INSTITUTO TECNOLÓGICO METROPOLITANO - ITM
FACULTAD DE INGENIERÍAS

DEPARTAMENTO DE ELECTRÓNICA Y TELECOMUNICACIONES

PROYECTO DE GRADO

METODOLOGÍA PARA MEJORAR LA CALIDAD DE MEDICIÓN DE MATERIAL PARTICULADO PM2.5 DE LA RED DE SENSORES DE BAJO COSTO DEL PROYECTO CIUDADANOS CIENTÍFICOS DEL VALLE DE ABURRÁ, UTILIZANDO TÉCNICAS DE APRENDIZAJE DE MÁQUINA.

MAGÍSTER EN AUTOMATIZACIÓN Y CONTROL INDUSTRIAL

LEÓN MAURICIO RIVERA MUÑOZ
MEDELLÍN, 2021

METODOLOGÍA PARA MEJORAR LA CALIDAD DE MEDICIÓN DE MATERIAL PARTICULADO PM2.5 DE LA RED DE SENSORES DE BAJO COSTO DEL PROYECTO CIUDADANOS CIENTÍFICOS DEL VALLE DE ABURRÁ, UTILIZANDO TÉCNICAS DE APRENDIZAJE DE MÁQUINA.

Autor: León Mauricio Rivera Muñoz

Tutor: Juan David Martínez Vargas

Tutor: Andrés Felipe Giraldo Forero

Departamento: Electrónica y Telecomunicaciones, ITM

Titulación: Magíster en Automatización y Control Industrial

Palabras clave: PM2.5, IoT, Redes de Sensores (WSNs), Estimación de información perdida, Factorización de Matrices (MF), Maching Learning, Deep Learning.

Resumen

Las redes de sensores (WSNs) de bajo costo, entre otras cosas, son implementadas para dar respuesta a la necesidad actual de conocer a fondo las dinámicas de la contaminación en entornos urbanos y rurales, y las consecuencias que el material particulado y los gases de efecto invernadero generan en la salud humana. Dichas redes se caracterizan por los bajos costos de operación y bajo consumo energético con respecto a las estaciones referencia usadas en la actualidad. Sin embargo, las redes de sensores de bajo costo son cuestionadas en gran medida por la calidad de los datos, debido a la cantidad de información perdida y las tecnologías usadas para determinar la concentración, ya que en la actualidad no se cuenta con protocolos de ajuste y aseguramiento de medida estandarizados.

Lo anterior evidencia la problemática de interés a resolver en este trabajo, para el cual, fueron usados los datos capturados durante el año 2018 de la red NOVA de sensores de bajo costo implementada en la ciudad de Medellín y el Valle de Aburrá. En consecuencia, este trabajo se centra en el planteamiento de una metodología de adaptación y aplicación de técnicas de factorización de matriz (MF), dadas las características de la técnica para estimar datos a partir de la información presente en el conjunto de datos, con el objetivo de mejorar la calidad de la misma, ya que actualmente la base de datos cuenta con 5 $\%$ de información anómala y aproximadamente el 40 $\%$ son no medidos.

Por lo tanto, en este trabajo se presentan los resultados obtenidos que se pueden distribuir en tres enfoques: el primero abarca todo el estudio del estado del arte sobre la aplicación de sensores de bajo costo para la vigilancia de la contaminación atmosférica, el uso técnicas de ajuste y la implementación de algoritmos para la estimación de información perdida, entre los que se encuentran las técnicas basadas en MF (Capítulo 1); el segundo se centra en el planteamiento, diseño, sintonización de un modelo de MF y la evaluación del mismo con respecto a técnicas mencionadas en la literatura, para la cual, fue usada una metodología de eliminación de datos aleatoria y de huecos con el fin de evaluar el error de ajuste entre el dato real y el eliminado artificialmente (Capítulos 2-4). Finalmente, en el tercer enfoque se presenta una propuesta de mejora a los resultados obtenidos, planteando una técnica de MF basada en redes neuronales denominada Deep Matrix Factorization (DMF), y el análisis comparativo de desempeño entre la técnica MF y la técnica de DMF con diferentes modificaciones, usando información espacio-temporales incorporadas por medio de Embedding Layers (Capítulo 5 y sección de conclusiones). De este trabajo se encontró que la técnica DMF presenta mejor desempeño respecto al modelo MF. Adicionalmente, la inclusión de información espacial al modelo DMF (DMF3) permite un mejor aprendizaje de la dinámica de la red, logrando un menor error de estimación que el logrado con el modelo DMF estándar (DMF1).

ELECTRONICS AND TELECOMMUNICATIONS DEPARTMENT, ITM

METHODOLOGY TO IMPROVE THE MEASUREMENT QUALITY OF PM2.5 PARTICULATE MATTER FROM THE LOW-COST SENSOR NETWORK OF THE ABURRÁ VALLEY SCIENTIFIC SCIENTISTS PROJECT, USING MACHINE LEARNING TECHNIQUES

Author: León Mauricio Rivera Muñoz

Supervisor: Juan David Martínez Vargas

Supervisor: Andrés Felipe Giraldo Forero

Department: Department of Electronics and Telecommunications

Degree: Magíster en Automatización y Control Industrial

Keywords: PM2.5, IoT, Wireless Sensor Networks (WSNs), Missing data estimation, Matrix Factorization (MF), Maching Learning, Deep Learning.

Abstract

Low-cost sensor networks (WSNs), among other things, are implemented to respond to the current need for an in-depth understanding of the dynamics of pollution in urban and rural environments, and the consequences of particulate matter and greenhouse gases on human health. These networks are characterized by low operating costs and low energy consumption compared to the reference stations used today.However, low-cost sensor networks are largely questioned by the quality of the data, due to the amount of information lost and the technologies used to determine the accumulation of pollutants, since currently there are no standardized protocols for adjustment and measurement assurance.

The above mentioned evidences the problem of interest to be solved in this work, for which, the data captured during the year 2018 from the NOVA network of low cost sensors implemented in the city of Medellin and the Aburrá Valley

were used. Consequently, this work is focused on the proposal of an adaptation methodology and application of matrix factorization (MF) techniques, given the characteristics of the technique to estimate data from the information present in the data set, with the aim of improving its quality, since currently the database has 5 % of anomalous information and approximately 40 % are not measured.

Therefore, this study presents the results obtained that can be distributed in three approaches. The first one covers the whole study of the state of the art on the application of low cost sensors for air pollution monitoring, the use of adjustment techniques and the implementation of algorithms for the estimation of lost information, among which are the techniques based on MF (Chapter 1). The second focuses on the approach, design, tuning, and evaluation of a MF model with respect to techniques mentioned in the literature, for which a gap and random data elimination methodology was used to evaluate the fit error between the actual data and the artificially eliminated data (Chapters 2-4). Finally, the third approach presents a proposal to improve the results obtained, proposing a MF technique based on neural networks called Deep Matrix Factorization (DMF), and the comparative analysis of performance between the MF technique and the DMF technique with different modifications, using spatial-temporal information incorporated by means of Embedding Layers (Chapter 5 and conclusions section). From this work it was found that the DMF technique presents better performance than the MF model. Additionally, the inclusion of spatial information to the DMF model (DMF3) allows a better learning of network dynamics, achieving a lower estimation error than that achieved with the standard DMF model (DMF1).

Dedicado a mi esposa Andrea Toro Henao
y a mi madre María Ceneida Muñoz Osorno,
que siempre me han apoyado incondicionalmente.

El autor

Agradecimientos

Agradezco principalmente a mi esposa y madre por la paciencia y poco tiempo con el que conté para ellas, pero que sin su apoyo este logro no seria posible.

Igualmente agradezco a mis asesores Andrés Felipe Giraldo Forero y Juan David Martínez Vargas y al ITM, por su acompañamiento, trasmisión de conocimiento y confianza para el desarrollo del presente trabajo.

Finalmente, agradezco a mis amigos, Andrés Mauricio Cárdenas Torres, Raúl Enrique Córdoba, Daniel Giraldo Guzmán, Jorge Mora y Pablo Hernández, por su gran amistad durante años, apoyo y motivación constante para la finalización de este trabajo.

Acrónimos

ANNs	Artificial Naural Networks
AQI	Air Quality Index
CO	Carbon Monoxide
CO_2	Carbon Dioxide
CH_4	Methane
DL	Deep Learning
DMF	Deep Matrix Factorization
EPA	Environmental Protection Agency
EVS	Explained Variance Score
GD	Gradient Descent
GPS	Global Positioning System
ICA	Índice de Calidad del Aire
IoT	Internet of Things
KNN	K-Nearest-Neighbor
MAPE	Mean Absolute Percentage Error
MC	Matrix Completion
MF	Matrix Factorization
ML	Machine Learning
MLP	Multilayer Perceptron

MOGPs	Multi-Output Gaussian Processes
MS	Mean substitution
NO_2	Nitrogen dioxide
NNLS	Non-Negative Least Squares
NNMF	Non-Negative Matrix Factorization
OMS	Organización Mundial de la salud
O_3	Ozone
$PM_{1,0}$	Particulate Matter (1 micrometers)
$PM_{2,5}$	Particulate Matter (2.5 micrometers)
PM_{10}	Particulate Matter (10 micrometers)
PMF	Probability Matrix Factorization
RF	Random Forest
RMSE	Root Mean Square Error
SIATA	Sistema de Alerta Temprana de Medellín y el Valle de Aburrá
SO_2	Sulfur Dioxide
SGD	Stochastic Gradient Descent
SVCA	Sistemas de Vigilancia de la Calidad de Aire
SVD	Singular Value Decomposition - $U\Sigma V^T$
TIC	Tecnologías de la Información y Comunicación
VOC	Volatile Organic Compounds
WHO	World Health Organization
WSNs	Wireless Sensor Networks

Índice

Índice de figuras

Índice de Tablas

Parte I

Introducción.

Capítulo 1

Introducción y visión general

Contenido

Sinopsis

Este capítulo contextualiza a partir de la mirada encontrada en la literatura las diferentes problemáticas abordadas por los autores en la implementación de sensores de bajo costo, presentando un resumen gráfico de las técnicas tratadas en la búsqueda del estado del arte que resuelven dichas problemáticas.

Lo anterior permite entender la actualidad de la temática del seguimiento e investigación de la contaminación atmosférica con la implementación de tecnologías IoT de bajo costo. Posibilitando el planteamiento de nuevas metodologías de mejora en los datos obtenidos por las WSNs como Ciudadanos Científicos, ya sea con técnicas de muestreo, calibración, ajuste, predicción o recuperación de información a partir de la dinámica de los mismos datos generados por la red, lo cual a futuro permita complementar las actuales redes de sensores de seguimiento de contaminación del Área Metropolitana (sensores y equipos con estándares y normativas de aseguramiento de la medida).

1.1. Introducción

De acuerdo con los datos entregados por la Organización Mundial de la Salud (WHO), el aumento de la población, la creciente demanda de recursos y la necesidad de transformarlos, generan un aporte significativo a la contaminación actual del aire, la cual es uno de los principales factores que generan afectaciones en la salud de las personas a corto y largo plazo dependiendo del grado de exposición y proximidad a puntos críticos de contaminación [56, 87].

La WHO determinó que para el año 2016 una de cada nueve muertes se encuentra relacionada con la contaminación del aire, además el 92 % de la población vive en áreas con niveles de contaminación que exceden los límites establecidos por la WHO [55]. Por lo tanto, se estima que el 23 % de las muertes a nivel mundial están relacionadas con la contaminación del medio ambiente. Esto representa 12,6 millones de muertes al año, donde a la región de Latinoamérica se le atribuye alrededor de 847.000 muertes por año y son los países en vía de desarrollos los que más aportan a las estadísticas anuales [86, 62, 42].

La contaminación atmosférica trae consecuencias a nivel local, regional y global y dependiendo del área de influencia se ven materializadas de la siguiente manera: a nivel local genera contaminación urbana, la cual en la actualidad atrae mayor interés en la comunidad en general [14], a nivel regional se evidencian problemáticas de lluvia ácida que afecta principalmente cultivos y ecosistemas, y a nivel global genera cambio climático o calentamiento global, siendo el más preocupante el generado como consecuencia de las actividades humanas [28, 81, 50]

Lo mencionado hasta el momento representa el contexto y mirada general sobre una de las grandes problemáticas a nivel mundial que se enlista como una prioridad actual por resolver, o por lo menos mitigar. Es por esto que el monitoreo de la calidad del aire es relevante, no solo para las personas que viven en

áreas urbanas, sino también en entornos rurales, debido a que allí se genera una parte importante de los contaminantes como consecuencia de las actividades antropogénicas [28, 81]. Por lo tanto, se deben buscar diferentes soluciones para asegurar la calidad del aire en dichos entornos y el primer paso es monitorear manteniendo buena cobertura para conocer y lograr un entendimiento del fenómeno a fondo [45, 69, 47, 6, 70].

En Colombia existe el Sistema de Vigilancia de la Calidad de Aire (SVCA), con la finalidad de monitorear y vigilar los niveles de emisión en el territorio nacional, los cuales se encuentran divididos en administraciones ambientales para cada zona del país [51]. En el caso de Medellín, la autoridad ambiental es el Área Metropolitana que a través del SIATA (Sistema de Alerta Temprana de Medellín y el Valle de Aburrá) administra el sistema de monitoreo y de alerta temprana, desplegado por gran parte del Valle de Aburrá para el desarrollo de cuatro tareas sustanciales: monitoreo, modelación, gestión y comunicación. Todo ello con miras a fortalecer la toma de decisiones y la intervención oportuna de los organismos de respuesta a partir de los datos generados [75].

Sin embargo, en Colombia la mayoría de los SVCA cuentan con limitaciones de cobertura espacial y flexibilidad, debido al costo de operación, consumo de energía y alto valor de equipos. Este es el caso del Valle de Aburrá, ya que en la actualidad de las 21 estaciones de monitoreo dispuestas en la zona, no todas miden material particulado (PM2.5 y PM10) reduciendo así la resolución espacial que provee la red para estas variables [75].

Por tal motivo y aprovechando la implementación de WSNs por medio del uso de tecnologías IoT (Internet of Things) [96, 20, 73], el SIATA, dentro de sus actividades de investigación implementó una WSN de bajo costo (Ciudadanos Científicos), ubicando nodos en viviendas y empresas distribuidas en el Valle de Aburrá con el fin de ampliar la cobertura del monitoreo en la zona, de manera tal que esto permita conocer mejor y en tiempo real el estado actual de la calidad del aire del Área Metropolitana.

Aunque el proyecto Ciudadanos Científicos mejora la resolución espacio- temporal de la medición de material particulado en la zona, durante el año 2018 aproximadamente el 45 % de la información generada por la red se encuentra representada por datos anómalos y datos vacíos, posiblemente causados por nodos fuera de operación y datos no medidos debido a condiciones energéticas, continuidad de la red de internet y robo de nodos. Adicionalmente, los dispositivos de bajo costo son propensos al ruido y se ven afectados por variaciones atmosféricas [14]. Además de lo anterior, las tecnologías de los sensores de bajo costo en su mayoría no cuentan con protocolos de aseguramiento de medida acreditados internacionalmente y no posibilita el aprovechamiento de los datos

por parte de las autoridades ambientales y para uso en investigación [22].

Por lo tanto, se hace necesario plantear metodologías de reconstrucción de señales para mejorar la calidad de la información generada por las redes de sensores de bajo costo lo suficientemente robustas para trabajar con gran cantidad de información perdida, y adicionalmente posibiliten la inclusión dentro de estas información temporal y geográfica lo cual permita disminuir el error de estimación aprendiendo mejor el espacio de baja dimensión para las representaciones de sensores y tiempo.

De esta manera se busca contribuir con modelos basados en Machicne Learning (ML) y Deep Learning (DL) para la reconstrucción de datos que permitan aprovechar las bondades de las WSNs de bajo costo en actividades de investigación, analítica y como complemento de las redes gubernamentales implementadas en la actualidad. Buscando así, mejorar la resolución y seguimiento de la medición de PM2.5 en zonas urbanas, rurales y de difícil acceso.

1.2. Estado del arte

En los últimos años se han desarrollado diferentes enfoques para tratar la temática de los sensores de bajo costo en la medición de gases contaminantes y material particulado. Debido a la presente búsqueda estos fueron agrupados por categorías como se muestra en la Figura 1.1.

De la estructura [1] se desprenden tres temas centrales que abarcan los investigadores en sus trabajos sobre la temática resumidos en: (a) Estudios sobre las tecnologías de sensores de bajo costo para la medición de contaminantes atmosféricos, con recomendaciones de selección y uso dependiendo de la aplicación y variables a medir (tema coyuntural para seleccionar la metodología de tratamiento de datos de acuerdo a tipo de sensor), (b) problemáticas de interés para los investigadores, entre las cuales se encuentran: predicción; ajuste; análisis temporal y espacial; y recuperación de información, siendo este último, el tema de interés para la presente trabajo de maestría y (c), con aplicaciones e implementaciones de sensores de bajo costo y de WSNs.

A continuación se hará referencia sobre cada uno de estos temas, iniciando por la identificación de ventajas y desventajas sobre algunas de las tecnologías implementadas para la medición de contaminantes atmosféricos, donde se fina-

[1] En azul se presenta el camino por donde es dirigida la búsqueda. En rojo se presenta la problemática de interés de la investigación, orientado la búsqueda sobre las diferentes técnicas y metodologías que permiten el ajuste y procesamiento de datos de las redes de sensores de bajo costo.

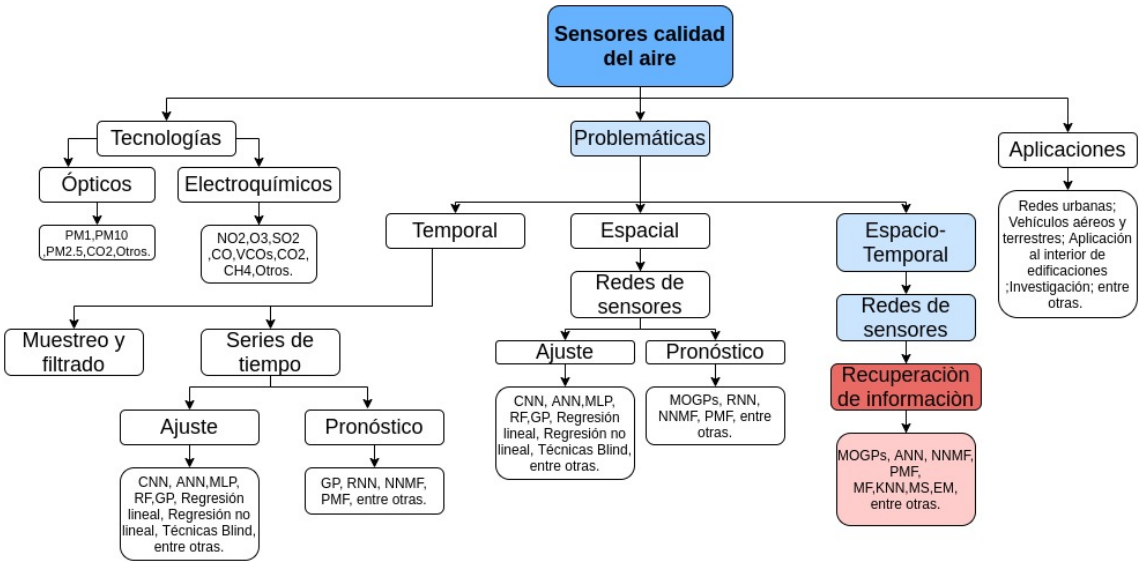


Figura 1.1: Estructura actual del estado del arte sobre los sensores de bajo costo, aplicaciones y técnicas implementadas para el procesamiento de señal

lizará mencionando las diferentes aplicaciones encontradas en la literatura, los problemas que resuelven en función del tratamiento de los datos y como estos han aportado a la solución de la problemática tratada en este trabajo.

1.2.1. Tecnologías sensores de bajo costo para medir calidad del aire

Actualmente existe una gran variedad de sensores de bajo costo para determinar la concentración de algún tipo de gas o material particulado, entre las variables de mayor interés se encuentran: gases reactivos NO_2, O_3, SO_2, CO y VOC_s; gases de efecto invernadero CO_2 y CH_4; y material particulado PM1, PM2.5 y PM10 [45]. Dependiendo del fenómeno físico aprovechado por el dispositivo deriva su tecnología, categorizados de la siguiente manera:

- Electroquímicos, pueden detectar gases por medio de reacciones químicas entre el gas objetivo y dos electrodos separados, generando variaciones de resistencia del elemento transductor [45]. Estos sensores se caracterizan por tener buena sensibilidad y rápida respuesta [4, 10]. Sin embargo, son altamente afectados por variables físicas como temperatura y humedad [48]. Adicionalmente cuentan con problemas de repetitividad [4] y alta reactividad con gases similares [16], [18]. Entre las variables medidas por

dichos sensores se encuentran: SO_2, NO, CO_2, O_3 y CO.

- Fotoionización, por medio de luz ultravioleta y el desprendimiento de electrones se excitan las moléculas del gas, lo cual, produce una corriente eléctrica que deriva con la medición de la concentración del gas [10, 49]. Estos sensores se caracterizan por su rápida respuesta, pero cuentan con alta reactividad a gases similares y mediciones altamente variables (Significant signal drift) [22, 48, 10]. Esta tecnología es implementada en la medición de compuestos orgánicos volátiles VOC [48].

- Semiconductores basados en óxidos metálicos, con la característica de variar la resistencia eléctrica superficial del material cuando la composición química del medio es alterado por el gas en cuestión [1]. Estos sensores cuentan con buena sensibilidad, pero son altamente afectados por la temperatura y humedad, y su respuesta es lenta llegando a varios minutos de estabilización para algunos gases [22, 38]. Las variables medidas con esta tecnología son: NO_2, O_3 ,CH_4 y CO.

- Ópticos, a través de un haz de luz detectan el tamaño de las partículas por medio de la dispersión de la luz al pasar la partícula o concentración de gas por el rayo infrarrojo [33]. Se caracterizan por tener buena respuesta y alta sensibilidad [22, 10]. Sin embargo, La respuesta del sensor depende de la forma de las partículas (índice de refracción), temperatura y humedad [33, 68, 1]. El modelo de conversión se basa en un modelo teórico [10] y presenta cambios en la respuesta del dispositivo debido al deterioro por uso [1]. Las variables medidas con esta tecnología son: PM10,PM2.5, PM1, CO y CO_2.

Por lo tanto, antes de pensar en un desarrollo o aplicación con sensores de bajo costo para medir la calidad del aire, es recomendable conocer las diferentes tecnologías, teniendo en cuenta especificaciones técnicas, limitaciones propias de cada uno de los dispositivos y si cumplen o no con las especificaciones de la aplicación [22]. Lo anterior debe ser tenido en cuenta en el momento de realizar una implementación o trabajo con los datos generados por los dispositivos, ya que de esto depende también la formulación y la estrategia para abordar el problema de tratamiento de los datos, procesamiento de la señal o ajuste de este tipo de dispositivos.

1.2.2. Características propias de los sensores, aplicaciones con sensores de bajo costo y metodologías para mejorar la calidad de medición

La idea general de las WSNs de bajo costo y los datos que éstas generan es potenciar la medición de variables ambientales debido al bajo costo de los equipos, fácil implementación y bajo consumo de energía con respecto a las plataformas actualmente utilizadas para el monitoreo ambiental [5].

De acuerdo con [2], las WSNs pueden ser implementadas en regiones o países de bajos recursos, permitiendo hacer frente al empeoramiento de la calidad del aire urbano y rural, con datos que posibiliten conocer la magnitud de los efectos en la salud asociados a la contaminación en las regiones; cerrando la brecha de información y dando paso a políticas y regulaciones de control, lo cual, ayude a crear una infraestructura para la gestión de las regiones en tiempo real [7].

De igual manera las WSNs se pueden aplicar en el desarrollo de equipamiento ligero debido al tamaño de los sensores y de la instrumentación que pueden ser ubicados tanto en vehículos aéreos autónomos, con el fin de abarcar grandes espacios con un solo dispositivo y conocer las concentraciones de los contaminantes a diferentes alturas [6] o en aplicaciones en interiores como es el caso de instituciones, colegios y empresas [64], donde son usados para estudiar fenómenos de acumulación de material particulado en interiores como consecuencia de las actividades desarrollas o por el diseño de la propia estructura [6, 41, 58].

Gracias a lo anterior y a la gran cantidad de aplicaciones posibles con las WSNs, varias de las investigaciones en la temática han enfocado sus esfuerzos en seguir protocolos y estándares de aseguramiento de medida, con el objetivo de que estas redes y sus aplicaciones pueden ser tenidas en cuenta por autoridades ambientales [15, 51], planteando alternativas que permitan mejorar las características y la calidad de los datos generados por los sensores, ya sea utilizando técnicas de filtrado para eliminar ruidos provenientes de la señal del sensor [41]; técnicas de reducción de datos y muestreo sin afectar la integridad de la información [5]; técnicas para la reconstrucción de señales [23]; o la aplicación de regresiones para compensar por medio de software las características de los dispositivos [22].

La generalidad encontrada en la revisión del estado del arte indica que sin importar la naturaleza de la aplicación y el uso que se le desee dar a los datos, los problemas a solucionar con las redes de bajo costo derivan en problemas de pronóstico (extrapolación) [57] y ajuste (interpolación) [98], siendo más evidente el interés por tratar el problema de ajuste aplicado a un sensor o redes de sensores, implementando como metodología de ajuste el uso de sensores referencia o

instrumentos patrón.

Actualmente las técnicas de ajuste y procesamiento de señal de los sensores de bajo costo van desde simples regresiones lineales, técnicas de aprendizaje de máquina (ML) y hasta complicados algoritmos matemáticos bastante sofisticadas que permiten el ajuste de datos de forma robusta, los cuales serán abordados mas adelante en el presenté estado del arte [22, 98, 83, 46].

Tal es el caso de [33] donde se concluye que las técnicas de regresión lineal en ambientes controlados se adaptan bien al problema de ajuste de los datos, logrando correlacionar un sensor óptico (Plantower PMS 1003/3003) para la medición PM2.5 con metodologías avaladas y equipos certificados usando métodos de referencia gravimétricos (FEM_s/ FRM_s), logrando un ajuste $R^2 = 0{,}88$ en un rango de 200 a 850 mg/m^2.

No obstante, para el trabajo realizado por [68] que dio continuidad a la investigación realizada por [33], se evaluó una red pequeña de sensores al aire libre durante 320 días. Los sensores mostraron una buena correlación con dispositivos referencia durante la temporada de invierno, obteniendo un de $R^2 = 0{,}8$. Sin embargo, para los meses de primavera y verano, épocas con temperatura y humedad altas, los sensores obtuvieron un $R^2 = 0{,}2$. Lo que significa que los métodos de regresión lineal pueden no ser lo suficientemente robustos como para describir las dinámicas de la contaminación, temperatura y humedad en un tiempo prolongado.

