.63% | +3.10% |
| CAM | +0.72% | +1.61% | +2.32% | +1.92% |

*Table 12. Difference between LSTM-ACO model and LSTM model predicting PM10 exceedances.*

| Stations | Precision | Accuracy | Recall | F1-Score |
|---|---|---|---|---|
| SAG | +1.71% | +1.18% | +0.93% | +1.46% |
| TLA | +1.59% | +1.08% | +0.68% | +1.37% |
| MER | +1.48% | +1.23% | +0.67% | +1.26% |
| XAL | +0.09% | +2.25% | +2.93% | +1.68% |
| HGM | +1.20% | +0.41% | +0.44% | +1.08% |
| CAM | +0.26% | +1.26% | +2.26% | +1.45% |

Table 11 and Table 12 show the impact in which the ACO algorithm improve an LSTM classification capacity, for the case evaluated in this thesis, by a significant delta that varies between 0.09% - 6.39% for precision, 0.41% - 9.28% for accuracy, 0.44% - 14.75% for recall, and 1.08% - 16.77% for F1-score. Summarizing, LSTM-ACO classifications are more accurate, precise, and sensitive than LSTM classification results.

Classification results obtained by [63-65] are described and compared with the ones obtained by the LSTM-ACO algorithm in Table 13.

*Table 13. LSTM-ACO results compared with other techniques.*

| Author | Year | Used techniques | Evaluation metrics | Author's results | LSTM-ACO results |
|---|---|---|---|---|---|
| Snezhana, et. al. | 2018 | CART | Accuracy | 80.4% | **94.9%** |
| Biancofiore, F. et. al. | 2017 | MLR | TN and TP percentage | 98% TN and 57% TP | **94.3% TN and 95.3% TP** |
| Dotse, S. et. al. | 2018 | Multiple computational techniques | Recall | 80% | **94.6%** |

In Table 13 we can observe how the presented work compares with other projects classifying exceedances for $PM_{2.5}$ and $PM_{10}$ in which we can observe that most of the evaluation metrics show an improvement in this work. In the case of Biancofiore there is a -3.7% difference between TN results, but a + 38.3% differencein TP meaning that it is highly possible, depending on the sample size, that accuracy,precision, recall, and F1-Score would get better results in the present work.

# VII. CONCLUSIONS

With the present work, a methodology was proposed, which consiste on taking raw $PM_x$ data, impute the data for all missing values. Based on this imputed data, the data is trained in order to obtain a series of LSTM models and finally improve the predicted $PM_x$ concentrations of the recurrent network using the Ant Colony Algorithm. The proposed LSTM-ACO model was tested through various evaluation metrics that demonstrated high repeatability and a very high correlation with the original $PM_x$ databases averaging a RMSE of 3.604 μg/m3, a MAE of 2.78 μg/m3 and a coefficient of determination of 0.976 taking into consideration the 6 stations for both $PM_{10}$ and $PM_{2.5}$. Similarly, the LSTM-ACO methodology to predict exceedances for $PM_{10}$ and $PM_{2.5}$ was tested through precision, accuracy, recall and F1-score that average 93.10%, 94.90%, 94.59% and 93.74% respectively, demonstrating a high repeatability as well. It is worth noting that the ACO getting improvements on the classification of exceedances of an averaged 2.57% in accuracy, 1.88% in precision, 3.58% in recall and 3.63% in F1-score, and reducing the error by around 13.00% in RMSE and 14.82% in MAE. Referring to the previousresults it is highly demonstrable that the hypothesis in this thesis is valid.

These results seem to indicate a highly reliable and generalized model to predict $PM_x$ concentrations. For future work, it may be pertinent to apply this methodology in different sequential (time-dependent) phenomenon and evaluate their results.

# VIII. REFERENCES

[1] Camarillo-Ramirez, P., Sánchez-López, A., Calva-Rosales, L., Perez-Vázquez, I. (2014). Análisis de datos de calidad del aire de la Zona Metropolitana de Valle de México mediante técnicas de agrupamiento. Computing Science, 72, 137–150. Doi:10.13053/RCS-72-1-11

[2] Miri, M., Ghassoun, Y., Dovlatabadi, A., Ebrahimnejad, A., Löwner, M. (2019) Estimate annual and seasonal PM1, PM2.5 and PM10 concentrations using land use regression model. Ecotoxicology and Environmental Safety, 174, 137–145. Doi:10.1016/j.ecoenv.2019.02.070

[3] Monn, C. (2001). Exposure assessment of air pollutants: a review on spatial heterogeneity and indoor/outdoor/personal exposure to suspended particulate matter, nitrogen dioxide and ozone. Atmospheric environment, 35(1), 1-32. Doi:10.1016/S1352-2310(00)00330-7.

[4] Nikoonahad, A., Naserifar, R., Alipour, V., Poursafar, A., Miri, M., Reza, H., Abdolahnejad, A., Nemati, S., Mohammadi, A. (2017). Assessment of hospitalization and mortality from exposure to PM10 using AirQ modeling in Ilam, Iran. Environmental Science and Pollution Research, 24 (27), 21791–21796. Doi:10.1007/s11356- 017-9794-7.

[5] WHO. (2021, Sept 22). Ambient (outdoor) air quality and health. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[6] Babatola, S. (2018). Global burden of diseases attributable to air pollution. Journal of Public Health in Africa, 9(3), 162-166. Doi:10.4081/jphia.2018.813

[7] North, C., Rice, M., Ferkol, T., Gozal, D., Hui, C., Jung, S., Kuribayashi, K., McCormack, M., Mishima, M., Monimoto, Y., Song, Y., Wilson, K., Kim, W., Fong, K. (2019). Air pollution in the Asia-Pacific Region. Respirology, 24(5), 484-491. Doi:10.1111/resp.13531

[8] Pérez-Cirera, V., Schmelkes, E., López-Corona, F., Carrera, O., García-Teruel, A., Teruel, G. (2016). Ingreso y calidad del aire en ciudades: ¿Existe una curva de Kuznets para las emisiones del transporte en la Zona Metropolitana del Valle de Mexico? El trimestre económico, 85(340). Doi:10.20430/ete.v85i340.717

[9] Vicente, A., Juan, P., Meseguer, S., Díaz-Avalos, C., Serra, L. (2018). Variability of PM10 in industrialized urban areas. New coefficients to establish significant differences between sampling points. Environmental Pollution, 234, 969-978. Doi:10.1016/j.envpol.2017.12.026

[10] Ostro, B., Chestnut, L., Vichit-Vadakan, N., Laixuthai, A. (1999). The impact of particulate matter on daily mortality in Bangkok. Air Waste Management Association, 49(9), 100–107. Doi: 10.1080/10473289.1999.10463875.

[11] Nikonahad, A., Khorshidi, A., Reza, H., Miri, M., Amarloei, A., Nourmoradi, H., Mohammadi, A. (2017). A time series analysis of environmental and metrological factors impact on cutaneous leishmaniasis incidence in an endemic area of Dehloran, Iran. Environmental Science and Pollution Research, 24(16), 14117-14123. Doi:10.1007/s11356-017-8962-0

[12] Mu, H., Otani, S., Okamoto, M., Yokoyama, Y., Tokushima, Y., Onishi, K., Hosoda, T., Kurozawa, Y. (2014). Assessment of Effects of Air Pollution on Daily Outpatient Visits using the Air Quality Index. Journal of Medical Sciences, 57(4), 133-136. PMID: 25901100.

