1. **Introduction**

In health analytics, preprocessing data is a crucial practice that enables effective analysis. In the absence of such methodology, the relevant decision makers are hindered at imperative operations such as policy formation. In this report, I will examine the different methods of data imputation that can be used to produce meaningful insights from datasets that contain missing values. By using air quality data that was collected for the Sebokeng area from January 2011 – February 2020, I will implement multivariate imputation & assess its impact on trends. The dataset is assumed to be missing at random, consequently the Iterative Imputer accessible in the scikit-learn library is the most effective in reducing bias & I will demonstrate why other methods such as kNN are inadequate.

1. **Method**
   1. **Dataset**

Legislations such as the National Framework for Air Quality Management are put in place for the government to ensure efficient air quality management (Department of Environmental Affairs, 2018). Therefore, five pollutants are examined in this study, namely Sulphur Dioxide (SO₂), Nitrogen Dioxide (NO₂), Ozone (O₃), Particular Matter (PM₂.₅ & PM₁₀). The concentration levels of these pollutants were collected & consequently the dataset consist of 3345 instances. Loading & inspecting the dataset are processes that were enabled through the use of the Pandas library. Subsequently, the tasks of imputing the data & creating a time series plot required that I utilize scikit-learn, numpy & matplotlib respectively.

* 1. **Missing values**

**Table 1.** Missingness in the dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Pollutant** | **Missing Count** | **% Missing** | **Non-missing Count** | **% Non-missing** |
| SO₂ | 1420 | 42.4% | 1925 | 57.6% |
| NO₂ | 1290 | 38.6% | 2055 | 61.4% |
| O₃ | 974 | 29.1% | 2371 | 70.9% |
| PM₂.₅ | 1492 | 44.6% | 1853 | 55.4% |
| PM₁₀ | 1518 | 45.4% | 1827 | 54.6% |

From table 1, it becomes clear that missing data is not evenly distributed across the pollutants. PM10 shows the largest proportion of missing values, closely followed by PM2.5, both with gaps in roughly four out of every ten records. Given that every pollutant has at least a quarter of its readings missing, removing rows with gaps would result in a substantial loss of information. This uneven pattern supports the assumption that data is missing at random, since gaps in one variable may be explained by the presence of other recorded variables. Studies on low-cost PM sensors further support this observation, finding that sensor performance degrades over time, increasing the proportion of failed or missing readings (Sayahi et al., 2022).

* 1. **Data description**

**Table 2.** Dataset distribution

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **SO₂** | **NO₂** | **O₃** | **PM₂.₅** | **PM₁₀** |
| **Count** | 1925 | 2055 | 2371 | 1853 | 1827 |
| **Mean** | 13.877048 | 25.350595 | 55.590021 | 31.673353 | 46.370742 |
| **Std** | 11.669311 | 10.510577 | 24.103645 | 18.867618 | 24.009341 |
| **Min** | 0.190387 | 0.000000 | 0.000000 | 3.634714 | 2.460333 |
| **P25** | 6.187676 | 18.360628 | 37.442500 | 19.661917 | 30.136500 |
| **P75** | 17.229557 | 30.367969 | 71.014333 | 38.294458 | 56.197313 |
| **Max** | 115.727693 | 115.438737 | 199.584250 | 288.647042 | 194.949870 |

