Article

Not peer-reviewed version

Imputation Bias in ARIMA Air Quality Models

Ejaz Hussain *, Yang Li , Atiqur Rahman Ahad

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1325.v1

Keywords: bias; air quality; ARIMA; forecasting; imputation; data analysis; predictive analysis; bias mitigation

 doi:10.20944/preprints202603.1325.v1

Article

Imputation Bias in ARIMA Air Quality Models

Ejaz Hussain *, Yang Li and Atiqur Rahman Ahad

Dept. of Computer Science, University of East London, London, UK

* Correspondence: u2342306@uel.ac.uk

Abstract

Missing data remains a pervasive challenge in air quality data analysis, where inappropriate imputation techniques can introduce hidden biases and compromise the reliability of time-series models such as AutoRegressive Integrated Moving Average (ARIMA). This paper examines the impact of linear interpolation and mean/median imputation on the performance of the ARIMA model and biases in the prediction of particulate matter 2.5 (PM2.5) concentration, together with a detailed analysis of ARIMA generated error metrics and their implications for the accuracy and reliability of the prediction. The findings reveal that package-default imputation significantly influences ARIMA forecasts, while mean/median imputation consistently delivers superior predictive performance, highlighting its robustness for handling missing environmental data. Moreover, imputation during the data transformation stage exerts a greater influence on model outcomes than methods applied at later analysis stages.

Keywords: bias; air quality; ARIMA; forecasting; imputation; data analysis; predictive analysis; bias mitigation

1. Introduction

Predicting air quality in urban environments has become a critical aspect of modern urban planning and public health initiatives, as evidenced by the increasing concerns surrounding air pollution in major cities around the world [1,2]. Accurate forecasting of pollutants such as particulate matter 2.5 (PM2.5) is crucial for effective environmental governance and the protection of public health. Pollutants within the atmosphere are complex and require sophisticated sensor based monitoring stations, despite their fundamental role in data acquisition, these stations frequently face issues such as missing data, which can significantly impair the reliability of predictive models [3]. The challenge itself stems from the intricate interplay of environmental factors and the inherent limitations of sensor technology, which can lead to gaps in crucial time-series datasets [2]. Similarly, data professionals use their own judgement in choosing appropriate imputation techniques, which can introduce hidden biases in time-series-based auto-regressive integrated moving average (ARIMA) models used for air quality prediction [4].

Bias is one of the significant challenges in time series data analysis, particularly when dealing with incomplete environmental datasets, as the selection of an imputation method can subtly influence the underlying data distribution and subsequent model outcomes [5]. When it comes to time-series forecasting, ARIMA is one of the most widely recognised and applied statistical models, offering robust capabilities to capture temporal dependencies and trends in air quality data [6]. However, the efficacy of ARIMA models is inherently linked to the completeness and quality of the input data, making the choice of imputation technique a critical factor in mitigating hidden biases that could skew air quality predictions [4]. This article investigates how different imputation strategies for handling missing environmental data can introduce subtle, yet significant, biases into ARIMA models, thereby affecting the accuracy and reliability of air quality predictions. This research focuses on quantifying these biases and proposing methodologies to mitigate their impact, ultimately improving the robustness of air quality forecasting.

 doi:10.20944/preprints202603.1325.v1

Finally, this article offers practical recommendations for data scientists and researchers to improve the precision and reliability of air quality predictions when dealing with incomplete datasets. This comprehensive analysis bridges the gap between theoretical imputation techniques and their practical implications, ensuring that air quality models are not only statistically sound but also ecologically relevant. This practical understanding and guidance not only minimise the risk of bias, but also enhances social value through informed decision-making for environmental policies and public health interventions required by central and local government bodies.

2. Related Work

The climate data landscape has seen significant transformation in recent years, as the scientific community has recognised the critical role that data play in understanding and addressing the challenges posed by climate change. For example, air pollution data have become indispensable for researchers and policy makers seeking to measure and mitigate the environmental consequences of human actions [7]. Another example is the use of advanced cutting-edge technologies such as remote sensing and internet of things (IoT) devices to gather unprecedented amounts of data on environmental conditions [7]. IoT devices can provide real-time information on factors such as temperature, humidity, and air quality, enabling more dynamic and responsive environmental monitoring [8,9]. Globally, climate data projected via IoT devices have now become widely available, providing researchers with a wealth of information to rigorously examine the climate systems of the Earth. At the same time, the abundance of air quality data has presented new challenges, including the most common issue of missing values, which can significantly hinder robust analysis and predictive modelling [4,10]. Data professionals often wrestle with the decision between deleting incomplete records or employing imputation techniques to fill the gaps, a decision that profoundly influences the integrity and representativeness of the dataset [11]. Indeed, the inability to analyse and handle missing data can substantially impede the air quality data analysis process. For example, incorrect imputation can lead to skewed distributions, altered temporal dependencies, and, ultimately, biassed air quality predictions [4,10].