Resultados similares fueron obtenidos por [64] y [58] en sus trabajos, donde fueron implementadas técnicas basadas en regresiones lineales obteniendo como resultado ajustes de $R^2 = 0{,}5377$ hasta $R^2 = 0{,}7847$.

Los resultados antes mencionados han despertado el interés de organizaciones gubernamentales como la Agencia de Protección Ambiental de los EE. UU (EPA) [16], quien implementó redes de sensores de PM2.5 y O_3 en nueve ubicaciones en el sur de California para evaluar su desempeño en condiciones ambientales variables. La validación de los dispositivos se realizó implementando técnicas de regresión lineal y el uso de dispositivos referencia, evidenciando nuevamente que el modelo planteado no es suficiente para responder a las variaciones ambientales, mostrando mayor error con la variable O_3, lo cual, respaldaría lo concluido por [68].

Por tal motivo, autores como [98] han incursionado en uso de algoritmos robustos de Machine Learning (ML) y Deep Learning (DL) que permiten obtener mejores resultados en el tratamiento de datos y que se adaptan mejor a los entornos no controlados, eliminando efectos propios de la dinámica, las variaciones del medio al que se ve expuesto el sensor y la pérdida masiva de información.

Para esto [98] adopto un modelo de bosques aleatorios (RF) de aprendizaje automático para resolver problemas de regresión a partir conjuntos de datos de entrenamiento, los cuales cumplen con las recomendaciones realizadas por la EPA obteniendo errores $<5\,\%$ para el CO_2, entre el 10 y 15 $\%$ para el CO y O_3, y aproximadamente el 30 $\%$ para el NO_2. Otro caso de aplicación de técnicas de ML y DL, es el tratado por [76], en su trabajo implementó técnicas de regresiones ortogonales y redes neuronales artificiales (ANN) y perceptron multicapa (MLP) con el objetivo de alcanzar la calidad de datos DQO de la Directiva Europea de Calidad del Aire (entre 25 y 30 $\%$ de incertidumbre) para O_3 y NO_2 con sensores ubicados en exteriores de zonas rurales. Como resultado se obtuvo que el NO_2 logró un mejor ajuste entre los sensores y las mediciones de referencia utilizando técnicas de aprendizaje supervisado.

Otras de las técnicas de ML usadas para el ajuste de las WSNs de bajo costo son las metodologías *blind calibration*, *collaborative calibration* y *transfer calibration* [39, 83, 13], las cuales implementan modelos basados en Factorización de Matriz (MF).Por otro lado, los métodos de *blind calibration* se han usado para la calibración de redes de sensores móviles de medición de gases contaminantes y en la estimación de observaciones no realizadas por los sensores de la red. Tal es el caso del enfoque presentado por [12], donde se implementa un modelo de Nonnegative Matrix Factorization (NNMF), demostrando ser robusto para un gran número de entradas perdidas (del 10 $\%$ al 90 $\%$) y un gran número de encuentros entre sensores calibrados y no calibrados, superando otros enfoques de *blind calibration* basados en Mínimos Cuadrados no Negativos ,más conocido por sus siglas en inglés como NNLS, y Matrix Completion (MC), el cual es un algoritmo altamente efectivo en la exploración plena de las características inherentes de los datos en el entorno para realizar recuperación de información no muestreada.

Tal es el caso del modelo MC propuesto por [90] denominado MC-Two-Phase del cual fueron incluidos tres algoritmos: un modelo de detección de fallos de estructura basado en el Análisis de Componentes Principales (PCA), una pre-interpolación espacial y una pre-interpolación temporal. Demostrando que el esquema propuesto aprovecha el modelo MC para integrar completamente los resultados obtenidos por los algoritmos incluidos en la metodología de recuperación de datos, el cual es efectivo incluso cuando algunas filas o columnas se encuentran completamente vacías. Así mismo [90] asegura que los modelos MC trabajan bien con datos de las WSNs que miden material particulado, ya que las características de la información es alimente dispersa y tienen altas correlaciones espacio-temporales, como es el caso de datos generados por sensores de temperatura, humedad, luz y PM2.5 [35, 9, 94]. En consecuencia y gracias a las virtudes visualizadas de los modelos basados en MF, estas novedosas técnicas tienen la ventaja que los ajustes de las medidas de la red pueden hacerse sin la

necesidad explícita de entornos controlados o el uso de un instrumentos estándar [3, 46] y que adicionalmente cuentan con la característica que los modelos de MF son ampliamente implementados en la recuperación de información [95, 74].

De acuerdo con lo anterior, en la literatura los problemas de recuperación de información cuentan gran cantidad de estudios y gran cantidad de algoritmos desarrollados para el tratamiento de datos en las WSNs, estos van desde técnicas estadísticas y de ML hasta la implementación de modelos de DL [89, 93]. Entre los métodos más populares se encuentran los modelos de interpolación global como Sustitución por la Media (MS) y de interpolación local basados por ejemplo en K-Nearest Neighbors (KNN). Un ejemplo de lo anterior es el trabajo realizado [25], el cual propuso un nuevo método para imputación de datos usando una similitud de grado relacional reducido denominado RKNN, el cual se puede entender como un método KNN mejorado para estimar iterativamente valores perdidos de datos. Sin embargo, es necesario señalar que la precisión de convergencia y el número máximo de iteraciones pueden afectar al rendimiento de la imputación del modelo KNN [25], donde adicionalmente se debe tener presente que las técnicas de interpolación de datos mencionadas anteriormente no visualizan bien las correlaciones espacio-temporales entre los sensores y se ven afectados en gran medida por la presencia de ruido y características no consistentes [90].

Es por lo anterior que técnicas como Procesos Gaussianos Multicanal (MOGPs) cuentan con gran aceptación a la hora de abordar problemas de recuperación de información en las WSNs, ya que aprovechan bien las correlaciones espacio-temporales y se adaptan bien a los datos generados por los de sensores de contaminación atmosférica, dado que el comportamiento de la contaminación es altamente acoplada [44, 11, 43] y gracias a su naturaleza Bayesiana, no solo realiza predicciones de datos perdidos, si no también, predicciones de los primeros momentos estadísticos, como lo son la media y la desviación estándar, determinando así un intervalo de confianza donde se encuentran las predicciones obtenidas [63, 88].

Lo anterior es posible siempre y cuando los datos generados no cuenten con pérdida de información significativa, en consecuencia de esto y debido al rápido progreso de las representaciones dispersas, los modelos basados en factorización de matriz se posicionan como una de las técnicas que se adaptan mejor a grandes cantidades de información perdida [91, 24, 90], usados en mayor parte para la generación de recomendaciones a partir de la poca o casi nula información suministrada por los usuarios [65, 37, 32].

Por otro lado, dichos métodos demuestran gran versatilidad, ya que pueden ser fácilmente adaptables a los datos con la inclusión de características espacio-

temporales usando Embedding Layers, las cuales transforman información no cuantificable, como palabras, en entradas de vectores, siendo estos componentes clave en el desempeño de los modelos de las redes neuronales profundas. Tal es el caso del trabajo desarrollado por [26], donde aplico Embedding Layers en una amplia gama de puntos de referencia en el procesamiento del lenguaje natural, obteniendo gran equilibrio entre el costo computacional y el rendimiento para una amplia gama de arquitecturas que van desde MLP hasta LSTM y Transformers. Adicionalmente, en la literatura también se encuentran adaptaciones con otras técnicas DL como ANNs y CNNs [91, 24, 97], dando paso a variantes de la técnica denominadas Deep Matrix Factorization (DMF), las cuales buscan con el desarrollo de estos modelos alternativas para mejorar la capacidad de aprender las representaciones de características desde cero [97] y al mismo tiempo contar con la interpretación matemática de la MF [67, 18].

Por lo tanto, los métodos basados en MF se han destacado muy bien en problemas de estimación y recuperación de información faltante de una base de datos dispersa [37, 85], el cual describe el problema actual con la información generada por el proyecto *Ciudadanos Científicos*. En este sentido para el presente trabajo buscamos aprovechar las virtudes de las técnicas MF señaladas anteriormente con la finalidad de llegar enfoques como el tratado por [23], donde fue implementado un modelo de factorización de matriz no negativa (NNMF) en la reconstrucción de señales, obteniendo soluciones cuya precisión es superior a lograda utilizando polinomios. Por otro lado, cuando el algoritmo MF se complementa con técnicas de regularización, el modelo es capaz de lograr rendimientos apropiados incluso en presencia de un gran número de datos faltantes [78, 27].

1.3. Objetivos

1.3.1. Objetivo principal

Diseñar una metodología que permita mejorar la calidad de medición de material particulado PM2.5 de la red de sensores de bajo costo del proyecto Ciudadanos Científicos en el Valle de Aburrá, utilizando algoritmos de aprendizaje de máquina.

1.3.2. Objetivos específicos

- Diseñar una función de optimización que permita a partir de la red de sensores, ajustar y completar datos faltantes con el fin de complementar las

metodologías de medición.

- Proponer un método para la sintonización de los parámetros del modelo garantizando un buen ajuste de los datos sin pérdida de generalización.
- Desarrollar un procedimiento con el fin de evaluar la calidad del ajuste dado por el modelo propuesto con respecto a lecturas de referencia.

1.4. Planteamiento del problema

La contaminación atmosférica trae consecuencias a nivel local, regional y global; entre las afectaciones más preocupantes se encuentran las afectaciones a la salud por exposiciones prolongadas a altas concentraciones de contaminación. Es por eso que en Colombia existen Sistemas de Vigilancia de la Calidad de Aire (SVCA) con la finalidad de monitorear y vigilar los niveles de emisión en el territorio nacional. Para Medellín la autoridad ambiental es el Área Metropolitana que a través del SIATA, administra las redes de monitoreo y vigilancia de la calidad del aire en el Valle de Aburrá.

Por lo tanto, Ciudadanos Científicos nace como un proyecto de la ciudad de Medellín y el Valle de Aburrá que implementa una red de aproximadamente 270 sensores de bajo costo para medir material particulado PM2.5 en los municipios pertenecientes al Área Metropolitana, con el objetivo de mejorar resolución espacial y a su vez informar a la ciudadanía sobre el estado de la contaminación en tiempo real.

Sin embargo, la red es implementada con sensores de bajo costo los cuales no cuentan con normativas y estándares para el aseguramiento de la medida. Adicionalmente durante el año 2018 el 45 % de datos generados por la red son faltantes y el 5 $\%$ de datos son anómalos, lo cual puede ser debido a las condiciones energéticas y de conectividad a internet donde son ubicados los nodos de sensores (viviendas de los ciudadanos del Área Metropolitana del Valle de Aburrá). En la Figura 1.2 se representa el porcentaje de datos faltantes de cada sensor y su distribución espacial a lo largo del Valle de Aburrá, donde se visualizan 6 categorías y se utilizaron 6 colores diferentes para determinar el nivel de perdida de sensores en cada nodo dispuesto en la red: 0 %-19 %, 20 %-39 %, 40 %-59 %, 60 %-79 %, 80 %-99 % y 100 % están representados por los colores: salmón claro, tomate, rojo, marrón, marrón silla y negro, respectivamente.

En este trabajo fue propuesta una metodología para completar datos faltantes generados por la red y mejorar la calidad de la medición a partir de la implementación de algoritmos basados en MF, integrando dentro de estos, parámetros de

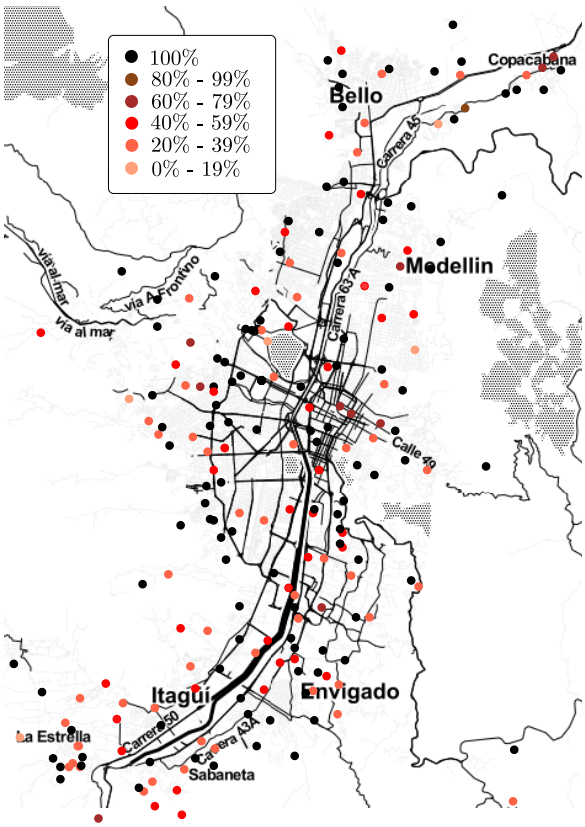


Figura 1.2: Comportamiento de los datos generados por la WSNs Ciudadanos Científicos

regularización, sesgo y características espacio-temporales que se adapten mejor al problema de datos perdidos (45 %) de la WSN de bajo costo implementada en el Valle de Aburrá.

Para esto se realizó una búsqueda de algoritmos basados en MF, donde fue evaluando el desempeño de estos siguiendo una metodología de eliminación de datos, para posteriormente comparar las estimaciones realizadas por el método y los datos reales eliminados artificialmente. Al algoritmo de mejor desempeño se le realizó una sintonización de parámetros por medio de una búsqueda de rejilla, realizando nuevamente una experimentación con los datos siguiendo la metodología de eliminación de datos planteada inicialmente y la comparación del mismo con algoritmos encontrados en la literatura para la estimación de información perdida como: MS, KNN y MOGPs. En este sentido en los experimentos realizados durante este trabajo, se busco principalmente evaluar los desempeños obtenidos por cada uno de los algoritmos principalmente para porcentajes de información pérdida del 40 % al 50 %, ya que estos representan el porcentaje

de datos faltantes generados por Ciudadanos Científicos durante el año 2018.

Finalmente fue planteada una modificación al modelo MF incluyendo redes neuronales (DMF) y el uso de variantes implementando Embedding Layers, donde se incluye información geográfica e información del estado del día y el día de la semana, de forma que, esto permitió adaptar los algoritmos al problema de datos perdidos y mejorar los resultados iniciales obtenidos con el algoritmo MF estándar.

1.5. Metodología

El presente trabajo trata de una investigación de tipo aplicada, ya que con esta se busca resolver un problema práctico sobre la calidad de los datos de una WSN de bajo costo para medir material particulado PM2.5. Por lo tanto, el objetivo de la misma será plantear una metodología por medio de la aplicación de técnicas de ML y DL, las cuales deben ser lo suficientemente robustas para solucionar problemas con gran cantidad de datos faltantes, donde sea incluida la dinámica de la red y las estimaciones sean realizadas con la información generada por los propios sensores de la red.

Por lo tanto en la Figura 1.3 se plantea el siguiente diagrama metodológico:

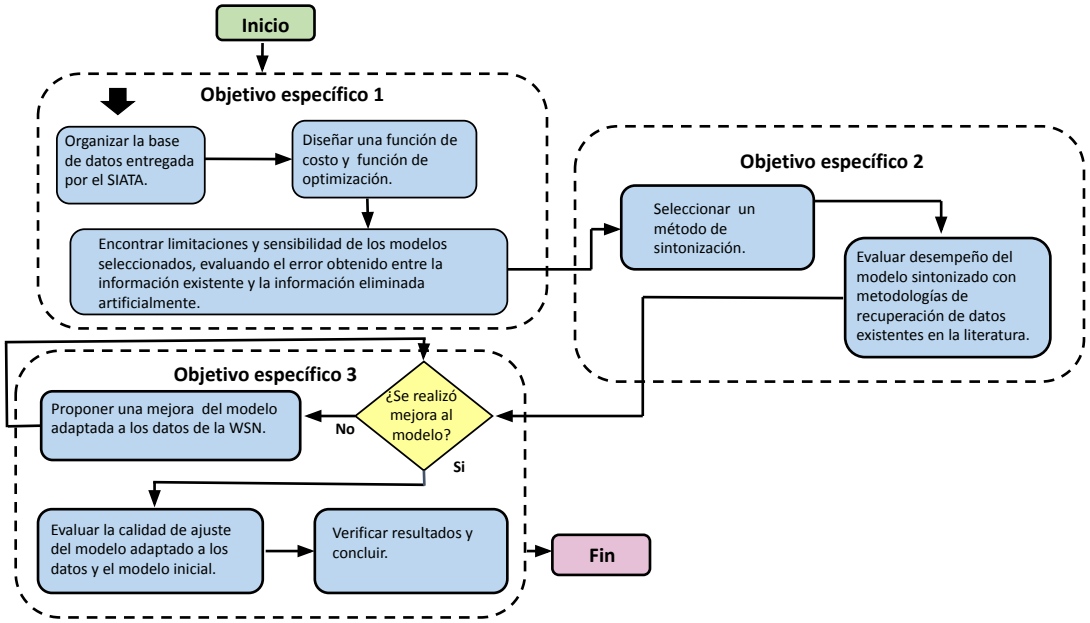


Figura 1.3: Diagrama metodológico

1.6. Resultados o contribuciones

Como fue descrito en el transcurso del documento, conocer el estado actual de la contaminación en zonas urbanas y rurales tiene como objetivo principal determinar el impacto a la salud y al medio ambiente generado por las actividades antropogénicas, lo cual permita desarrollar alternativas que van desde la prevención, la planeación de ciudad y ordenamiento territorial para mitigar dicha problemática de salud.

Para esto es necesario implementar WSNs de bajo costo que apunten a mejorar la resolución espacial en las zonas de interés y que a su vez cuenten con menores costos de operación y de implementación, ya que las redes de monitoreo actuales se caracterizan por el costo significativo de equipos y de consumo energético.

Por lo tanto, las contribuciones de este trabajo se lograron con la escritura de dos artículos científicos sobre la implementación de técnicas ML para mejorar la calidad de los datos desde el punto de vista de recuperación de información perdida. En este sentido, el primer artículo consta de una publicación de resultados del planteamiento del algoritmo MF con el desarrollo del proceso de sintonización de parámetros y la evaluación del mismo con algoritmos encontrados en la literatura MS, KNN y MOGPs (Revista IAENG categorizada como Q2). Para el segundo artículo fueron presentados los resultados obtenidos del planteamiento del modelo MF con redes neuronales DMF y la inclusión de información espaciotemporal usando Embedding Layers, obteniendo con esto mejores adaptaciones del modelo a los datos y menores errores en las predicciones generadas para la WSN Ciudadanos Científicos (En proceso de sometimiento a la revista Measurement categorizada como Q1).

Finalmente, con la finalización de este trabajo se logró seguir aportando al estado del arte en la temática de la implementación de WSNs de bajo costo, haciendo uso de técnicas computacionales basadas ML y DL, lo cual permitirá a los investigadores contar con mejor calidad de los datos y de esta manera aprovecharlos mejor en sus investigaciones.

Parte II

Desarrollo experimental.

Capítulo 2

Base de datos Ciudadanos Científicos

Contenido

Sinopsis

En este capítulo se da desarrollo al primer objetivo del proyecto, donde se inicia identificando los tipos de datos perdidos, abordando una metodología de adecuación de la base de datos y de eliminación de datos anómalos aplicando criterios estadísticos, lo que permitirá posteriormente la aplicación de modelos basados en MF y la reconstrucción de información faltante.

2.1. Base de datos

2.1.1. Nubes Ciudadanos Científicos

Ciudadanos Científicos es un sistema de monitoreo de variables de material particulado PM2.5 y PM10 desplegado por todo el Valle de Aburrá, creado como un proyecto local de ciencia, educación y tecnología, el cual permite a los ciudadanos ser partícipes dentro de una red urbana de calidad del aire de carácter informativo y de levantamiento de información con fines académicos y de uso colectivo o personal [75].

Actualmente cada nodo de la red (Figura 2.1) se encuentra conformado por un sensor de polvo PMS1003; un sensor de temperatura y humedad SHT1X; un dispositivo GPS y una Raspberry Pi encargada de dar conectividad y despliegue de datos a una base de datos y a un geovisor, el cual, se puede visualizar a través de la página[1] o descargando la app de Ciudadanos Científicos. Por lo tanto, los datos son tomados por el nodo y desplegados a la base de datos cada minuto, publicando información actualizada cada media hora y que puede ser descargada a través del sitio web[2] [75].

(a) Red distribuida por el Valle de Aburrá (b) Nodo ciudadanos científicos

Figura 2.1: Distribución de la red por la ciudad de Medellín y el resto del Área Metropolitana del Valle de Aburrá

Por otro lado, en la actualidad la red cuenta con 270 estaciones desplegadas por todo el Valle de Aburrá, abarcando los municipios de Medellín, Envigado, Sabaneta, Bello, entre otros; siendo esta una de las redes de sensores de bajo costo más grandes implementadas en la región.

[1] www.siata.gov.co.
[2] http://datosabiertos.metropol.gov.co.

Otro aspecto importante de la WSN Ciudadanos Científicos es la forma como se encuentran organizados los datos, por lo que fue necesario realizar una depuración de los mismos, dado que los datos obtenidos durante el año 2018 se encontraban conformados por archivos individuales para cada nodo, con información propia de humedad, temperatura, PM2.5, PM10 y datos de coordenadas geográficas.

En este sentido entonces, el primer procedimiento para depuración de la infracción consistió en agrupar en una matriz los datos de cada nodo correspondientes a la variable PM2.5 (fila) y su respectiva medición durante el año (columna). Finalmente la información fue organizando por hora según los establecido en el protocolo para el monitoreo de la calidad del aire en Colombia [51] (ver Figura 2.2).

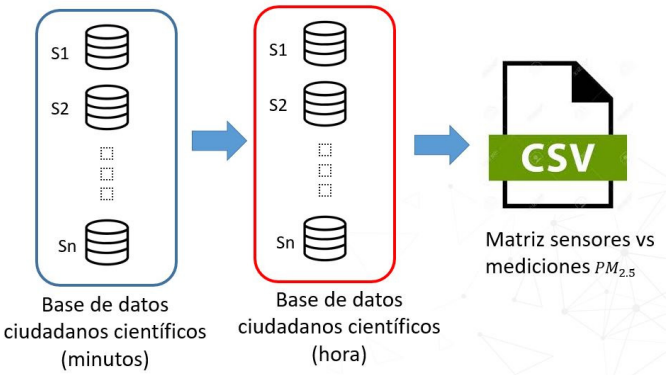


Figura 2.2: Procedimiento inicial para la depuración de la base de datos [51]

2.1.2. Índice de calidad del aire

De acuerdo con el protocolo para el monitoreo y seguimiento de la calidad del aire [51], el índice de calidad de aire (ICA) o en inglés AQI, permite comparar los niveles de contaminación del aire a los que se ven expuestos los territorios. Dicho índice es usado por Ciudadanos Científicos y los SVCA dispuestos en Colombia.

Este índice corresponde a una escala numérica identificada por colores, relacionado dicha escala con los efectos de la contaminación en la salud. Para esto son tenidos en cuenta los contaminantes que son monitoreados y las características de los combustibles que se distribuyen en todo el país, por lo que es asignado un valor adimensional que varía entre 0 y 500 y que se basa en información reportada por la US-EPA [82].

Por lo tanto, para la depuración y limpieza de la base de datos (datos anóma-

los) se usó el ICA, con la finalidad de lograr la mejor calidad de los datos posible y la menor inclusión de sesgo en los mismos. En la Figura 2.3 [3] se presentan los rangos cualitativos del rango de la escala del ICA usados para evaluar la calidad del aire en la ciudad y el Área Metropolitana.

ATRIBUTOS DEL IBOCA			D RANGOS DE CONCENTRACIÓN Y TIEMPO DE EXPOSICIÓN PARA CADA CONTAMINANTE					
A Rangos numéricos	B Estado de calidad del aire	C Estado de actuación y respuesta	PM10, 24h (µg/m3)	PM2.5, 24h (µg/m3)	0_3, 8h (µg/m3)	CO, 8h (µg/m3)	SO_2, 1h (µg/m3)	NO_2, 1h (µg/m3)
0-10	FAVORABLE	Prevención	(0 -54)	(0 -12)	(0 -116) [0-59]	(0 -5038) [0.0-4.4]	(0 -93) [0-35]	(0 -100)
10,1 - 20	MODERADA	Prevención	(55 -154)	(12.1 -35.4)	(117 -148) [60-75]	(5039 -10762) [4.5-9.4]	(94 -198) [36-75]	(101 -188)
20,1 - 30	REGULAR	Alerta Amarilla	(155 -254)	(35.5 -55.4)	(149 -187) [76-95]	(10763 -14197) [9.5-12.4]	(199 -486) [76-185]	(189 -67) [101-360]
30,1 - 40	MALA	Alerta Naranja	(255 -354)	(55.5 -150.4)	(188 -226) [96-115]	(14198 -17631) [12.5-15.4]	(487 -797) [186-304]	(678 -1221) [361-649]
40,1 - 60	MUY MALA	Alerta Roja	(355 -424)	(150.5 -250.4)	(227 -734) [116-374]	(17632 -34805) [15.5-30.4]	(798 -1538) [305-604]	(1221 -2349) [650-12491]
60,1 - 100	PELIGROSA	Emergencia	(425 -604)	(250.5 -500.4)	(734 -938) [374-938]	(34806 -57703) [30.5-50.4]	(1584 -2630) [605-1004]	(2350 -3853) (1250-2049]

Figura 2.3: Índice de calidad de aire según [51] y [82]

2.1.3. Patrones de datos faltantes

La imputación de datos en el área del análisis de datos es una práctica recurrente ya que de su correcta implementación asegura el mejor desempeño de los algoritmos implementados, y por consiguiente los resultados obtenidos a partir de estos [66].

En un trabajo de investigación es probable que encuentren diferentes razones para la pérdida de información, como es el caso de una red IoT comunitaria que depende de diferentes factores como: deterioro natural de los sensores instalados, conectividad a internet, continuidad del servicio energético, ubicación geográfica, robo de estaciones, entre otros. Por lo tanto, la forma y los patrones que estos conforman son realmente útiles para tener una comprensión básica de los datos y como estos deben ser manejados [66]. Los patrones de datos faltantes dentro de una red IoT pueden ser:

- Patrón MCAR (Missing Completely at Random): en la práctica este patrón de datos faltantes es muy poco probable que se de. Lo anterior se debe a que la existencia o ausencia del dato es completamente independiente a

[3] Imagen tomada de https://gobernanzadelaire.uniandes.edu.co/.

los parámetros de interés y las observaciones [66]. En el caso de la base de datos Ciudadanos Científicos, es como si los datos perdidos se dieran de forma consistente por el fallo de un grupo de sensores que durante el año dejaron de medir en un tiempo indeterminado. Lo anterior podría ser posible para el caso de la red, ya que varios sensores pueden fallar por pérdida de conexión o de energía durante varios instantes en el año. Sin embargo, no concuerda con instantes de saturación de la red, condiciones climáticas o ruidos inducidos en el sensor.

