The University of Tehran Press

Pollution

Print ISSN: 2383-451X
Online ISSN: 2383-4501

https://jpoll.ut.ac.ir/

Comparative Study of Imputation Techniques for Missing Value Estimation in Particulate Matter 2.5 µm Time Series

Anibal Flores[1]✉ | Hugo Tito-Chura[1] | Alejandro Ecos-Espino[2] | Eduardo Flores-Quispe[3]

1. Departamento de Ingeniería de Sistemas e Informática, Universidad Nacional de Moquegua, Moquegua, Peru
2. Departmento de Ciencias Básicas, Universidad Nacional de Moquegua, Moquegua, Peru
3. Departmento de Ingeniería Ambiental, Universidad Nacional de Moquegua, Moquegua, Peru

Article Info

Article type:
Research Article

Article history:
Received: 20 March 2024
Revised: 28 June 2024
Accepted: 28 August 2024

Keywords:
Missing Value Estimation, Short-Gaps, Moving Averages, Interpolation, Deep Learning

ABSTRACT

Particulate matter 2.5 µm (PM2.5) or less in diameter is one of the most important air pollutants owing to its harmful effects on health. However, the measured data of PM2.5 in air quality monitoring networks may have large missing values owing to equipment failure. We conducted a comparative study of imputation techniques for missing value estimation in PM2.5, which was regularly measured in the air quality monitoring network in Lima City, Peru. Lima is the second most polluted city in South America. In this regard, various imputation techniques were implemented, among them, moving averages-based approaches (e.g., Autoregressive Integrated Moving Average ARIMA, Exponentially Weighted Moving Average EWMA, Linear Weighted Moving Average LWMA, and Local Average of Nearest Neighbors LANN), interpolation-based models (e.g., spline), and deep learning-based methods (e.g., Long Short-Term Memory LSTM, Bidirectional LSTM, Gated Recurrent Unit GRU, and Bidirectional GRU) to estimate missing values in PM2.5 time series. For experimentation, a dataset of 11822 h was used, considering 80% for training and the remaining 20% for testing. The results in terms of RMSE, MAPE, and R2 demonstrated that for different configurations of short-gaps of missing values, the techniques based on moving averages yielded better results than those based on deep learning. Among the moving average-based techniques, ARIMA was the best model for estimating missing values in PM2.5 time series, and the MAPE values ranged from 0.0005% to 11.6522%.

Cite this article: Flores, A., Tito-Chura, H., Ecos-Espino, A. & Flores-Quispe, E. (2024). Comparative Study of Imputation Techniques for Missing Value Estimation in Particulate Matter 2.5 µm Time Series. *Pollution*, 10 (4), 1117-1127. https://doi.org/10.22059/poll.2024.374178.2296

DOI: https://doi.org/10.22059/poll.2024.374178.2296

INTRODUCTION

Peru is a South American country with the worst air quality (Colorado, 2019). Thus, Lima is the second capital with the worst air quality after Santiago de Chile. A total of 12 cities in Peru have air pollution levels above those recommended (Republica, 2023) by the World Health Organization (WHO): Lima, Arequipa, Ica, Trujillo, Ucayali, Ancash, Apurimac, Cajamarca, Cusco, Madre de Dios, Moquegua, and Piura.

According to the World Bank, 79.6% of the energy consumed in Peru is derived from fossil fuels (RumboMinero, 2022), with an increasing trend. Another cause is aridity and lack of vegetation, which makes it easier for air currents to raise polluting particles. Air pollution, specifically PM2.5 and PM10, cause respiratory infections (Bu *et al.*, 2021), throat irritation (Wyer *et al.*, 2022)though there have been several studies which indicate that NH3 has a direct effect on the respiratory health of those who handle livestock. These health impacts can include

*Corresponding Author Email: *afloresg@unam.edu.pe*

a reduced lung function, irritation to the throat and eyes, and increased coughing and phlegm expulsion. More recent studies have indicated that agricultural NH3 may directly influence the early on-set of asthma in young children. In addition to the potential direct impact of ammonia, it is also a substantial contributor to the fine particulate matter (PM2.5, cough, heart disease (Oh *et al*., 2023), stroke (Chen *et al*., 2022), and lung cancer (Huang *et al*., 2017).

Thus, the study of particulate matter 2.5 µm (PM2.5) time series is significant because, based on them, prediction models can be implemented, allowing the corresponding authorities to make decisions to safeguard the health of the population (Priya & Khanaa, 2023)air pollution is becoming a very big issue in metropolitan cities due to the availability of vehicles. This air pollution has spread various breathing-related diseases to people vigorously in recent years. People are affected much more severely due to the lack of knowledge and updates about the air pollution level in the specific area and dates. To safeguard the people and to provide enough knowledge about the air pollution level, many researchers have developed many air qualities during COVID-19 in prediction systems using various classifiers. Even though no system resolves this issue due to the rapid growth of the population, automobile usage, and industries. To fulfil the current requirements, this paper proposes an Intelligent Air Quality During COVID-19 Prediction and Monitoring System (IAQPMS.

However, before implementing the predictive models, reviewing the quality of the data acquired by the monitoring stations is necessary, which in most cases presents missing values (NAs) with different gap sizes, small, medium, and large, owing to various reasons, such as power outages, equipment malfunctions (Moritz & Bartz-Beielstein, 2017), and human errors. Most models based on machine and deep learning require data with no missing values. A simple treatment that is commonly performed involves eliminating records with missing values, but this can lead to the loss of valuable information. Therefore, finding the most appropriate technique/model to estimate these types of values is crucial.