Furthermore, biassed air quality modelling can propagate errors in policy decisions, undermining efforts to mitigate environmental risks and protect public health [12]. The existing literature extensively explores various imputation methods, ranging from simple statistical approaches to complex machine learning algorithms, each with its own assumptions and suitability depending on the nature of the missing-ness and the characteristics of the time series [13]. However, a comprehensive understanding of how these choices specifically introduce and propagate bias within the autoregressive integrated moving average framework for air quality prediction remains less thoroughly investigated.

The field of air quality prediction has seen extensive research, with various models ranging from statistical approaches to advanced machine learning and deep learning tech- niques being applied [1]. ARIMA forecasting models are foundational for time-series analysis, offering robust capabilities to capture temporal dependencies and trends in air quality data [2]. However, the efficacy of ARIMA models is inherently linked to the completeness and quality of the input data, making the choice of imputation technique a critical factor in mitigating hidden biases that could skew air quality predictions [14].

ARIMA statistical methods can be implemented through various open-source packages. Although these packages facilitate forecasting models; their performance is critically dependent on the quality and completeness of the time series data, some of the packages include functions to tackle missing values, for example, the 'forecast' package in R provides functionalities for handling gaps, although its effectiveness is limited for longer periods of contiguous missing-ness [15].

Air quality data often contains significant gaps due to sensor malfunctions or maintenance, which can reach up to 5% of the total data set, thus requiring sophisticated gap-filling techniques to maintain the integrity of the time-series analysis [19]. In such scenarios, where the missing-ness is substantial, the limitations of simpler imputation methods become apparent, as they often fail to

 doi:10.20944/preprints202603.1325.v1

capture the complex pollution dynamics or accurately represent peak pollution hours over multi-day gaps [16]. The existing literature highlights various imputation techniques, from classic mean to median interpolation and more advanced statistical methods, such as linear regression imputation, each possessing distinct advantages and disadvantages depending on the nature and extent of missing data [17].

The article explores the most common imputation methods and evaluates their impact on the performance of ARIMA models for the prediction of air quality. Specifically, this study investigates how different imputation techniques, when applied to environmental datasets with high missing information rates, can inadvertently introduce systematic bias into forecasts generated by ARIMA models, ultimately compromising the reliability of air quality assessments.

3. Materials and Methods

3.1. Air Quality Dataset

This article specifically addresses bias challenges by evaluating the performance of ARIMA models under various imputation strategies, aiming to identify methods that minimise bias and enhance prediction accuracy. The data set comprises an hourly air quality measurement of PM2.5 concentrations collected over a five-year period from January 1, 2019 to December 31, 2023, from a monitoring station located in Wren Close, London, United Kingdom. This particular data set is ideal for examining the effects of various imputation techniques on time-series forecasting due to its inherent seasonal and daily patterns, as well as the presence of varied missing data. The selected monitoring site is managed by Air Quality England on behalf of the London Borough of Newham, and the data was accessed and extracted from their publicly available open-source data platform [18].

3.2. Proposed Methodology

The details of the methodological framework employed in this study involved various steps and they are as follows.

- Data Exploration and Transformation;
- Implementation of the baseline ARIMA Model;
- Implementation of Package's Imputation Technique;
- Implementation of Mean/Median Imputation Technique.

3.2.1. Data Exploration and Transformation

Initially, raw hourly PM2.5 concentration data was extracted, followed by a thorough exploratory data analysis to identify preliminary trends, seasonality, and the extent of missing values. The preliminary data analysis revealed that PM2.5 has in total 43,824 hourly observations, of which 12,664 were missing, representing approximately 28.9% of the total data set, a significantly high rate that requires robust imputation strategies for subsequent modelling. The Figure 1 below illustrates the distribution of missing values in the data set, highlighting periods of consecutive missing observations and their potential impact on the continuity of the time series.