[13] W. H. Organization. (2006). WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005: summary of risk assessment. Occupational and Environmental Health 6–10. https://apps.who.int/iris/handle/10665/69477

[14] Cal. Env. Prot. Agency. (2021). Inhalable Particle Matter and Health (PM2.5 and PM10), Res. Div. of California Air Resources Board, Outdoor Air Quality Standards. https://ww2.arb.ca.gov/es/resources/inhalable-particulate-matter-and-health#:~:text=Those%20with%20a%20diameter%20of,Therefore%2C%20PM2.

[15] Vincent, D. (2019). Airbourne particulate matter and their health effects, Enc. Of the Env. 2 – 3. https://www.encyclopedie-environnement.org/en/health/airborne-particulate-health-effects/

[16] Kaul, V., Enslin, S., Gross, S. (2020). The history of artificial intelligence in medicine. Gastrointestinal Endoscopy, 7(20), 34466-7. Doi:10.1016/j.gie.2020.06.040

[17] Greenhill, D., Edmunds, D. (2019). A primer of AI in Medicine. Techniques in Gastrointestinal Endoscopy, 22(2), 85-89. Doi: 10.1016/j.tgie.2019.150642

[18] Pathania, M., Kumar, V. (2019). Overview of artificial intelligence in medicine. Journal of Family Medicine and Primary Care, 8(7), 23-28. Doi:10.4103/jfmpc.jfmpc_440_19

[19] Pandl, K., Thiebes, S., Schmidt-Kraepelin, M., Sunyaev, A. (2020). On the convergence of Artificial Intelligence and Distributed Ledger Technology: A Scoping Review and Future Research Agenda. IEEE Access, 8, 57075-57095.

[20] Bughin, J., Seong, J., Manyika, J., Chui, M., Joshi., L. (2020). Notes From the AI Frontier: Modeling the Impact of AI on the World Economy. McKinsey G. Ins. https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Notes%20from%20the%20frontier%20Modeling%20the%20impact%20of%20AI%20on%20the%20world%20economy/MGI-Notes-from-the-AI-frontier-Modeling-the-impact-of-AI-on-the-world-economy-September-2018.pdf

[21] Keith, A. (2021, December 3). Early History of Machine Learning, IFAC-PapersOnline, 53(2), 1385-1390. https://www.dataversity.net/a-brief-history-of-machine-learning/

[22] Rosenblatt, F. (1957). The Perceptron, A Perceiving and Recognizing Automaton, Project Para Report, 85(460).

[23] Bush, R., Mosteller, F. (1951). A mathematical model for simple learning. Psychological Review, 58(5), 313-323. Doi: 10.1037/h0054388

[24] Wang, Y., Velswamy, K., Huang, B. (2017). A Long-Short Term Memory Recurrent Neural Network Based Reinforcement Learning Controller for Office Heating Ventilation and Air Conditioning Systems. Processes, 5(3), 46. Doi: 10.3390/pr5030046

[25] Hua, Y., Zhao, R., Li, Z., Chen, L., Liu, H., Zhang, Z. (2019). Deep Learning with Long Short-Term Memory for Time Series Prediction. Internet of Things and Sendors Networks, 57(6), 114-119. Doi:10.1109/MCOM.2019.1800155.

[26] Prokuryakov., Y. (2017). Intelligent system for time series forecasting. Procedia Computer Science, 103, 363–369. Doi: 10.1016/j.procs.2017.01.122

[27] Kumar, A., Dash, P., Dash, R., Bisoi, R. (2017). Forecasting financial time series using a low complexity recurrent neural network and evolutionary learning approach. Journal of King Saud University - Computer and Information Sciences, 29(4), 536-552. Doi: 10.1016/j.jksuci.2015.06.002

[28] Chen, K., Zhou, Y., Dai, F. (2015). A LSTM-based method for stock returns prediction: A case study of China stock market. IEEE International Conference on Big Data. Doi: 10.1109/BigData.2015.7364089

[29] Yixiang, X., Chunning, Y. (2020). Fault recognition of electric servo steering gear based on long and short-term memory neural network. International Conference on Unmanned Aircraft Systems. Doi: 10.1109/ICUAS48674.2020.9213882

[30] Hochreiter, S., Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. Doi: 10.1162/neco.1997.9.8.1735

[31] J. Lipton, Z and Berkowitz, C. Elkan. (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning, arXiv preprint arXiv: 1506.00019.

[32] Gu, F., Ma, B., Guo, J., Summers, P., Hall, P. (2017). Internet of things and Big Data as potential solutions to the problems in waste electrical and electronic equipment management: An exploratory study. Waste Management, 68, 434-448. Doi:10.1016/j.wasman.2017.07.037.

[33] Faris, H., Al-Zoubi, A., Heidari, A., Aljarah, M., Mafarja, I., Hassonah, M., Fujita, H. (2019). An intelligent system for spam detection and identification of the most relevant features based on evolutionary Random Weight Networks. Information Fusion, 48, 67–83. Doi:10.1016/j.inffus.2018.08.002.

[34] Liu, Y., Yang, C., Wu, C., Sun, Q., Bi, W. (2019). Threshold changeable secret image sharing scheme based on interpolation polynomial. Multimedia Tools and Applications, 78, 18653–18667. Doi: 10.1007/s11042-019-7205-4

[35] Zhu, B., Ma, S., Xie, R., Chevallier, J., Wei, Y. (2018). Spectra and Empirical Mode Decomposition: A Multiscale Event Analysis Method to Detect the Impact of Economic Crises on the European Carbon Market. Computational Economics, 52(1), 105–121. Doi:10.1007/s10614-017-9664-x

[36] Zhao, D., Liu, L., Yu, F., Heidari, A., Wang, M., Oliva, D., Chen, H. (2020). Ant Colony Optimization with Horizontal and Vertical Crossover Search: Fundamental Visions for Multi-threshold Image Segmentation. Expert Systems with Applications 114122. Doi:10.1016/j.eswa.2020.114122.

[37] Nguyen, B., Xue, B., Zhang, M. (2020). A survey on swarm intelligence approaches to feature selection in data mining. Swarm and Evolutionary Computation, 54. Doi:10.1016/j.swevo.2020.100663.

[38] Beni, G., Wang, J. (1989). Swarm Intelligence in cellular robotic systems. Proceedings of the NATO Advanced Workshop on Robots and Biological Systems, 102, 703–712. Doi:10.1007/978-3-642-58069-7_38

[39] Schranz, M., Di Caro, G., Schmickl, T., Elmenreich, F., Arvin, W., Sekercioglu, A., Sende, M. (2020). Swarm Intelligence and Cyber-Physical Systems: Concepts, Challenges and Future Trends, Swarm and Evolutionary Computation 100762. Doi:10.1016/j.swevo.2020.100762

[40] Chakraborty, A., Kar, A. (2017). Swarm Intelligence: A Review of Algorithms. Modeling and Optimization in Science and Technologies, 479–494. Doi:10.1007/978-3-319-50920-4_19

[41] Deng, J., Xu, W., Zhao, X. (2019). An Improved Ant Colony Optimization Algorithm Based on Hybrid Strategies for Scheduling Problem. IEEE Access, 7. Doi:10.1109/ACCESS.2019.2897580.