- Patrón MAR (Missing at Random): Dicho patrón implica que los datos faltantes se dan de forma aleatoria y no dependen del patrón de comportamiento de los registros generados sin información. En este caso se puede decir que la ausencia de datos adquiere un comportamiento similar al obtenido de lazar una moneda justa [66]. Para el caso de la red Ciudadanos Científicos es poco probable obtener dicho comportamiento en los datos, dado que en la Figura 2.6 las mediciones obtenidas durante todo el año 2018 no presentan un comportamiento aleatorio.

- Patrón MNAR (Missing Not at Random): Cuando los datos no son ni MCAR ni MAR, se dice que el patrón de datos perdidos es MNAR [66]. En la práctica es común que se presenten situaciones en que los datos faltantes no siguen un patrón completamente aleatorio (MCAR) y tampoco aleatorio (MAR). En este caso supondremos que para el caso de la red de Ciudadanos Científicos se cumple que el patrón de datos faltantes es MNAR, debido a que ninguno de los anteriores patrones se acomoda completamente al comportamiento de los datos faltantes generados por la red en el año 2018.

2.1.4. Detección de datos anómalos

Los datos anómalos o atípicos son un problema presente en la recolección de datos de sistemas IoT [31], como es el caso de la WSNs Ciudadanos Científicos, donde dichos datos son provenientes de observaciones que se desvían del resto de conjunto de datos [53, 44]. Por lo tanto, estas desviaciones se pueden dar debido a ruidos electromagnéticos, inconsistencia energética o saturación del sensor causada por exposición excesiva de contaminación que pueden afectar los resultados al aplicar modelos estadísticos o modelos de aprendizaje de máquina en los datos [21].

Por lo tanto, es importante la detección y depuración de estos valores, ya sea para eliminarlos o asumir que la medición no ocurrió y posteriormente estimar

su valor por medio de un algoritmo de recuperación de datos faltantes [21]. El objetivo finalmente es atenuar los efectos de los datos atípicos, garantizando introducir el mínimo sesgo posible en futuros análisis de la información generada.

En este caso, el análisis del cuarto momento estadístico o Kurtosis es un método eficiente para detección de valores anómalos, independientemente de la proporción de éstos y la presencia de correlación entre las variables consideradas en los datos. La Kurtosis entonces, da una idea sobre el peso de las colas de distribución, lo que en consecuencia determina que tan frecuentes son las desviaciones extremas sobre el valor de la media [53].

Otro análisis interesante para determinar si existen datos atípicos es encontrar el tercer momento estadístico o asimetría. En otras palabras, en una distribución de datos la asimetría tendrá un valor de cero si la media es cero y la desviación estándar se hace uno (distribución normal). Para valores negativos de asimetría indicará que los datos están sesgados hacia la izquierda, y para valores positivos de asimetría estos indicarán datos sesgados hacia la derecha como es el caso de las observaciones generadas por Ciudadanos Científicos durante el año 2018. Lo anterior tiene sentido, ya que la media de contaminación se encuentra alrededor de 13 $\mu g/m^3$ y posee una cola con valores físicamente imposibles para el Valle de Aburrá o que no son muy frecuentes durante el año (ver Figura 2.4) [82].

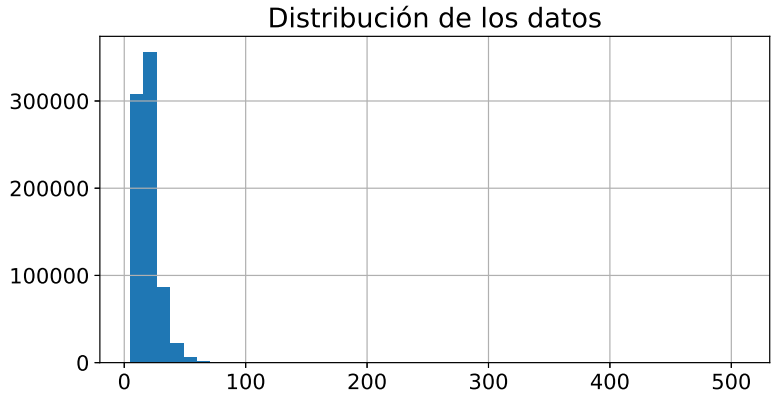


Figura 2.4: Distribución de los datos no depurados (atípicos) con $k_3 = 48{,}15$ y skewness$= 3{,}35$

Por lo tanto, al realizar el análisis de Kurtosis en los datos generados por la WSN se obtiene que estos cuentan con valores de hasta 500 $\mu g/m^3$ de material particulado PM2.5 (Figura 2.4), indicando que estos valores se encuentran por fuera de la media de los datos y una asimetría (skewness) positiva, da a entender

que la cola de datos anómalos se encuentran al lado derecho de la agrupación. Este resultado muestra que algunos datos sobrepasan por mucho los valores presentados en Medellín en épocas de contingencia, llegando a valores que corresponderían a una alerta morada según el índice AQI (ver Figura 2.3). Lo anterior hace necesario realizar limpieza de datos que no representan sentido alguno con los niveles de contaminación generados en Valle de Aburrá, evitando con esto, inducir al mínimo ruidos y obtener resultados deficientes en el momento de implementar los algoritmos para la recuperación de la información perdida o no generada.

2.1.5. Depuración de datos anómalos

Se procedió con la eliminación de datos atípicos de acuerdo con lo observado en la Figura 2.4 y el uso del criterio de sensibilidad donde se encuentran la mayor ocurrencia de las observaciones con mediciones acordes a los datos generados por la red [82].

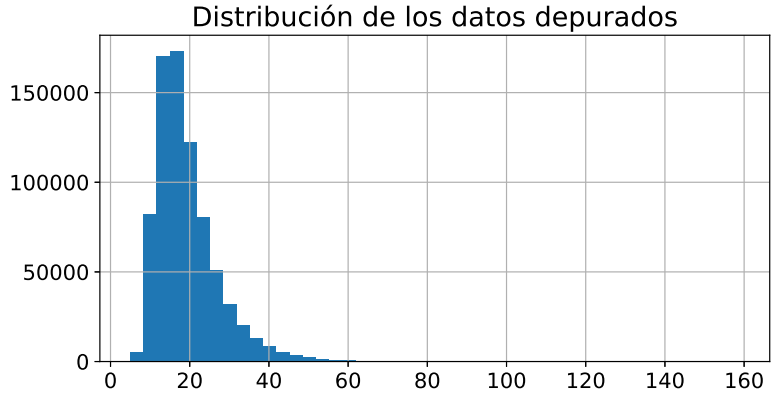


Figura 2.5: Distribución de los datos depurados con $k_3 = 7{,}26$ y skewness$= 1{,}89$

Por lo tanto, se realizó la eliminación de datos atípicos aplicando el criterio por percentiles, los cuales dan una medida de posición que indica una vez organizados los datos de mayor a menor, el valor de la medición de material particulado PM2.5 que se encuentra por debajo de cierto porcentaje dado en las observaciones de un grupo de datos.

En el caso de la Figura 2.4 se observa que los datos atípicos aunque representan valores muy por encima de los que sería una medida de calidad de aire en el Valle de Aburrá, estos representan un porcentaje muy bajo con respecto a las observaciones obtenidas por la red durante el año 2018, por lo que un criterio

de p_{99} resulto suficiente para eliminar los datos atípicos. En el procedimiento de depuración se asumió la no medición de los valores atípicos y se incluyeron en la base de datos con valores NaN (ver Figura 2.5)

Finalmente, se eliminaron 140 nodos de sensores que no presentaron mediciones durante el año (ver Figura 1.2) dado que estos no aportan información valiosa a la red y generan ruido innecesario. Sin embargo en la Figura 2.6, se puede verificar que la red conformada por 130 sensores cuenta con gran cantidad de información perdida (espacios en negro), lo cual presenta un reto significativo para los algoritmos evaluados en los siguientes capítulos.

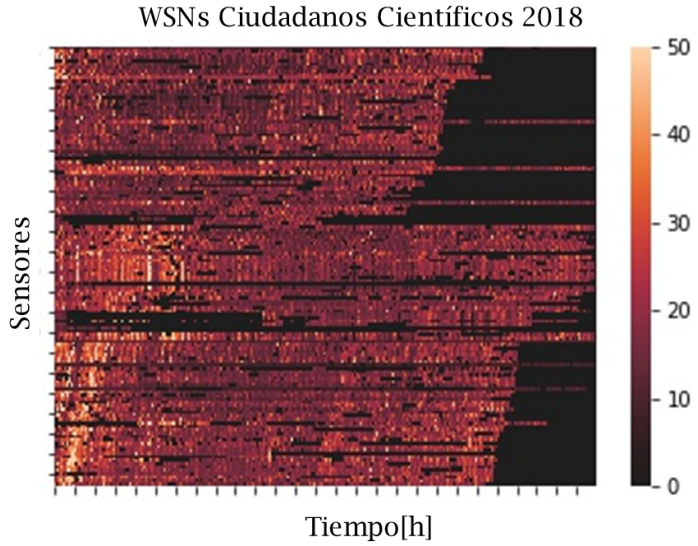


Figura 2.6: Mapa de calor de la red con información depurada de acuerdo con [51]

Capítulo 3

Planteamiento de una función de optimización para recuperar datos perdidos de una WSN

Contenido

Sinopsis

El desarrollo de este capítulo busca plantear una función de optimización basada en MF para la estimación de datos perdidos y el planteamiento de variaciones de la misma para evaluar el método que mejor se adapte a los datos de la WSN incluyendo parámetros como: α (paso de entrenamiento de del SGD), λ (parámetro de regularización) y k (espacio de factor latente del conjunto de dimensionalidad) inicialmente usados en la literatura. Para esto fue tomada una ventana de 24 horas donde se eliminó artificialmente y de forma aleatoria información del 10 $\%$ al 90 $\%$.

Las métricas de desempeño y evaluación de error RMSE, EVS y MAPE fueron implementadas para evaluar los resultados obtenidos por las funciones de optimización desarrolladas ante diferentes porcentajes de información eliminada. Adicionalmente se realizó la evaluación de entrenamiento obtenido con la implementación de algoritmos de optimización Momentum, RMSprop y SGD, los cuales son ampliamente usados en el estado del arte para el entrenamiento de algoritmos de aprendizaje de máquina.

3.1. Índices para la evaluación de rendimiento

3.1.1. RMSE

El RMSE o Error Cuadrático Medio es un índice de desempeño de algoritmos de ML más implementadas en la literatura. Esta métrica mide la cantidad de error existente entre un conjunto de datos. Por lo tanto, la métrica entrega la distancia o desviación del error medido entre un valor estimado y un valor observado [72, 77]. La expresión para el cálculo del índice RMSE se define en la Eq.(3.1) y se presenta a continuación:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(m_{st} - \widehat{m}_{st})^2}{n}} \tag{3.1}$$

donde, n es el número de observaciones; m_{st} representa las observaciones; y $\widehat{m}_{st}$ la estimación obtenida por el algoritmo de aprendizaje ML implementado.

3.1.2. MAPE

El Error Porcentual Absoluto Medio es otra métrica de desempeño bastante implementada para evaluar el comportamiento de los algoritmos de ML. Este indicador mide el tamaño del error en magnitudes porcentuales, haciéndolo un indicador fácil de interpretar, con respecto al RMSE [72, 77]. La expresión para el cálculo del índice MAPE se presenta a continuación en la Eq.(3.2):

$$MAPE = \frac{\sum_{i=1}^{n} \frac{(m_{st} - \widehat{m}_{st})}{m_{st}}}{n} \tag{3.2}$$

3.1.3. EVS

La variación explicada o en inglés Explained Variance Score (EVS) da una idea de la variación o la dispersión obtenida por cada modelo para cada una de las series de tiempo. En otras palabras, entrega un proporción del ajuste dado por cada modelo matemático reflejando la calidad de la regresión (entre más alto mejor se ajusta el modelo). A la proporción de varianza explicada se le llama coeficiente de determinación y es igual al coeficiente de correlación R^2, donde la mejor puntuación posible es 1.0 cuando el ajuste del modelo es exacto [59].

3.2. Factorización de matriz (MF)

3.2.1. Algoritmo MF

De acuerdo con el estado del arte, el algoritmo MF se utiliza ampliamente en la estimación de datos faltantes [78, 37]. Por lo tanto, la aplicación del método en el problema de datos perdidos por WSNs Ciudadanos Científicos, se basa entonces en generar un modelo que mapee tanto los sensores de material particulado, como las medidas por hora a un espacio de factor latente del conjunto de dimensionalidad k. De tal manera que las interacciones entre sensores y medidas se modelen como productos internos en ese espacio.

En este sentido, cada medida en un tiempo t se asocia con un vector $\mathbf{q}_t \in \Re^k$, donde los elementos de $\mathbf{q}_t$ representan las medidas que son obtenidas por todos los sensores para el instante de tiempo t, y cada sensor s está asociado con un vector $\mathbf{p}_s \in \Re^k$, en el cual los elementos de $\mathbf{p}_s$ representan todas las medidas que se obtienen del sensor m durante el período de tiempo analizado.

Los productos punto resultantes $\mathbf{p}_s^T\mathbf{q}_t$, capturan la interacción entre los sensores para diferentes momentos de tiempo. Como resultado es posible obtener la estimación de una medida $\widehat{m}_{st}$ de un sensor $\mathbf{p}_s$ por $\mathbf{q}_t$, si calculamos el producto de punto de sus vectores:

$$\widehat{m}_{st} = \mathbf{p}_s^T\mathbf{q}_t = \sum_{i=1}^{k} p_{si} \cdot q_{it} \tag{3.3}$$

El desafío consiste en encontrar el mapeo de cada medida y cada sensor en los vectores factoriales $\mathbf{p}_s$ y $\mathbf{q}_t$. Para este propósito, inicialmente los vectores factoriales se asignan al azar, y luego los errores cuadrados e_{st}^2 entre las medidas estimadas $\widehat{m}_{st}$ y las medidas verdaderas m_{st} se calculan de acuerdo con la Eq.(3.4), tratando de reducir el error en cada iteración.

$$e_{st}^2 = (m_{st} - \widehat{m}_{st})^2 = \left(m_{st} - \sum_{i=1}^{k} p_{si} \cdot q_{it} \right)^2 \tag{3.4}$$

Por lo tanto, la función de error al cuadrado que se muestra en la Eq.(3.4) puede minimizarse utilizando sólo el conjunto de medidas conocidas. Como resultado, podemos reescribir la función de costo como en la ecuación Eq.(3.5)

$$\min_{\mathbf{p}_s, \mathbf{q}_t} \left(m_{st} - \sum_{i=1}^{k} p_{si} \cdot q_{it} \right)^2 \tag{3.5}$$

Para minimizar la Eq.(3.5) seguimos el algoritmo de descenso de gradiente estocástico SGD. La idea es saber en qué dirección deben modificarse los valores de $\mathbf{p}_s$ y $\mathbf{q}_t$. Para ello, debemos calcular el gradiente de cada uno de estos valores, sin embargo los valores desconocidos se tratan como faltantes, lo que lleva a una función objetivo dispersa. En otras palabras, la Eq.(3.5) debe diferenciarse con respecto a $\mathbf{p}_s$ y $\mathbf{q}_t$ por separado, teniendo en cuenta sólo los valores conocidos.

$$\frac{\partial f}{\partial \mathbf{p}_s} = -2e_{st}\mathbf{q}_t \tag{3.6}$$

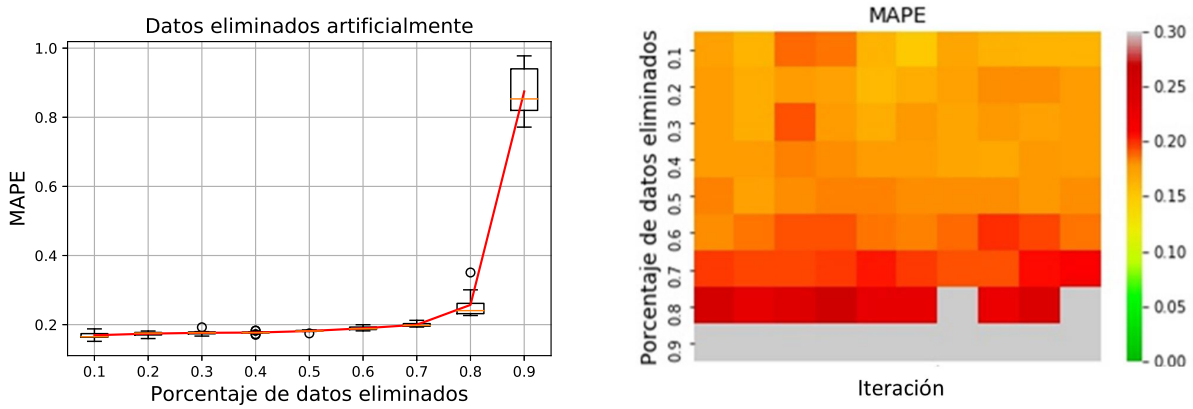
$$\frac{\partial f}{\partial \mathbf{q}_t} = -2e_{st}\mathbf{p}_s \tag{3.7}$$

Las reglas para encontrar nuevos valores de $\mathbf{p}_s$ y $\mathbf{q}_t$, que de ahora en adelante se denotarán con $\rightarrow$ en la actualización de los parámetros. Entonces, la

forma de calcular los valores de $\overrightarrow{\mathbf{p}_s}$ y $\overrightarrow{\mathbf{q}_t}$, pueden obtenerse a partir de las ecuaciones de gradiente dadas en Eqs.(3.27-3.28). Las actualizaciones de muestran a continuación:

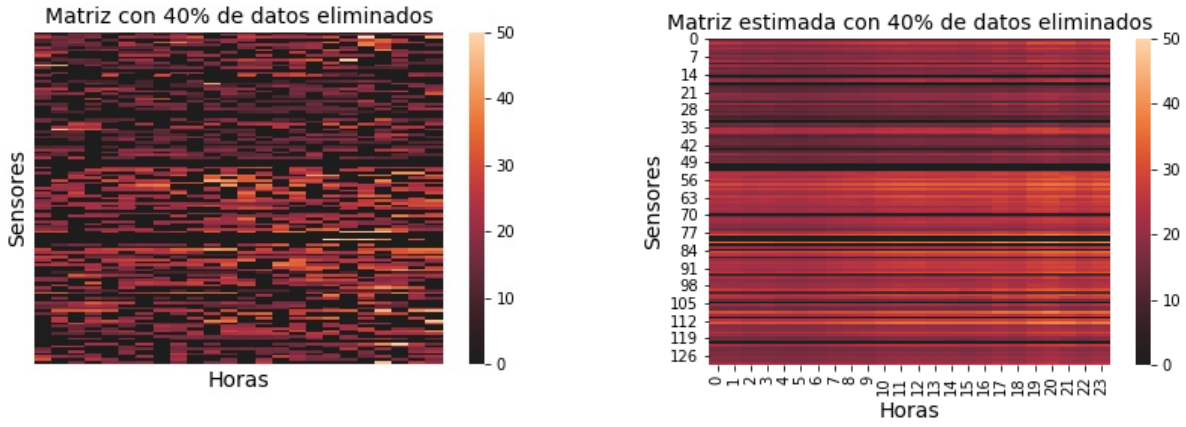
$$\overrightarrow{\mathbf{p}_s} = \mathbf{p}_s - 2\alpha e_{st}\mathbf{q}_t \tag{3.8}$$

$$\overrightarrow{\mathbf{q}_t} = \mathbf{q}_t - 2\alpha e_{st}\mathbf{p}_s \tag{3.9}$$



(a) (b)

Figura 3.1: Índice MAPE para diferentes porcentajes de datos eliminados



(a) Datos eliminados (b) Datos estimados por MF

Figura 3.2: Ventana de tiempo de 24 horas

El planteamiento anterior fue aplicado a una matriz de 24 horas (sin información perdida) correspondiente al 15 de febrero del 2018, de la cual se eliminaron porcentajes de datos iniciando del 10 $\%$ hasta completar el 90 $\%$ de datos eliminados. Este procedimiento se realizó con la finalidad de evaluar por medio del índice de desempeño MAPE la sensibilidad del algoritmo ante porcentajes bajos y altos de información perdida. Los parámetros usados en el entrenamiento del algoritmo son: $k = 4$,seleccionado de forma libre; $\alpha = 0{,}01$, seleccionado de acuerdo con [37] y 100 iteraciones, también seleccionadas de forma libre. En la Figura 3.1a y en la Figura 3.2 se presentan los resultados obtenidos por el algoritmo MF propuesto, mostrando el error MAPE obtenido para cada porcentaje de información eliminada artificialmente y su respectiva reconstrucción de la ventana de 24 horas analizada para el 40 % de información faltante, debido a que esta es la información perdida promedio presentada por la red durante el año 2018.

3.2.2. Algoritmo MF + Regularización

Dado que el objetivo es generalizar las medidas obtenidas de manera que se recuperen los datos que faltan, es esencial evitar el sobre entrenamiento y al mismo tiempo reducir al mínimo el error. Ruslan y Andriy [52] propusieron una regularización Eq.(3.10) de los parámetros de aprendizaje basada en fundamentos probabilísticos.

$$f(\mathbf{p}_s, \mathbf{q}_t) = (m_{st} - \mathbf{p}_s^T \mathbf{q}_t)^2 + \lambda(||\mathbf{p}_s||^2 + ||\mathbf{q}_t||^2) \tag{3.10}$$

Siendo λ el parámetro de regularización que se utiliza para controlar las magnitudes de los vectores factoriales $\mathbf{p}_s$ y $\mathbf{q}_t$, los cuales den una aproximación apropiada de m_{st}. Por lo tanto, la función de error cuadrado que se muestra en la Eq.(3.4), puede minimizarse utilizando la función de costo determinada en la Eq.(3.11).

$$\min_{\mathbf{p}_s, \mathbf{q}_t, b_s, b_t} (m_{st} - \mathbf{p}_s^T \mathbf{q}_t)^2 + \lambda(||\mathbf{p}_s||^2 + ||\mathbf{q}_t||^2) \tag{3.11}$$

Aplicando el algoritmo de descenso del gradiente SGD en la Eq.(3.11). Se obtiene el siguiente resultado:

$$\frac{\partial f}{\partial \mathbf{p}_s} = -2(m_{st} - \mathbf{p}_s^T \mathbf{q}_t)\mathbf{q}_t + 2\lambda \mathbf{p}_s \tag{3.12}$$

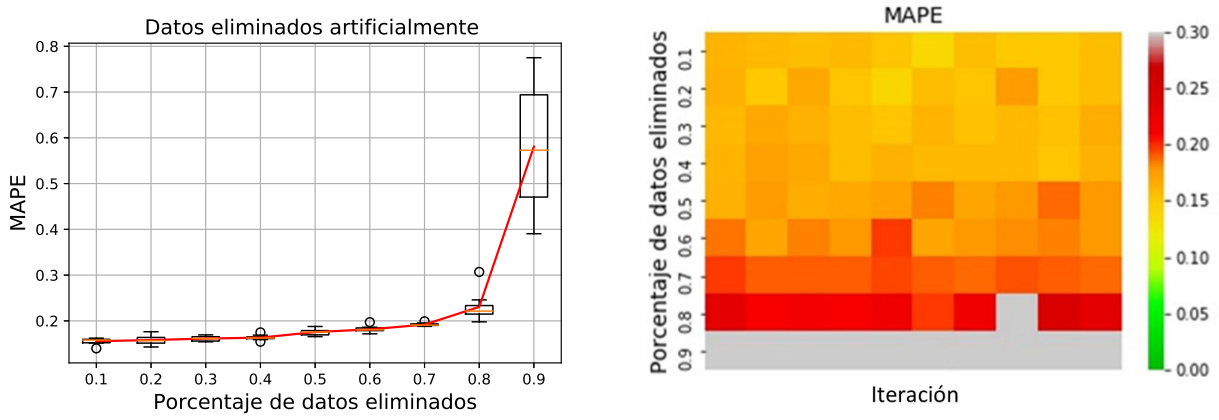
$$\frac{\partial f}{\partial \mathbf{q}_t} = -2(m_{st} - \mathbf{p}_s^T \mathbf{q}_t)\mathbf{p}_s + 2\lambda \mathbf{q}_t \tag{3.13}$$

Eso quiere decir entonces, que las reglas para los nuevos valores $\mathbf{p}_s$ y $\mathbf{q}_t$ producto del entrenamiento serán:

$$\overrightarrow{\mathbf{p}_s} = \mathbf{p}_s - 2\alpha(\lambda\mathbf{p}_s - (m_{st} - \mathbf{p}_s^T\mathbf{q}_t)\mathbf{q}_t) \tag{3.14}$$

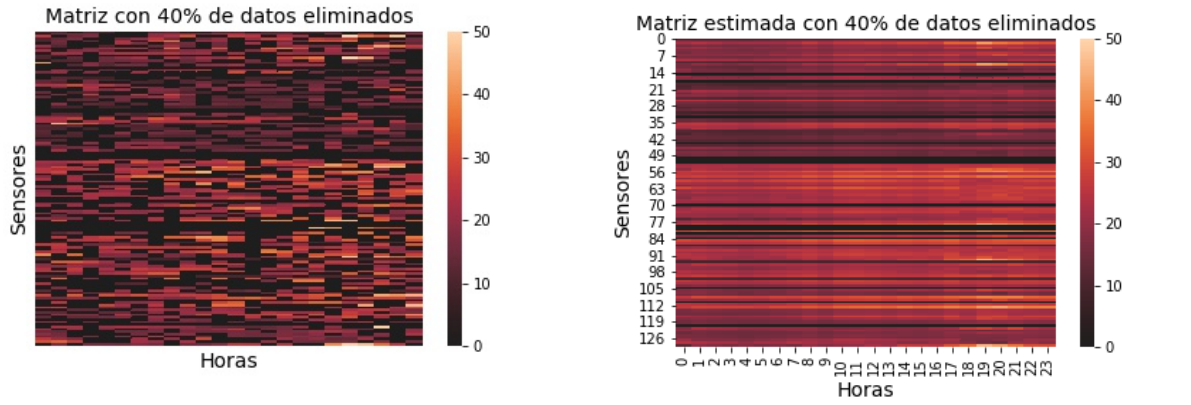
$$\overrightarrow{\mathbf{q}_t} = \mathbf{q}_t - 2\alpha(\lambda\mathbf{q}_t - (m_{st} - \mathbf{p}_s^T\mathbf{q}_t)\mathbf{p}_s) \tag{3.15}$$

El planteamiento anterior fue aplicado a una matriz de 24 horas (sin información perdida) correspondiente al 15 de febrero del 2018, de la cual se eliminaron porcentajes de datos iniciando del 10 % hasta completar el 90 % de datos eliminados. Este procedimiento se realizó con la finalidad de evaluar por medio del índice de desempeño MAPE la sensibilidad del algoritmo ante porcentajes bajos y altos de información perdida. Los parámetros usados en el entrenamiento del algoritmo son: $k = 4$, seleccionado de forma libre; $\alpha = 0{,}01$ y $\lambda = 0{,}1$, seleccionados de acuerdo con [37]; y 100 iteraciones, también seleccionadas de forma libre. En la Figura 3.3 y en la Figura 3.4 se presentan los resultados obtenidos por el algoritmo MF + Regularización, mostrando el error MAPE obtenido para cada porcentaje de información eliminada artificialmente y su respectiva reconstrucción de la ventana de 24 horas analizada para el 40 % de información perdida, debido a que esta es la información faltante promedio presentada por la red durante el año 2018.