 doi:10.20944/preprints202603.1325.v1

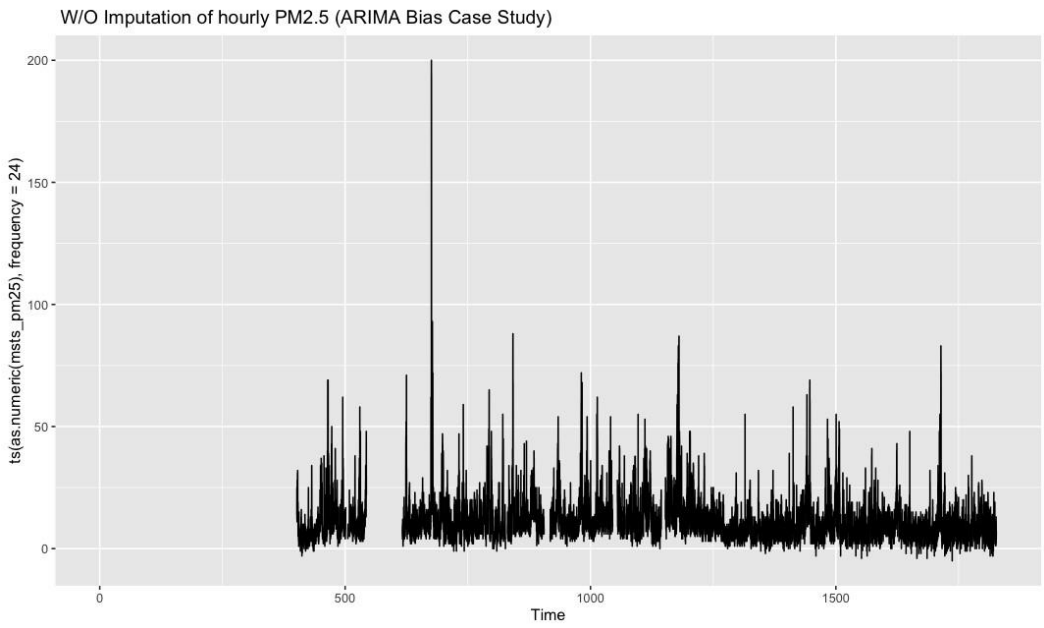

Figure 1. The figure illustrates missing data in PM2.5 concentration which statistically shows that 12,664 out of 43,824 hourly reads are missing Subsequently, the data set was subjected to several preprocessing transformations, including the elimination of extraneous pollutant variables and the detection and removal of outliers by statistical techniques. These procedures were implemented to maintain the integrity of the data and prepare the data for subsequent time-series analysis and the application of various imputation methods.

3.2.2. Implementation of Baseline ARIMA Model

A baseline ARIMA model was established using pre-processed PM2.5 data to serve as a comparative benchmark for evaluating the impact of different imputation strategies on model performance. This foundational model, built on optimally chosen ARIMA parameters, provides a reference point to quantify improvements or degradations in predictive accuracy attributable to different missing data handling techniques.

Table 1 below presents the performance metrics of ARIMA model 1, including a mean error of 0.002, which indicates a minor positive bias in the predictions. The root mean square error of 3.64 reflects the typical scale of prediction errors, alongside a mean absolute squared error of 2.50 and a residual standard deviation of 13.27. These indicators affirm the model's proficiency in capturing underlying data patterns while revealing opportunities for enhancement via sophisticated imputation strategies.

Table 1. Key breakdown of ARIMA baseline model 1 output.

Error Metric	Model Score
Mean Error (ME)	0.002
Root Mean Square Error (RMSE)	3.64
Mean Absolute Error (MAE)	2.50
Mean Absolute Scaled Error (MASE)	0.29
Autocorrelation Function at leg 1 (ACF1)	0.001
Sigma^2	13.27

The Figure 2 below illustrates a visual representation of the 7 day forecast of the ARIMA model, where the blue line shows 168 hours of predicted PM2.5 concentrations. The black line represents the actual concentration levels of PM2.5 observed.

 doi:10.20944/preprints202603.1325.v1

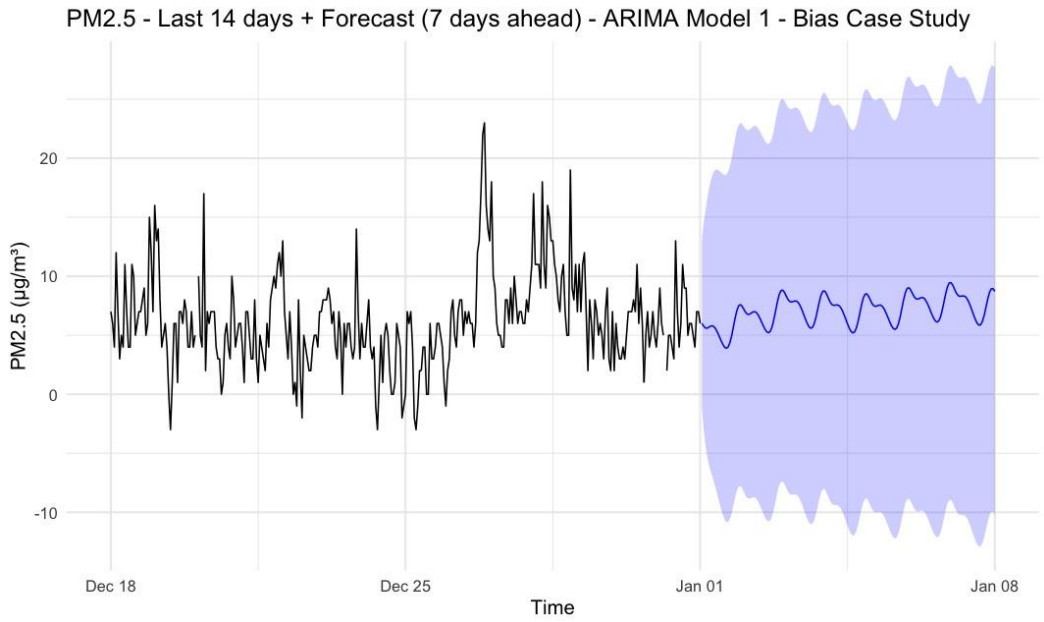

Figure 2. The figure illustrates 7-day future forecasting of PM2.5 concentration using ARIMA Model 1.