[42] Dorigo, M., Maniezzo, V., Coloni, A. (1996). Ant system: Optimization by a colony of cooperating agents. Transactions on Systems, Man, and Cybernetics, 26(1), 29-41. Doi: 10.1109/3477.484436

[43] Wang, Z., Xing, H., Li, T., Yang, Y., Qu, R., Pan, Y. (2016). A modified Ant Colony Optimization Algorithm for Network Coding Resource Minimization. IEEE Transactions on Evolutionary Computation 3(20), 325-342. Doi:10.1109/TEVC.2015.2457437

[44] Ning, J., Zhang, Q., Zhang, C., Zhang, B. (2018). A best-path-updating information-guided, Information Sciences, 433(434), 142-162. Doi:10.1016/j.ins.2017.12.047

[45] Ismkhan, H. (2017). Effective heuristics for ant colony optimization to handle large-scale problems. Swarm and Evolutionary Computation, 32, 140-149. Doi:10.1016/j.swevo.2016.06.006

[46] Li, Y., Soleimani, H., Zohal, M. (2019). An improved ant colony optimization algorithm for the multi-depot green vehicle routing problem with multiple objectives. Journal of Cleaner Production, 227, 1161–1172. Doi:10.1016/j.jclepro.2019.03.185

[47] Ghosh, M., Guha, R., Sarkar, R., Abraham, A. (2019). A wrapper-filter feature selection technique based on ant colony optimization. Neural Computing and Applications, 32, 7839-7857. Doi:10.1007/s00521-019-04171-3.

[48] Shtein, A., Karnieli, A., Katra, R., Raz, I., Levy, A., Lyapustin, I., Dorman, M., Broday, D., Kloog, I. (2018). Estimating daily and intra-daily PM10 and PM2.5 in

Israel using a spatio-temporal hybrid modeling approach. Atmospheric Environment 191, 142–152. Doi:10.1016/j.atmosenv.2018.08.002

[49] Abdullah, S., Ismail, M., Abdul, N., Najah, A. (2018). Modelling particulate matter (PM10) concentration in industrialized area: A comparative study of linear and nonlinear algorithms. Journal of Engineering and Applied Sciences, 13(20), 8227-8235. http://dspace.uniten.edu.my/jspui/handle/123456789/11544

[50] Feng, R., Gao, H., Luo, K., Fan, J. (2020). Analysis and accurate prediction of ambient PM2.5 in China using Multilayer Perceptron. Atmospheric Environment, 232. Doi:10.1016/j.atmosenv.2020.117534.

[51] Shen, J., Valagolam, D., McCalla, S. (2020). Prophet forecasting model: a machine learning approach to predict the concentration of air pollutants (PM2.5, PM10, O3, NO2, SO2, CO) in Seoul, South Korea. PeerJ 8 (9961). Doi:10.7717/peerj.9961

[52] Han, L., Zhao, J., Gao, Y., Gu, K., Xin, Z., Zhang, J. (2020). Spatial distribution characteristics of PM2.5 and PM10 in Xi'an City predicted by land use regression models, Sustainable Cities and Society 61. Doi: 10.1016/j.scs.2020.102329

[53] Miri, M., Ghassoun, Y., Dovlatabadi, A., Ebrahimnejad, A., Löwner, M. (2019). Estimate annual and seasonal PM1, PM2.5 and PM10 concentrations using land use regression model. Ecotoxicology and Environmental Safety, 174, 137–145. Doi:10.1016/j.ecoenv.2019.02.070

[54] Pak, J., Ma, U., Ryu, K., Ryom, U., Kyongsok, C., Pak, U. (2019). Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China. Science of the Total Environment, 699(10), 133561. Doi:10.1016/j.scitotenv.2019.07.367

[55] Zhang, J., Peng, Y., Ren, B., Li, T. (2021). PM2.5 Concentration Predicition Based on CNN-BiLSTM and Attention Mechanism, Algorithms, 14(7), 208-223. Doi:10.3390/a14070208

[56] Collazo-Cuevas, J., Aceves-Fernández, M., Gorrostieta-Hurtado, E., Pedraza-Ortega, J., Sotomayor-Olmedo, A., Delgado-Rosas, M. (2010). Comparison Between Fuzzy C-means Clustering and Fuzzy Clustering Subtrative in Urban Air Pollution. CONIELECOMP, 174-179. Doi: 10.1109/CONIELECOMP.2010.5440772

[57] Sotomayor-Olmedo, A., Aceves- Fernández, M., Gorrostieta-Hurtado, E., Pedraza-Ortega, C., Ramos-Arreguín, J., Vargas-Soto, E. (2013). Forecast Urban Air Polltuion in Mexico City by Using Support Vector Machines: A Kernel Performance Approach. Int. Journal of Intelligence Science, 3, 126-135. Doi:10.4236/ijis.2013.33014

[58] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, Computation and Language (2014).

[59] Becerra-Rico, J., Aceves-Fernández, M., Esquivel-Escalante, K., Pedraza-Ortega, J. (2020). Airborne particle pollution predictive model using Gated Recurrent Unit (GRU) deep neural networks. Earth Science Informatics, 13, 821-834. Doi:10.1007/s12145-020-00462-9

[60] Aceves-Fernández, M., Domínguez-Guevara, R., Pedraza-Ortega, J., Vargas-Soto. E. Evaluation of Key Parameters Using Deep Convolutional Neural Networks for Airborne Pollution (PM10) Prediction, Discrete Dynamics in Nature and Society (2020). Doi: 10.1155/2020/2792481

[61] Cabrera-Hernández, M., Aceves-Fernández, M., Ramos-Arreguín, J., Vargas-Soto, J., Gorrostieta-Hurtado, E. (2019). Parameters Influencing the Optimization Process in Ariborne Particles PM10 Using a Nuero-Fuzzy Algorithm Optimized with Bacteria Foraging (BFOA). Int. Journal of Intelligence Science, 9, 67-91. Doi:10.4236/ijis.2019.93005

[62] Ordoñéz-De León, B., Aceves-Fernández, M., Fernández-Fraga, S., Ramos-Arreguín, J., Gorrostieta-Hurtado, E. (2019). An improved particle swarm

optimization (PSO): method to enhance modelingof airborne particulate matter (PM10). Evolving Systems, 11, 615-624. Doi: 10.1007/s12530-019-09263-y

[63] Snezhana G., Gocheva-Ilieva, M. (2018). PM10 Prediction and Forecating Using CART: A Case Study for Pleven, Bulgaria. International Journal of Environmental and Ecological Engineering, 12(9), 572-577. Doi: 10.5281/zenodo.1474475

[64] Dotse, S., Iskandar, M., Dagar, L., De Silva, L. (2017). Application of computational intelligence techniques to forecast daily PM10 exceedances in Brunei Darussalam. Atmospheric pollution research, 9(2), 358-368. Doi:10.1016/j.apr.2017.11.004