(a) (b)

Figura 3.3: Índice MAPE para diferentes porcentajes de datos eliminados



(a) Datos eliminados (b) Datos estimados por MF + Regularización

Figura 3.4: Ventana de tiempo de 24 horas

3.2.3. Algoritmo MF + Bias

De acuerdo con [36], el desempeño del algoritmo puede mejorar si es añadido un nuevo parámetro de sesgo. Lo cual le permite al método modelar las variaciones en las medidas causadas por un sensor específico o por ciertos momentos de tiempo donde se dio una observación del sensor. En este sentido, el nuevo parámetro se puede adicionar a la función de costo como se muestra en Eq.(3.16)

$$f(\mathbf{p}_s, \mathbf{q}_t, b_s, b_t) = (m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t)^2 \tag{3.16}$$

Donde μ denota el promedio general de las observaciones, mientras que los parámetros b_s y b_t corresponden al error de compensación de los sensores y a las derivaciones observadas a lo largo del tiempo respectivamente. Eso quiere decir entonces que la función de optimización estará dada en Eq.(3.17).

$$\min_{\mathbf{p}_s, \mathbf{q}_t, b_s, b_t} (m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t)^2 \tag{3.17}$$

Por lo tanto, para minimizar la Eq.(3.17) se implementa el algoritmo SGD, para determinar en qué dirección deben modificarse los valores de $\mathbf{p}_s$, $\mathbf{q}_t$, b_s y b_t. Por lo tanto, el propósito es calcular el gradiente en cada uno de estos parámetros.

$$\frac{\partial f}{\partial \mathbf{p}_s} = -2(m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t)\mathbf{q}_t \tag{3.18}$$

$$\frac{\partial f}{\partial \mathbf{q}_t} = -2(m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t)\mathbf{p}_s \tag{3.19}$$

$$\frac{\partial f}{\partial b_s} = -2(m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t) \tag{3.20}$$

$$\frac{\partial f}{\partial b_t} = -2(m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t) \tag{3.21}$$

Entonces la reglas de actualización para los nuevos parámetros están determinadas por:

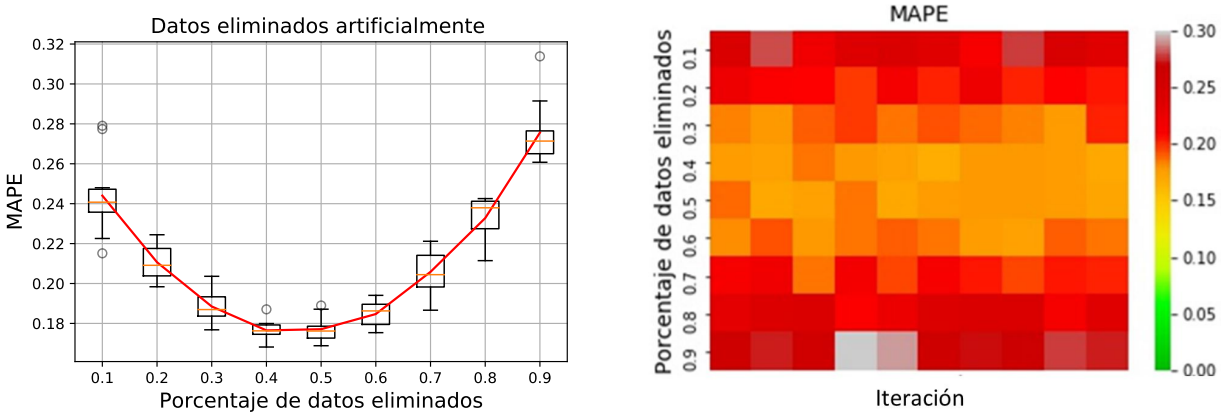
$$\overrightarrow{\mathbf{p}_s} = \mathbf{p}_s - 2\alpha(-(m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t)\mathbf{q}_t) \tag{3.22}$$

$$\overrightarrow{\mathbf{q}_t} = \mathbf{q}_t - 2\alpha(-(m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t)\mathbf{p}_s) \tag{3.23}$$

$$\overrightarrow{\mathbf{b}_s} = b_s - 2\alpha(-m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t) \tag{3.24}$$

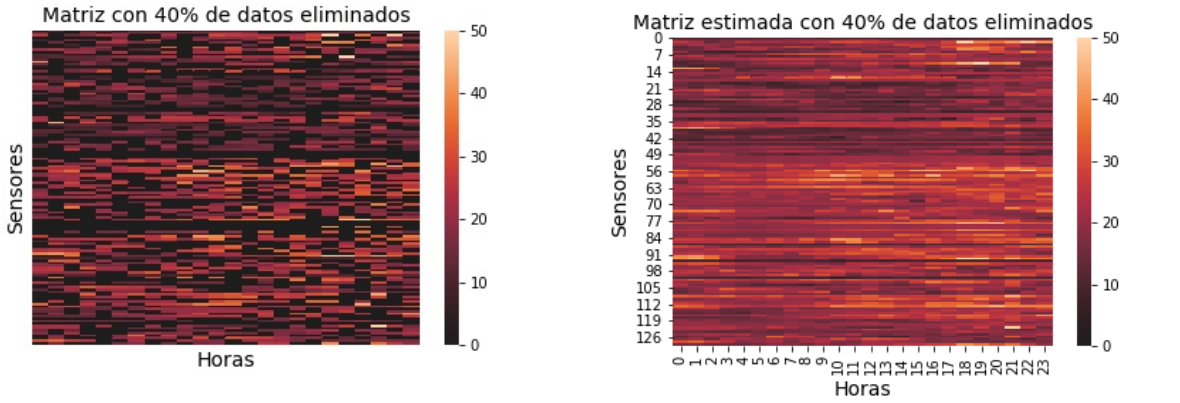
$$\overrightarrow{\mathbf{b}_t} = b_t - 2\alpha(-m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t) \tag{3.25}$$

Continuando con la experimentación para matriz de 24 horas (sin información perdida) correspondiente al 15 de febrero del 2018, de la cual se eliminaron porcentajes de datos iniciando del 10 $\%$ hasta completar el 90 $\%$ de datos eliminados. Este procedimiento se realizó con la finalidad de evaluar por medio del índice de desempeño MAPE la sensibilidad del algoritmo ante porcentajes bajos y altos de información perdida. Los parámetros usados en el entrenamiento del algoritmo son: $k = 4$, seleccionado de forma libre; $\alpha = 0{,}01$, seleccionado de acuerdo con [37]; y 100 iteraciones, también seleccionadas de forma libre. En la Figura 3.5 y en la Figura 3.6 se presentan los resultados obtenidos por el algoritmo MF + Bias, mostrando el error MAPE obtenido para cada porcentaje de información eliminada artificialmente y su respectiva reconstrucción de la ventana de 24 horas analizada para el 40 % de información perdida, debido a que esta es la información faltante promedio presentada por la red durante el año 2018.



(a) (b)

Figura 3.5: Índice MAPE para diferentes porcentajes de datos eliminados



(a) Datos eliminados (b) Datos faltantes estimados por MF + Bias

Figura 3.6: Ventana de tiempo de 24 horas

3.2.4. Algoritmo MF + Regularización + Bias

Para finalizar el diseño de la función de costo del algoritmo MF se opta por realizar una última modificación, combinando los aportes que pueden generar al entrenamiento los parámetros de regularización y sesgo. Esto quiere decir entonces, que la función de costo a minimizar se puede plantear como se muestra a continuación:

$$\min_{\mathbf{p}_s,\mathbf{q}_t,b_s,b_t} (m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T\mathbf{q}_t)^2 + \lambda(||\mathbf{p}_s||^2 + ||\mathbf{q}_t||^2 + b_s^2 + b_t^2) \tag{3.26}$$

Por lo tanto el cálculo de los gradientes para cada uno se los parámetros será:

$$\frac{\partial f}{\partial \mathbf{p}_s} = -2(m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t)\mathbf{q}_t + 2\lambda \mathbf{p}_s \tag{3.27}$$

$$\frac{\partial f}{\partial \mathbf{q}_t} = -2(m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t)\mathbf{p}_s + 2\lambda \mathbf{q}_t \tag{3.28}$$

$$\frac{\partial f}{\partial b_s} = -2(m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t) + 2\lambda b_s \tag{3.29}$$

$$\frac{\partial f}{\partial b_t} = -2(m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t) + 2\lambda b_t \tag{3.30}$$

Entonces las reglas de actualización para cada parámetro de la función de costo son:

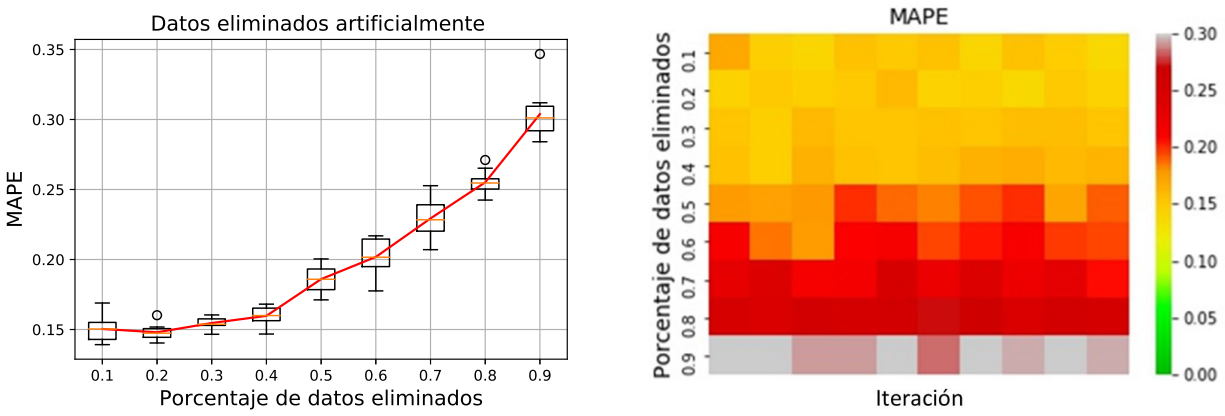
$$\overrightarrow{\mathbf{p}_s} = \mathbf{p}_s - 2\alpha(\lambda \mathbf{p}_s - (m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t)\mathbf{q}_t) \tag{3.31}$$

$$\overrightarrow{\mathbf{q}_t} = \mathbf{q}_t - 2\alpha(\lambda \mathbf{q}_t - (m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t)\mathbf{p}_s) \tag{3.32}$$

$$\overrightarrow{\mathbf{b}_s} = b_s - 2\alpha(\lambda b_s - m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t) \tag{3.33}$$

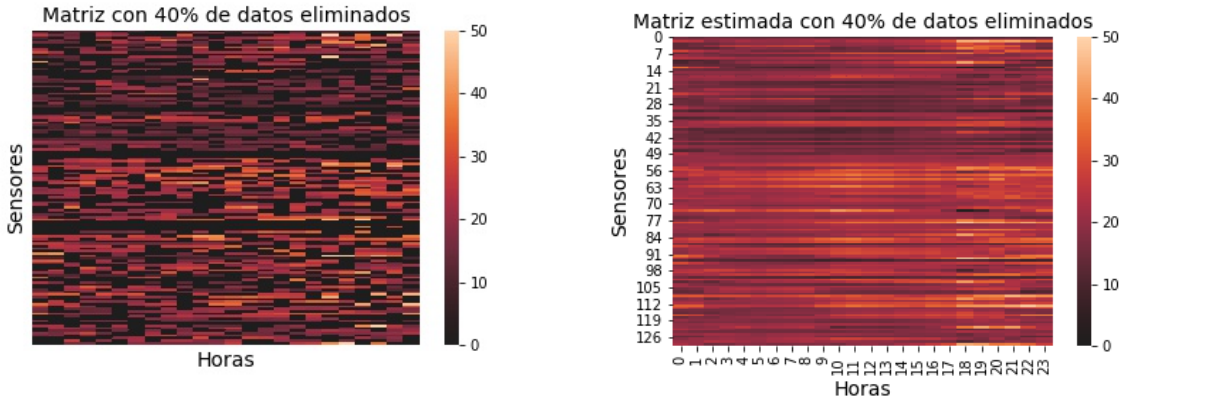
$$\overrightarrow{\mathbf{b}_t} = b_t - 2\alpha(\lambda b_t - m_{st} - \mu - b_s - b_t - \mathbf{p}_s^T \mathbf{q}_t) \tag{3.34}$$

Finalmente se realiza la evaluación de desempeño del algoritmo planteado en la ventana de 24 horas seleccionada previamente (ver Figuras 3.7 y 3.8). Los parámetros usados para el entrenamiento del algoritmo son: $k = 4$, seleccionado de forma libre; $\alpha = 0{,}01$ y $\lambda = 0{,}1$, seleccionados de acuerdo con [37]; y 100 iteraciones, también seleccionadas de forma libre.



(a) (b)

Figura 3.7: Índice MAPE para diferentes porcentajes de datos eliminados



(a) Datos eliminados

(b) Datos faltantes estimados por MF + Regularización + Bias

Figura 3.8: Ventana de tiempo de 24 horas

3.2.5. Comparación entre las funciones de costo propuestas

En la Tabla 3.1 y en la Figura 3.9 se presenta un comparativo del desempeño de cada una de las variantes del algoritmo MF propuestos. Se observa que del 40 % al 50 % de información eliminada artificialmente los algoritmos presentaron un desempeño similar. Sin embargo, el modelo MF + Regularización + Bias mostró mejores resultados para cantidades de información perdida superiores al 70 % , lo cual muestra un resultado que se adapta mejor a la problemática visualizada en la WSN Ciudadanos Científicos, ya que de acuerdo la Figura 1.2 la red

de sensores cuenta con nodos que alcanzan dichos porcentajes.

Por lo tanto, en la siguiente subsección se dará continuidad con el modelo MF + Regularización + Bias para la evaluación de algoritmos de optimización basados en gradiente, los cuales más adelante serán sometidos en sus parámetros a un proceso de sintonización. Todo lo anterior se analiza con miras en obtener el mejor rendimiento para la estimación de datos perdidos de la WSN de bajo costo Ciudadanos Científicos.

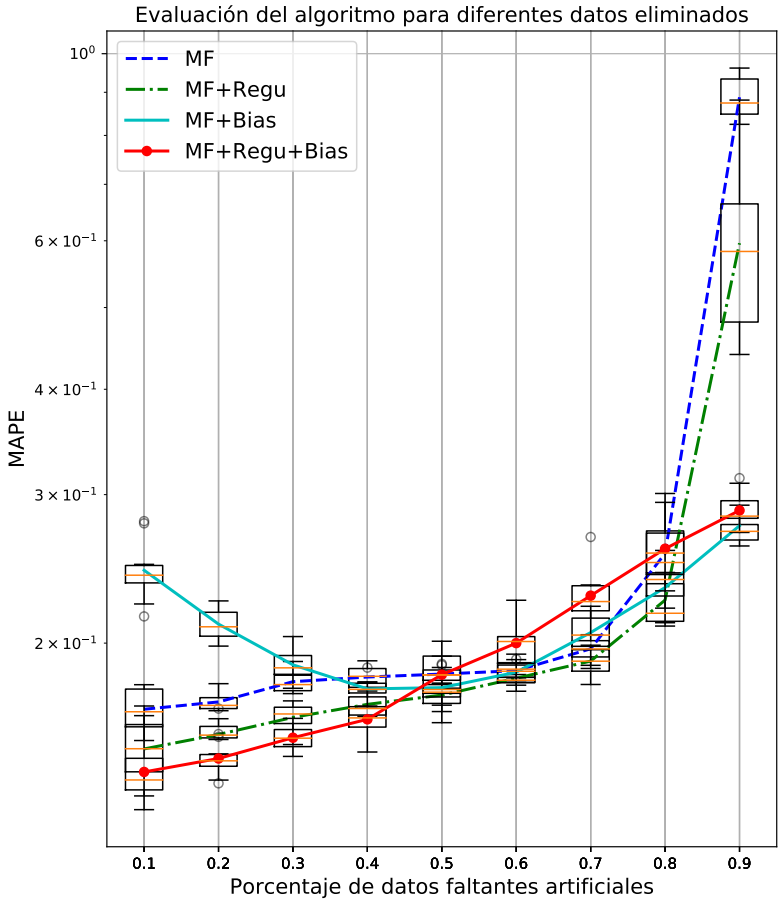


Figura 3.9: Comparación de resultados obtenidos para los algoritmos MF propuestos

Función de costo	Datos eliminados	MAPE
MF	$40\,\%$	$0{,}182313 \pm 0{,}005270$
MF + Regularización	$40\,\%$	$0{,}169205 \pm 0{,}005528$
MF + Bias	$40\,\%$	$0{,}162425 \pm 0{,}006055$
MF + Regularización + Bias	$40\,\%$	$0{,}160133 \pm 0{,}005376$
MF	$50\,\%$	$0{,}183795 \pm 0{,}003906$
MF + Regularización	$50\,\%$	$0{,}176536 \pm 0{,}007555$
MF + Bias	$50\,\%$	$0{,}183492 \pm 0{,}010862$
MF + Regularización + Bias	$50\,\%$	$0{,}183631 \pm 0{,}007446$

Tabla 3.1: Resumen de resultados de diferentes funciones de costo para valores del 40 % y 50 % de datos eliminados artificialmente.

3.3. Algoritmos de optimización basados en GD

El descenso de gradiente es el método de optimización más popular para encontrar el mínimo de funciones [72, 77]. Son ampliamente implementados en los modelos de ML y más concretamente para encontrar el mínimo error obtenido en una función de pérdida, como por ejemplo, la diseñada en Eq.(3.26).

Actualmente existen gran variedad de algoritmos basados en GD que plantean modificaciones o mejoras al desempeño del mismo. Para el abordaje de este trabajo se realizó la evaluación del algoritmo de SGD con los algoritmos de optimización Momentum y RMSprop, en la función de costo diseñada en la sección 3.2 que mejor desempeño presentó con los datos de la WSN de bajo costo Ciudadanos Científicos (ver la subsección 3.2.4). Dicha evaluación se realizó como parte de la exploración y diseño, no solo de una función de costo, sino también, elegir un algoritmo que mejor entrene el modelo de MF seleccionado.

3.3.1. Momentum

El algoritmo Momentum es una variación de SGD planteado inicialmente por [61]. Momentum determina que en lugar de depender del gradiente actual para actualizar los parámetros, este es sustituido por la velocidad V, la exponencial de la media móvil de los gradientes actuales y las iteraciones realizadas por el algoritmo. Por lo tanto, el algoritmo Momentum se desarrolla minimizando la función de costo que mejores resultados presento en evaluaciones anteriores (ver subsección 3.2.4 Eq.(3.26)), obteniendo las expresiones que se muestran a continuación:

$$\overrightarrow{\mathbf{p}_s} = \mathbf{p}_s - \alpha V_{\mathbf{p}_s} \tag{3.35}$$

$$\overrightarrow{\mathbf{q}_t} = \mathbf{q}_t - \alpha V_{\mathbf{q}_t} \tag{3.36}$$

$$\overrightarrow{\mathbf{b}_s} = \mathbf{b}_s - 2\alpha V_{\mathbf{b}_s} \tag{3.37}$$

$$\overrightarrow{\mathbf{b}_t} = \mathbf{b}_t - 2\alpha V_{\mathbf{b}_t} \tag{3.38}$$

Finalmente, para para suavizar la trayectoria del gradiente y conservar el efecto de los mismos, se plantea mantener el parámetro de sensibilidad β cercano y menor a 1. En consecuencia se configura $\beta = 0{,}9$, tal y como es sugerido en la literatura [61]. Por lo tanto, la velocidad para cada parámetro de la función de costo se actualiza como se muestra a continuación:

$$V_{\mathbf{p}_s} = \beta V_{\mathbf{p}_s} + (1-\beta)\frac{\partial f}{\partial \mathbf{p}_s} \tag{3.39}$$

$$V_{\mathbf{q}_t} = \beta V_{\mathbf{q}_t} + (1-\beta)\frac{\partial f}{\partial \mathbf{q}_t} \tag{3.40}$$

$$V_{\mathbf{b}_s} = \beta V_{\mathbf{b}_s} + (1-\beta)\frac{\partial f}{\partial \mathbf{b}_s} \tag{3.41}$$

$$V_{\mathbf{b}_t} = \beta V_{\mathbf{b}_t} + (1-\beta)\frac{\partial f}{\partial \mathbf{b}_t} \tag{3.42}$$

3.3.2. RMSprop

Por sus siglas en inglés RMSprop (Root Mean Square Propagation) se plantea como una variación del algoritmo de optimización AdaGrad, donde en lugar de realizar la acumulación de los gradientes se utiliza el concepto de ventaneo, donde solo son tenidos en cuenta los gradientes más recientes.

Esta variación de algoritmo de descenso de gradiente de aprendizaje adaptativo fue propuesta por [8] y plantea que en lugar de tomar la suma acumulada de los gradientes cuadrados toma su media móvil. Por lo tanto, el planteamiento de

RMSprop se desarrolla minimizando la función de costo que mejores resultados presento en evaluaciones anteriores (ver subsección 3.2.4 Eq.(3.26)), obteniendo las expresiones que se muestran a continuación:

$$\overrightarrow{\mathbf{p}_s} = \mathbf{p}_s - \frac{\alpha}{\sqrt{S_{\mathbf{p}_s} + \epsilon}} \frac{\partial f}{\partial \mathbf{p}_s} \tag{3.43}$$

$$\overrightarrow{\mathbf{q}_t} = \mathbf{q}_t - \frac{\alpha}{\sqrt{S_{\mathbf{q}_t} + \epsilon}} \frac{\partial f}{\partial \mathbf{q}_t} \tag{3.44}$$

$$\overrightarrow{\mathbf{b}_s} = \mathbf{b}_s - \frac{\alpha}{\sqrt{S_{\mathbf{b}_s} + \epsilon}} \frac{\partial f}{\partial \mathbf{b}_s} \tag{3.45}$$

$$\overrightarrow{\mathbf{b}_t} = \mathbf{b}_t - \frac{\alpha}{\sqrt{S_{\mathbf{b}_t} + \epsilon}} \frac{\partial f}{\partial \mathbf{b}_t} \tag{3.46}$$

En consecuencia, el algoritmo busca encontrar el tamaño adecuado de paso para cada iteración, estimando el tamaño de actualización de cada una de las variables de forma independiente. De acuerdo con las Eqs.(3.43-3.46), se adapta la tasa de aprendizaje dividiendo por la raíz del parámetro $S_{\mathbf{x}_x}$ y por el gradiente de la función de costo. Por lo tanto, el promedio exponencial cuadrado de los gradientes para cada parámetro se muestra a continuación:

$$S_{\mathbf{p}_s} = \beta S_{\mathbf{p}_s} + (1 - \beta) \left[\frac{\partial f}{\partial \mathbf{p}_s} \right]^2 \tag{3.47}$$

$$S_{\mathbf{q}_t} = \beta S_{\mathbf{q}_t} + (1 - \beta) \left[\frac{\partial f}{\partial \mathbf{q}_t} \right]^2 \tag{3.48}$$

$$S_{\mathbf{b}_s} = \beta S_{\mathbf{b}_s} + (1 - \beta) \left[\frac{\partial f}{\partial \mathbf{b}_s} \right]^2 \tag{3.49}$$

$$S_{\mathbf{b}_t} = \beta S_{\mathbf{b}_t} + (1-\beta)\left[\frac{\partial f}{\partial \mathbf{b}_t}\right]^2 \tag{3.50}$$

De igual manera que en la Sección 3.3.1, para suavizar la trayectoria del gradiente y conservar el efecto de los mismos, se plantea mantener el parámetro de sensibilidad $\beta = 0{,}9$ y $\epsilon = 10^{-6}$ sugeridos por la literatura [8].

3.3.3. Evaluación de los algoritmos de optimización propuestos para el entrenamiento de los datos

Luego de evaluar la función de costo Eq.(3.26) en cada uno de los algoritmos de optimización planteados, se verifica en la Tabla 3.2 que el algoritmo SGD para el $40\,\%$ de datos eliminados tiene mejor desempeño sobre los algoritmos Momentum y RMSprop. Por lo tanto, dado que el interés se centra en analizar detalladamente el desempeño para el 40 % de datos eliminados, dado que estos representan aproximadamente la información total perdida por Ciudadanos Científicos durante el año 2018, en el siguiente capítulo de sintonización de parámetros se dará continuidad con el uso de SGD para el entrenamiento y la búsqueda de parámetros de sintonía del modelo MF + Regularización + Bias tratado en la subsección 3.2.4.

A continuación en la Tabla 3.2 y en la Figura 3.10, se presenta el resultado de desempeño obtenido por los algoritmos de optimización evaluados [1] en el modelo MF + Regularización + Bias:

Función de costo	Datos eliminados	MAPE
MF + Regularización + Bias:SGD	$40\,\%$	$0{,}160133 \pm 0{,}005376$
MF + Regularización + Bias:Momentum	$40\,\%$	$0{,}160875 \pm 0{,}006145$
MF + Regularización + Bias:RMSprop	$40\,\%$	$0{,}161107 \pm 0{,}005191$
MF + Regularización + Bias:SGD	$50\,\%$	$0{,}183631 \pm 0{,}007446$
MF + Regularización + Bias:Momentum	$50\,\%$	$0{,}176541 \pm 0{,}008928$
MF + Regularización + Bias:RMSprop	$50\,\%$	$0{,}185030 \pm 0{,}007780$

Tabla 3.2: Resumen de resultados obtenidos para la evaluación de algoritmos de optimización Momentum, RMSprop y SGD.