The process of estimating missing values is known as imputation, and, according to the state of the art, several techniques exist for this process. Among them, autoregressive integrated moving average (ARIMA) (Moritz, 2021), spline interpolation (Moritz & Bartz-Beielstein, 2017), Stineman interpolation (Moritz & Bartz-Beielstein, 2017), exponential weighted moving average (EWMA) (Moritz, 2021), linear weighted moving average (LWMA) (Moritz, 2021), local average of nearest neighbors (LANN) (Flores *et al*., 2019)”ISSN”:”21565570”,”abstract”:”The imputation of time series is one of the most important tasks in the homogenization process, the quality and precision of this process will directly influence the accuracy of the time series predictions. This paper proposes two simple algorithms, but quite powerful for univariate time series imputation process, which are based on the means of the nearest neighbors for the imputation of missing data. The first of them Local Average of Neighbors Neighbors (LANN, multiple imputation by chained equations (MICE) (Peker & Kubat, 2021), K-nearest neighbors (KNN), long short-term memory (LSTM) (Sak *et al*., 2014)we explore LSTM RNN architectures for large scale acoustic modeling in speech recognition. We recently showed that LSTM RNNs are more effective than DNNs and conventional RNNs for acoustic modeling, considering moderately-sized models trained on a single machine. Here, we introduce the first distributed training of LSTM RNNs using asynchronous stochastic gradient descent optimization on a large cluster of machines. We show that a two-layer deep LSTM RNN where each LSTM layer has a linear recurrent projection layer can exceed state-of-the-art speech recognition performance. This architecture makes more effective use of model parameters than the others considered, converges quickly, and outperforms a deep feed forward neural network having an order of magnitude more parameters.”,”author”:[{“dropping-particle”,””,”family”:

”Sak”,”given”:”Haşim”,”non-dropping-particle”:””,”parse-names”:false,”suffix”:””},{“dropping-particle”:””,”family”:”Senior”,”given”:”Andrew”,”non-dropping-particle”:””,”parse-names”:false,”suffix”:””},{“dropping-particle”:””,”family”:”Beaufays”,”given”:”Françoise”,”non-dropping-particle”:””,”parse-names”:false,”suffix”:””}],”container-title”:”Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH”,”id”:”ITEM-1”,”issued”:{“date-parts”:[[“2014”]]},”title”:”Long short-term memory recurrent neural network architectures for large scale acoustic modeling”,”type”:”paper-conference”},”uris”:[“http://www.mendeley.com/documents/?uuid=4b74cfe0-e638-3127-ba9f-87329bc8f067”]}],”mendeley”:{“formattedCitation”:”(Sak et al., 2014, and gated recurrent unit (GRU) (Chung *et al*., 2014)we focus on more sophisticated units that implement a gating mechanism, such as a long short-term memory (LSTM. ARIMA, EWMA, LWMA, and LANN are moving average techniques, spline and Stineman are techniques based on interpolation, and LSTM and GRU are techniques based on deep learning. These are much more recent than the previous ones, whose main difference lies in the amount of data for their training to obtain satisfactory results.

According to the literature, studies on PM2.5, in Lima City, mostly address the correlation between PM2.5 volumes and various variables, such as cardiorespiratory emergency cases (V. Tapia *et al*., 2020)(Vilma Tapia *et al*., 2020), asthma emergency cases (Vu *et al*., 2021), maternal exposure (V. L. Tapia *et al*., 2020)034 births from 2012 to 2016, at three public hospitals. We used estimated daily PM2.5 from a newly created model developed using ground measurements, satellite data, and a chemical transport model. Exposure was assigned based on district of residence (n = 39, overweight/obesity, and covid-19 cases (Vasquez-Apestegui *et al*., 2021). Other studies have addressed this analytically, using trends (Reátegui-Romero *et al*., 2021)(Rojas *et al*., 2022), and volumes (Reátegui-Romero *et al*., 2018). No studies have analyzed the estimation of missing or NA values for any monitoring station in Lima City. This served as the motivation for conducting a comparative study of imputation techniques for the PM2.5 time series from the city of Lima, Peru.

In this study, ARIMA, spline interpolation, EWMA, LWMA, LANN, LSTM, BiLSTM, GRU, and BiGRU were implemented for short-gaps of missing values, with our main contribution being the comparative study of imputation techniques to estimate short-gaps of missing values in PM2.5 time series of Lima City.

The remainder of the paper is organized as follows: Section 2 presents the methodology where the experiments are described; Section 3 describes the results; and finally, the conclusions of the study are presented.

MATERIAL AND METHODS

The dataset

The dataset was downloaded from openaq.org, and corresponds to the Ceres station which is located in the Ate district of Lima city in Peru, with latitude -12.028694 and longitude: -76.927056 at 300 masl. It contains hourly data from 2021-12-24 07:00 :00 hours to 2024-02-29 00:00:00 hours with a total of 11,822 available records. The Ceres monitoring station can be seen in Fig.1 in orange color and the 103 number.

It is important to highlight that from 2021-12-24 the dataset should contain 19,146 records, however, there were only 11,822 records, which means that 7,324 records were not recorded, they represented missing values. In this study, only the available data were used. Fig.2 shows the available data plotted.