[65] Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G., Di Carlo, P. (2017). Recursive neural network model for analysis and forecast of PM10 and PM2.5. Atmospheric pollution research, 8(4), 1-8. Doi:10.1016/j.apr.2016.12.014

[66] Ramírez-Montañez, J., Fernández, M. A., Arriaga, A. T., Arreguín, J. R., Calderón, J. S. (2019). Evaluation of a Recurrent Neural Network LSTM for the Detection of Exceedances of Particles PM10. International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE). Doi:10.1109/ICEEE.2019.8884516

[67] RAMA. (2020). Databases - Automatic Air Monitoring Network. www.aire.cdmx.gob.mx/default.php?opc=\%27aKBh\%27

[68] Tellez, M., López, D., Fernández, P. (2018). La movilidad en la Ciudad de México: Impactos, conflictos y oportunidades. Academia, 73-86. Doi:10.14350/sc.07

[69] Rubin, D. (1987). Multiple Imputation for Nonresponsive in Surveys. Department of Statistics, Harvard University, John Wiley & Sons.

[70] White, I., Royston, P., Wood, A. (2010). Multiple Imputation using chained equations: Issues and guidance for practice. Statistics in Medicine, 30(4), 377-399. Doi:10.1002/sim.4067

[71] Wulff, J., Jeppesen, L. (2017). Multiple imputation by chained equations in praxis: Guidelines and review. Electronic Journal of Business Research Methods, 15(1), 41-56.

[72] Resche-Rigon, M., White, I. (2016). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. Statistical Methods in Medical Research, 27(6), 1– 16. Doi:10.1177/0962280216666564.

[73] Beesley, L., Taylor, J. (2019). A stacked approach for chained equations multiple imputation incorporating the substantive model. Biometrics. Doi:10.1111/biom.13372.

[74] Chen, K., Zhou, Y., Dai, F. (2015). A LSTM-based method for stock returns prediction: A case study of China stock market. International Conference on Big Data, 2823-2824. Doi:10.1109/BigData.2015.7364089

[75] Paniri, M., Bagher, M., Nezamabadi-pour, H. (2020). Ant colony optimization for multi-objective optimization problems. Knowledge-Based Systems, 192(15). Doi:10.1016/j.knosys.2019.105285.

[76] Cheng, J., Zhang, G., Li, Z., Li, Y. (2012). Multi-objective ant colony optimization based on decomposition for bi-objective traveling salesman problems. Soft computing, 16(4), 597-614. Doi:10.1007/s00500-011-0759-3

[77] Dorigo, M., Stützle, T. (2010). Ant Colony Optimization: Overview and Recent Advances. International Series in Operations Research & Management Science, 272, 311-351. doi:10.1007/978-1-4419-1665-5-8.

[78] Wang, Y., Lu, W. (2018). Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. IOP Conference

Series: Materials Science and Engineering, 324(1). Doi:10.1088/1757-899X/324/1/012049

[79] Zhang, D. (2016). A Coefficient of Determination for Generalized Linear Models. The American Statistician, 71(4), 310–316. Doi:10.1080/00031305.2016.1256839

[80] Chai, T., Draxler, R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7, 1247-1250. Doi: 10.5194/gmd-7-1247-2014

[81] Zhou, L., Zhao, P., Wu, D., Cheng, C., Huang, H. (2018). Time series model for forecasting the number of new admission inpatients. BMC Medical Informatics and Decision Making, 18(1), 1-11. Doi: 10.1186/s12911-018-0616-8

[82] Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis on the Tehran stock exchange. IEEE access, 8, 150199-150212. Doi: 10.1109/ACCESS.2020.3015966

**A QUIEN CORRESPONDA:**

La que suscribe, Directora de la Facultad de Lenguas y Letras, hace **C O N S T A R** que

**KURI MONGE GERARDO JAVIER**

Presentó y acreditó el **Examen de Comprensión de Textos en Inglés** efectuado el día dieciocho de octubre de dos mil veintiuno.

Se extiende la presente a petición de la parte interesada, para los fines escolares y legales que le convengan, en el Campus Aeropuerto de la Universidad Autónoma de Querétaro, el día veintinueve de noviembre de dos mil veintiuno.

A t e n t a m e n t e,
"Enlazar Culturas por la Palabra"

**DRA. ADELINA VELÁZQUEZ HERRERA**

**AVH/japa*CL*FLL-C.-2335**



Campus Aeropuerto, Anillo Vial Fray Junípero Serra S/N, Querétaro, Qro. C.P. 76140
Tel. 442 192 12 00 Dirección Ext. 61010, Secretaría Administrativa Ext.61300, Posgrado Ext. 61140,
Licenciatura Ext.61070, Centro de Lenguas Ext.61050, Secretaría Académica Ext.61100 y Planeación Ext.61110

2021 International Conference on Information Technology (ICIT) | 978-1-6654-2870-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIT52682.2021.9491649

# Capability of a Recurrent Deep Neural Network Optimized by Swarm Intelligence Techniques to Predict Exceedances of Airborne Pollution (PMx) in Largely Populated Areas

Gerardo Javier Kuri-Monge*, Marco Antonio Aceves-Fernández*, Julio Alberto Ramírez-Montañez*, Jesús Carlos Pedraza-Ortega*
*Facultad de Ingeniería,
Universidad Autónoma de Querétaro, 76010, Querétaro, México

Email: marco.aceves@uaq.mx (Marco Antonio Aceves-Fernández)
Corresponding author: Marco Antonio Aceves-Fernández

***Abstract*—The lack of air quality affects population´s health exposed to it. This makes it a topicof current interest. There are different pollutantsthat contribute to this problem, such as particulate matter generated mainly by industrial development and traffic flow. The World Health Organization stipulates air quality guidelines globally based in their risk assessment which allows certain airborne pollution. This paper proposes a methodology to improve the prediction of exceedances of PM10 and PM2.5 made by a recurrent long-term/short term memory (LSTM) network using the Ant Colony Optimization (ACO) algorithm. Getting improvements of an averaged 2.57% in accuracy, 1.88% in precision, 3.58% in recall and 3.63% in F1-score using as reference the averaged results obtained with the LSTM network.**

***Index Terms*—Air pollution, Ant Colony Optimization, Recurrent Neural Network, Predictive models**

## I. INTRODUCTION

### *A. Air Quality*

By seeking our comfort and development in society as human beings we have developed technological advances that have facilitated transportation, daily habits and the manufacture of various products. Just as these technological advances increase, so does environmental deterioration which significantly threatens our health and current development [1]. Among this environmental deterioration, one of the most challenging issues is the atmospheric pollution. By atmospheric pollution we can refer to the presence in the air of substances or compounds in an amount that involves discomfort or risk to the health of the population exposed to it. This paramount importance affects all countries causing 4.2 million of premature deaths in rural and urban areas only in 2016, deaths from lung cancer, respiratory infections, strokes, ischemic heart disease and obstructive pulmonary diseases [2]. Particulate matter (PM) has become a relevant subject of research between these pollutants due to PM10 (particulate matter having an effective aerodynamic diameter smaller than 10 $\mu$m) and its high correlation to the in- crease in hospital admissions for lung and heart disease [3]. PM2.5 (particulate matter having an effective aerodynamic diameter smaller than 2.5 $\mu$m) impacts more negatively population´s health exposed to it than PM10 since it penetrates more deeply in the respiratory system due to its smaller size [4].