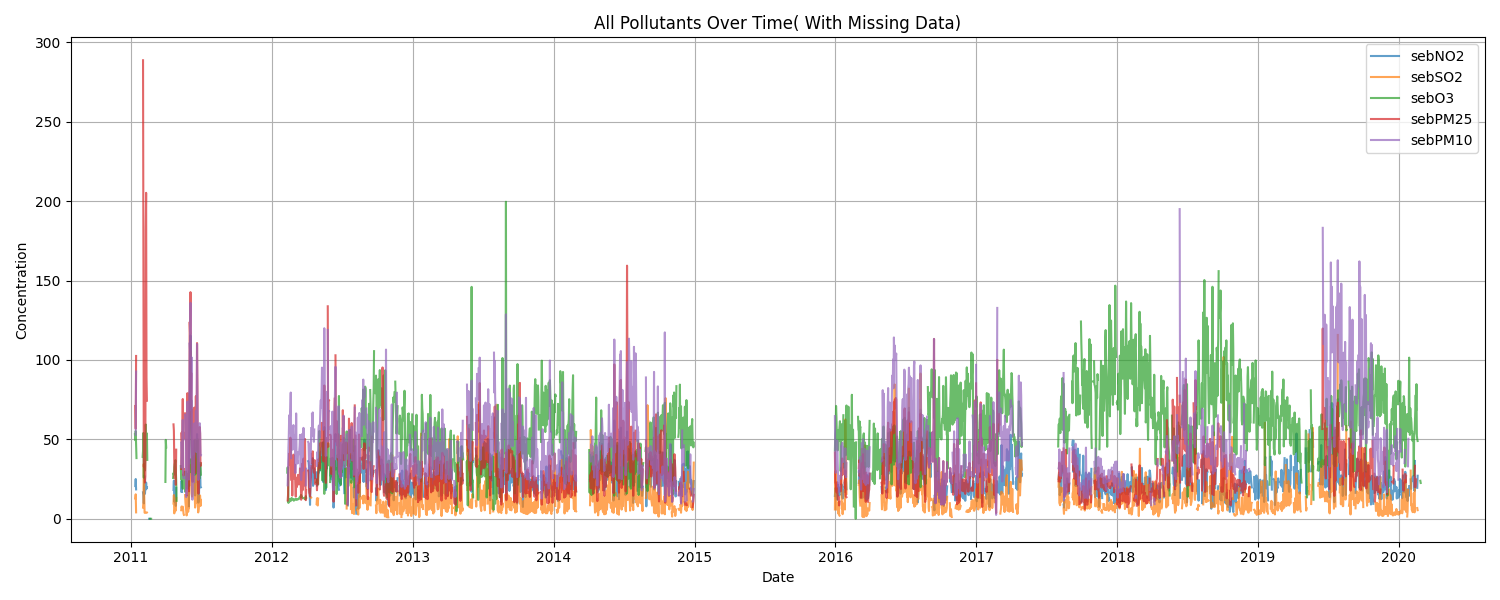
Table 2 presents the mean, median, 25th percentile (P25), 75th percentile (P75), minimum and maximum concentration levels, calculated using only the non-missing values. In the case of SO₂, the average concentration is relatively low compared to the maximum recorded value, which may suggest occasional high spikes. NO₂ shows a similar pattern, with the upper quartile (P75) noticeably above the median, hinting at a skewed distribution towards higher concentrations. O₃ levels tend to be more balanced, with the interquartile range smaller in proportion to the total range. PM₂.₅ and PM₁₀ both have high maximum values compared to their medians, suggesting episodes of elevated particulate matter levels. These descriptive statistics will be useful for two reasons: firstly, they provide a baseline for comparing the data before and after imputation, and secondly, they help to detect any unusual patterns or extreme readings that may require further examination during the analysis.

1. **Results**

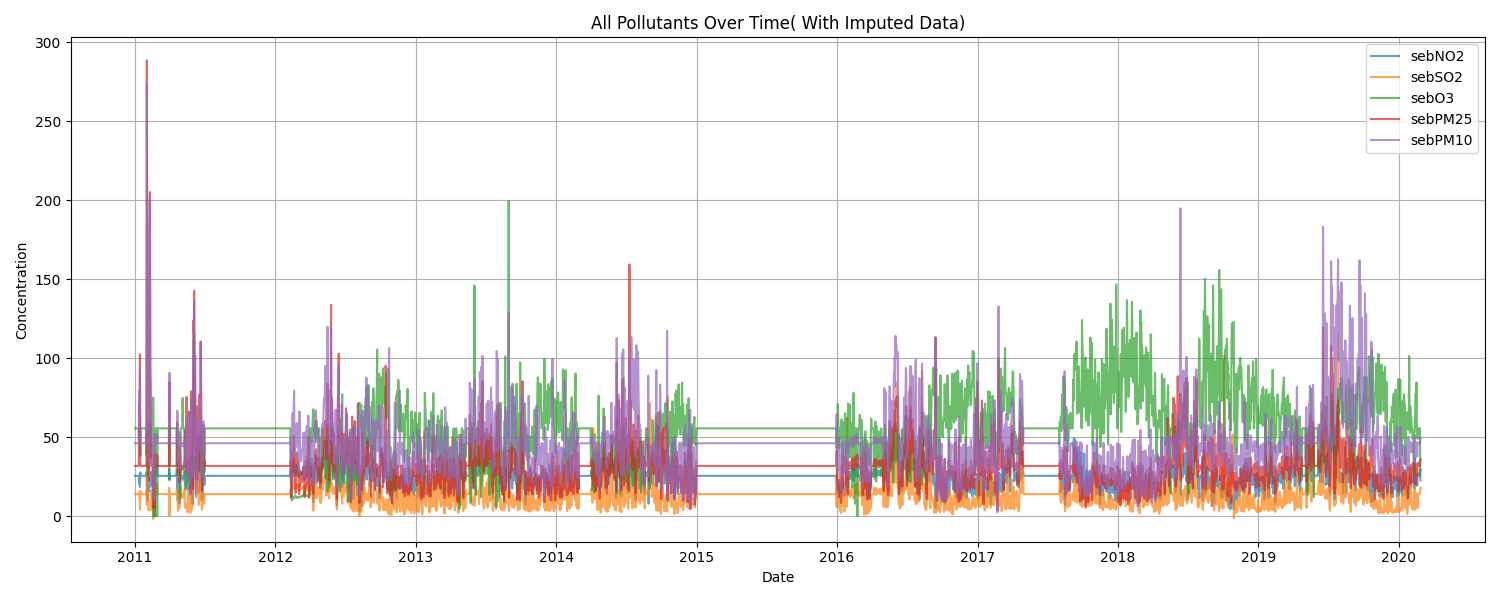
For imputing the missing values in the Sebokeng dataset, I employed the Iterative Imputer method, which is an implementation of Multivariate Imputation by Chained Equations (MICE). This approach was selected over simpler methods such as mean or median imputation, and alternatives like k-Nearest Neighbors (kNN), due to its ability to leverage relationships between variables and produce less biased estimates. Mean or median imputation, while straightforward, can underestimate variability and distort correlations in the data. Similarly, kNN imputation, although more sophisticated, relies heavily on the similarity between samples and may not fully capture the multivariate nature of missingness in this dataset (Yoon et al., 2019; Waljee et al., 2013).