[1] Los parámetros usados en el entrenamiento del algoritmo fueron: $k = 4, \alpha = 0{,}01, \lambda = 0{,}1$ y 100 iteraciones

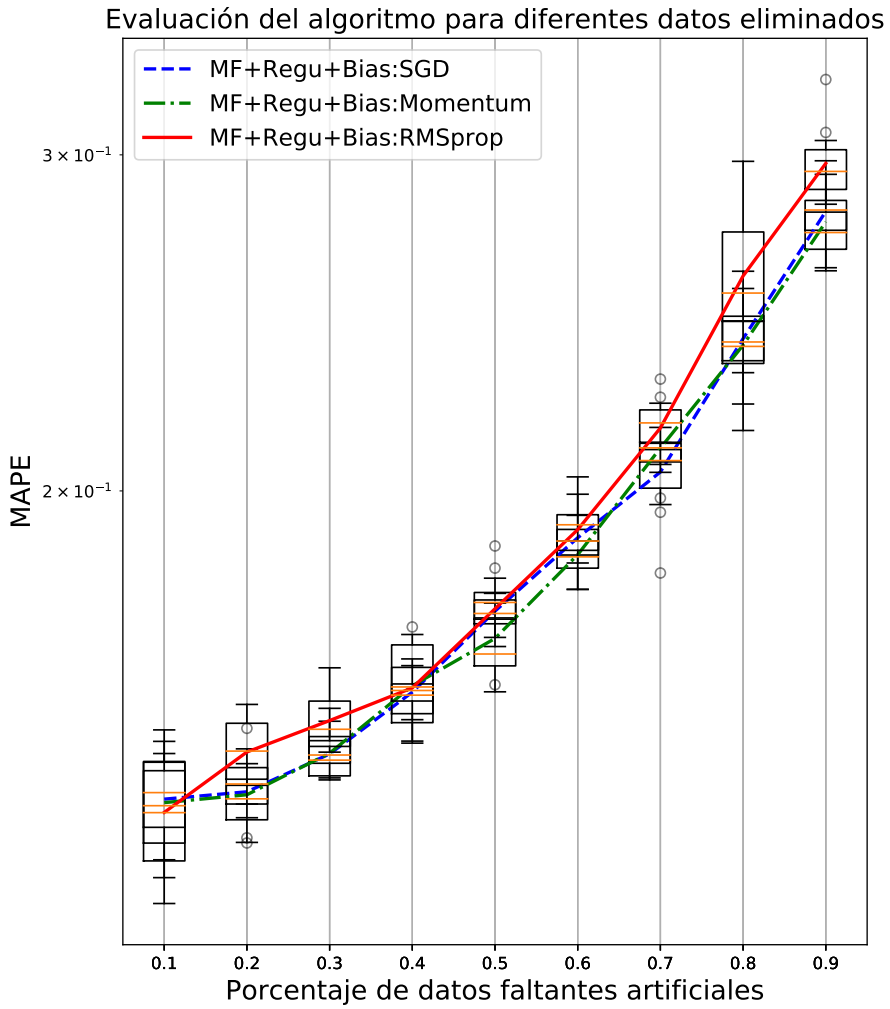


Figura 3.10: Evaluación algoritmos SGD, Momentum y RMSprop para la función de costo Eq.(3.26)

Capítulo 4

Ajuste de parámetros algoritmo MF + Regu + Bias: SGD

Contenido

Sinopsis

Este capítulo da tratamiento a la sintonización de parámetros de la función de costo y el algoritmo de optimización del modelo MF + Regularización + Bias (ver subsección 3.2.4 Eq.(3.26)) que presentó mejores resultados con los parámetros usados en la literatura (Capítulo 3). La evaluación de desempeño se realizó comparando los resultados obtenidos por la MF + Regularización + Bias sintonizada y algoritmos encontrados en la literatura MS, KNN [90] y MOGPs [44, 11, 43], comúnmente usados en problemas de estimación de datos perdidos e interpolación no paramétrica. Se implementó un enfoque general donde se analiza toda la red (todas las series de tiempo), y uno particular donde se toma un sensor para un análisis más detallado.

Lo anterior permitirá identificar el desempeño del algoritmo frente a otras técnicas y el planteamiento de mejoras que se adapten a los datos de calidad de

aire PM2.5, con el fin de obtener el menor error posible y sin pérdida de generalización para parte del método.

4.1. Sintonía de parámetros

De acuerdo con el algoritmo encargado de generar el modelo MF (MF + Regularización + Bias) en este intervienen múltiples parámetros, los cuales son responsables del tiempo de entrenamiento y del rendimiento del mismo. Sin embargo, de acuerdo con el enfoque matemático desarrollado en la Sección 3.2 en la Eq(3.26) y en las evaluaciones preliminares, nos centramos únicamente en el cambio de α, tasa de entrenamiento; λ, parámetro de regularización y k, dimensión del espacio de factores latentes. Debido a que éstos tienen una mayor influencia en el modelo y es por lo tanto esencial tener una etapa de ajuste para estos parámetros. Tradicionalmente se utiliza una búsqueda de cuadrículas con este fin en la que cada punto de la cuadrícula es el producto de 5 repeticiones de los valores donde se presenten los mejores rendimientos del modelo MF seleccionado.

Para esto se genera una red que emplea una secuencia de crecimiento exponencial para los valores de $\lambda = \{0{,}0001, 0{,}001, 0{,}01, 0{,}1, 1\}$ y una secuencia de crecimiento lineal para los valores de $k = \{8, 12, 16, 20, 24\}$. Por otro lado, el valor α se ajusta dinámicamente a medida que se realizan las interacciones del algoritmo. Como resultado de esto las actualizaciones descritas en las Eqs.(3.31-3.34), el valor α varían de acuerdo a la siguiente expresión:

$$\alpha = \frac{1}{k_\alpha},\ k_\alpha = 1, 2, 3, ..., \text{iterations} \tag{4.1}$$

El enfoque anterior se realizó con la finalidad de simplificar el espacio de búsqueda generado por la combinación de los parámetros λ y k, ya que α interviene en la velocidad de aprendizaje. Por los tanto se decidió emplear pasos amplios en las primeras iteraciones y pasos finos para la final del entrenamiento.

Con el fin de implementar el método en la estimación y reporte del día inmediatamente anterior, seleccionamos una ventana de tiempo de 24 horas al azar, correspondiente al día 15 de febrero de 2018. Finalmente, como medida para la evaluación de los resultados se implementó el índice de desempeño MAPE. La Figura 4.1 muestra los resultados de la aplicación de la búsqueda en la cuadrícula, donde se evidencia que para el valor de $\lambda = 0{,}001$ y $k = 20$, se presenta el

mejor desempeño del algoritmo. Lo anterior es consistente con [84] donde se afirma que la dimensión del espacio latente sería de $min(s, m)$. En este caso el valor de k de ahora en adelante sera seleccionado para que coincida con el número de mediciones por hora de la ventana de tiempo seleccionada en el experimento.

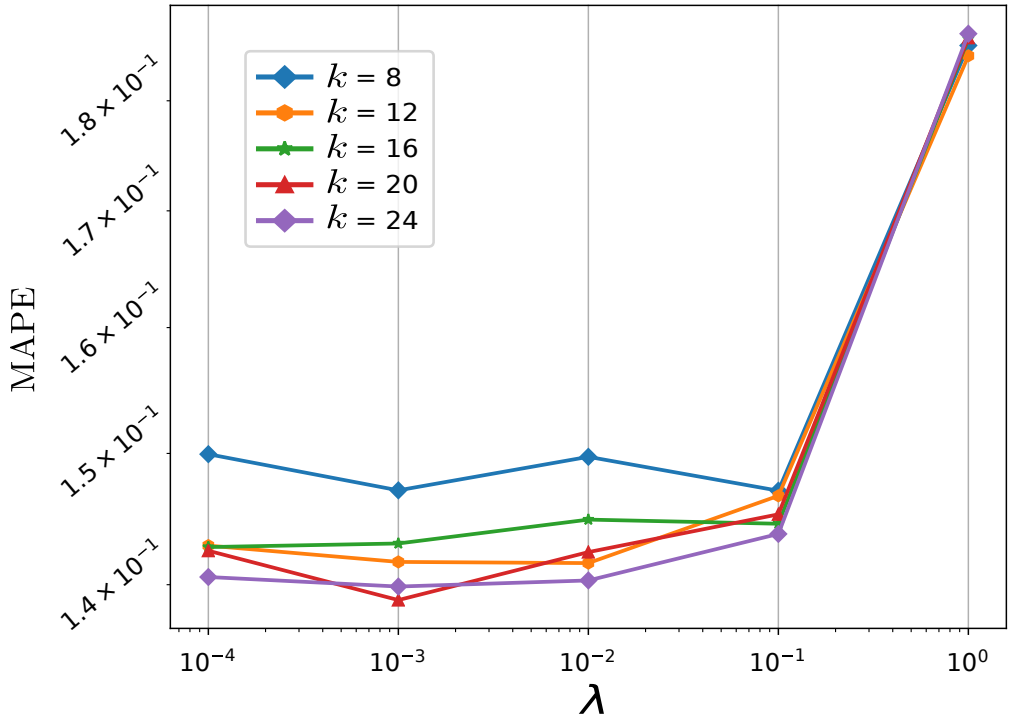


Figura 4.1: Evaluación de parámetros de sintonía

Continuando con la evaluación del algoritmo de MF para completar los datos faltantes se decidió tomar la misma ventana de tiempo de 24 horas de los experimentos descritos en la sección 3.2 y en la subsección 4.1.

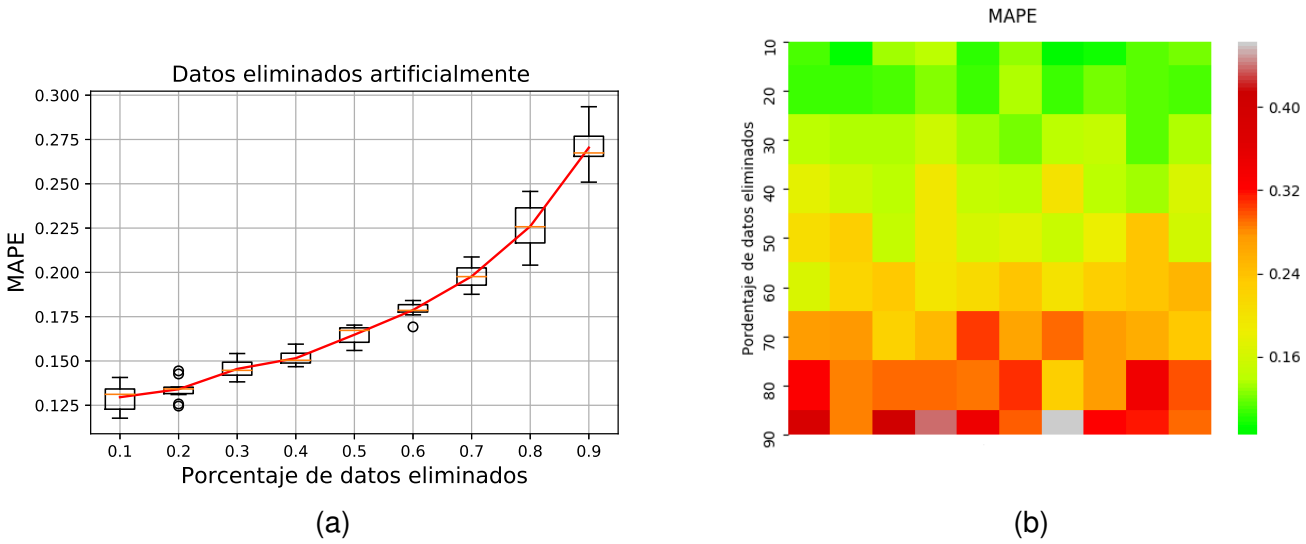
En este caso, el objetivo es evaluar la sensibilidad del algoritmo para diferentes porcentajes de datos perdidos, para ello algunos datos se eliminan aleatoriamente hasta dejar un porcentaje de información deseado. Para definir cada conjunto de datos se eliminó 10 % de la información hasta obtener un conjunto de datos con 90 % de datos eliminados, eliminando en cada paso 10 % de la información. Para evaluar la variabilidad del método en cada porcentaje, se evaluó el algoritmo varias veces, con la finalidad de medir el rendimiento en cada conjunto de datos, se calcula el índice MAPE a partir de los datos originales y los datos estimados por el modelo MF (MF + Regularización + Bias: SGD). Los re-

Función de costo	Datos eliminados	MAPE
MF + Regularización + Bias	40 %	0,160133 ± 0,005376
MF + Regularización + Bias(S)	40 %	0,153811 ± 0,005492
MF + Regularización + Bias	50 %	0,183631 ± 0,007446
MF + Regularización + Bias(S)	50 %	0,168182 ± 0,006682

Tabla 4.1: Comparación entre el modelo MF sintonizado y no sintonizado para el 40 % y 50 % de datos eliminados artificialmente.

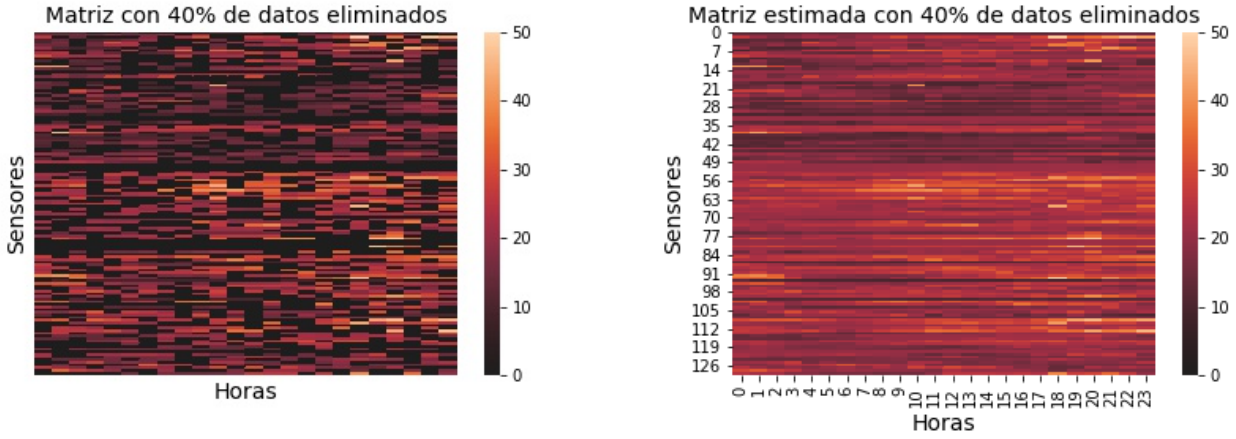
sultados se presentan en la Figura.4.2 utilizando un diagrama de cajas y en la la Figura.4.3 se presenta la reconstrucción de la ventana de tiempo para el 40 % de información eliminada artificialmente, lo cual representa el promedio de perdida de información en los datos generados por Ciudadanos Científicos.

Lo anterior muestra una mejoría en el algoritmo MF + Regu + Bias: SGD respecto a los resultados obtenidos antes de ser sintonizado (ver Tabla 4.1).



(a) (b)

Figura 4.2: Evaluación del error ante diferentes porcentajes de datos faltantes con parámetros de ajuste



(a) Datos eliminados (b) Datos faltantes estimados

Figura 4.3: Estimación de datos faltantes MF + Regu + Bias: SGD y parámetros ajustados

4.2. Comparación de desempeño del modelo MF sintonizado vs algoritmos de la literatura

La comparación entre MF + Regularización + Bias (S) [1] con respecto a los algoritmos MS, KNN y MOGPs se desarrolla a partir de dos enfoques: En el primero se presenta un enfoque general donde se analiza toda la red (todas las series de tiempo), y en el segundo se analiza uno enfoque particular donde se toma un sensor para un análisis más detallado (sensor de periferia).

Las configuraciones de parámetros realizadas en los algoritmos implementados serán los mismos para cada uno de los enfoques mencionados anteriormente. Para el modelo MF fueron usados los parámetros obtenidos en los experimentos de la subsección 4.1, donde se eligieron para los parámetros de entrenamiento que presentaron mejor desempeño. Como resultado λ se establece en 0,001, mientras que α se ajusta dinámicamente, de acuerdo con la Eq.4.1. Sin embargo, para analizar el método en periodos de tiempo más amplios, tomamos una ventana de tiempo de 15 días equivalente a 360 horas. La ventana comprende la fecha del 15 de febrero al 3 de marzo de 2018. En consecuencia, se produce un cambio en la dimensión de los datos de $\Re^{130\times24}$ a $\Re^{130\times360}$, forzando un cambio de k a 130, debido a que las dimensiones del espacio latente no deben exceder las dimensiones del espacio original ($min(s, m)$) [84].

[1] Modelo MF sintonizado usando como algoritmo de entrenamiento SGD

Para el entrenamiento del modelo de MOGPs fueron seleccionaron 4 sensores vecinos para cada nodo en particular, los cuales se eligieron de acuerdo con la información geográfica de cada sensor. El modelo implementado que mejor resultados obtuvo con la base de datos está conformado por un Kernel de mezcla espectral multivariado o MOSM (Multi-Output Spectral Mixture kernel), una función de costo de log-verosimilitud (NLL) y una función de optimización L-BFGS-B, tal y como es planteado por [11]. La técnica KNN se sintonizó con una búsqueda de rejilla, evaluando diferentes números de vecinos $k_n = \{3, 4, 5, 6, 7, 8, 9\}$ obteniendo el mejor resultad para $k_n = 5$, donde se observó mejor desempeño ante efectos de ruido y datos no consistentes (datos faltantes). Finalmente, la técnica MS se llevó a cabo usando el promedio para el mismo instante de tiempo de todos los sensores de la red.

4.2.1. Enfoque de evaluación general

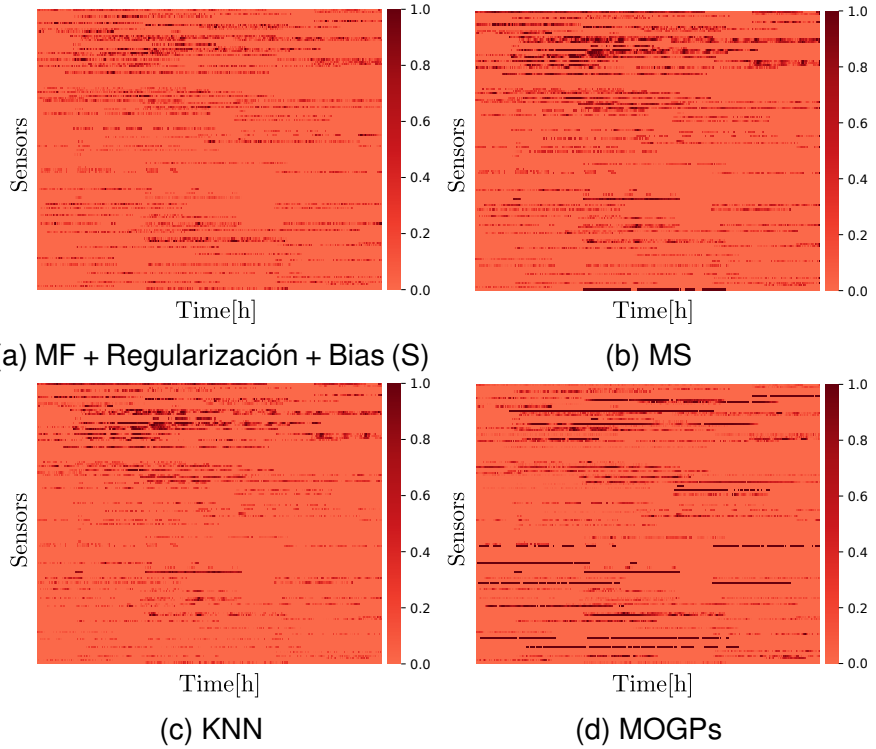
En consecuencia para el primer enfoque (general), los grupos de datos se eliminaron siguiendo un patrón de lotes. Para ello, fueron descartadas lecturas consecutivas a lo largo del tiempo del mismo sensor de forma aleatoria. Los tamaños de los lotes van desde 1 hora a 120 horas, que equivalen a un tercio de la medición del sensor durante la ventana de tiempo seleccionada. En este sentido, se busca visualizar cómo el método cambia su capacidad de estimar los datos eliminados artificialmente ante un escenario más realista de pérdida de información por la red (Sección 2.1.3).

Por ejemplo: deficiencias en la conexión a Internet donde el nodo se encuentra conectado; fallas de energía en el punto de conexión; deterioro natural del dispositivo (sensor); robo del nodo; entre otros. Lo anterior puede ocurrir en cualquier nodo de la red en períodos determinados por: horas, días, semanas y meses.

La Figura 4.4 muestra los mapas de calor con los errores obtenidos para las técnicas comparadas, donde para calcular los errores fue implementada la métrica MAPE que da una idea del error obtenido de forma porcentual, donde los tonos oscuros [2] representan grandes porcentajes de error, mientras que los tonos claros representan pequeños porcentajes de error.

Por otro lado, en la Figura 4.5 se presenta también un resultado con una

[2]Si bien los MOGPs pueden ser eficientes en la recuperación de datos, el resultado obtenido en la Figura 4.4 demuestra que dichos algoritmos cuentan con limitaciones cuando los datos son altamente dispersos, al punto que no permite la convergencia del método. Explicando entonces los resultados adversos representados por las manchas oscuras generadas por el mapa de calor en Figura 4.4d



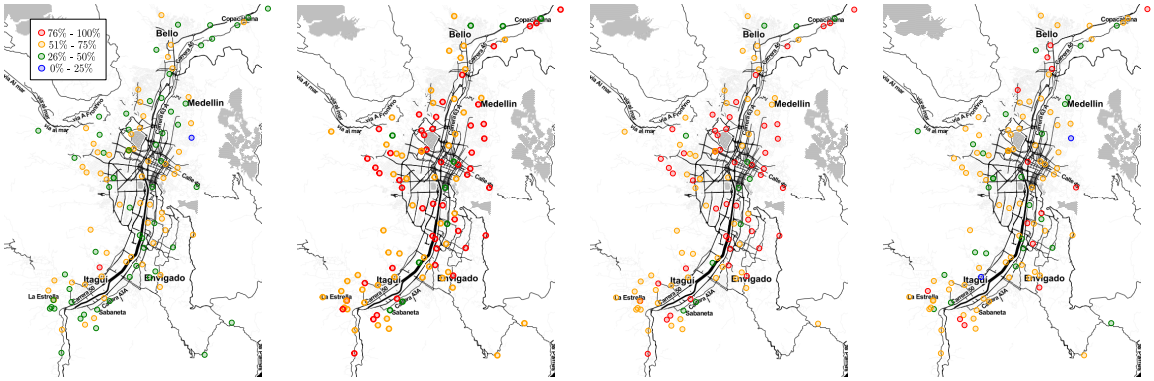
(a) MF + Regularización + Bias (S) (b) MS

(c) KNN (d) MOGPs

Figura 4.4: Aplicación MF + Regularización + Bias (S) para una ventana de 15 días con $40\,\%$ de datos faltantes y comparación de desempeño con MS, KNN y MOGP

mirada espacial de la evaluación MAPE, donde se logra apreciar de forma un poco mas clara el desempeño de los algoritmos evaluados para cada uno de los sensores de la red Ciudadanos Científicos. En este sentido, fue determinado el mayor ($100\,\%$) y el menor ($0\,\%$) error obtenido entre los 4 métodos y de esta manera generar un rango de escala distribuida en 4 colores: azul para un valor entre 0 y $25\,\%$; verde para un valor entre 26 y $50\,\%$; naranja para un valor entre $51\,\%$ y $75\,\%$; y finalmente rojo para un valor entre $76\,\%$ y $100\,\%$.

Finalmente, como complemento del enfoque general se realiza un análisis individual por nodo, llevando a cabo una evaluación del error obtenido por cada método comparado y sumando un punto al algoritmo de mejor rendimiento para cada nodo en particular. Los resultados en la Figura.4.6 muestran claramente el rendimiento de cada algoritmo, para lo cual se utilizaron los siguientes métodos de evaluación de error: RMSE, es una medida de la distancia entre el dato real y el estimado, dando mayor importancia a los errores más altos obtenidos. MAPE es un método de interpretación más simple, ya que este da una idea del por-



(a) Matrix factorization. (b) Mean substitution. (c) KNN. (d) MOGPs.

Figura 4.5: Distribución espacial del rendimiento obtenido mediante el método de evaluación del error MAPE.

centaje de error obtenido en cada una de las estimaciones realizadas por cada algoritmo implementado y que apunta al promedio de los errores obtenidos; y EVS visualiza la variación o la dispersión obtenida por el modelo para cada una de las series de tiempo. En otras palabras, entrega un proporción del ajuste dado por cada modelo matemático reflejando la calidad de la regresión (entre más alto mejor se ajusta el modelo) [59].

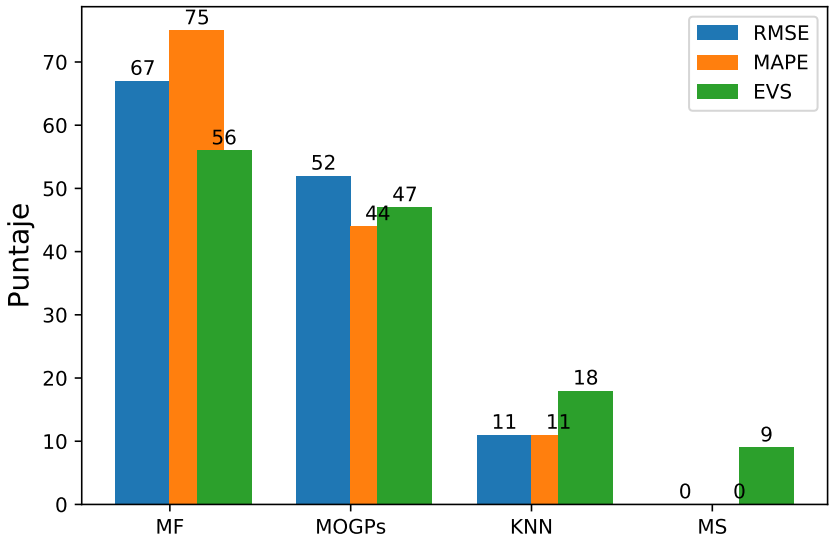


Figura 4.6: Desempeño de los algoritmos implementados para diferentes medidas de evaluación

De acuerdo con la Figura 4.6 los resultados obtenidos muestran que, si bien los MOGPs son robustos para la interpolación y extrapolación de información a partir de señales similares (uso de sensores vecinos), el método presenta limitaciones cuando, tanto el sensor a recuperar información como los sensores vecinos cuentan gran información perdida, dificultando entonces la aplicación del método, lo cual da como resultado la imposibilidad de convergencia del algoritmo. Por otro lado, los MOGPs para su funcionamiento dependen altamente de una adecuada selección de Kernel [43], haciendo generalmente una tarea difícil la selección del modelo que mejor se adapte a la problemática planteada, por lo que en este aspecto los algoritmos MF ganan por su simplicidad tanto en desempeño computacional como de ajuste de parámetros.