An exceedance of a particle is determined when such particle exceeds some defined standard. The

air quality guidelines for particulate matter stipulated by the World Health Organization or WHO defines a daily average of 50 $\mu$g/m3 and 20 $\mu$g/m3 annually for PM10 and a daily average of 25 $\mu$g/m3 and 10 $\mu$g/m3 annually for PM2.5 as the permitted value of each, any value above the aver- ages mentioned is considered an exceedance [5].

### *B. Health problems related to environmental pollution*

The human being is exposed to atmospheric pollution since it can enter the organism through different routes, mainly through the respiratory route and ingestion [4]. The exhibition of the population to these pollutants in the air can cause serious physiological effects, which include serious illness and even death [3]. Even if the concentration of these air pollutants stay below international standards there is direct correlation between these concentrations and deterioration of pulmonary functions that could lead to some degree of asthma [6].

### *C. Recurrent Neural Networks*

One of the most powerful instruments within artificial intelligence refers to Artificial Neural Networks or ANN's. ANN's are mathematical models built based on biological neural networks. This means that the fundamental units of ANN's are artificial neurons [7]. ANN's are specially useful in modeling complex phenomena where the presence of non-linear relationships between variables is common [8].

The Long-Short Term Memory or LSTM algorithm is an algorithm that belongs to the recurrent neural networks or RNN's. The RNN's refer to neural networks that take as an input their previous state, this means that the neural network will have two inputs, the new information entered in the network and its previous state, shown in Figure 1. With this model we are able to have short-term memory in the neural network [9]. These neural networks have applications in sequential predictions, that is, predictions that depend on a temporal variable.

As shown in Figure 1 the LSTM model introduces a block of internal memory, composed of simple blocks connected in a specific way, each of them described as follows [10] :

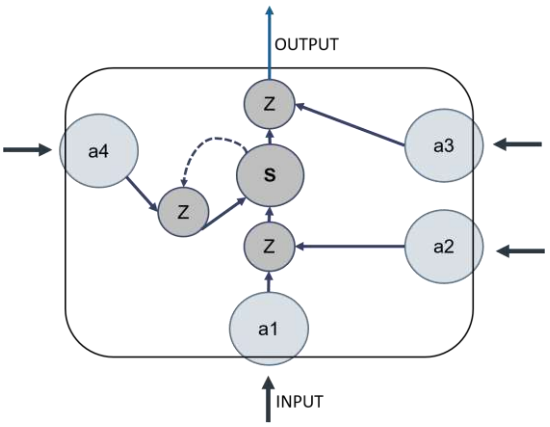



Fig. 1. Basic LSTM Model. Adapted figure from [10]

- Input node: represented as "a1" weights the input values.
- Input gate: represented as "a2" acts a block memory control.
- Internal state: represented as "S" prevents the error from increasing.
- Oblivion gate: represented as "a4" provides a method by which the network can adjust to the content of its internal state.
- Output gate: The value that the model shown outputs is the internal state multiplied by the node represented as "a3".

### *D. Swarm Intelligence*

One of the problem-solving areas covered by artificial intelligence is swarm intelligence. Swarm intelligence refers to the ability that arises from the interaction of simple units capable of processing information [11]. This concept suggests multiplicity, stochasticity, randomness and disorder for problem solving. There are various models that follows this concept with different logical approaches, although having in common the interaction of their processing units.

Among the concept of swarm intelligence, there are algorithms based on animal behaviors that demonstrate a social intelligence but not particular intelligence as an individual. Within these algorithms we can highlight the algorithm of the ant colony that is used for problems of optimization, ant colony optimization algorithm or ACO [12]. This algorithm is based on the pheromone paths

that real ants deposit and follow, this behavior is simulated through simple units that process information (artificial ants) which interact with each other through artificial pheromone. This algorithm works as follows, artificial ants randomly construct traces of pheromones through the possible solutions in the optimization problem combinatorial. This trace of artificial pheromones' weight or concentration is inversely related to the solution's cost which is the one we are trying to reduce by optimizing the problem's solution. The increase is the pheromones' concentration will have as a consequence a tendency of the ants to take that route or solution, the traces not taken by the ants will have a tendency of losing their pheromones through the process of evaporation through time [13]. In Figure 2, a graphic descrip- tion of ants' behavior is shown.

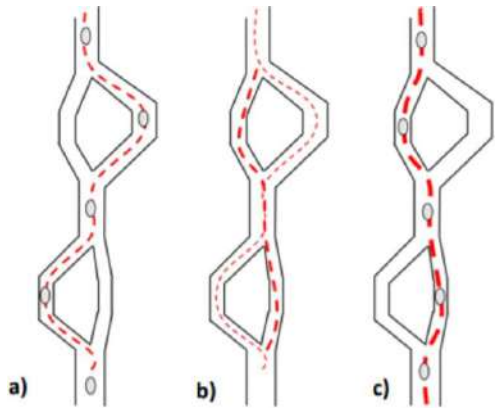



Fig. 2. Ants' behavior in ACO algorithm and their pheromone traces. (own authorship)

In Figure 2 it can be seen how the first iteration (a) the artificial ant chooses a route or solution randomly, but in such a way that the algorithm iterates (b) the weight of the pheromone traces, in Figure 2 represented by the thickness of the dotted red line, the most optimal route changes and gains more and more weight. In (c) it is shown how the pheromone trail of less optimal paths disappear completely. This algorithm belongs to metaheuristic algorithms which refer to algorithms that are design to solve combinatorial problems.

## II. Material and Methods

### A. Materials

The metropolitan area of the valley of Mexico has a constant atmospheric monitoring network called Automatic Atmospheric Monitoring Network (or RAMA for its acronym in Spanish). There are 24 stations that belong to this net- work. Each of these stations register the concen- trations of different pollutants, including PM10 and PM2.5, among others [14]. This database is maintained and updated by RAMA.

The stations used in this work were chosen taking into account two considerations:

- The availability of PM10 and PM2.5 datafrom 2012 to 2019.
- The available data has a maximum of 30% missing data during the whole evaluated period. Otherwise, the modeling and prediction may be biased.

Based in these considerations 6 stations were chosen. These stations are San Agustín (SAG), Tlanepantla (TLA), Merced (MER), Xalostoc (XAL), Camarones (CAM) and Hospital General de México (HGM). The locations of these stations are shown in Figure 3 painted in yellow.

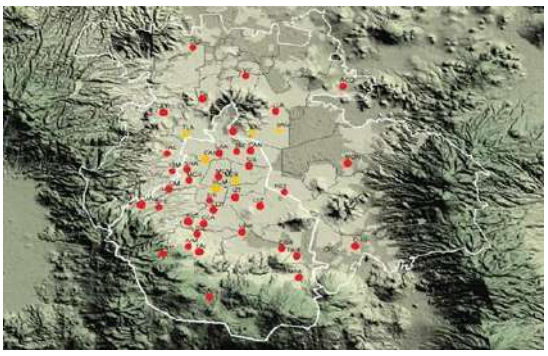


Fig. 3. All stations available in RAMA (red), stations used in this work (yellow) [14].