Data preparation

Data were downloaded in files of 1,000 records, and the first step was to join them in just one file. Once the records were in a single file, the next step was to organize them into two partitions, training 80% equivalent to 9,456 records, and the remaining 20% for testing equivalent to 2,366 records. Fig. 3 shows training and testing data.

The test data was structured with three simple strategies, the first inserts missing values (NA) every three items, the second every two items, and the third every 1 item. Each strategy generates 25%, 33%, and 50% of missing values respectively. A clearer appreciation is shown in Table 1.

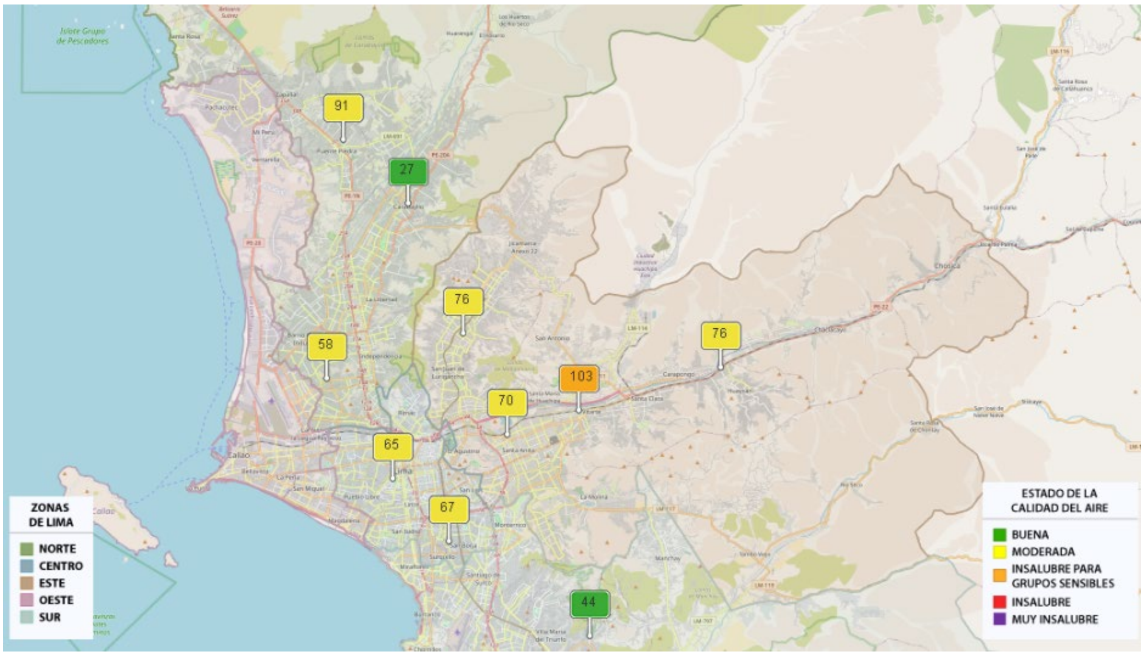

Fig. 1. The air monitoring stations of Lima city

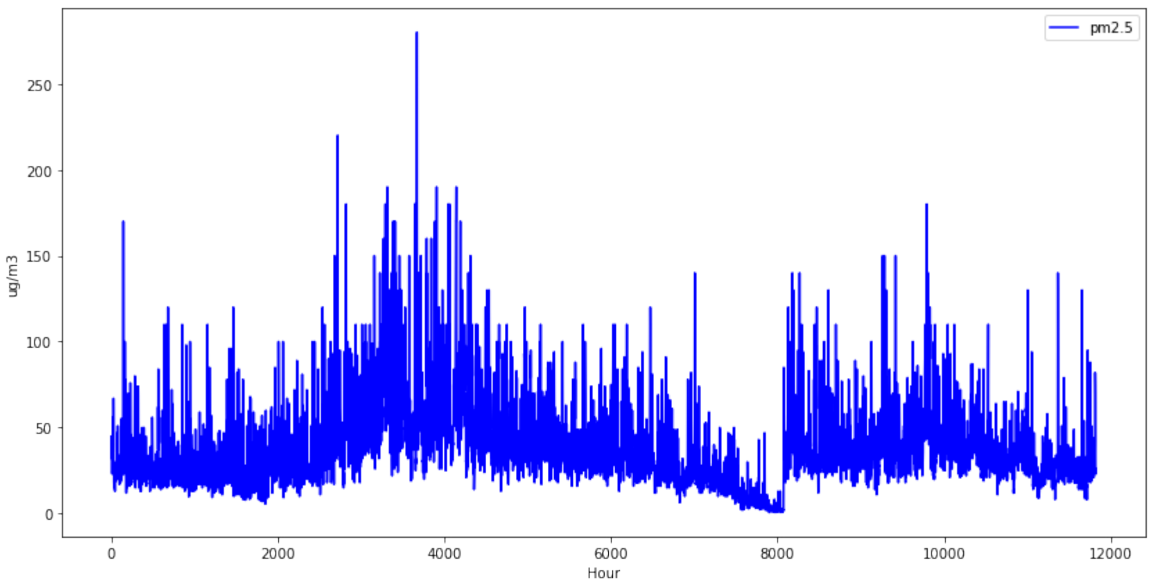

Fig. 2. Available Pm2.5 time series of Ceres Station in Lima city

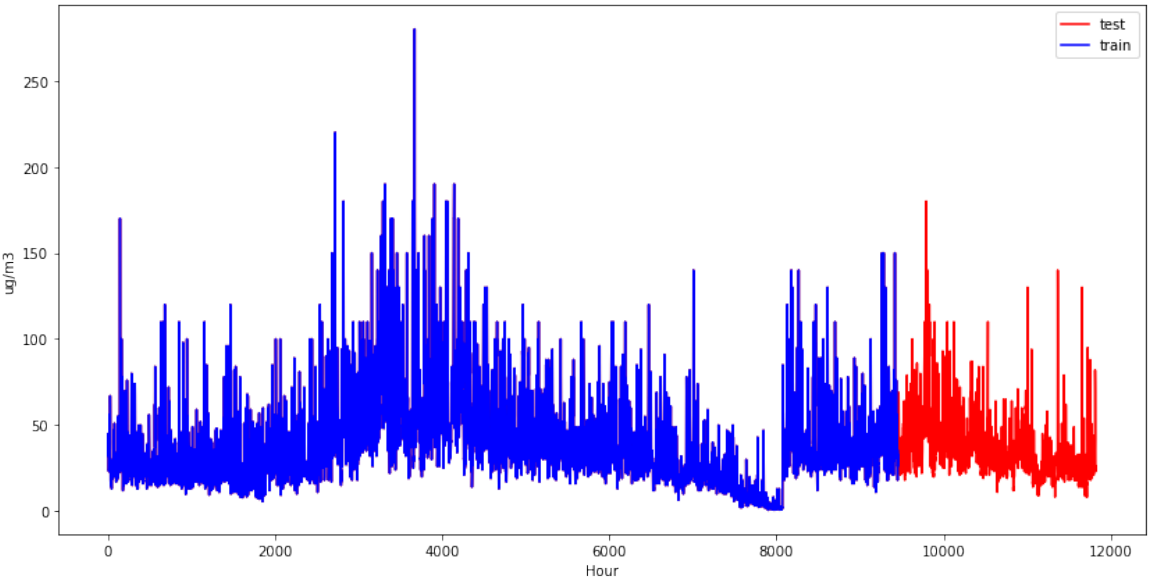

Fig. 3. Training and testing data.

Table 1. Strategies to generate missing values (NA)

Strategy	NA Percentage	Time Series
1	25%	[50, 'NA', 45, 30, 33, 'NA', 67, 58, 51, 'NA', 47, 41, 32, 'NA', 29, 28, 30, 'NA', 28, 34, 35,…]
2	33%	[50, 'NA', 45, 30, 'NA', 69, 67, 'NA', 51, 47, 'NA', 41, 32, 'NA', 29, 28, 'NA', 27, 28, 'NA', 35,…]
3	50%	[50, 'NA', 45, 'NA', 33, 'NA', 67, 'NA', 51, 'NA', 47, 'NA', 32, 'NA', 29, 'NA', 30, 'NA', 28,…]

NA stands for Not Available values

Table 2. Parameters for Moving Average-based techniques and Spline interpolation

Technique	Function	Parameters
ARIMA	na_kalman	model=auto.arima'
Spline	na_interpolation	option='spline'
EWMA	na_ma	k=8, weighting='exponential'
LWMA	na_ma	K=2, weighting='linear'
LANN	na_ma	k=2, weighting='simple'

Implementation of Techniques/Models

In this phase, different imputation techniques were implemented. ARIMA, Spline, EWMA, LWMA, and LANN were implemented through the imputeTS library(Moritz, 2021) in RStudio. The respective parameters are shown in Table 2.