Lo anterior puede corroborarse en los resultados obtenidos en la Figura 4.4 ya que de acuerdo con el mapa de calor, para el modelo de MOGPs se obtuvieron zonas (series de tiempo) donde el método no generó ninguna estimación (zonas oscuras) y en la Figura 4.6, se puede observar que el modelo es superado por la técnica MF en todas las medidas de desempeño analizadas, reflejando así inconsistencias entre los datos reales y las estimaciones no generadas (errores altos debido a la no convergencia).

4.2.2. Enfoque de evaluación particular

Para el enfoque particular se seleccionó un sensor de la red que no mostró datos faltantes durante la ventana de tiempo analizada. La selección del sensor corresponde al Nodo 5 ubicado en el municipio de Caldas (periferia). Posteriormente se eliminó el 40 $\%$ de los datos del nodo lo que corresponde a 144 medidas. La eliminación se realizó en 6 intervalos de 24 horas cada uno, buscando recrear patrones largos de información perdida que podrían presentarse en los nodos de la red.

Modelo	RMSE	MAPE
MS	$5{,}8359$	$0{,}6755$
KNN	$4,0978$	$0,3226$
MOGPs	$0{,}9989$	$0{,}0875$
MF + Regularización + Bias (S)	$0{,}8558$	$0{,}0780$

Tabla 4.2: Errores RMSE y MAPE obtenidos por los algoritmos evaluados para el enfoque particular

Los resultados del experimento se reportan en la Figura 4.7, Figura 4.8 y

Figura 4.9; donde los datos reales están representados por la línea azul; los datos eliminados se representan con puntos rojos; para la MF están representados por la línea punteada naranja; y finalmente para MOGPs, KNN y MS se utilizaron las líneas punteadas en verde. Adicionalmente, se calcularon los valores de MAPE y RMSE para cada técnica utilizando solo los datos eliminados artificialmente, lo que dio lugar a los resultados mostrados en la Tabla 4.2.

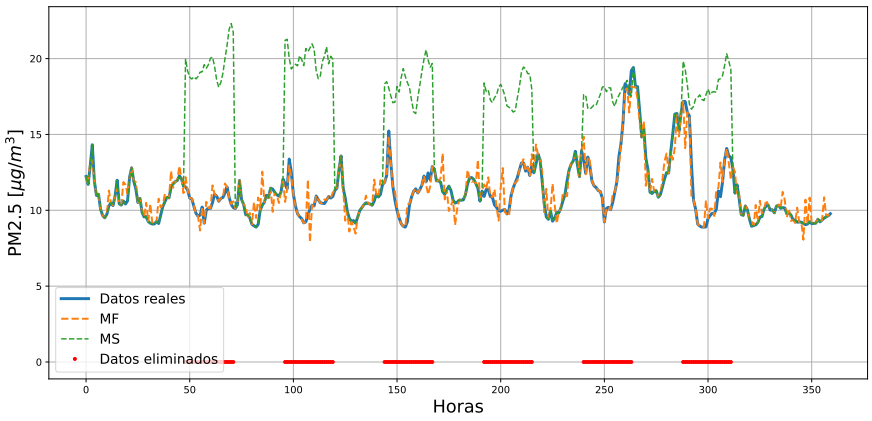


Figura 4.7: Comparación MF y MS para 15 días de medición del Nodo 5

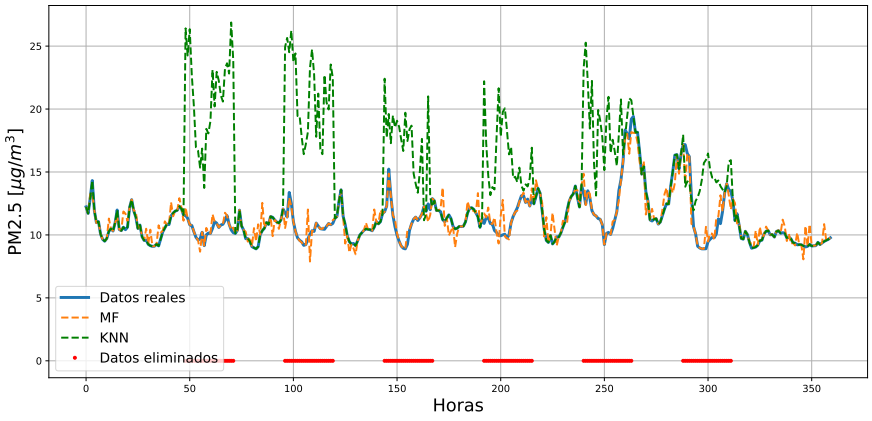


Figura 4.8: Comparación MF y KNN para 15 días de medición del Nodo 5

En consecuencia de la Figura 4.6 y los resultados obtenidos en la Figura 4.7, Figura 4.8 y Figura 4.9, se puede visualizar que los resultados obtenidos son

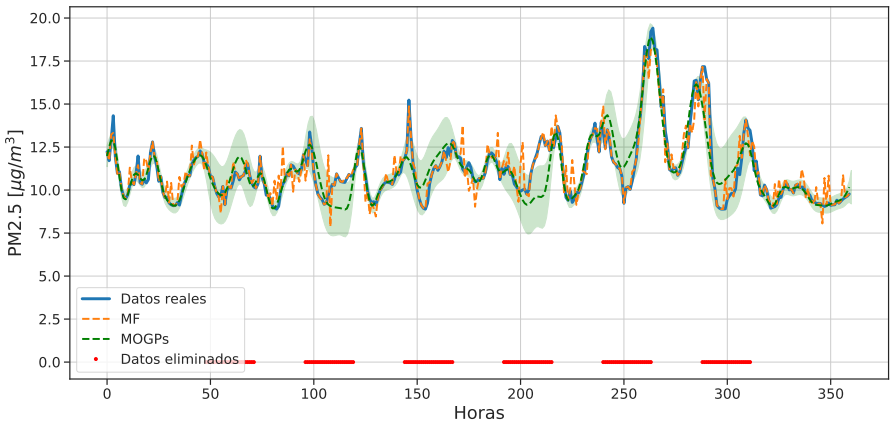


Figura 4.9: Comparación MF y MOGPs para 15 días de medición del Nodo 5

consecuentes con la literatura [90], dado que las técnicas MF y MOGPs mostraron un desempeño ampliamente superior a los obtenidos con las técnicas KNN y MS que fallan cuando la información perdida que se extiende por grandes periodos de tiempo. Adicionalmente, según detalla el estado del arte, el modelo MF cuando se incluyen parámetros de regularización y sesgo genera buenos resultados con información altamente dispersa [65, 37, 32] y el modelo MOGPs al tener en cuenta correlaciones espacio-temporales le permite entonces generar estimaciones e intervalos de confianza de los datos recuperados a partir de sensores vecinos [44, 11, 43].

Por último, es interesante destacar los resultados parejos obtenidos por MF y MOGPs en la métrica EVS, lo cual puede explicar que las técnicas evaluadas logran ajustar bien los datos de la WSN Ciudadanos Científicos, dando una idea de la similitud en la variación o dispersión obtenida por cada modelo para cada una de las series de tiempo.

Capítulo 5

Algoritmo DMF como método de mejora para la recuperación de datos perdidos por la WSN Ciudadanos Científicos

Contenido

Sinopsis

Este capítulo da tratamiento al planteamiento del modelo DMF como mejora de la arquitectura del modelo MF desarrollado y sintonizado en los capítulos 3 y 4, y la evaluación de desempeño del mismo en la estimación de datos faltantes de una WSN de bajo costo. Para esto se plantea inicialmente una arquitectura novedosa ampliamente desarrollada en la literatura basada en la implementación de redes neuronales denominada DMF. Por lo tanto, se realiza el planteamiento de tres variaciones del modelo DMF, los cuales contengan características espacio-temporales, a través del uso de Embedding Layers para la codificación de la información.

Para esto se busca realizar una evaluación de desempeño por medio de la implementación de la prueba estadística de Friedman, para verificar la existencia de alguna diferencia significativa entre los métodos evaluados, y la implementación de la prueba pos hoc de Nemenyi, donde será seleccionada la variación del modelo que mejores resultados presente.

Lo anterior busca darle paso al planteamiento de una herramienta efectiva para el trabajo con arquitecturas que implícitamente incluyan la capacidad de las ANNs de aprender representaciones de características desde cero. Lo cual conlleve a una mejor estimación de la información perdida durante el proceso de generación de datos de la red y la medición de material particulado PM2.5. En consecuencia, en este capítulo se busca combinar la interpretación y el sentido matemático de la MF y el poder de aprendizaje contenido en las redes neuronales profundas (DNNs).

5.1. Deep Matriz Factorization DMF

5.1.1. Embedding Layers

Un Embedding Layer es un módulo de red neuronal profunda implementado generalmente en problemas de procesamiento natural del lenguaje [71] y filtros colaborativos (generación de recomendaciones) [24]. Las Embedding Layers resuelven el problema de codificación de las redes neuronales profundas (One hot encoding), ya que las representaciones latentes en los modelos clásicos se encuentran conformadas por representaciones dispersas, principalmente por vectores compuestos en su mayoría por valores cero. En cambio, las Embedding layers cambian dichos vectores dispersos por proyecciones de las características en espacios vectoriales continuos [71]. En consecuencia, estos resultan más

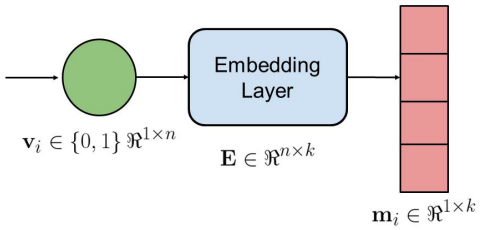


Figura 5.1: Descripción general de una embedding layer

eficientes computacionalmente.

Otro aspecto importante que motiva el uso de Embeddings Layers para resolver el problema planteado en este trabajo sobre las WSNs, se debe a que estas permiten decidir la dimensionalidad del espacio latente k que se le asigna a cada índice d (característica) [60, 79]. Lo que permite a su vez incluir información relevante dentro de los algoritmos que no cuentan con representación matemática, como por ejemplo: día de la semana; estado del día; información geográfica; tiempo en el que generó un dato; índice del sensor; etc. Por lo tanto, la aplicación de Embedding Layers en los modelos DMF propuestos posibilitan entonces, capturar eficientemente la dimensionalidad espacial y temporal asociadas al comportamiento de cada sensor usando estas como entradas del modelo.

En la Fig.5.1 se visualiza la arquitectura general de una Embedding Layer., esta codifica los índices d asociados a cada observación n conocida. En la Eq.(5.1) se presenta en un vector one-hot $\mathbf{v}_i \in \{0,1\}\,\Re^d | i = 1, ..., n$, donde d será la dimensión de la Embedding Layer y v_{ij} es un numero natural que representa cada una de las posiciones del vector $\mathbf{v}_i$.

$$v_{ij} = \begin{cases} 1 & if \;\; i = j \\ 0 & others \end{cases} \tag{5.1}$$

En este sentido, la Embedding Layer irá aprendiendo los mejores valores de $\mathbf{e}_i \in \Re^k$, que representan la codificación de cada característica, conforme se mueve el gradiente del modelo de optimización usado por la DMF. Para esto la matriz embedding $\mathbf{E} \in \Re^{k\times d}$ inicia la primera iteración con valores aleatorios, muestreando cada una de las posiciones como una distribución Gaussiana $\mathbf{E} \sim \mathcal{N}(0,1)$ [40]. Por lo tanto, al finalizar el entrenamiento, se obtiene una salida que codifica las representaciones en el espacio latentes de dimensión k de cada entrada de dimensión d ingresado a la DMF. En consecuencia la representación matemática del vector $\mathbf{v}_i$ será $\mathbf{e}_i$ y se puede desarrollar de acuerdo con la Eq.(5.2):

$$\mathbf{e}_i = \mathbf{E}\mathbf{v}_i = [e_{i1} \ ... \ e_{ik}]^T \tag{5.2}$$

Lo anterior se cumple siempre y cuando $v_{ij} = 1$. Esto quiere decir entonces, que la expresión de la Eq.(5.2) es diferenciable y puede ser integrada en el "back-propagationrespecto a la función de costo de la red neuronal. Esta integración se puede entender de acuerdo con el siguiente desarrollo:

$$\frac{\partial J}{\partial E_{ij}} = \sum_k \frac{\partial J}{\partial \mathbf{e}_k} \frac{\partial \mathbf{e}_k}{\partial E_{ij}} \tag{5.3}$$

La Eq.(5.3) es 0 si $n \neq i$. También $\frac{\partial \mathbf{e}_k}{\partial E_{nj}} = 0$ si $k \neq j$ y 1 para el resto llegando a la expresión mostrada en la Eq.(5.4)

$$\frac{\partial J}{\partial E_{nj}} = \frac{\partial J}{\partial \mathbf{e}_j} \tag{5.4}$$

En otras palabras:

$$\frac{\partial J}{\partial \mathbf{E}} = \begin{bmatrix} 0 & \frac{\partial J}{\partial \mathbf{e}_1} & 0 \\ \vdots & \vdots & \vdots \\ 0 & \frac{\partial J}{\partial \mathbf{e}_k} & 0 \end{bmatrix} \tag{5.5}$$

De esta manera solo se debe rellenar la columna i^{th} de la Encodding Layer con el gradiente descendente de la red DMF.

5.1.2. Modelo DMF propuesto (DMF1)

DMF se basa en un modelo de variables latentes no lineales, en el que las variables latentes son mucho menores que las variables observadas [17]. Por lo tanto, DMF es capaz de recuperar información no muestreada en una matriz de datos con características no lineales en sus estructuras [18, 92]. La arquitectura de red mostrada en la Fig.5.2 se basa principalmente en los trabajos realizados por [17] y [92]. Sin embargo, al planteamiento de las entradas que alimentan la red, fue adicionada una capa de Embedding Layer por cada parámetro de información incluido en el modelo (temporal, espacial o espacio-temporal), dado que se quiere una red que se adapte a los datos generados por la WSN Ciudadanos Científicos y adicionalmente cuente con las ventajas señaladas en la subsección 5.1.1.

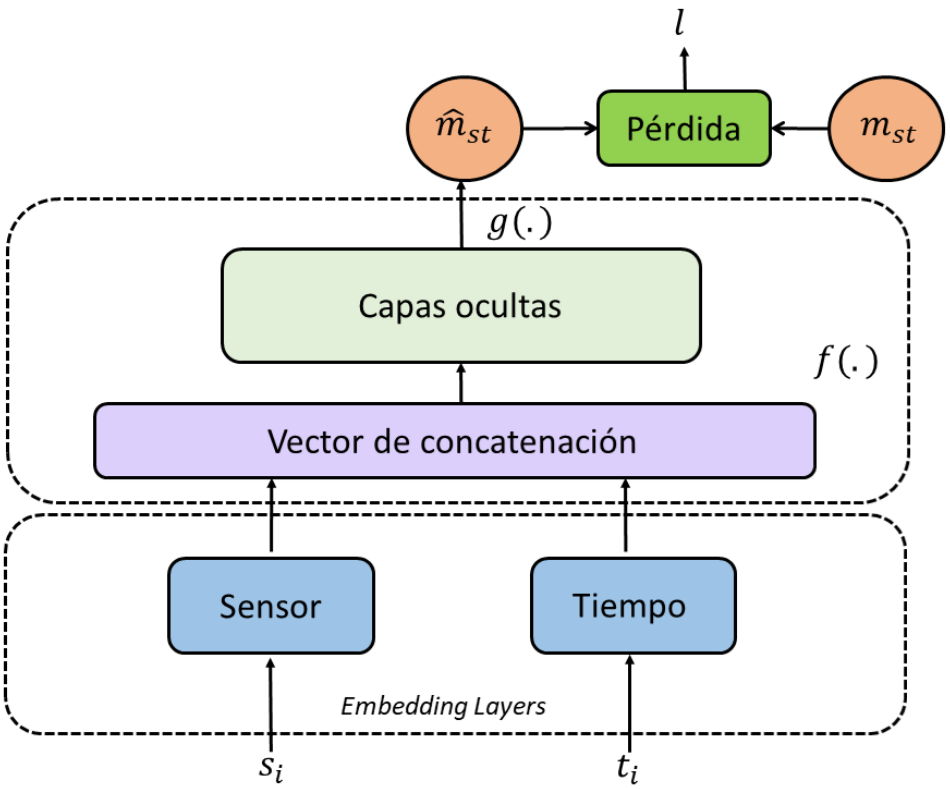


Figura 5.2: Arquitectura del modelo DMF (DMF1) planteado de acuerdo con [92]

Suponga $f(.)$ como la función no lineal que mapea el vector Embedding $\mathbf{e}_i$, que represente las posiciones observadas Ω de la matriz dispersa $\mathbf{M}$ para el sensor s y el tiempo t. Por lo tanto, para el entrenamiento del modelo las entradas pueden ser mapeadas fácilmente resolviendo el siguiente problema:

$$\min_{\Theta} \sum_{i=1}^{n} \left(m_{st} - f\left(\Theta, \mathbf{e}_i\right)\right)^2, (s,t) \in \Omega \tag{5.6}$$

En este sentido y de acuerdo con el planteamiento realizado en [17], el problema DMF1 puede ser representado como:

$$\begin{aligned} &f\left(\Theta, \mathbf{e}_i\right) = g^{(h+1)} \\ &\left(g^{(h)}\left(..g^{(1)}\left(\mathbf{e}_i, \Theta^{(1)}\right), ..\Theta^{(h)}\right), \Theta^{(h+1)}\right) \end{aligned} \tag{5.7}$$

Donde $\mathbf{W}$ denota la matriz de pesos en cada capa; $\mathbf{b}_i$ el vector de bias; $g(.)$ denota la función de activación Sigmoide;$\Theta^{(j)}$ representa los hiperparámetros de

la red $\Theta^{(j)} = \{\mathbf{W}^{(j)}, \mathbf{b}_i^{(j)}\}$, $g^{(j)}(t, \Theta^{(j)}) = g^{(j)}(\mathbf{W}^{(j)}, \mathbf{b}_i^{(j)})$, $j = 1, 2, .., h+1$; y h es el número de capas ocultas de la red.

De acuerdo con [17], el modelo presentado en Eq.(5.7) es llamado finalmente como DMF y para el caso de experimentación sera nombrado como DMF1. Por lo tanto, el modelo buscará reconstruir las representación no observadas $\widehat{m}_{st}, (s,t) \in \overline{\Omega}$ de la matriz $\mathbf{M}$.

5.1.3. Función de pérdida y algoritmo de optimización

Como ya se ha venido desarrollando en el transcurso de este trabajo, otro de los componentes clave en la definición de los algoritmos de ML y DL es definir la función objetivo para la optimización del modelo según los datos observados y la retroalimentación no observada. En este caso, fue implementada la función de costo del error cuadrado planteado en la Eq.(3.4) del capítulo 3. Sin embargo, para la función de optimización se planteó la implementación del modelo Adam, este algoritmo es un modelo de estimación de momento adaptativo, planteado por primera vez en [34] y que en la actualidad es el método más usado para el entrenamiento de algoritmos de DL, como redes neuronales recurrentes (RNN) y redes neuronales convolucionales (CNN).

De acuerdo con [34], este es un método de optimización de gradiente descendente que plantea una combinación del impulso (Momentum) y RMSprop descrito por la Eq.(5.8):

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\hat{S}_t + \epsilon}} \tag{5.8}$$

donde,

$$\hat{V}_t = \frac{V_t}{1 - \beta_1^t} \tag{5.9}$$

$$\hat{S}_t = \frac{S_t}{1 - \beta_2^t} \tag{5.10}$$

$$V_t = \beta_1 V_{t-1} + (1 - \beta_1)\frac{\partial l}{\partial w_t} \tag{5.11}$$

$$S_t = \beta_2 S_{t-1} + (1 - \beta_2)\left[\frac{\partial l}{\partial w_t}\right]^2 \tag{5.12}$$

Con parámetros de sensibilidad $\beta_1 = 0{,}9$, $\beta_2 = 0{,}99$ y $\epsilon = 10^{-8}$, sugeridos por la literatura [34].

5.1.4. Comparación algoritmo DMF1 y MF

Dando continuidad al desarrollo experimental abordado en la sección 4.2 se realizó una experimentación rápida bajo un enfoque general y otro particular usados en la sección 4.2 manteniendo de este el análisis del nodo 5. Como resultado del estudio general se presenta el siguiente gráfico de barras que cuantifica la cantidad de nodos donde cada método evaluado presentó menor error (calificación por desempeño en cada uno de los nodos de la red):

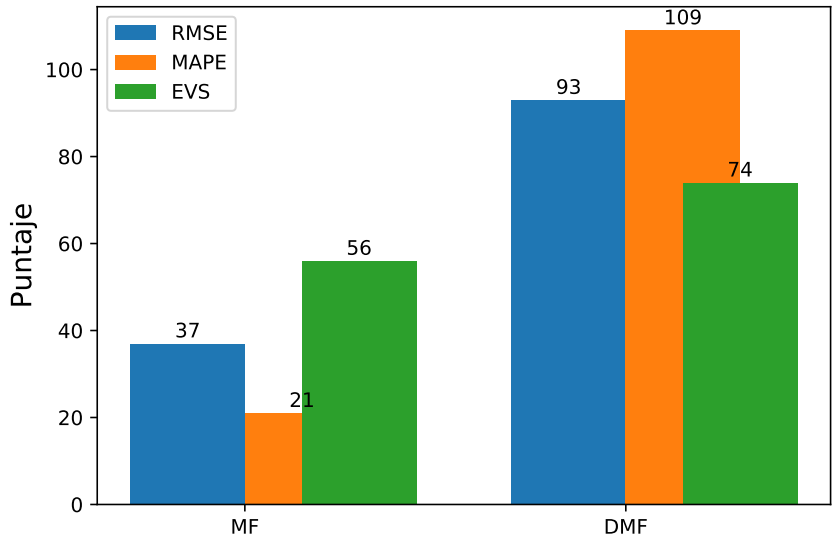


Figura 5.3: Desempeño de los algoritmos DMF y MF

Por otro lado, del análisis particular entre DMF (DMF1) y MF para el nodo 5 de la WSN (ver Figura 5.4), se observa un desempeño similar visualmente, lo cual es consecuente dado que obtener errores más bajos se hace cada vez más difícil (ver Tabla 5.1). Sin embargo del análisis general en la Figura 5.3, es claro que el modelo DMF1 es superior al modelo MF desarrollado y sintonizado en el Capítulo 4.

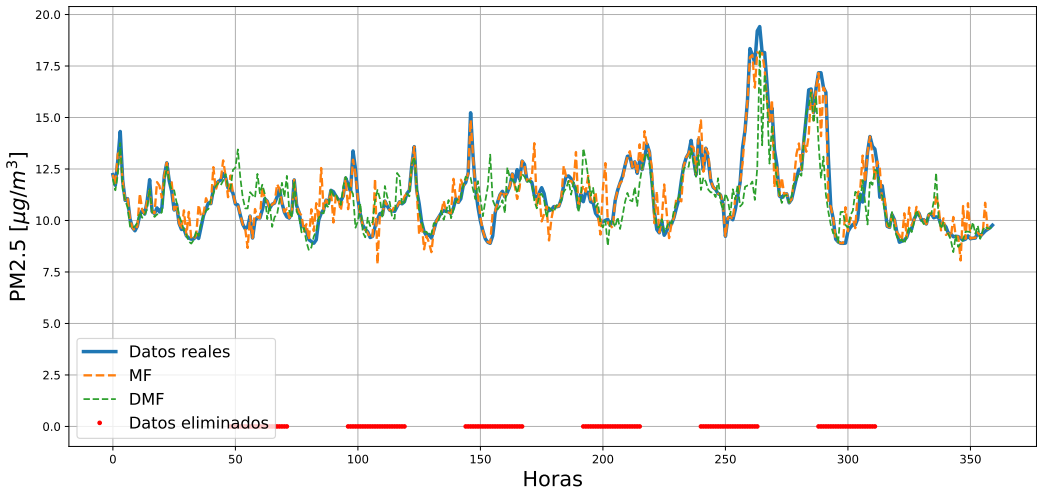


Figura 5.4: Comparación de desempeño entre el modelo DMF (DMF1) y el algoritmo MF + Regularización + Bias (S) usando el nodo 5 de la WSN

Modelo	RMSE	MAPE
MF + Regularización + Bias (S)	0,8558	0,0780
DMF1	0,8037	0,0640

Tabla 5.1: Errores obtenidos por los algoritmos DMF1 vs MF + Regularización + Bias (S) para el Nodo 5 de la WSN

5.2. Planteamiento de arquitecturas DMF con características espacio - temporales y evaluación de desempeño

5.2.1. Prueba de hipótesis nula (H_o) test de Friedman

El test de Friedman es una prueba de hipótesis nula (H_o) no paramétrica implementada frecuentemente para evaluar diferencias significativas entre algoritmos de ML. Este test se plantea para evaluar varios métodos frente a diferentes bases de datos (test sobre rankings).

Para el caso de la WSNs Ciudadanos Científicos esta prueba permitirá evaluar los métodos de DMF planteados para diferentes series de tiempo proceden-

tes de distintos sectores del Área Metropolitana de Valle de Aburrá [19, 80].

En este sentido la H_o plantea que no hay diferencias significativas entre los algoritmos para los conjuntos de datos considerados. Eso quiere decir entonces, que los rankings promedios deberían ser equivalentes e iguales al ranking promedio si todos empatan siempre [19]. Para esto se evalúa la siguiente expresión:

$$X_F^2 = \frac{12N}{k_m(k_m+1)}\left[\sum_j R_j^2 - \frac{k_m(k_m+1)^2}{4}\right] \tag{5.13}$$

donde N representa el número de bases de datos (nodos de sensores) ; k_m es número de métodos DMF a evaluar y R_j es el ranking del j-ésimo método a evaluar.