3.2.3. Implementation of Package's Imputation Technique

ARIMA models are commonly implemented using open-source packages, such as the 'forecast' package in R, which provides functionalities for fitting and forecasting while including rudimentary handling of missing values. In this methodological step, we explored the package's default imputation mechanisms, particularly the 'na.interp' function, which performs basic linear interpolation to assess their influence on ARIMA model performance without external imputation. This approach generated an imputed data set that enables later evaluation of its effects on model accuracy and potential biases, especially in the results and discussion sections. The Table 2 below summarises the error metrics of ARIMA model 2.

Table 2. Key breakdown of ARIMA baseline model 2 output.

Error Metric	Model Score
Mean Error (ME)	0.00
Root Mean Square Error (RMSE)	3.35
Mean Absolute Error (MAE)	2.5
Mean Absolute Scaled Error (MASE)	0.25
Autocorrelation Function at leg 1 (ACF1)	0.0007
Sigma^2	11.21

The Figure 3 below illustrates the ARIMA model 2 forecast for 7 days, with the interpolated PM2.5 concentrations in green line and the actual observed values in black line, visually confirming the performance of the model on the pre-processed data.

 doi:10.20944/preprints202603.1325.v1

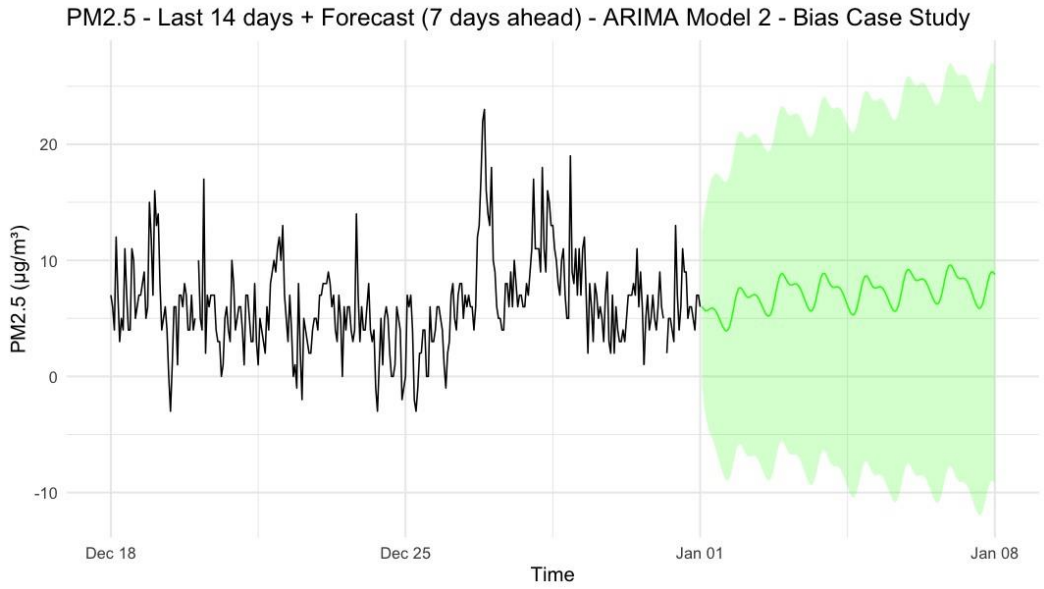

Figure 3. The figure illustrates 7-day future forecasting of PM2.5 concentration using ARIMA Model 2.

3.2.4. Implementation of Mean/Median Imputation Technique

During the implementation of mean or median imputation strategies, missing values in the PM2.5 time series were replaced with the global median of the entire data set, respectively. This straightforward approach offers a simplistic solution to fill data gaps during the data transformation phase. Subsequently, a new variable was created incorporating these imputed values to facilitate direct comparison with the baseline and other imputed ARIMA model 1. The Table 3 below summarises the error metrics for ARIMA model 3.

Table 3. Key breakdown of ARIMA baseline model 3 output.