MICE works by iteratively modelling each variable with missing values as a function of the other variables in the dataset (Waljee et al., 2013). In each iteration, the imputer fits a regression model to predict missing values in one variable using the observed values from the others. This process is repeated for a set number of iterations, in this case ten to refine the imputations progressively. The default regression model used is Bayesian Ridge regression, which provides a robust framework by incorporating regularization and uncertainty in parameter estimates (Jakobsen et al., 2017). MICE is particularly suitable because its iterative, multivariate nature allows it to utilize available information across variables to produce more accurate and consistent imputations, thereby reducing potential bias and improving the statistical quality of the completed dataset (Jakobsen et al., 2017).

**Figure 1.** Gaps in Data



**Figure 2.** Imputed Data



**Table 3.** Imputed Dataset Distribution

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **SO₂** | **NO₂** | **O₃** | **PM₂.₅** | **PM₁₀** |
| **Count** | 3345 | 3345 | 3345 | 3345 | 3345 |
| **Mean** | 13.997828 | 25.547520 | 55.631861 | 31.826846 | 46.181893 |
| **Std** | 9.044333 | 8.396836 | 20.316783 | 14.643527 | 19.120029 |
| **Min** | -1.639424 | 0.000000 | 0.000000 | 3.634714 | 2.460333 |
| **P25** | 8.710722 | 21.125247 | 43.847417 | 24.594458 | 36.456864 |
| **P75** | 15.186909 | 27.566048 | 63.566500 | 33.854667 | 48.066132 |
| **Max** | 115.727693 | 115.438737 | 199.584250 | 288.647042 | 273.435889 |

In the original dataset as it can be seen in figure 1, discontinuous traces with frequent gaps, particularly evident in PM₁₀ during 2016 and NO₂ between 2013 and 2014, reflected the challenges of missingness of the data. Following imputation, these gaps were seamlessly filled without introducing artificial spikes or distorting seasonal patterns, as demonstrated by the continuous curves in Figure 2. Notably, the imputed values for PM₂.₅ adhered to expected seasonal trends, with higher concentrations during winter months (June–August), consistent with increased particulate matter from domestic heating and stagnant atmospheric conditions. This alignment with known environmental patterns underscores the validity of the MAR assumption, where missingness could be explained by observed variables such as temperature or other pollutant levels. Consequently, extreme pollution episodes remained intact post-imputation, ensuring that high-impact events retained their significance for public health studies. However, a modest reduction in variability was observed, as evidenced by the narrower concentration bands for O₃ and the decreased standard deviations reported in Table 3. This smoothing effect suggests that while the completed dataset is robust for trend analysis, we should remain cautious when interpreting short-term fluctuations.