5.2.2. Diagramas de distancia critica usando test de Nemenyi

La comparación de desempeño entre algoritmos evaluados para una o varias bases de datos (DB) nos ayudan a determinar el rendimiento significativo, y por lo tanto, cuál de ellos obtiene mejor desempeño. Sin embargo, en muchos artículos presentados en la literatura también es necesario realizar pruebas de significancia estadística o pruebas post-hoc, como lo es el test de Nemenyi y el uso de diagramas de Distancia Crítica (CD) [30].

A grandes rasgos el test de Nemenyi obtiene la diferencia promedio entre dos métodos, esto quiere decir que esta prueba post-hoc se basa en la creación de un umbral por encima del cual, la diferencia entre las dos o varias medias aritméticas será significativa y por debajo del cual, esa diferencia no será estadísticamente significativa. Por lo tanto, el objetivo del diagrama de CD usando el test de Nemenyi será mostrar las comparaciones en pares y ubicar posicionalmente el ranking de los mejores métodos, dando una idea del nivel de significancia en relación al número de pruebas estadísticas realizadas simultáneamente sobre un conjunto de datos (test de comparaciones múltiples) [54, 30].

Eso quiere decir que al rechazar la H_o es posible aplicar el test post-hoc de Nemenyi, con el cual será posible determinar que dos métodos son significativamente diferentes, si sus rankings promedios difieren al menos en la distancia crítica de Nemenyi. Esta puede ser obtenida de acuerdo con la siguiente expresión:

$$CD = q_\alpha \sqrt{\frac{k_m(k_m - 1)}{6N}} \tag{5.14}$$

5.2.3. Arquitecturas DMF propuestas con características espacio - temporales

Debido a las facilidades de adaptación de las ANNs para aprender las representaciones desde cero y la adición de Embedding Layers para la inclusión de información espacial y temporal [91, 24, 97]. Se presenta el diseño de tres modelos de DMF, los cuales permitan complementar el modelo MF con dicha información y en consecuencia, resolver la problemática planteada para el presente trabajo, donde buscamos aprovechar las virtudes de las técnicas MF señaladas anteriormente con la finalidad de llegar a enfoques como el tratado por [23] para la reconstrucción de señales con gran cantidad de información perdida, como es el caso de la WSN de bajo costo Ciudadanos Científicos.

Día Semana	Codificación
Lunes	1
Martes	2
Miércoles	3
Jueves	4
Viernes	5
Sábado	6
Domingo	7

(a) Día de la semana.

Estado Día	Codificación
Madrugada	1
Mañana	2
Tarde	3
Noche	4

(b) Estado del día.

Tabla 5.2: Codificación temporal para la adición de características al modelo DMF2

En la Figura 5.5 se presenta la modificación planteada a la red DMF1 con la inclusión de características temporales, las cuales son decodificadas como se muestra en la Tabla 5.2. Donde se espera con la inclusión de estas, tener presente durante el entrenamiento del modelo, las variaciones de las contaminación de material particulado PM2.5 en cada uno de los días de la semana y en cada uno de los momentos del día.

Por ejemplo, es de esperarse que el material particulado sea más alto un día lunes comparado con un domingo (día festivo). Igualmente se espera que el comportamiento de la contaminación varíe drásticamente entre la madrugada del día lunes, el medio día y la tarde del mismo lunes.

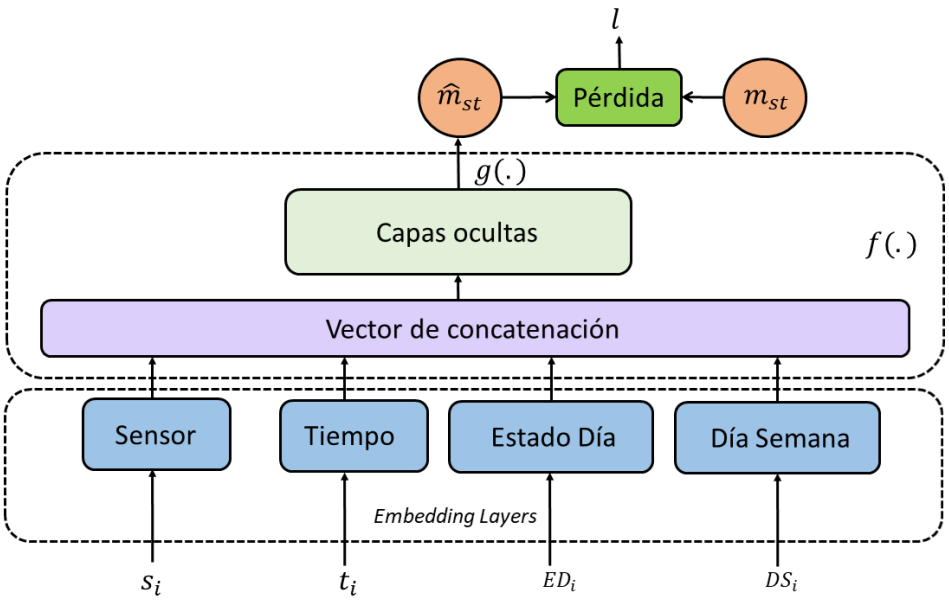


Figura 5.5: Modelo DMF2 con la inclusión de información temporales (días de la semana y estado del día)

Siguiendo con el planteamiento de modificaciones al modelo DMF1, en la Figura 5.6 se presenta el diseño de dos Redes Embedding para la inclusión de características espaciales (latitud y longitud) procedentes de cada nodo de la WSNs Ciudadanos Científicos.

Dado que las Redes Embedding solo reciben valores enteros en su entrada, se resuelve para la construcción de los nuevos índices realizar una división del Valle de Aburrá en una cuadrícula discreta (ver Figura 5.7), de forma tal, que dentro de cada cuadro se ubiquen los sensores con coordenadas espaciales similares.

El planteamiento de la cuadrícula de la Figura 5.7 se realiza dado que en la ciudad las dinámicas de la contaminación difieren de acuerdo a la ubicación del nodo. Por ejemplo, se espera que un nodo ubicado al lado de una avenida concurrida o ubicado en una zona céntrica del Valle de Aburrá, genere datos de mediciones de contaminación más altos, que un nodo ubicado en la periferia del Valle donde las avenidas poseen menos flujo de personas y de vehículos.

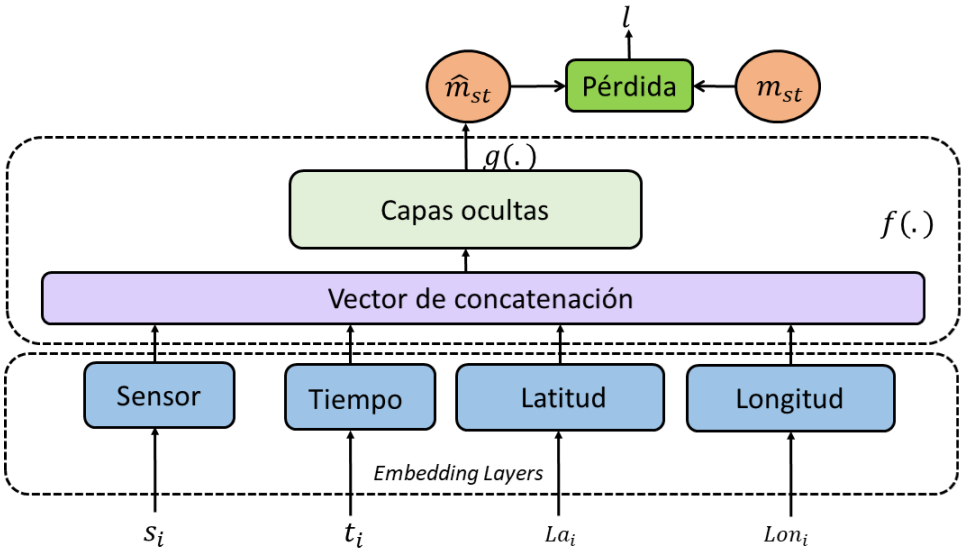


Figura 5.6: Modelo DMF3 con la inclusión de información espaciales (longitud y latitud)

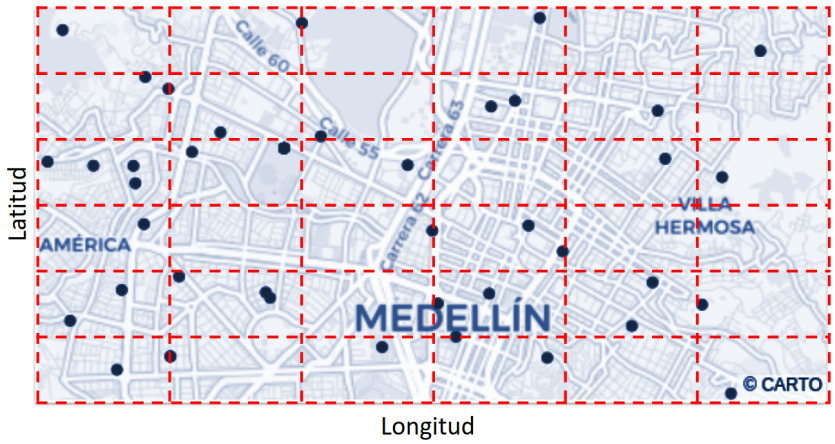


Figura 5.7: Cuadrícula espacial

Finalmente, en la Figura 5.8 se presenta el último planteamiento de modificación al modelo DMF1. Para dicho modelo fueron incluidos tanto los índices presentados en la Tabla 5.2 como los índices generados por la cuadrícula discreta presentada en la Figura 5.7. Por lo tanto, se espera que la inclusión de ambas características al modelo, ayuden a la hora de realizar las estimaciones de los datos faltantes generados por la WSN de bajo costo Ciudadanos Científicos.

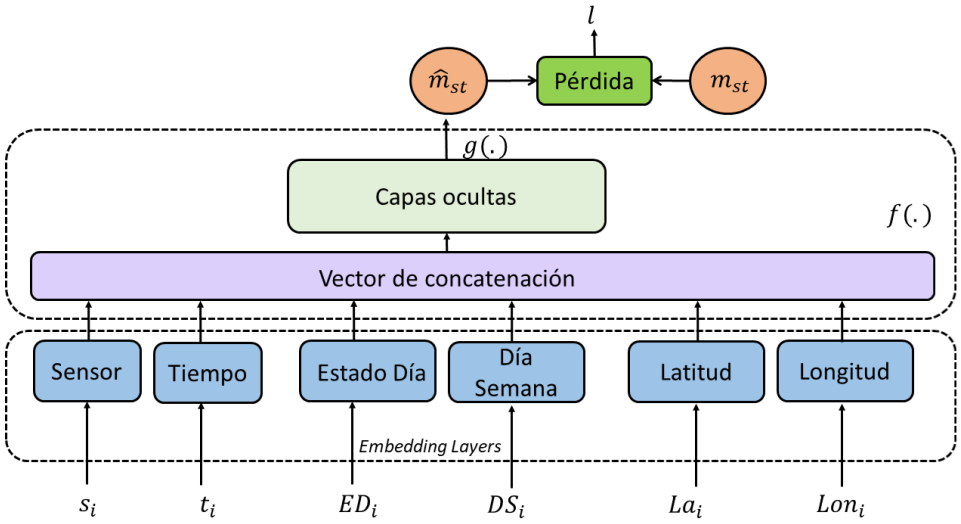


Figura 5.8: Modelo DMF4 con la inclusión de información espacio - temporales

5.2.4. Metodología de evaluación para el desempeño de las arquitecturas DMF propuestas

La evaluación de desempeño de los algoritmos DMF propuestos se realizó analizando el comportamiento de cada modelo en todas las series de tiempo generadas por los nodos de la WSN (se asumen las series de tiempo como bases de datos), siguiendo como metodología la evaluación por rankings. Para esto se da una posición a cada modelo de acuerdo al error generado entre un dato eliminado de forma aleatoria por huecos y el dato real generado por la red.

Sin embargo, el desempeño de los algoritmos fue definido por una serie de pruebas estadísticas para determinar la existencia de diferencias entre el desempeño de cada variación del modelo propuesto (test de Friedman) y que tanto difiere una técnica de otra (test de Nemenyi).

De acuerdo a lo anterior, para la aplicación de cada una de las pruebas estadísticas fue necesario transformar la base de datos generada por la WSN. En la Figura 5.9 se plantean los experimentos con la información obtenida con la red durante el año 2018, para esto fueron seleccionados los primeros 6 meses del año dado que en estos se encuentra la mayor cantidad de información perdida por la red y posibilitaba la experimentación (eliminación de información artificialmente), de forma tal, que fuera posible eliminar el 40 % de la información y de esta manera evaluar el desempeño de cada algoritmo DMF propuesto.

Por lo tanto, fueron creadas 6 bases de datos con información real de la WSN,

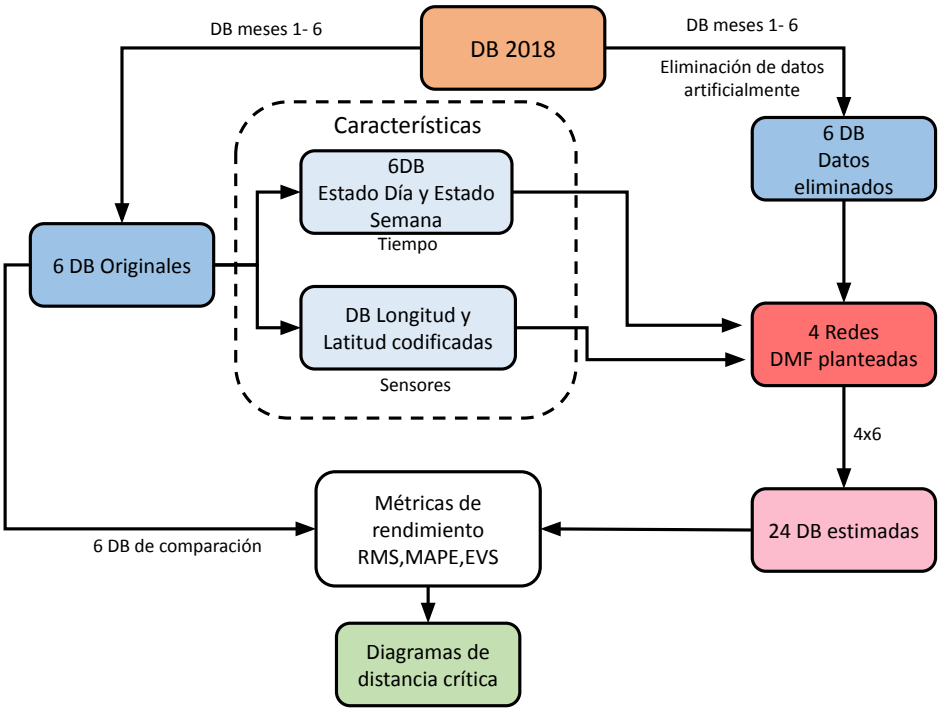
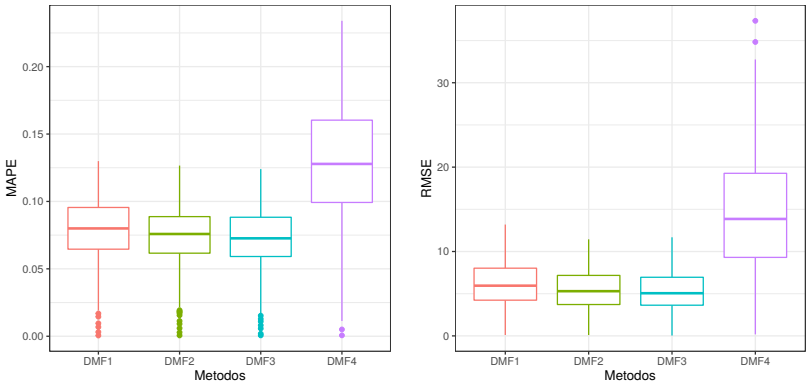


Figura 5.9: Diseño experimental para la evaluación de los algoritmos DMF planteados

6 bases de datos con información eliminada por huecos, 6 bases de datos con la información temporal codificada para cada mes y una base de datos con la información espacial de cada sensor. La información generada sirvió para evaluar la información de cada mes con los 4 modelos DMF propuestos, de las cuales fueron generadas 24 bases de datos con los resultados obtenidos por cada método DMF evaluado (estimación de información eliminada).

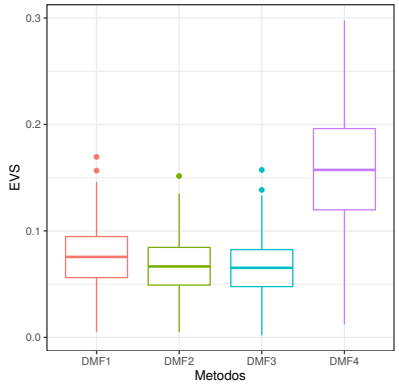
Finalmente, el desempeño del error en la estimación de información de cada algoritmo, se cotejó con las 6 bases de datos con información real, evaluando el desempeño de cada método con los índices de desempeño RMSE, MAPE y EVS. Consecuente con lo anterior, los errores generados en cada mes fueron promediados, generando 3 bases de datos (una por índice de desempeño) de tamaño 130×4, donde, 130 representa cada sensor de la red y 4 cada uno de los modelos DMF evaluados.

En la Figura 5.10 se puede observar la variación de los errores obtenidos por cada modelo DMF planteado ante cada ítem de desempeño analizado. Por ultimo, los datos contenidos dentro de cada matriz fueron convertidos a valores



(a) Diagrama de cajas para el método MAPE

(b) Diagrama de cajas para el método RMSE



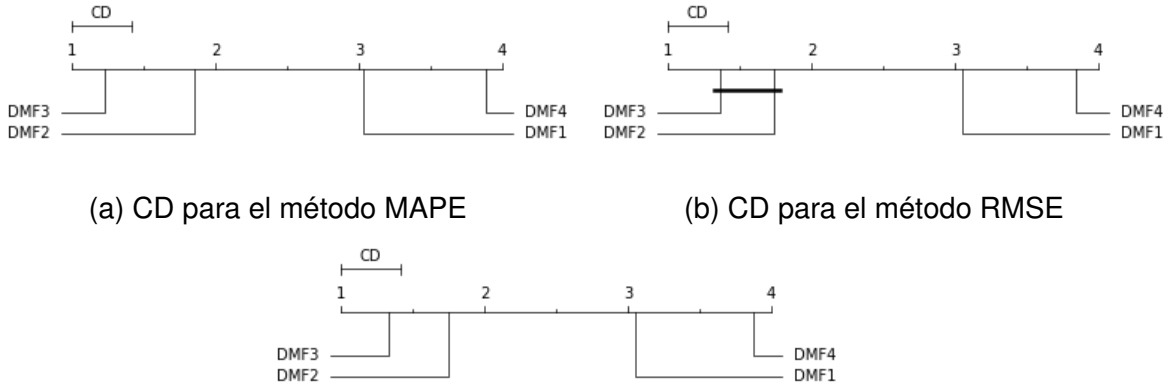
(c) Diagrama de cajas para el método EVS

Figura 5.10: Desviación de los errores obtenidos por cada método DMF evaluado

de 1 a 4, lo que quiere decir que el modelo con menor error le es asignado un valor de 1 y al peor de los modelos se le asigna un valor de 4.

5.2.5. Resultados evaluación de desempeño

De acuerdo con lo mencionado en las secciones anteriores, se realizó el test de Friedman para determinar si los métodos evaluados son estadísticamente diferentes. Esto quiere decir que se asume como H_o: Los métodos son equivalentes.Dado que el estadístico de Friedman es sumamente inflexible, utilizamos la variante de Iman and Davenport [29] donde un estadístico F corregido que se distribuye según la distribución F con $\nu_1 = 3$ and $\nu_2 = 387$ puede derivar de $k_m - 1$ grados de libertad del nominador y ν_1 mientras $(k_m - 1)(N - 1)$ grados de liber-



(a) CD para el método MAPE

(b) CD para el método RMSE

(c) CD para el método EVS

Figura 5.11: Diagramas de distancia crítica aplicando el test de Nemenyi

tad para el denominador ν_2. Siendo k_m y N el número de métodos y sensores respectivamente. En consecuencia y teniendo en cuenta un nivel de significación $\alpha = 0,05$ es posible encontrar el valor crítico $\chi^2_F = 2,63$. Por lo tanto, tras calcular los valores estadísticos de las tres medidas de regresión, se obtienen los siguientes valores de F-score $512,73$, $705,78$ y $585,23$ para RMSE, MAPE y EVS respectivamente, donde se dice entonces se rechaza la hipótesis nula H_o dado que $\chi^2_F <$ F-score.

Finalmente, usando la Eq.(5.14) y con $q_{0,05} = 2{,}569$ se obtienen las representaciones gráficas presentadas en la Figura 5.11, las cuales entregan un valor de CD $= 0{,}4114$. En consecuencia de esto, de los diagramas gráficos se obtiene la comparación de cuatro algoritmos DMF, entre los cuales se comparan los resultados obtenidos por el algoritmo DMF1 basado en [17, 92] y las tres variaciones planteadas acordes con las 130 bases de datos evaluadas (nodos de sensores), donde se incluye información espaciales (latitud y longitud), información temporal (estado del día y día de la semana) y la combinación espacio-temporal antes mencionadas.

Dicha prueba se realizó para tres métodos de regresión usadas en la literatura: MAPE, RMSE y EVS. De la cual se obtiene para la gráfica MAPE que el método DMF3 fue el mejor algoritmo frente al método base (DMF1) y las otras variaciones propuestas DMF2 y DMF4. Resultado similar se obtuvo con la métrica EVS, donde también se puede concluir que existe una diferencia significativa entre el algoritmo DMF3 y los algoritmos DMF 1, DMF2 y DMF4. Sin embargo, para la métrica RMSE, se obtiene que no existe diferencia significativa entre DMF3 y DMF2, agrupadas por la línea gruesa, pero sí existe diferencia significativa entre

los modelos agrupados (DMF3 y DMF2) con los modelos DMF1 y DMF4.

De acuerdo con lo anterior, se puede concluir que para la métricas MAPE y EVS, el algoritmo DMF3 tiene una probabilidad del 95 $\%$ de obtener un mejor desempeño respecto a los métodos DMF1, DMF2 y DMF4 para las 130 bases de datos evaluadas, dado que para la presentación de resultados el test se realizó con un α de 0,05.

5.3. Estimación de datos perdidos usando DMF3

Como experimento final se realiza una muestra de estimación de datos faltantes usando el modelo DMF3 que mejor resultado presentó en evaluación realizada en la subsección 5.2.5. Para esto fueron seleccionados nodos de sensores de forma aleatoria entre los meses evaluados: Enero, Febrero, Marzo, Abril, Mayo y Junio, usando una longitud de ventana de 720 horas (un mes). En las Figuras 5.12,5.13,5.14,5.15,5.16 y 5.17, en rojo se presentan las gráficas de la serie de tiempo sin recuperación de información y en azul se presentan las formas de señal generadas por el modelo DMF3 para la estimación de los datos no observados por la WSN en uno o varios instantes de tiempo.

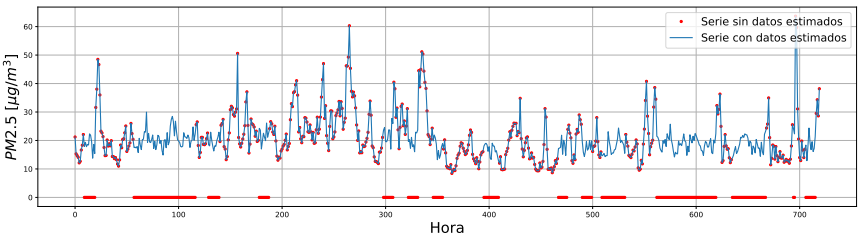


Figura 5.12: Nodo 10 mes de Enero 2018

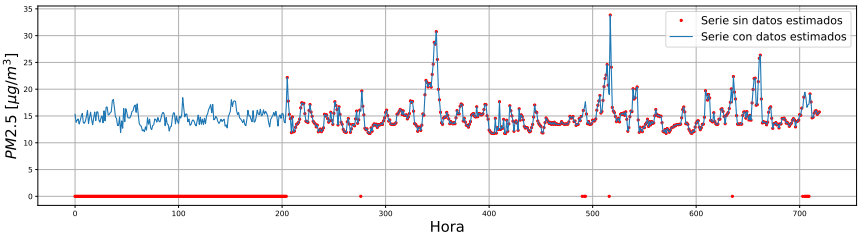


Figura 5.13: Nodo 108 mes de Abril 2018

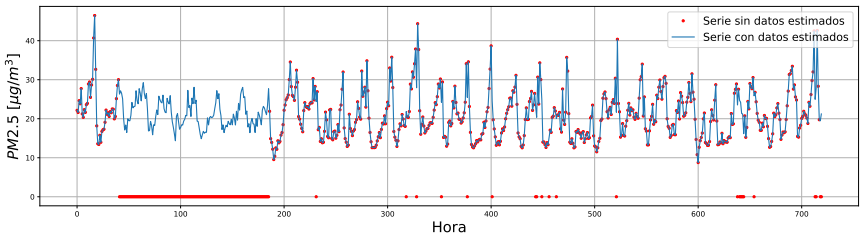


Figura 5.14: Nodo 52 mes de Junio 2018

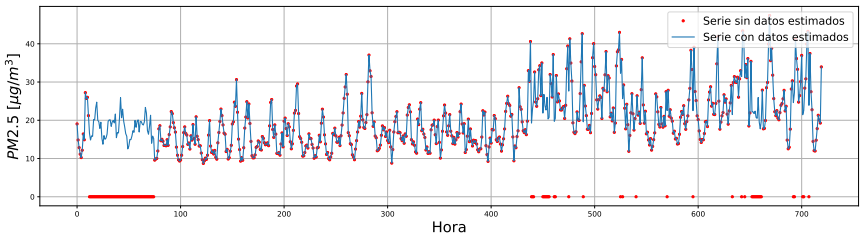


Figura 5.15: Nodo 34 mes de Mayo 2018

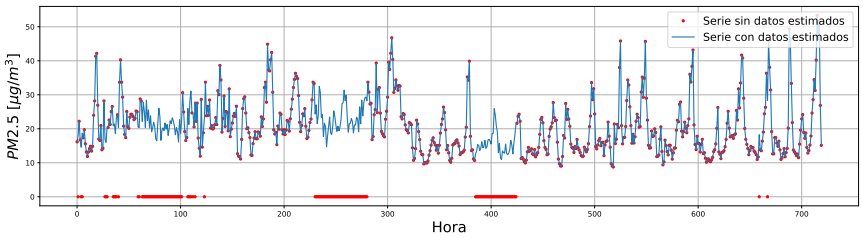


Figura 5.16: Nodo 12 mes de Enero 2018

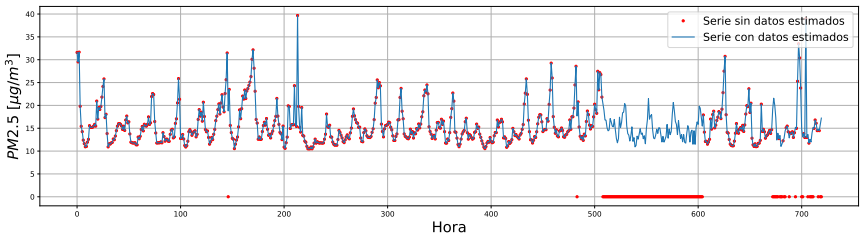


Figura 5.17: Nodo 49 mes de Febrero 2018

Parte III

Conclusiones y líneas futuras.