The databases for these stations available in RAMA give the concentration in $\mu$g/m3 which is captured every hour. The period chosen for this work is from 2012 to 2019 given that the stations chosen meet the requirements mentioned above only during this period. The ideal size of each of these databases should be 70,080. Nevertheless, all of the stations have a certain percentage of missing

data which are in the range of 13.29% to 26.35%. Since it is necessary for the LSTM algorithm to be trained by a database without missing data an imputation algorithm needed to be implemented.

*B. Methodology*

To solve the missing data problem the Multivariate Imputation by Chained Equations or MICE algorithm was applied to the data base. Consider the *n*-dimensional vector *x refering to the database* combining the $n_o$ -dimensional vector $x_o$ of captured values and the $n_m$-dimensional vector $x_m$ of missing values. For each variable *x* with missing values, in the first step the missing values are initialized with a simple imputation where they are replaced by the mean of the captured values. In the second step, consider one variable *y* and linear-regress the observed values in vector $y_o$ on the other variables in an imputation model. Inthe third step, the missing values in vector $y_m$are replaced with predictions from the estimated imputation model and, when *y* is subsequentlyused as an independent variable in the imputation model for another variable, both the observed and these imputed values will be used. In the fourth step, the previous two steps are repeatedfor each variable with missing data to complete a cycle, and multiple cycles are then performed with the imputations being updated at each one until convergence [15] .

Once the database is full with real and imputated data it is ready to train the LSTM model design and construction which is based in the author's work [10]. The first step is to normalize the data into a scale from -1 to 1, this normalization intends to facilitate training in the LSTM model by decreasing the non-linearity of the data.

The network used for modeling consists in:

- Layer 1 consists of 50 LSTM neurons, which will take the first 50 data and expand them to feed the second layer, generating a record of them.
- Layer 2 is the hidden layer that consists in 256 LSTM neurons.
- Layer 3 is a simple neuron, which based on the previous recorded data will generate anew value, successively.

The predictive neural network of the model consists in:

- Layer 1 each neuron receives a value from the input data vector which generates an output response.
- Layer 2 receives the results of the first network and generates a classification, given that initially the days with exceedances are known. The network uses a continuous re- gression to adjust its weights and obtain the expected result.

The complete model of the LSTM predictive network is shown in Figure 4 where the first three layers correspond to the modeling network and layer 4 and 5 correspond to the classification network.

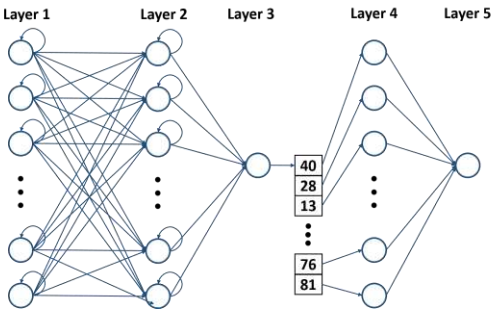



Fig. 4. Complete predictive model [10].

The construction of the ACO algorithm needs to consider that the search space in which theant colony will look for the optimized model isa matrix of *n* by *m*, *n* being the quantity of LSTM algorithms trained and *m* being the vector sizethat describes the days' concentration prediction from the model. The ACO algorithm has to select one of these daily concentrations given by one of the multiple LSTM models to have as a result an optimized LSTM-ACO model.

To accomplish these the cost matrix must be initialized. The cost matrix refers to a matrix that describes how "distantly" distributed the nodes in the problem are, these "distance" refers to the variable trying to be minimized or maximized [16]. For this work the variable being optimized is the difference between the evaluated value with the centroid of the *n* LSTM-given values, this means that ants will tend to go where the concentration of these values given by the LSTM algorithm is higher, the cost matrix was generated using the

following equations:

$$y_{ij} = \left(\frac{1}{n} \times \sum_{j=1}^{n} x_{ij}\right) - x_{ij} \quad (1)$$

$$cost\ matrix = \begin{bmatrix} y_{ij} & \cdots & y_{1k} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nk} \end{bmatrix} \quad (2)$$

Where $y_{ij}$ refers to the cost value of the *i*-th row and the *j*-th column. In this work the columns refer to the day being evaluated and rows refer to the prediction of the LSTM model being evaluated. *n* refers to the quantity of LSTM models included in the algorithm and *k* refers to the quantity of days that compose the prediction.

The pheromone matrix needs to be initialized next, this matrix describes how pheromones are distributed in each of these nodes. The pheromone matrix is initialized with a single value chosen for all the matrix or with custom values chosen by the user, this method is used if the user wants to give ants a preference of a solution. For this work the pheromone matrix is going to be initialized with 1 single value *tau* calculated with the following equation:

$$tau = 10 \times \left(\frac{1}{n} \times \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij}\right) \quad (3)$$

Once the matrices have been initialized the initialization of the ant colony is next, in which each ant is going to refer to a solution of the problem where the ant will be constructing the solution by deciding which node it is going to choose next. This decision is made using the probability described in the following equation:

$$P_{ij} = \frac{\tau_{ij}^{\alpha}\, \eta_{ij}^{\beta}}{\sum \tau_{ij}^{\alpha}\, \eta_{ij}^{\beta}} \quad (4)$$

Where $P_{ij}$ refers to the probability of traveling the path between node *i* and node j and $\tau_{ij}$ refers to the concentration of pheromones between node *i* and node *j*, this pheromone concentration is obtained from pheromone matrix and its update in each iteration. Where $\eta_{ij}$ refers to the feasibility between node *i* and node *j*. This feasibility is obtained through the reciprocal value of the cost matrix describing the cost between node i and node *j*. Also $\alpha$ refers to the weight that pheromone concentration will be given and $\beta$ refers to the weight that feasibility will be given [13].

When the ants have the solutions fully constructed the pheromone matrix is updated through the following equation:

$$\tau_{ij} = \frac{L_f}{a} \quad (5)$$

Where $L_f$ refers to the summation of all costs in the trajectory that are part of the solution of ant *f*.

$$\tau^{f}_{ij} = (1 - \rho)\, \tau_{ij} + \sum_{f=1}^{g} \tau^{f}_{ij} \quad (6)$$

Where $\rho$ refers to the evaporation rate of the pheromones and *g* to the quantity of ants.

A flowchart of the entire methodology is shown in Figure 5. Where the sections describing preprocessing of data and LSTM methodology where taken from [10].

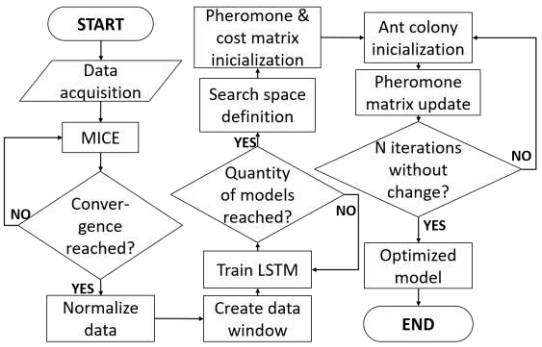



Fig. 5. Methodology used.