The imputation process led to modest upward shifts in the mean for most pollutants as illustrated in table 3. For example, SO₂ increased from 13.877 to 13.9978 , and NO₂ from 25.3506 to 25.5475. PM₂.₅ rose from 31.6734 to 31.8268, while PM₁₀ showed a smaller increase from 46.3707 to 46.1819, indicating that imputed values tended to align with slightly higher readings observed in neighbouring time points. Standard deviations decreased across pollutants, such as SO₂ from 11.6693 to 9.0443 and PM₂.₅ from 18.8676 to 14.6435, suggesting that extreme short-term fluctuations were smoothed during imputation. Quartile values (P25, P75) generally shifted upward, most notably for PM₂.₅ and PM₁₀, consistent with the MAR assumption where higher readings in correlated pollutants influenced the predicted values. Importantly, the maxima remained unchanged across all pollutants, showing that the procedure did not artificially create extreme values. Overall, the multivariate imputation preserved the overall distributional shape while slightly reducing spread, which helps maintain trend integrity and reduces bias in subsequent statistical analyses (Jakobsen et al., 2017).

1. **Discussion**

Notably, extreme peaks were preserved in the imputed dataset. The maximum recorded concentrations for each pollutant remained unchanged, ensuring that the dataset still captures severe pollution episodes that may have significant public health implications. However, because imputed values are generated using statistical models rather than actual measurements, they may not perfectly replicate the true variability of the environment. In cases where missingness coincided with extreme events, there remains a risk that imputation produced conservative estimates, potentially underestimating peak intensity.

A key strength of this approach is the creation of a complete, continuous dataset, which greatly improves the reliability of statistical analyses and predictive modelling. For example, health authorities can now perform more robust time-series analyses linking pollutant concentrations to hospital admissions or respiratory disease incidence, without losing statistical power due to missing data. Nevertheless, a limitation is that the imputed values are model-dependent and could introduce bias if the underlying missingness mechanism deviates from the assumed MAR (Missing At Random) condition.

Overall, the iterative, multivariate nature of the imputation preserved the dataset’s core distributional properties while improving completeness, enabling stronger and more reliable environmental health assessments. These improvements, however, should be interpreted alongside an understanding of the method’s assumptions and potential biases to ensure that subsequent public health decisions are well-informed and contextually appropriate.

1. **Conclusion**

The iterative imputation of the Sebokeng air quality dataset successfully addressed missing values while preserving the core statistical characteristics of the original data. The process produced a complete and continuous time series, enabling more robust and reliable trend analyses. Modest increases in mean values and reductions in standard deviations indicate that the imputation smoothed short-term fluctuations without eliminating significant pollution peaks, thereby retaining the dataset’s ability to capture critical air quality events.

This enhanced completeness strengthens the potential for accurate environmental and public health modelling, particularly in studies linking pollutant exposure to health outcomes. However, results must be interpreted with an understanding of the MAR assumption and the possibility of bias if the missingness mechanism deviates from it. While imputation offers substantial analytical benefits, it is not a substitute for high-quality, uninterrupted monitoring. Overall, the completed dataset represents a significant improvement in data quality and usability, supporting more informed decision-making for air quality management and public health protection in Sebokeng.

**REFERENCES:**

Department of Environmental Affairs (2018) *National Framework for Air Quality Management in the Republic of South Africa*, Notice No. 1144, Government Gazette No. 41996. Available at: [https://www.gpwonline.co.za](https://www.gpwonline.co.za/) (Accessed: [2025-08-10]).

Jakobsen, J.C., Gluud, C., Wetterslev, J. and Winkel, P., 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Medical Research Methodology*, 17(1), p.162. <https://doi.org/10.1186/s12874-017-0442-1>.

Kebalepile, M. M., Dzikiti, L. N., & Voyi, K. (2024). Using Diverse Data Sources to Impute Missing Air Quality Data Collected in a Resource-Limited Setting. *Atmosphere, 15*(3), 303. <https://doi.org/10.3390/atmos15030003>.

Sayahi, T., Butterfield, A., Kelly, K.E. and Gaillardon, P.E. (2022) ‘Long-term performance of low-cost PM sensors for air quality monitoring’, *arXiv*. Available at: <https://arxiv.org/abs/2210.14759> (Accessed: 9 August 2025).

Waljee, A.K., Mukherjee, A., Singal, A.G., Zhang, Y., Warren, J., Balis, U., Marrero, J. and Higgins, P.D.R., 2013. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, 3(8), e002847. <https://doi.org/10.1136/bmjopen-2013-002847>.

Yoon, J., Jordon, J. and van der Schaar, M., 2019. Gain: Missing data imputation using generative adversarial nets. *Proceedings of the 36th International Conference on Machine Learning*, 97, pp.5689-5698.