Error Metric	Model Score
Mean Error (ME)	0.00
Root Mean Square Error (RMSE)	3.10
Mean Absolute Error (MAE)	1.86
Mean Absolute Scaled Error (MASE)	0.28
Autocorrelation Function at leg 1 (ACF1)	0.001
Sigma^2	9.62

The Figure 4 below illustrates the ARIMA model 3 forecast for 7 days, with the median PM2.5 concentrations added in the orange line and the actual observed values in the black line, visually confirming the model's improved performance on the pre-processed data.

 doi:10.20944/preprints202603.1325.v1

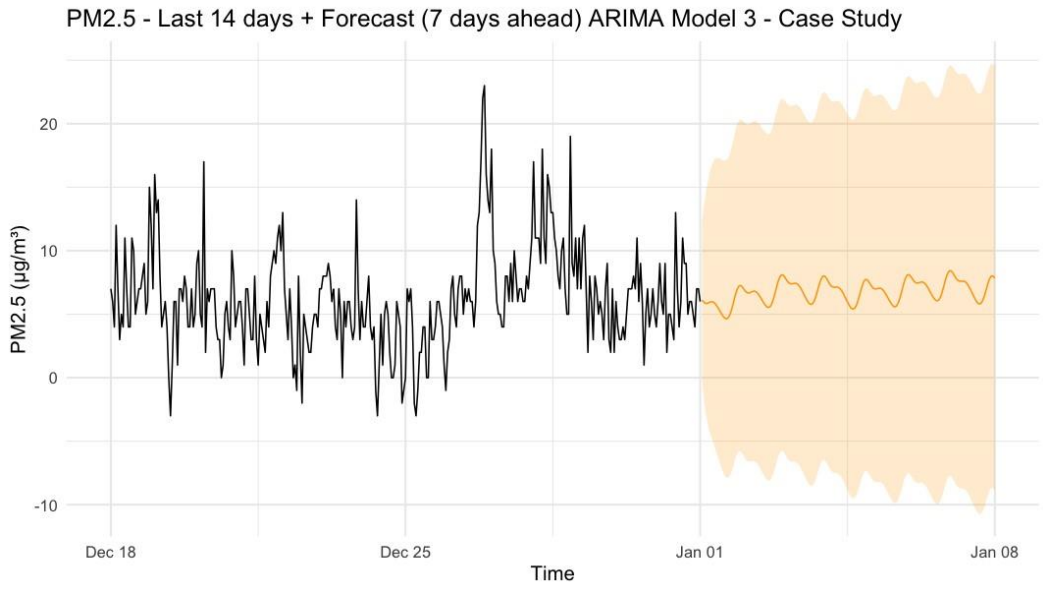

Figure 4. The figure illustrates 7-day future forecasting of PM2.5 concentration using ARIMA Model 3.

4. Results & Discussion

In this section, the results will be outlined and explained, reflecting the methodology proposed in the above subsection 3.2. All results were obtained using the open-source tool RStudio in conjunction with the R programming language and the relevant libraries.

4.1. Results

All 3 ARIMA models have projected different results, a direct quantitative comparison of key performance indicators, such as ME, RMSE, MAE, MASE, and ACF1, across the different ARIMA models to determine which imputation method yields the most accurate and reliable forecasts. The Table 4 below summarises the statistical parameters obtained for each model.

Table 4. Statistical analysis of all 3 ARIMA models outputs.

Error Metric	ARIMA 1	ARIMA 2	ARIMA 3
ME	0.001949125	0.0000844	-0.0003585713
RMSE	3.640386	3.346055	3.101502
MAE	2.500321	2.187955	1.861751
MASE	0.2934355	0.2586301	0.2842862
ACF1	0.001194734	0.0007530299	0.001449699

4.2. Validation Based Discussion

Upon examination of the statistical comparative analysis, the ME for all three ARIMA models is close to zero, indicating negligible systematic bias in the predictions. The forecast accuracy improves progressively from ARIMA model 1 to ARIMA model 3, with ARIMA model 3 demonstrating the lowest RMSE and MAE values, thus indicating superior predictive performance. In contrast, scaled accuracy metrics such as MASE favour ARIMA model 2, while ACF1 shows a marginally better fit for ARIMA model 3. Overall, while ARIMA model 3 provides the most accurate forecasts in absolute terms and ARIMA model 2 minimises bias and relative error, ARIMA model 1 performs comparatively weaker across most evaluation criteria.

 doi:10.20944/preprints202603.1325.v1

The study findings using visual interpretation of the Figure 5 provide a comparative analysis of PM2.5 concentrations over a 7-day forecast period for three ARIMA models.

This visualisation Figure 5 displays the baseline data, the linearly interpolated data set, and the median-imputed data set, allowing a comprehensive assessment of the impact of each imputation strategy on the forecast trajectories compared to the actual observed values.

In the visuals, the blue line represents ARIMA model 1, the green line represents ARIMA model 2, and the orange line represents ARIMA model 3.