### C. Evaluation

To evaluate the LSTM-ACO model and successfully compare it with the previous LSTM model, the RMSE, accuracy, precision, recall and F1-score were implemented. The comparison was made between the resulting LSTM-ACO model and the mean of the measurements between the *n* LSTM models trained. The RMSE represents the standard deviation between actual and predicted values as the following equation describes:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (YM_i - YR_i)^2} \quad (7)$$

Where $YM_i$ refers to the $i$-th element of the prediction model and $YR_i$ refers to the $i$-th element of the real data.

RMSE is a representation of how close distance between the prediction and real values are, this value will be used to decide the optimal number of LSTM models to use as the search space in the ACO algorithm.

A classification evaluation was implemented in which a confusion matrix is calculated with 4 different values that are described below:

- True positive or TP: Which is interpreted as predicted positive and it's true.
- False positive or FP: Which is interpreted as predicted negative and it's true.
- True negative or TN: Which is interpreted as predicted positive and it's false.
- False negative or FN: Which is interpreted as predicted negative and it's false.

These measurements fit right with this classification problem because it may have 1 of 2 possible classifications, exceedance or not exceedance, 50 $\mu g/m^3$ daily permisible for PM10 and 25 $\mu g/m^3$ daily permisible for PM2.5. Each of the following equations describe the measurements implemented for this classification problem:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$Accuracy = \frac{TP + TN}{Total} \tag{10}$$

$$F1\_score = \frac{2 * TP}{(2 * TP + FP + FN)} \tag{11}$$

## III. Results and Discussion

As mentioned in the materials section above, the only stations considered for this work were the ones with less than a 30% of data missing in both particles evaluated, PM2.5 and PM10. After applying the MICE algorithm the days with exceedances following the WHO guidelines were calculated. Initial missing data and days with exceedances are shown in Table I per particle and station.

TABLE I
INITIAL MISSING DATA AND EXCEEDANCES PER STATION

| Station | Initial missing data | | Days with Exc. | |
|---|---|---|---|---|
| | PM2.5 | PM10 | PM2.5 | PM10 |
| SAG | 26.35% | 26.35% | 29.33% | 38.94% |
| TLA | 19.27% | 19.24% | 35.96% | 35.96% |
| MER | 13.64% | 13.29% | 43.27% | 49.52% |
| XAL | 18.28% | 18.21% | 47.95% | 67.64% |
| HGM | 16.49% | 16.49% | 33.04% | 19.69% |
| CAM | 24.02% | 24.02% | 47.09% | 42.30% |

To decide the optimal number of LSTM models that make up the ACO search space an experiment was implemented. This experiment consists in run the ACO algorithm using a search space of 3 LSTM models and then comparing the RMSE of the LSTM-ACO model to the mean RMSEof the LSTM models. Subsequently, 1 LSTM will be concatenated to the search space and the process will be repeated. The objective of this experimentation is to find a correlation between the quantity of LSTM that compose the search space with the RMSE, these results are shown in Figure 6.

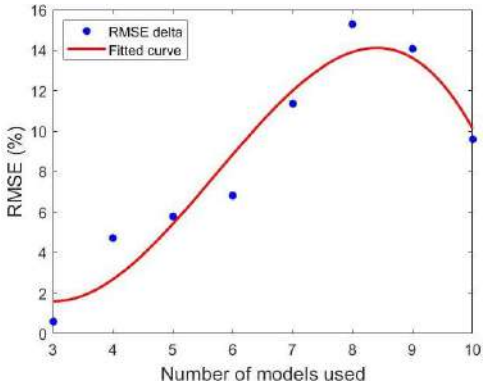



Fig. 6. RMSE improvement percentage vs number of LSTM models that compose the search space.

In Figure 6 the data shown represents the difference between the LSTM-ACO model RMSE and the mean RMSE of the LSTM models being evaluated. The number LSTM models that this experiment proves to be the optimal is 8. This experiment was implemented using the PM10 data from the Merced station which was chosen because it is the station with less missing PM10 data, the missing PM10 data represented 13.29%

from the chosen range, 2012 to 2019.

Once the quantity of LSTM models was defined, an experiment to find an efficient combination of parameters for the ACO algorithm was developed. The definition of efficiency considered in the experiment was to reach the least RMSE difference between the obtained model and the imputated data using the least amount of time. This experiment was repeated 25 times to observe repeatability through the standard deviation of the RMSE results per combination of parameters, each of the results shown in Table II is the mean of these 25 repetitions. The experiment consisted in combining values for the number of ants (g in Equation (6)) in the algorithm and for the evaporation rate ( $\rho$in Equation (6) ), the weight of the pheromone concentration ($\alpha$in Equation (4) ) and the weight of feasibility ($\beta$in Equation (4) ) were defined as 1 because there was no interest in giving the algorithm any bias towards any of them. The results considering a quantity of ants of5,10 and 25, and the evaporation rate of 1%,5% and 10% are shown in Table 2, these test was developed using PM10 data of station in Merced. The stop condition considered for the ants is to reach a 100 iterations without finding a better solution than the best one found.

TABLE II
EVAPORATION RATES AND ANT QUANTITIES TESTED

| No. | $\rho$= 1% | | | $\rho$= 5% | | | $\rho$= 10% | | |
|---|---|---|---|---|---|---|---|---|---|
| | t(s) | rms e | std. | t(s) | *rms e* | std. | t(s) | rms e | std. |
| 5 | 14.89 | 4.40 | 0.011 | 10.78 | 4.39 | 0.002 | 7.46 | 4.39 | 0.00 |
| 10 | 25.42 | 4.39 | 0.006 | 21.38 | 4.39 | 0.002 | 16.85 | 4.39 | 0.00 |
| 25 | 59.84 | 4.39 | 0.007 | 46.28 | 4.39 | 0.002 | 30.84 | 4.39 | 0.00 |

Results in Table II presented $\rho$= 10% and ants =5 as the most efficient combination of parameters by the less RMSE reached, 4.39, in the least amount of time, 7.46 seconds. The repeatibility of this combination of parameters is considered as high as the standard deviation of the 25 RMSE's obtained is 0.001.

Once the 8 LSTM models per particle and station were trained they were evaluated using precision (Equation (8)), accuracy (Equation (10)), recall (Equation (9) ) and f1-score (Equation (11) ). The mean results of the 8 LSTM models are shown in Table III (PM2.5) and Table IV (PM10).