The techniques based on deep learning e.g. LSTM, BiLSTM, GRU, and BiGRU were implemented using the Tensorflow 2.9.0 library in Python language considering the hyperparameters shown in Table 3.

Each model had 4 layers, the first three were of model type and the last layer was Dense with sigmoid as the activation function. The models were compiled with loss function = 'mse', optimizer = 'adam', and 100 epochs.

The results achieved by each model are shown in the Results and Discussion section.

Evaluation

The implemented techniques were evaluated through three well-known metrics including Root Mean Squared Error (RMSE)(1) which measures the error between the observed and predicted data. Mean Absolute Percentage Error (MAPE)(2), which measures the error in a similar way to RMSE, but expressed in percentage terms, is more useful than the previous ones if it is necessary to compare results from different datasets. Finally, the correlation between the actual data and the estimated/predicted data is measured by R^2(3).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(Pi - Oi\right)^2}{n}} \tag{1}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\left(O_i - P_i\right)}{O_i} * 100\right| \tag{2}$$

$$R^2 = (\sum_{i=1}^{n}\left(O_i - \bar{O}\right)\left(P_i - \bar{P}\right) / \sum_{i=1}^{n}\left(O_i - \bar{O}\right)^2 \sum_{i=1}^{n}\left(P_i - \bar{P}\right)^2)) \tag{3}$$

RESULTS AND DISCUSSION

This section describes the results that were achieved after the modeling phase.

According to Table 4, for 25% of missing values (NA), the best models, in terms of RMSE, MAPE, and R^2 were ARIMA, Spline interpolation, and LWMA with RMSEs of 7.1746 µm/m^3, 7.2646 µm/m^3, 8.3427 µm/m^3; MAPEs of 9.7689%, 10.8286%, 11.5104%; R^2s of 0.8278, 0.8316, and 0.7672 respectively. A graphical view of the observed vs predicted data can be seen in Fig. 4. Predicted data correspond to the best moving averages technique (ARIMA), and the best deep learning model (BiLSTM).