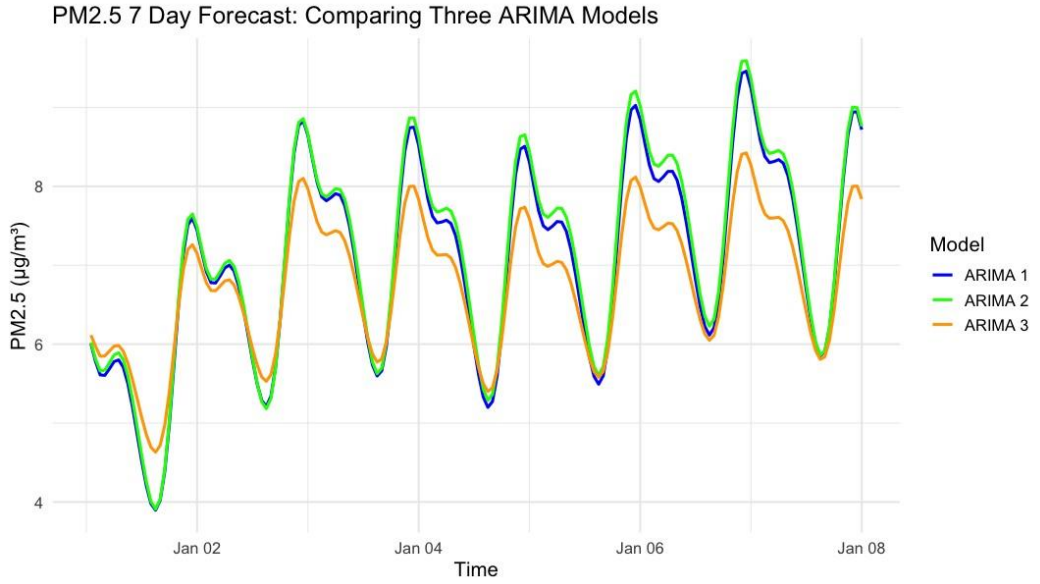

Figure 5. The figure illustrates a comparison chart of all three ARIMA models.

In addition to statistical and visual analysis, a deeper dive is to examine confidence interval bounds (CIB) between the three ARIMA models to further evaluate the certainty and precision of their predictions across different imputation methods. The visual image 6 confirms that the CIB are narrowest for ARIMA model 3, indicating improved predictive reliability and reduced uncertainty attributable to the median imputation.

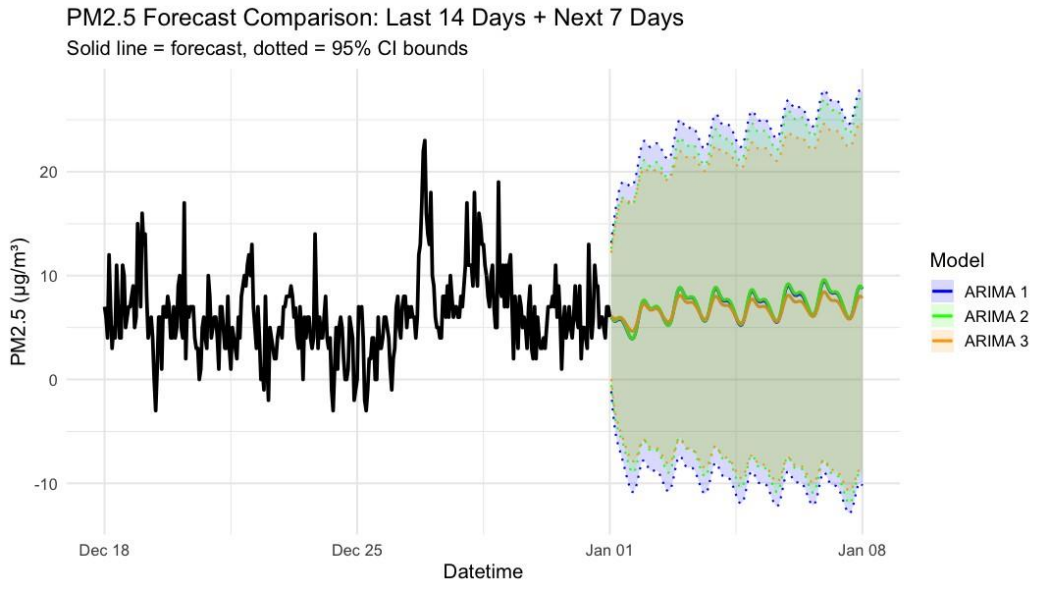

Figure 6. A visual chart comparing three ARIMA models with 95% CIB.

 doi:10.20944/preprints202603.1325.v1

Following the CIB 6 findings above, statistically ARIMA model 3 has a lower peak (8.4 μg/m³) compared to other ARIMA model peaks (9.6 μg/m³). Finally, another visual 7 was generated to illustrate the comparison of error metrics, with a line chart that highlights the variance in ACF1, MAE, MASE and RMSE values in all three ARIMA models, thereby offering an intuitive understanding of each model's predictive accuracy and bias with respect to the different imputation techniques. In visual form, the blue line represents ARIMA model 1, the green line represents ARIMA model 2, and the orange line represents ARIMA model 3, demonstrating that ARIMA model 3 consistently exhibits the lowest error metrics across all parameters tested.