TABLE III
EVALUATION METRICS OF LSTM MODELS PREDICTING PM2.5 EXCEEDANCES.

| Station | Precision | Accuracy | Recall | F1-score |
|---|---|---|---|---|
| SAG | 83.77% | 84.95% | 75.41% | 73.39% |
| TLA | 89.87% | 91.87% | 89.42% | 88.63% |
| MER | 88.74% | 91.38% | 93.47% | 90.47% |
| XAL | 84.27% | 86.20% | 91.68% | 86.43% |
| HGM | 93.12% | 93.52% | 88.08% | 89.81% |
| CAM | 94.13% | 93.08% | 91.50% | 92.41% |

TABLE IV
EVALUATION METRICS OF LSTM MODELS PREDICTING PM10 EXCEEDANCES.

| Station | Precision | Accuracy | Recall | F1-score |
|---|---|---|---|---|
| SAG | 93.97% | 95.46% | 94.75% | 94.22% |
| TLA | 91.81% | 94.47% | 93.61% | 92.47% |
| MER | 94.07% | 93.48% | 93.02% | 93.35% |
| XAL | 91.93% | 91.41% | 96.31% | 93.81% |
| HGM | 95.06% | 96.85% | 89.23% | 91.71% |
| CAM | 93.90% | 95.32% | 95.72% | 94.58% |

The LSTM-ACO model was built by the ACO algorithm with the characteristics mentionedabove using the 8 LSTM models as the ant's search space and minimizing the cost in Equa-tion (1). The improvement in the results of the evaluation metrics applied to the LSTM-ACO model are shown in Table V (PM2.5) and Table IV(PM10). The results shown are taken by a simple difference between the LSTM-ACO result and the LSTM result.

TABLE V
DIFFERENCE BETWEEN LSTM-ACO MODEL AND LSTM MODEL PREDICTING PM2.5 EXCEEDANCES.

| Station | Precision | Accuracy | Recall | F1-score |
|---|---|---|---|---|
| SAG | +6.39% | +9.28% | +14.75% | +16.77% |
| TLA | +1.57% | +3.67% | +7.25% | +5.35% |
| MER | +1.21% | +1.65% | +0.97% | +1.67% |
| XAL | +4.12% | +6.60% | +6.18% | +6.45% |
| HGM | +2.26% | +0.63% | +3.63% | +3.10% |
| CAM | +0.72% | +1.61% | +2.32% | +1.92% |

## IV. CONCLUSIONS

With the present work, a methodology was proposed to take raw PMx data, impute the database to complete it, based on this imputed data traina series of LSTM models and finally improve the predicted classifications of the recurrent network

TABLE VI
DIFFERENCE BETWEEN LSTM-ACO MODEL AND LSTM MODEL PREDICTING PM10 EXCEEDANCES.

| Station | Precision | Accuracy | Recall | F1-score |
|---|---|---|---|---|
| SAG | +1.71% | +1.18% | +0.93% | +1.46% |
| TLA | +1.59% | +1.08% | +0.68% | +1.37% |
| MER | +1.48% | +1.23% | +0.67% | +1.26% |
| XAL | +0.09% | +2.25% | +2.93% | +1.68% |
| HGM | +1.20% | +0.41% | +0.44% | +1.08% |
| CAM | +0.26% | +1.26% | +2.26% | +1.45% |

using the Ant Colony Algorithm. The proposed LSTM-ACO model was tested through various evaluation metrics that demonstrated high repeatibility in precision, accuracy, recall and F1-score that average 93.10%, 94.90%, 94.59% and 93.74% respectively. These averages were calculated taking into consideration both particles for every station evaluated in this work, which seems to indicate that a highly reliable, generalized PMx model to predict exceedances may be achieved.

To improve this predicted classifications it would be helpful to carry out a characterization of the whole methodology evaluating different parameter combinations. To make a complete characterization it is proposed to apply this methodology in different phenomenons which have a sequential nature in their and evaluate their results.

## REFERENCES


[1] A. N. Arredondo, "Control de la contamincación atmosférica en la Zona Metropolitana del Valle de México," *Estudios demográficos y urbanos*, vol. 34, no. 3, 2019. [Online]. Available: 10.24201/edu.v34i3.1806

[2] WHO, "Ambient (outdoor) air quality and health." May 2018.

[3] X. Zhu, H. Qiu, L. Wang, Z. Duan, H. Yu, R. Deng, Y. Zhang, and L. Zhou, "Risks of hospital admissions from a spectrum of causes associated with particulate matter pollution," *Science of the Total Environment*, vol. 656, pp. 90–100, 2019. [Online]. Available: 10.1016/j.scitotenv.2018.11.240

[4] C. Wang, L. Feng, and K. Chen, "The impact of ambient particulate matter on hospital outpatient visits for respiratory and circulatory system disease in an urban Chinese population," *Science of the Total Environment*, vol. 666, no. 20, pp. 672–679, 2019. [Online]. Available: 10.1016/j.scitotenv.2019.02.256

[5] W. H. Organization, "WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide," pp. 6–10.

[6] C. Ubilla and K. Yohanssen, "Contaminación atmosférica: Efectos en la salud respiratoria de un niño," *Revista Médica Clínica Condes*, vol. 28, no. 1, pp. 111–118, 2017. [Online]. Available: :10.1016/j.rmclc.2016.12.003

[7] G. Yang and X. Wang, "Artificial Neural Networks for Neuroscientists: A Primer," *Neuron*, vol. 107, no. 6, pp. 1048–1070, 2020. [Online]. Available: 10.1016/j.neuron.2020.09.005

[8] C. Zhou, Q. Wang, and C. Zhou, "Photocatalytic degradation of antibiotics by molecular assembly porous carbon nitride: Activity studies and artificial neural networks modeling," *Chemical Physics Letter* , vol. 750, 2020. [Online]. Available: 10.1016/j.cplett.2020.137479

[9] H. Sak and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," *Google*, 2014.

[10] J. Ramírez-Montañez, M. A. Fernández, A. T. Arriaga, J. R. Arreguín, and G. S. Calderón, "Evaluation of a Recurrent Neural Network LSTM for the Detection of Exceedances of Particles PM10," in *International Conference on Electrical Engineering, Conputing Science and Automatic Control (CCE), September 11-13, Mexico City, Mexico*, September 2019.

[11] M. Thrun and A. Ultsch, "Swarm intelligence for self-organized clustering," *Artificial Intelligence*, vol. 290, 2021. [Online]. Available: 10.1016/j.artint.2020.103237

[12] J. Uthayakumar, N. Metawa, K. Shankar, and S. Lakshmanaprabu, "Financial crisis prediction model using ant colony optimization," *International Journal of Information Management* , vol. 50, pp. 538–556, 2020. [Online]. Available: 10.1016/j.ijinfomgt.2018.12.001

[13] M. Dorigo and T. Stützle, " Ant Colony Optimization: Overview and Recent Advances," *International Seriesin Operations Research & Management Science* , vol. 272, pp. 311–351, 2010. [Online]. Available: 10.1007/978-1-4419-1665-5-8

[14] RAMA, "Databases - Automatic Air Monitoring Network." 2021. [Online]. Available: www.aire.cdmx.gob.mx/default.php?opc=%27aKBh%27

[15] L. Li, C. G. Prato, and Y. Wang, "Ranking contributors to traffic crashes on montainous freeways from an incomplete dataset: A sequential approach of multivariate imputation by chained equations and random forrest classifier. ," *Accident Analysis and Prevention*, vol. 146, no. 105744, 2020. [Online]. Available: 10.1016/j.aap.2020.105744

[16] M. Paniri, M. Bagher, and H. Nezamabadi-pour, "Ant colony optimization for multi-objective optimization problems," *Knowledge-Based Systems*, vol. 192, no. 15, 2020. [Online]. Available: 10.1016/j.knosys.2019.105285