Table 3. Hyperparameters for deep learning models

Technique	Hyperparameters
LSTM	[30,30,30,1] activation=’relu’, dropout=’0.1’, learning_rate=’1e-3’
BiLSTM	[30,30,30,1] activation=’relu’ dropout=’0.1’, learning_rate=’ 1e-3’
GRU	[30,30,30,1] activation=’relu’ dropout=’0.1’, learning_rate=’ 1e-3’
BiGRU	[30,30,30,1] activation=’relu’ dropout=’0.1’, learning_rate=’ 1e-3’

Table 4. Results on test data with 25% of missing values

Model	RMSE	MAPE	R^2
ARIMA	**7.1746**	**9.7689**	**0.8278**
SPLINE	**7.2646**	**10.8286**	**0.8316**
EWMA	8.6002	12.7682	0.7546
LWMA	**8.3427**	**11.5104**	**0.7672**
LANN	8.7725	12.1900	0.7426
LSTM	10.5402	21.4211	0.6975
BILSTM	9.7553	14.8956	0.6946
GRU	10.4131	16.2035	10.4091
BIGRU	10.4092	15.3246	0.6792

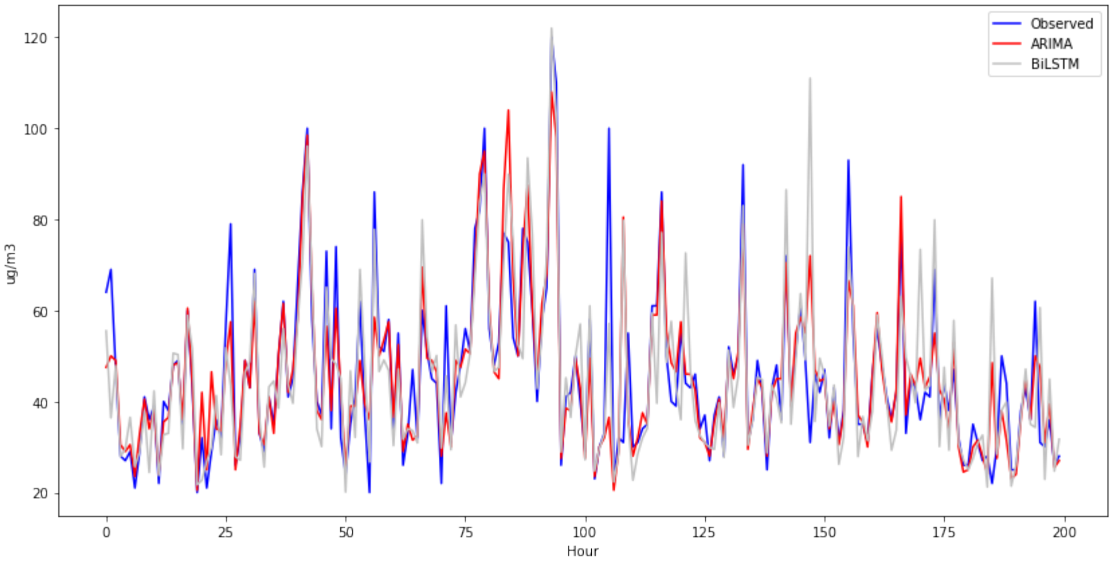

Fig. 4. Predicted values. ARIMA vs BiLSTM (25% of missing values)

Table 5. Results on test data with 33% of missing values

Model	RMSE	MAPE	R^2
ARIMA	**8.0642**	**11.8268**	**0.7726**
SPLINE	8.8064	**12.2399**	**0.7686**
EWMA	**8.5643**	13.4677	0.7423
LWMA	**8.5524**	**12.3646**	**0.7473**
LANN	8.7510	12.8295	0.7341
LSTM	10.4083	21.5205	0.7000
BILSTM	9.6255	13.8829	0.6947
GRU	10.4269	14.8330	0.6967
BIGRU	10.5230	14.3942	0.6715

According to Table 5, for 33% of missing values, the best models, in terms of RMSE were ARIMA, LWMA, and EWMA with RMSEs of 8.0642 $\mu m/m^3$, 8.5524 $\mu m/m^3$, and 8.5643 $\mu m/m^3$ respectively. In terms of MAPE and R^2 were: ARIMA, Spline, and LWMA with MAPEs of 11.8268%, 12.2399%, 12.3646, and R^2s of 0.7726, 0.7686, and 0.7473 respectively.

A graphical view of the observed and predicted data can be seen in Fig. 5.

According to Table 6, for 50% of missing values, the best models in terms of RMSE are ARIMA, LWMA, and LANN with similar RMSEs of 7.3623 $\mu m/m^3$. In terms of MAPE the same models with MAPEs of 10.0192%, 10.0197%, and 10.0198% respectively. Finally, in terms of R^2, similar to RMSE, the three models presented the same R^2 of 0.8247.

A graphical view of the observed and predicted data can be seen in Fig. 6.

According to the results obtained, it can be seen that for short-gaps of missing values, in different configurations (25%, 33%, and 50%) the best model to estimate this type of values was ARIMA, which is within the set of moving average-based models. The best Deep learning-based model was BiLSTM which is in sixth place.