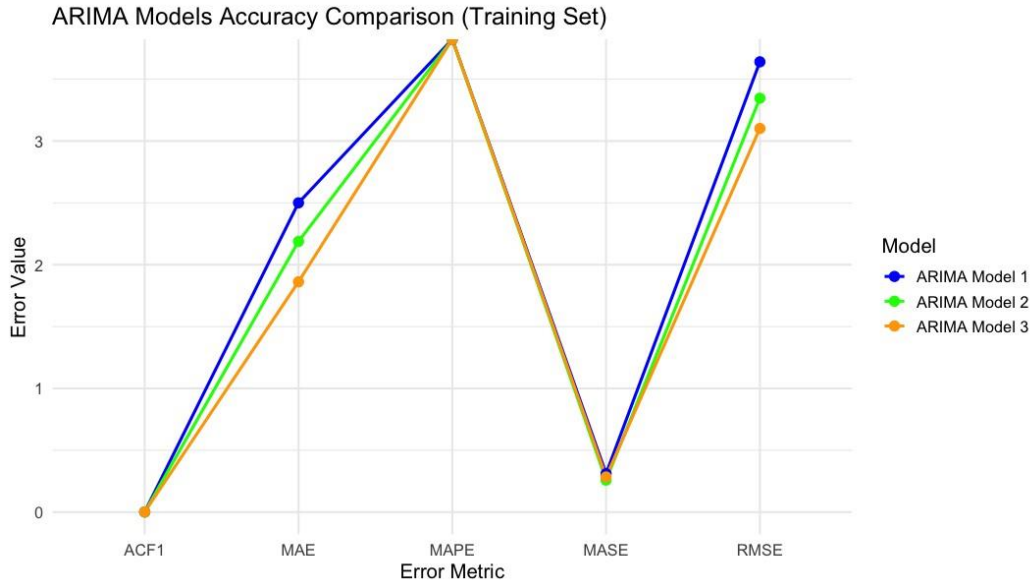

Figure 7. A visual chart comparing error metrics for three ARIMA models and shows the statistical difference between each error value.

5. Conclusions

The paper concludes by highlighting key insights into biases in imputation techniques for advanced data science practices, while showcasing the distinct results from each deployed ARIMA model. In particular, median imputation consistently delivered superior performance, particularly for PM2.5 forecasting, underscoring its effectiveness in handling missing data in environmental time series analysis. Additionally, the findings demonstrate that the imputation techniques applied during the data transformation stage significantly impact the performance of the model compared to the methods used in the later stages of the AQ data analysis. The results of all three ARIMA models illustrate how bias can infiltrate the modelling process. Thus, best practices, such as equitable data aggregation, rigorous validation, model transparency, and appropriate algorithm selection, are essential to mitigate bias and improve predictive accuracy in air quality forecasting. Based on these findings, future research should prioritise the development of a comprehensive air quality bias mitigation framework that integrates various AQ datasets and evaluates an expanded range of imputation techniques to further refine bias-reduction strategies.

Funding: This research received no external funding.

Data Availability Statement: The research data has been extracted using an open-source online platform, supported by Air Quality England [18], the online platform allows researchers to pick and select relevant air quality monitoring station, for this instance Wren Close monitoring station was selected for ARIMA modelling.

Acknowledgments: During the preparation of this manuscript/study, the author used RStudio Version 2025.05.1+513 for the purposes of air quality data analysis, transformation and ARIMA based statistical

 doi:10.20944/preprints202603.1325.v1

modelling. The author also used Overleaf online platform for the preparation of the paper write-up. The authors have reviewed and edited the output and take full responsibility for the content of this publication.".

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AQ	Air Quality
ARIMA	AutoRegressive Integrated Moving Average
PM2.5	Particulate Matter 2.5
AQE	Air Quality England
ME	Mean Error
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
MASE	Mean Absolute Scaled Error
ACF1 CIB	Autocorrelation Function at leg 1 Confidence Interval Bounds
OS	Open-Source
IoT	Internet of Things