Conclusiones y líneas futuras

5.4. Conclusiones

- Se diseña una función de optimización que permite a partir de la red de sensores ajustar y completar datos faltantes con el fin de complementar las metodologías de medición. Para esto en los capítulos 3 y 4 se presenta un modelo basado en MF (algoritmos de ML) para la estimación de datos faltantes en una WSN de bajo costo que mide material particulado PM2.5. De la experimentación realizada con los algoritmos evaluados y con los datos de la WSN dispersa, se encontró que el modelo MF + Regularización + Bias (sintonizado) presenta un rendimiento superior a los resultados obtenidos con las técnicas MS y KNN (ver Tabla 4.2). Adicionalmente, se puede concluir que los MOGPs son métodos con gran rendimiento para la estimación de datos faltantes, sin embargo, fallan cuando la cantidad de información proveniente de los nodos vecinos es altamente dispersa, lo que imposibilita la convergencia del modelo (ver Figuras 4.4 y 4.5). En consecuencia, de la Figura 4.6 se puede deducir que el modelo MF + Regularización + Bias (sintonizado) supera al modelo MOGPs, dado que este logra convergencia con gran cantidad de información faltante logrando con esto el cumplimiento del primer objetivo específico planteado.

- Por otro lado, fue propuesto un método para la sintonización de los parámetros del modelo garantizando un buen ajuste de los datos sin pérdida de generalización. Para esto en el capítulo 4 se realizó la exploración de la mejor configuración de parámetros para el modelo de optimización y un análisis de sensibilidad del algoritmo MF ante diferentes porcentajes de datos faltantes. En la Tabla 4.1 se reconoce claramente que el modelo MF + Regularización + Bias (sintonizado) presenta una mejora de desempeño frente a los resultados obtenidos en el capítulo 3. Con lo anterior, se puede concluir que los modelos MF responden al problema de recuperación de datos para la WSN bajo costo Ciudadanos Científicos, manteniendo un

buen ajuste de los datos sin pérdida de generalización con la inclusión de parámetros de bias y regularización [37]. Esto puede ser corroborado en la Tabla 4.2), donde el modelo MF + Regularización + Bias (Sintonizado) presentó menores errores de estimación con diferentes patrones de información perdida en la reconstrucción de las señales (presentan diferentes perfiles de contaminación y ubicación en el Valle de Aburrá) respecto a los algoritmos evaluados del estado del arte.

- Luego de lo anterior fue desarrollado un procedimiento con el fin de evaluar la calidad del ajuste dado por el modelo propuesto con respecto a lecturas de referencia. Por lo tanto se planteó un procedimiento para la evaluación de los modelos comparados, realizando eliminación aleatoria y por huecos de información existente de acuerdo con los patrones de datos tratados en la subsección 2.1.3 y desarrollados en los experimentos realizados en los capítulos 3, 4 y 5. Se implementaron los métodos de evaluación de error MAPE, RMSE y EVS para valorar el error de ajuste de cada modelo implementado (ver Figuras 4.6 y 5.3), test estadísticos de H_o y pruebas post-hoc (ver Figura 5.11). De las Tablas 4.2 y 5.1 se puede concluir de la experimentación realizada, que los modelos basados en MF presentan mejores resultados respecto a los modelos MS, KNN y MOGPs descritos en la literatura. Encontrando que la aplicación de algoritmos MF implementando redes neuronales (DMF), representa claramente un enfoque que permite mejorar la capacidad de aprender las representaciones de características desde cero [97] y al mismo tiempo contar con la interpretación matemática de la MF [67, 18].

  Lo anterior se logra al plantear modelos DMF usando Embedding Layers que permiten incluir información que se adapta a los datos generados por la WSN Ciudadanos Científicos, representando una mejora de desempeño significativa comparada con los algoritmos MF estándar (ver Figura 5.3). Esto se puede corroborar en los enfoques DMF2 (inclusión de información temporal) y DMF3 (información espacial), donde se evidencia que para la red Ciudadanos Científicos es más relevante la información espacial que la temporal (ver Figura 5.11). Sin embargo con la experimentación realizada, se encontró que la inclusión de información espacio-temporal no represento mejoras en el desempeño del modelo para la estimación de información perdida (DMF4). Por lo tanto, queda abierto el planteamiento de nuevas configuración de red que permitan mejorar el desempeño del algoritmo, ya sea incluyendo nueva información (información atmosférica) o la inclusión de información espacial y temporal.

- Finalmente, se puede concluir que el modelo DMF3 es un buen candida-

to para la solución de problemas de estimación de información perdida y la reconstrucción de señales de PM2.5, logrando con esto el cumplimiento del objetivo principal planteado en este trabajo “Diseñar una metodología que permita mejorar la calidad de medición de material particulado PM2.5 de la red de sensores de bajo costo del proyecto Ciudadanos Científicos en el Valle de Aburrá, utilizando algoritmos de aprendizaje de máquina”. Lo anterior se puede verificar en la subsección 5.3 donde se observa que las predicciones realizadas por el modelo, cuando se cuenta con patrones de huecos prolongados en el tiempo, buscan mantener el periodo de la señal (ver Figura 5.14), lo cual es un resultado esperado, ya que al aumentar el tiempo de predicción aumenta el error de estimación, y por lo tanto, no se espera que el modelo DMF3 sea capaz de estimar picos máximos o mínimos de la señal, que de acuerdo con las señales vistas en la subsección 5.3 podrían representar eventos espontáneos o atípicos en los perfiles de contaminación de la ciudad.

5.5. Lineas futuras

Luego del desarrollo del presente trabajo se plantean 3 escenarios generales de trabajo futuro:

- El primero consiste en seguir evaluando la base de datos Ciudadanos Científicos con las variaciones existentes en la literatura de modelos MF, planteando dentro de estos, la inclusión de información de variables ambientales como: velocidad del viento, temperatura y humedad. Este desarrollo queda abierto ya que no se consideró para el alcance del trabajo y de acuerdo con la literatura, son factores que afectan considerablemente la respuesta de los sensores de material particulado PM2.5 usados por la WSN Ciudadanos Científicos.
- En el segundo escenario se plantea la implementación de técnicas MF para el ajuste de sensores de la red Ciudadanos Científicos. Para esto se propone usar metodologías Blind Calibration o Transfer Calibration, ya que de acuerdo con la literatura, dichas técnicas permiten ajustar los sensores usando la dinámica de la propia red, posibilitando la inclusión dentro del modelo características temporales, espaciales y ambientales. Lo cual conlleve a un ajuste lo suficientemente robusto como para mejorar la respuesta de los sensores ante ambientes y condiciones no controladas.

- Finalmente, para la tercera línea de trabajo se plantea el uso de los resultados de este trabajo, estimación de datos no muestreados, en la aplicación de técnicas de predicción como Redes Neuronales Recurrentes (RNN) y modelos LSTM. Avanzando un paso adelante dentro del campo del procesamiento de señales, proponiendo metodologías que permitan estimar episodios de contaminación, los cuales sirvan como complemento en temas de investigación como: pronóstico de eventos de contaminación, planeación de ciudad, toma de decisiones en los planes de ordenamiento territorial, reportes a la ciudadanía, etc.

Bibliografía

[1] Erika von Schneidemesser Alastair C. Lewis and Richard E. Peltier. Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications. Technical report, World Meteorological Organization, 2018.

[2] A. Kofi Amegah. Proliferation of low-cost sensors. What prospects for air pollution epidemiologic research in Sub-Saharan Africa? *Environmental Pollution*, 241:1132–1137, 2018.

[3] Jose M. Barcelo-Ordinas, Messaud Doudou, Jorge Garcia-Vidal, and Nadjib Badache. Self-calibration methods for uncontrolled environments in sensor networks: A reference survey. *Ad Hoc Networks*, 2019.

[4] Ronan Baron and John Saffell. Amperometric gas sensors as a low cost emerging technology platform for air quality monitoring applications: A review. *ACS Sensors*, 2(11):1553–1566, 2017. PMID: 29025261.

[5] Juan Botero-Valencia, Luis Castano-Londono, David Marquez-Viloria, and Mateo Rico-Garcia. Data reduction in a low-cost environmental monitoring system based on LoRa for WSN. *IEEE Internet of Things Journal*, PP(X):1, 2018.

[6] Andrés M. Cárdenas, León M. Rivera, Beatriz L. Gómez, Germán M. Valencia, Hernán A. Acosta, and Juan D. Correa. Short communication: Pollution-and-greenhouse gases measurement system. *Measurement*, 129:565 – 568, 2018.

[7] Nuria Castell, Hai-ying Liu, Franck R Dauge, and Mike Kobernus. Supporting sustainable mobility using mobile technologies and personalized environmental information: The citi-sense- mob approach in oslo, norway. *Environmental Informatics*, (June), 2018.

[8] Mahesh Chandra Mukkamala and Matthias Hein. Variants of RMSProp and Adagrad with Logarithmic Regret Bounds. Technical report, 2017.

[9] Jie Cheng, Qiang Ye, Hongbo Jiang, Dan Wang, and Chonggang Wang. Stcdg: An efficient data gathering algorithm based on matrix completion for wireless sensor networks. *Wireless Communications, IEEE Transactions on*, 12:850–861, 02 2013.

[10] E. S. Cross, L. R. Williams, D. K. Lewis, G. R. Magoon, T. B. Onasch, M. L. Kaminsky, D. R. Worsnop, and J. T. Jayne. Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements. *Atmospheric Measurement Techniques*, 10(9):3575–3588, 2017.

[11] T. de Wolff, A. Cuevas, and F. Tobar. MOGPTK: The Multi-Output Gaussian Process Toolkit. *arXiv e-prints*, page arXiv:2002.03471, 2020.

[12] Clément Dorffer, Matthieu Puigt, Gilles Delmaire, and Gilles Roussel. Blind calibration of mobile sensors using informed nonnegative matrix factorization. In Emmanuel Vincent, Arie Yeredor, Zbyněk Koldovský, and Petr Tichavský, editors, *Latent Variable Analysis and Signal Separation*, pages 497–505, Cham, 2015. Springer International Publishing.

[13] Clement Dorffer, Matthieu Puigt, Gilles Delmaire, and Gilles Roussel. Nonlinear mobile sensor calibration using informed semi-nonnegative matrix factorization with a Vandermonde factor. *Proceedings of the IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2016-Septe, 2016.

[14] ONU enviromental World Meteorological Organization, IGAC. *Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications*. Organization, World Meteorological, Geneva 2„ 2018.

[15] EPA. Leyes y normas: el proceso de reglamentación.

[16] EPA. Spatial and Temporal Trends of Air Pollutants in the South Coast Basin Using Low Cost Sensors. Technical report, EPA, 2018.

[17] Jicong Fan and Jieyu Cheng. Matrix completion by deep matrix factorization. *Neural Networks*, 98:34 – 41, 2018.

[18] Jennifer Flenner, Blake Hunter, J Flenner, and B Hunter. A Deep Non-Negative Matrix Factorization Neural Network. Technical report.

[19] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[20] Jorge E. Gómez, Fabricio R. Marcillo, Freddy L. Triana, Victor T. Gallo, Byron W. Oviedo, and Velssy L. Hernández. IoT for ENVIRONMENTAL VARIABLES in URBAN AREAS. In *Procedia Computer Science*, 2017.

[21] Steven J. Hadeed, Mary Kay O'Rourke, Jefferey L. Burgess, Robin B. Harris, and Robert A. Canales. Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of The Total Environment*, 730:139140, 2020.

[22] Gayle S.W. Hagler, Ronald Williams, Vasileios Papapostolou, and Andrea Polidori. Air Quality Sensors and Data Adjustment Algorithms: When Is It No Longer a Measurement?, may 2018.

[23] Cecile Hautecoeur and Francois Glineur. Nonnegative matrix factorization over continuous signals using parametrizable functions. *Neurocomputing*, 2020.

[24] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat Seng Chua. Neural collaborative filtering. In *26th International World Wide Web Conference, WWW 2017*, pages 173–182. International World Wide Web Conferences Steering Committee, 2017.

[25] Yun He and De chang Pi. Improving knn method based on reduced relational grade for microarray missing values imputation. *IAENG International Journal of Computer Science*, 43(3):356–362, 2016.

[26] Oleksii Hrinchuk, Valentin Khrulkov, Leyla Mirvakhabova, Elena Orlova, and Ivan Oseledets. Tensorized embedding layers for efficient model compression, 2020.

[27] Qi Huang, Xuesong Yin, Songcan Chen, Yigang Wang, and Bowen Chen. Robust nonnegative matrix factorization with structure regularization. *Neurocomputing*, 2020.

[28] IDEAM. *La Variabilidad Climática Y El Cambio Climático En Colombia*. IDEAM, 2018.

[29] Ronald L Iman and James M Davenport. Approximations of the critical region of the fbietkan statistic. *Communications in Statistics-Theory and Methods*, 9(6):571–595, 1980.

[30] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, 2019.

[31] Ivan Izonin, Natalia Kryvinska, Roman Tkachenko, and Khrystyna Zub. An approach towards missing data recovery within iot smart system. *Procedia Computer Science*, 155:11 – 18, 2019. The 16th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2019),The 14th International Conference on Future Networks and Communications (FNC-2019),The 9th International Conference on Sustainable Energy Information Technology.

[32] Martin Jakomin and Tomas Curk. Simultaneous incremental matrix factorization for streaming recommender systems. *Expert Systems with Applications*, page 113685, 2020.

[33] K. E. Kelly, J. Whitaker, A. Petty, C. Widmer, A. Dybwad, D. Sleeth, R. Martin, and A. Butterfield. Ambient and laboratory evaluation of a low-cost particulate matter sensor. *Environmental Pollution*, 2017.

[34] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[35] L. Kong, D. Jiang, and M. Wu. Optimizing the spatio-temporal distribution of cyber-physical systems for environment abstraction. In *2010 IEEE 30th International Conference on Distributed Computing Systems*, pages 179–188, 2010.

[36] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[37] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix recommender techniques for factorization systems. pages 42–49, 2009.

[38] Prashant Kumar, Lidia Morawska, Claudio Martani, George Biskos, Marina Neophytou, Silvana Di Sabatino, Margaret Bell, Leslie Norford, and Rex Britter. The rise of low-cost sensing for managing air pollution in cities. *Environment International*, 75:199 – 205, 2015.

[39] Byung-tak Lee, Seung-chul Son, and Kyungran Kang. A Blind Calibration Scheme Exploiting Mutual Calibration Relationships for a Dense Mobile Sensor Network. *IEEE Sensors Journal*, 14(5):1518–1526, 2014.

[40] Adam Lerer, Ledell Wu, Jiajun Shen, Timothée Lacroix, Luca Wehrstedt, Abhijit Bose, and Alexander Peysakhovich. Pytorch-biggraph: A large-scale graph embedding system. *CoRR*, abs/1903.12287, 2019.

[41] Jiayu Li, Haoran Li, Yehan Ma, Yang Wang, Ahmed A. Abokifa, Chenyang Lu, and Pratim Biswas. Spatiotemporal distribution of indoor particulate matter concentration with a low-cost sensor network. *Building and Environment*, 127(November 2017):138–147, 2018.

[42] Chul-Hee Lim, Jieun Ryu, Yuyoung Choi, Seong Woo Jeon, and Woo-Kyun Lee. Understanding global pm2.5 concentrations and their drivers in recent decades (1998–2016). *Environment International*, 144:106011, 2020.

[43] Haitao Liu, Jianfei Cai, and Yew-Soon Ong. Remarks on multi-output gaussian process regression. *Knowledge-Based Systems*, 144:102 – 121, 2018.

[44] X. Liu, T. Xi, and E. Ngai. Data modelling with gaussian process in sensor networks for urban environmental monitoring. In *2016 IEEE 24th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 457–462, 2016.

[45] S.C. Candice Lung, Rod Jones, Christoph Zellweger, Ari Karppinen, Michele Penza, Tim Dye, Christoph Hüglin, Zhi Ning, Roland Leigh, David Hagan, Olivier Laurent, Greg Carmichael, Gufran Beig, Ron Cohen, Eben Cross, Drew Gentner, Michel Gerboles, Sean Khan, Pierpaolo Mudu, Xavier Querol Carceller, Giulia Ruggeri, Kate Smith, and Oksana Tarasova. *Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications*. Number May. World Meteorological Organization, 2018.

[46] Balz Maag, Zimu Zhou, and Lothar Thiele. A Survey on Sensor Calibration in Air Pollution Monitoring Deployments. *IEEE Internet of Things Journal*, 5(6):4857–4870, 2018.

[47] Mandana Mazaheri, Samuel Clifford, Bijan Yeganeh, Mar Viana, Valeria Rizza, Robin Flament, Giorgio Buonanno, and Lidia Morawska. Investigations into factors affecting personal exposure to particles in urban microenvironments using low-cost sensors. *Environment International*, 120(January):496–504, 2018.

[48] M.I. Mead, O.A.M. Popoola, G.B. Stewart, P. Landshoff, M. Calleja, M. Hayes, J.J. Baldovi, M.W. McLeod, T.F. Hodgson, J. Dicks, A. Lewis, J. Cohen, R. Baron, J.R. Saffell, and R.L. Jones. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmospheric Environment*, 70:186 – 203, 2013.

[49] Aleixandre Manuel; Gerboles Michel. Review of small commercial sensors for indicative monitoring of ambient gas. *Chemical Engineering Transactions*, 2012.

[50] Janos Mika, Peter Forgo, Laszlo Lakatos, Andras B. Olah, Sandor Rapi, and Zoltan Utasi. Impact of 1.5 K global warming on urban air pollution and heat island with outlook on human health effects, 2018.

[51] MINAMBIENTE. Protocolo para el monitoreo y seguimiento de la calidad del aire. Technical report, Ministerio de Ambiente, Vivienda y Desarrollo Territorial, Bogotá D.C, 2008.

[52] Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2008.

[53] Steffen Moritz and Thomas Bartz-Beielstein. imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, 9(1):207–218, 2017.

[54] P. Nemenyi. *Distribution-free Multiple Comparisons*. Princeton University, 1963.

[55] World Health Organization. Ambient (outdoor) air pollution. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health, may 2018.

[56] World Health Organization. The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, may 2018.

[57] Min-Bin Park, Tae-Jung Lee, Eun-Sun Lee, and Dong-Sool Kim. Enhancing source identification of hourly PM2.5 data in Seoul based on a dataset segmentation scheme by positive matrix factorization (PMF). *Atmospheric Pollution Research*, 2019.

[58] Sameer Patel, Jiayu Li, Apoorva Pandey, Shamsh Pervez, Rajan K. Chakrabarty, and Pratim Biswas. Spatio-temporal measurement of indoor particulate matter concentrations using a wireless network of low-cost sensors in households using solid fuels. *Environmental Research*, 2017.

[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[60] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge

tracing. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 505–513. Curran Associates, Inc., 2015.

[61] Boris Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4:1–17, 12 1964.

[62] Zihan Ran, Yanpeng An, Ji Zhou, Jingmin Yang, Youyi Zhang, Jingcheng Yang, Lei Wang, Xin Li, Daru Lu, Jiang Zhong, Huaidong Song, Xingjun Qin, and Rui Li. Subchronic exposure to concentrated ambient pm2.5 perturbs gut and lung microbiota as well as metabolic profiles in mice. *Environmental Pollution*, page 115987, 2020.

[63] Carl Edward Rasmussen. LNAI 3176 - Gaussian Processes in Machine Learning. Technical report, 2003.

[64] Mariusz Rogulski. Indoor PM10 concentration measurements using low-cost monitors in selected locations in Warsaw. In *Energy Procedia*, 2018.

[65] B Ross. *Non-Negative Matrix Factorization Techniques and Optimizations*. Springer, 2008.

[66] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[67] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6655–6659, 2013.

[68] T. Sayahi, A. Butterfield, and K. E. Kelly. Long-term field evaluation of the Plantower PMS low-cost particulate matter sensors. *Environmental Pollution*, pages 932–940, 2019.

[69] Philipp Schneider, Nuria Castell, Matthias Vogt, Franck R. Dauge, William A. Lahoz, and Alena Bartonova. Mapping urban air quality in near real-time using observations from low-cost sensors and model information. *Environment International*, 106(December 2016):234–247, 2017.

[70] Jalpa Shah and Biswajit Mishra. Iot-enabled low power environment monitoring system for prediction of pm2.5. *Pervasive and Mobile Computing*, 67:101175, 2020.

[71] Amit Kumar Sharma, Sandeep Chaurasia, and Devesh Kumar Srivastava. Sentimental short sentences classification by using cnn deep learning model with fine tuned word2vec. *Procedia Computer Science*, 167:1139 – 1147, 2020. International Conference on Computational Intelligence and Data Science.

[72] Chetan Shetty, Sowmya B J, Seema Shedole, and K. Srinivasa. *Air pollution control model using machine learning and IoT techniques*. 01 2019.

[73] Chetan Shetty, B.J. Sowmya, S. Seema, and K.G. Srinivasa. Chapter eight - air pollution control model using machine learning and iot techniques. In Pethuru Raj and Preetha Evangeline, editors, *The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases*, volume 117 of *Advances in Computers*, pages 187 – 218. Elsevier, 2020.

[74] Weiwei Shi, Yongxin Zhu, Philip S. Yu, Tian Huang, Chang Wang, Yishu Mao, and Yufeng Chen. Temporal Dynamic Matrix Factorization for Missing Data Prediction in Large Scale Coevolving Time Series. *IEEE Access*, 4(c):6719–6732, 2016.

[75] SIATA. Geoportal. https://siata.gov.co, feb 2020.

[76] Laurent Spinelle, Michel Gerboles, Maria Gabriella Villani, Manuel Aleixandre, and Fausto Bonavitacola. Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide. *Sensors and Actuators, B: Chemical*, 215:249–257, 2015.

[77] Abdulhamit Subasi. Chapter 3 - machine learning techniques. In Abdulhamit Subasi, editor, *Practical Machine Learning for Data Analysis Using Python*, pages 91 – 202. Academic Press, 2020.

[78] Gábor Takács, Istvan Pilaszy, Bottyan Nemeth, and Domonkos Tikk. Matrix Factorization and Neighbor Based Algorithms for the Netflix Prize Problem Categories and Subject Descriptors. *Proceedings of the 2008 ACM conference on Recommender systems - RecSys '08*, pages 267–274, 2008.

[79] Ashay Tamhane, Sagar Arora, and Deepak Warrier. Modeling contextual changes in user behaviour in fashion e-commerce. In Jinho Kim, Kyuseok Shim, Longbing Cao, Jae-Gil Lee, Xuemin Lin, and Yang-Sae Moon, editors, *Advances in Knowledge Discovery and Data Mining*, pages 539–550, Cham, 2017. Springer International Publishing.

[80] Maksim Terpilowski. scikit-posthocs: Pairwise multiple comparison tests in python. *The Journal of Open Source Software*, 4(36):1169, 2019.

[81] UNAL and IDEAM. *CAUSAS DE DEGRADACIÓN FORESTAL EN COLOMBIA: DEGRADACIÓN: una primera aproximación*. IDEAM, Bogotá D.C, 2018.

[82] US EPA. Air Quality Index (AQI) Basics.

[83] Emmanuel Vincent, Arie Yeredor, Zbyněk Koldovský, and Petr Tichavský. Blind Calibration of Mobile Sensors Using Informed Nonnegative Matrix Factorization. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9237(March 2018), 2015.

[84] Jie Wang and Jun Zhang. Addressing accuracy issues in privacy preserving data mining through matrix factorization. *ISI 2007: 2007 IEEE Intelligence and Security Informatics*, pages 217–220, 2007.

[85] Yu Xiong Wang and Yu Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353, 2013.

[86] WHO. *Guías de calidad del aire de la OMS relativas al material particulado, el ozono, el dióxido de nitrógeno y el dióxido de azufre*. World Health Organization, 2006.

[87] World Health Organization (WHO). Evolution of who air quality guidelines: past, present and future. Technical report, WHO Regional Office for Europe, 2017.

[88] C Williams. Prediction with Gaussian processes. *Learning in graphical models*, pages 599–621, 1999.

[89] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. Chapter 2 - input: Concepts, instances, attributes. In Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal, editors, *Data Mining (Fourth Edition)*, pages 43 – 66. Morgan Kaufmann, fourth edition edition, 2017.

[90] Kun Xie, Xueping Ning, Xin Wang, Dongliang Xie, Jiannong Cao, Gaogang Xie, and Jigang Wen. Recover Corrupted Data in Sensor Networks: A Matrix Completion Solution. *IEEE Transactions on Mobile Computing*, 16(5):1434–1448, 2017.

[91] Hong-Jian Xue, Xin-Yu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Deep Matrix Factorization Models for Recommender Systems *. Technical report, 2017.

[92] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Deep matrix factorization models for recommender systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3203–3209, 2017.

[93] Xin-She Yang. 8 - neural networks and deep learning. In Xin-She Yang, editor, *Introduction to Algorithms for Data Mining and Machine Learning*, pages 139 – 161. Academic Press, 2019.

[94] Q. Yuan, Z. Liu, J. Li, S. Yang, and F. Yang. An adaptive and compressive data gathering scheme in vehicular sensor networks. In *2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*, pages 207–215, 2015.

[95] Alireza Zaeemzadeh, Mohsen Joneidi, Behzad Shahrasbi, and Nazanin Rahnavard. Missing spectrum-data recovery in cognitive radio networks using piecewise constant Nonnegative Matrix Factorization. *Proceedings - IEEE Military Communications Conference MILCOM*, 2015-December:238–243, 2015.

[96] Andrea Zanella, Nicola Bui, Angelo Castellani, Lorenzo Vangelista, and Michele Zorzi. Internet of things for smart cities. *IEEE Internet of Things Journal*, 2014.

[97] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives, feb 2019.

[98] Naomi Zimmerman, Albert A. Presto, Sriniwasa P.N. Kumar, Jason Gu, Aliaksei Hauryliuk, Ellis S. Robinson, Allen L. Robinson, and R. Subramanian. A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring. *Atmospheric Measurement Techniques*, 2018.