Although the classic models outperformed the deep learning models in the different configurations of missing values, according to the literature, a model is excellent if it has a MAPE lower than 10%, considering this, ARIMA is the one model that achieved this rating

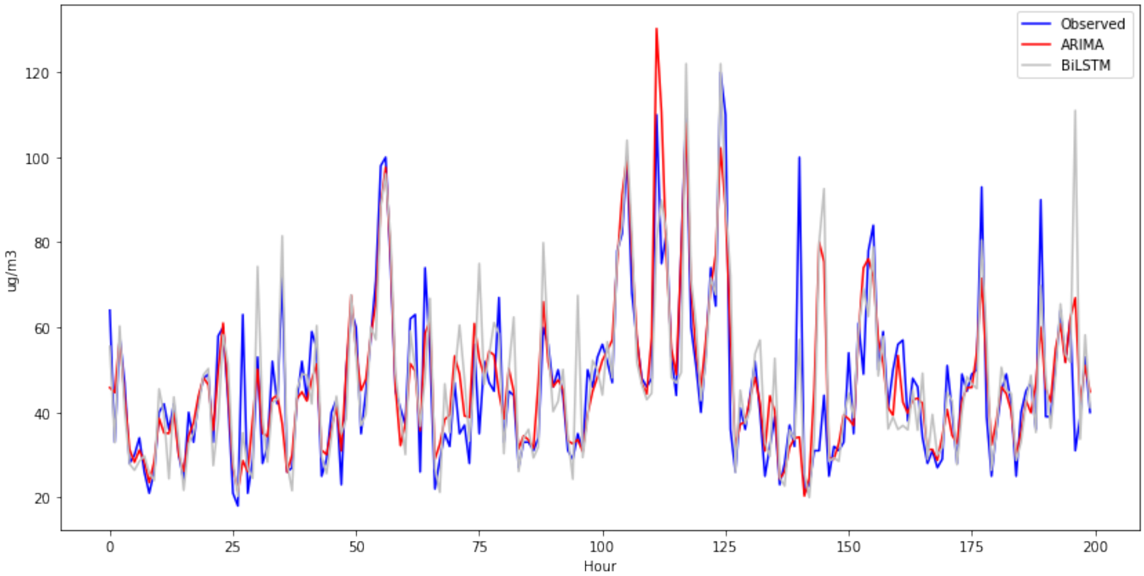

Fig. 5. Predicted values. ARIMA vs BiLSTM (33% of missing values)

Table 6. Results on test data with 50% of missing values

Model	RMSE	MAPE	R^2
ARIMA	**7.3623**	**10.0192**	**0.8247**
SPLINE	7.4536	10.5615	0.8235
EWMA	8.0967	12.2998	0.7924
LWMA	**7.3623**	**10.0197**	**0.8247**
LANN	**7.3623**	**10.0198**	**0.8247**
LSTM	10.3925	21.5006	0.7141
BILSTM	9.4765	14.1246	0.7149
GRU	9.9464	15.7523	0.7199
BIGRU	9.7511	14.6528	0.7132

for the configuration of 25% of missing values with MAPE of 9.7689%. This means that the implemented models still need to be improved.

It is important to highlight that the implemented models based on moving averages in this study only used test data to estimate the different configurations of missing values, while the deep learning-based models used the training data for their respective training, even so, they were surpassed by the first ones.

According to the literature, for various regression tasks, deep learning models outperformed moving average models, but these require a greater amount of data. As reported in different works, hourly data of 2 or 3 years was insufficient to achieve good results.

Table 7 shows the results obtained by different techniques or models for PMPM2.5 time series imputation.

In terms of RMSE, which is not the most appropriate metric to compare the results of different datasets, however, it was appreciated that the best result of the related works was reported by (Alkabbani *et al.*, 2022)timely air quality index (AQI with an RMSE of 3,756 **µ**m/m^3 using the machine learning technique known as Random Forest. The second-best work was reported by (Flores *et al.*, 2023) with an RMSE of 6.8369 **µ**m/m^3 using deep learning and polynomial

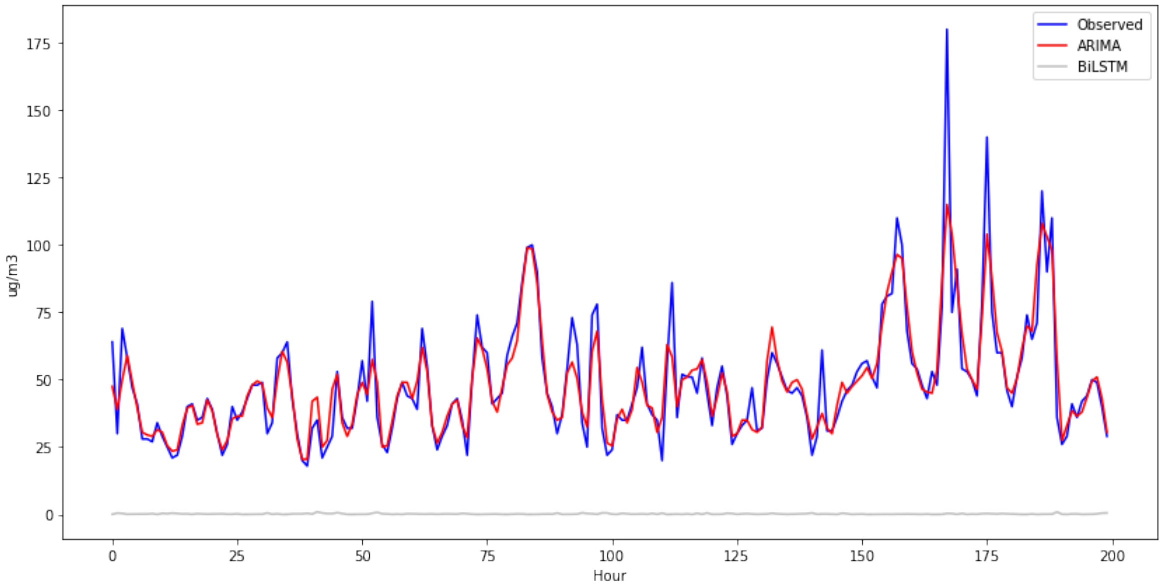

Fig. 6. Predicted values. ARIMA vs BiLSTM (50% NAs)