References

1. Mehmood K, Bao Y, Cheng W, et al (2022) Predicting the quality of air with machine learning approaches: Current research priorities and future perspectives. Journal of Cleaner Production 379:134656. https://doi.org/10.1016/j.jclepro.2022.134656
2. Liu X, Ngai EC -H., Zachariah D (2021) Scalable Belief Updating for Urban Air Quality Modeling and Prediction. ACM/IMS Transactions on Data Science/ACMIMS Transactions on Data Science 2:1. https://doi.org/10.1145/3402903
3. Yan S, O'Connor DJ, Wang X, et al (2025) Comparative Analysis of Machine Learning-Based Imputation Techniques for Air Quality Datasets with High Missing Data Rates. 1:. https://doi.org/10.1109/cietes63869.2025.10995064
4. Hua VM, Nguyen TL, Dao M-S, et al (2024) The impact of data imputation on air quality prediction problem. PLoS ONE 19:. https://doi.org/10.1371/journal.pone.0306303
5. Silibello C, D'Allura A, Finardi S, et al (2015) Application of bias adjustment techniques to improve air quality forecasts. Atmospheric Pollution Research 6:928. https://doi.org/10.1016/j.apr.2015.04.002
6. Qin, S., Liu, F., Wang, J., Sun, B. (2014). Analysis and forecasting of the particulate matter (PM) concentration levels over four major cities of China using hybrid models. Atmospheric Environment, 98, 665. https://doi.org/10.1016/j.atmosenv.2014.09.046
7. Blair, G.S., Henrys, P.A., Leeson, A., Watkins, J., Eastoe, E., Jarvis, S.G., Young, P.J.: Data science of the natural environment: A research roadmap. Frontiers in Environmental Science 7, Article 121 (2019). https://doi.org/10.3389/fenvs.2019.00121
8. Dosemagen, S., Williams, E.: Data usability: The forgotten segment of environmental data workflows. Frontiers in Climate 4, Article 785269 (2022). https://doi.org/10.3389/fclim.2022.785269
9. Okafor, N.U., Alghorani, Y., Delaney, D.: Improving data quality of low-cost IoT sensors in environmental monitoring networks using data fusion and machine learning approach. ICT Express 6(3), 220–228 (2020). https://doi.org/10.1016/j.icte.2020.06.004
10. Kim, T.-S., Kim, J.-H., Yang, W., Lee, H., Choo, J.: Missing value imputation of time-series air-quality data via deep neural networks. Int. J. Environ. Res. Public Health 18(22), 12213 (2021). https://doi.org/10.3390/ijerph182212213
11. Jaoudé, A.A., Abou, A., Qamber, I.S., Al-Hamad, M.Y., Turabieh, H., Sheta, A., Braik, M., Kovacˇ-Andric', E., Saroha, S., Aggarwal, S.K., Rana, P., Sanjeev, K., Deneshkumar, V., Kannan, K.S., Niraikulathan, S.M., Mado, I., Aman, Z., Ezzine, L., Erraoui, Y., ... Moussami, H.E.: Forecasting in mathematics: Recent advances,

 doi:10.20944/preprints202603.1325.v1

new perspectives and applications. In: IntechOpen eBooks. IntechOpen (2020). https://doi.org/10.5772/intechopen.87892
12. Phan, T.-T.-H.: Elastic matching for classification and modelisation of incomplete time series. HAL (Le Centre pour la communi- cation scientifique directe) (2018). https://tel.archives-ouvertes.fr/tel-02001195
13. Ribeiro, S.M., de Castro, C.L.: Missing data in time series: A review of imputation methods and case study. Learning and Nonlinear Models 20(1), 31 (2022). https://doi.org/10.21528/lnlm-vol20-no1-art3
14. Rahman, N.H.A., Lee, M.H.: Artificial neural network forecasting performance with missing value imputations. IAES Int. J. Artif. Intell. 9(1), 33–39 (2020). https://doi.org/10.11591/ijai.v9.i1.pp33-39
15. Hoffman, S.: Estimation of prediction error in regression air quality models. Energies 14(21), 7387 (2021). https://doi.org/10.3390/en14217387
16. Safarov, R., Shomanova, Z., Nossenko, Y., Kopishev, E., Bexeitova, Z., Kamatov, R.: Filling gaps in PM2.5 time series: A broad evalu- ation from statistical to advanced neural network models. PLoS ONE 20(8) (2025). https://doi.org/10.1371/journal.pone.0330211
17. Libasin, Z., Fauzi, W.S.W.M., Ul–Saufie, A.Z., Idris, N.A., Mazeni, N.A.: Evaluation of single missing value imputa- tion techniques for incomplete air particulates matter (PM10) data in Malaysia. Pertanika J. Sci. Technol. 29(4) (2021). https://doi.org/10.47836/pjst.29.4.46
18. AQE: About Air Quality England. (2020). https://www.airqualityengland.co.uk/about
19. Ramadan, M.S., Abuelgasim, A., Hosani, N.A.: Advancing air quality forecasting in Abu Dhabi, UAE using time series models. Frontiers in Environmental Science 12 (2024). https://doi.org/10.3389/fenvs.2024.1393878