Table 7. Results of related work

Work	Technique	Data	Frequency	Metric	Value
(Yuan *et al.*, 2018)	LSTM	30,700	Hourly	RMSE	17.78
(Belachsen & Broday, 2022)	KNN	140,256	HalfHourly	NMAE	[0.21 – 0.26]
(Saif-ul-Allah *et al.*, 2022)	GRU	2,514	Daily	RMSE	10.60
(Alkabbani *et al.*, 2022)	Random Forest		Daily	RMSE	3.756
(Lee *et al.*, 2023)	GAIN	26,281	Hourly	R^2	0.895
(Flores *et al.*, 2023)	Deep learning and Polynomial Interpolation	56,424	Hourly	RMSE	6.8369
				R^2	0.6962
This work	ARIMA	11,822	Hourly	RMSE	7.1746
				R^2	0.8278

interpolation. And, this work using ARIMA was located in the third position with an RMSE of 7.1746 µm/m^3.

In terms of R^2, which is better than RMSE for comparing results of different datasets, but unfortunately, few related works reported results with R^2. The best result was reported by (Lee *et al.*, 2023) with an R^2 of 0.895, followed by ARIMA of this study with an R^2 of 0.8278 in the set of 25% of missing values.

CONCLUSION

According to the results obtained, it can be stated that for imputation of short-gaps in hourly PM2.5 time series of fewer than 3 years, it is more appropriate or pertinent to use techniques based on moving averages such as ARIMA, EWMA, LWMA, LANN or techniques based in interpolation such as Spline, instead of resorting to more complex techniques such as those based on Deep learning including LSTM, BiLSTM, GRU and BiGRU.

GRANT SUPPORT DETAILS

The present research did not receive any financial support.

CONFLICT OF INTEREST

The authors declare that there is not any conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/ or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

LIFE SCIENCE REPORTING

No life science threat was practiced in this research.

REFERENCES

Alkabbani, H., Ramadan, A., Zhu, Q., & Elkamel, A. (2022). An Improved Air Quality Index Machine Learning-Based Forecasting with Multivariate Data Imputation Approach. Atmosphere, 13(7). https://doi.org/10.3390/atmos13071144

Belachsen, I., & Broday, D. M. (2022). Imputation of Missing PM2.5 Observations in a Network of Air Quality Monitoring Stations by a New kNN Method. Atmosphere, 13(11). https://doi.org/10.3390/atmos13111934

Bu, X., Xie, Z., Liu, J., Wei, L., Wang, X., Chen, M., & Ren, H. (2021). Global PM2.5-attributable health burden from 1990 to 2017: Estimates from the Global Burden of disease study 2017. Environmental Research, 197. https://doi.org/10.1016/j.envres.2021.111123

Chen, Z., Liu, P., Xia, X., Wang, L., & Li, X. (2022). The underlying mechanism of PM2.5-induced ischemic stroke. In Environmental Pollution (Vol. 310). https://doi.org/10.1016/j.envpol.2022.119827

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Gated Recurrent Neural Networks on Sequence Modeling. ArXiv.

Colorado, M. (2019). Perú es el país con la peor calidad de aire y Santiago la capital más contaminada de Latinoamérica. France 24, 1–3. https://www.france24.com/es/20190313-medio-ambiente-calidad-aire-contaminacion

Flores, A., Tito-Chura, H., Centty-Villafuerte, D., & Ecos-Espino, A. (2023). Pm2.5 Time Series Imputation with Deep Learning and Interpolation. Computers, 12(8). https://doi.org/10.3390/computers12080165

Flores, A., Tito, H., & Silva, C. (2019). Local average of nearest neighbors: Univariate time series imputation. International Journal of Advanced Computer Science and Applications, 10(8). https://doi.org/10.14569/ijacsa.2019.0100807

Huang, F., Pan, B., Wu, J., Chen, E., & Chen, L. (2017). Relationship between exposure to PM2.5 and lung cancer incidence and mortality: A meta-analysis. Oncotarget, 8(26). https://doi.org/10.18632/oncotarget.17313

Lee, Y. S., Choi, E., Park, M., Jo, H., Park, M., Nam, E., Kim, D. G., Yi, S. M., & Kim, J. Y. (2023). Feature extraction and prediction of fine particulate matter (PM2.5) chemical constituents using four machine learning models. Expert Systems with Applications, 221. https://doi.org/10.1016/j.eswa.2023.119696

Moritz, S. (2021). imputeTS. In The R Journal (Vol. 9, Issue 1).

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time series missing value imputation in R. R Journal, 9(1). https://doi.org/10.32614/rj-2017-009

Oh, J., Choi, S., Han, C., Lee, D. W., Ha, E., Kim, S., Bae, H. J., Pyun, W. B., Hong, Y. C., & Lim, Y. H. (2023). Association of long-term exposure to PM2.5 and survival following ischemic heart disease. Environmental Research, 216. https://doi.org/10.1016/j.envres.2022.114440

Peker, N., & Kubat, C. (2021). A hybrid modified deep learning data imputation method for numeric data sets. International Journal of Intelligent Systems and Applications in Engineering, 9(1). https://doi.

org/10.18201/ijisae.2021167931

Priya, S. A., & Khanaa, V. (2023). An Intelligent Air Quality During COVID-19 Prediction and Monitoring System Using Temporal CNN-LSTM. In EAI/Springer Innovations in Communication and Computing: Vol. Part F274. https://doi.org/10.1007/978-3-031-23683-9_31

Reátegui-Romero, W., Sánchez-Ccoyllo, O. R., Andrade, M. de F., & Moya-Alvarez, A. (2018). PM2.5 Estimation with the WRF/Chem Model, Produced by Vehicular Flow in the Lima Metropolitan Area. Open Journal of Air Pollution, 07(03). https://doi.org/10.4236/ojap.2018.73011

Reátegui-Romero, W., Zaldivar-Alvarez, W. F., Pacsi-Valdivia, S., Sánchez-Ccoyllo, O. R., Garciá-Rivero, A. E., & Moya-Alvarez, A. (2021). Behavior of the Average Concentrations As Well As Their PM10 and PM2.5 Variability in the Metropolitan Area of Lima, Peru: Case Study February and July 2016. International Journal of Environmental Science and Development, 12(7). https://doi.org/10.18178/ijesd.2021.12.7.1341

Republica, G. La. (2023). Perú es el país con peor calidad de aire de Sudamérica. https://especial.larepublica.pe/la-republica-sostenible/2023/09/14/peru-es-el-pais-con-peor-calidad-de-aire-de-sudamerica-1225756

Rojas, F. J., Pacsi-Valdivia, S., & Sánchez-Ccoyllo, O. R. (2022). Simulación computacional e influencia de las variables meteorológicas en las concentraciones de PM10 y PM2.5 en Lima Metropolitana. Información Tecnológica, 33(3). https://doi.org/10.4067/s0718-07642022000300223

RumboMinero. (2022). El 79% de su consumo de energía provino de fuentes de combustibles fósiles en 2021. https://www.rumbominero.com/usa-internacionales/consumo-energia-combustibles-fosiles-2021/

Saif-ul-Allah, M. W., Qyyum, M. A., Ul-Haq, N., Salman, C. A., & Ahmed, F. (2022). Gated Recurrent Unit Coupled with Projection to Model Plane Imputation for the PM2.5 Prediction for Guangzhou City, China. Frontiers in Environmental Science, 9. https://doi.org/10.3389/fenvs.2021.816616

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. https://doi.org/10.21437/interspeech.2014-80

Tapia, V. L., Vasquez, B. V., Vu, B., Liu, Y., Steenland, K., & Gonzales, G. F. (2020). Association between maternal exposure to particulate matter (PM2.5) and adverse pregnancy outcomes in Lima, Peru. Journal of Exposure Science and Environmental Epidemiology, 30(4). https://doi.org/10.1038/s41370-020-0223-5

Tapia, V., Steenland, K., Sarnat, S. E., Vu, B., Liu, Y., Sánchez-Ccoyllo, O., Vasquez, V., & Gonzales, G. F. (2020). Time-series analysis of ambient PM2.5 and cardiorespiratory emergency room visits in Lima, Peru during 2010–2016. Journal of Exposure Science and Environmental Epidemiology, 30(4). https://doi.org/10.1038/s41370-019-0189-3

Tapia, Vilma, Steenland, K., Vu, B., Liu, Y., Vásquez, V., & Gonzales, G. F. (2020). PM2.5exposure on daily cardio-respiratory mortality in Lima, Peru, from 2010 to 2016. Environmental Health: A Global Access Science Source, 19(1). https://doi.org/10.1186/s12940-020-00618-6

Vasquez-Apestegui, B. V., Parras-Garrido, E., Tapia, V., Paz-Aparicio, V. M., Rojas, J. P., Sanchez-Ccoyllo, O. R., & Gonzales, G. F. (2021). Association between air pollution in Lima and the high incidence of COVID-19: findings from a post hoc analysis. BMC Public Health, 21(1). https://doi.org/10.1186/s12889-021-11232-7

Vu, B. N., Tapia, V., Ebelt, S., Gonzales, G. F., Liu, Y., & Steenland, K. (2021). The association between asthma emergency department visits and satellite-derived PM2.5 in Lima, Peru. Environmental Research, 199. https://doi.org/10.1016/j.envres.2021.111226

Wyer, K. E., Kelleghan, D. B., Blanes-Vidal, V., Schauberger, G., & Curran, T. P. (2022). Ammonia emissions from agriculture and their contribution to fine particulate matter: A review of implications for human health. In Journal of Environmental Management (Vol. 323). https://doi.org/10.1016/j.jenvman.2022.116285

Yuan, H., Xu, G., Yao, Z., Jia, J., & Zhang, Y. (2018). Imputation of missing data in time series for air pollutants using long short-term memory recurrent neural networks. UbiComp/ISWC 2018 - Adjunct Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2018 ACM International Symposium on Wearable Computers. https://doi.org/10.1145/3267305.3274648