Homework 4

Topics on Optimization and Machine Learning

Roberto Meroni
email: roberto.meroni@estudiantat.upc.edu

June 2024

Contents

1 Introduction

In Internet of Things (IoT) architecture, it is not uncommon to encounter missing data. This data loss can occur due to various reasons, such as sensor malfunctions, communication issues, or power outages. Instead of discarding the incomplete datasets, we aim to evaluate the feasibility of accurately estimating missing data in IoT sensor networks, maintaining the integrity and usefulness of the data.

We will generate random missing values in a dataset comprising measurements of O_3 from eight different sensors placed in Barcelona, and attempt to estimate the missing data with two main categories of methods:

- Time-series forecasting models to predict missing values based on past data from the same sensor.

- Multivariate models that utilize data from other sensors to reconstruct the missing values.

Finally, we will assess the accuracy of the reconstruction methods using metrics such as R-squared (R^2) and Root Mean Square Error (RMSE), aiming to identify the most effective methods for reconstructing missing IoT sensor data.

2 Univariate Models

2.1 Polynomial Interpolation

Missing input values: Percentage of Missings

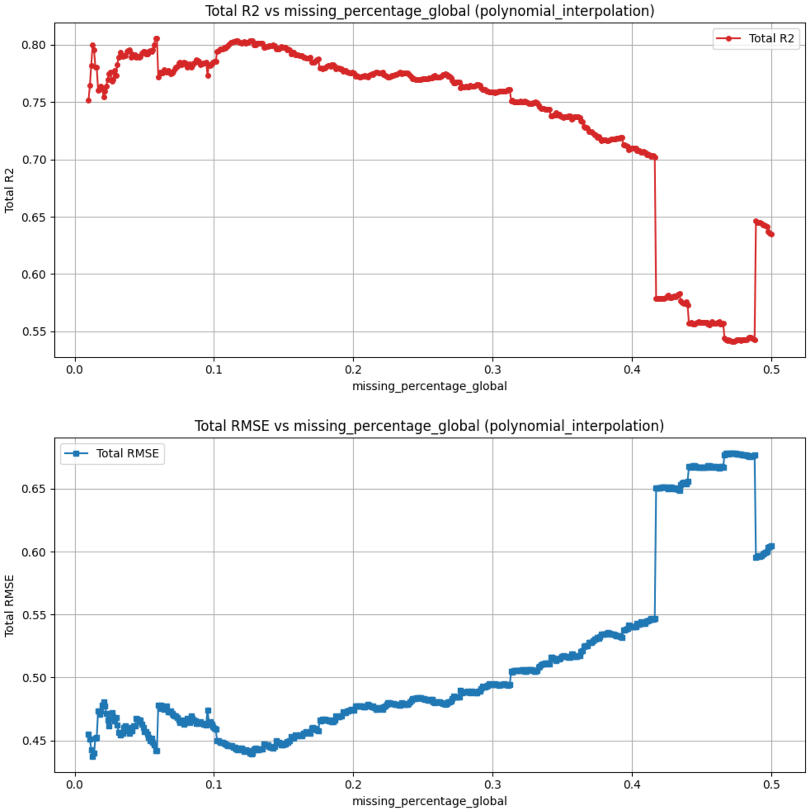

Figure 1: Performance of the model for different values of missing inputs

The accuracy of the model significantly decreases as the percentage of missing values increases. The ”step” observed for high missing percentage values is possibly due to the manner in which the missings are produced, which is also constrained by the parameter burst_length.

Missing input values: Bursts Length

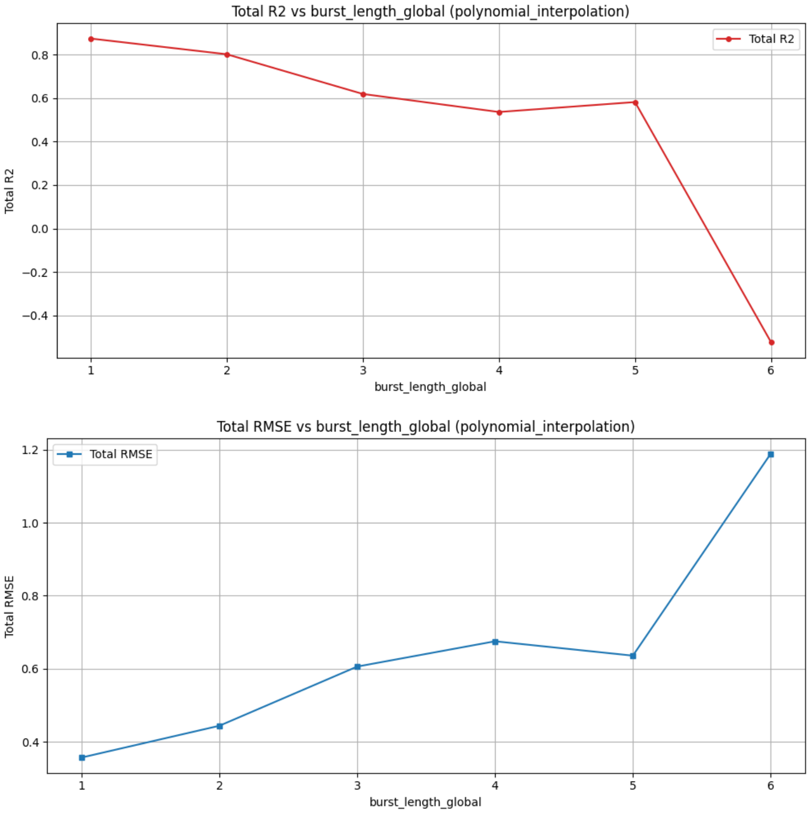

Figure 2: Performance of the model for different lengths of missing inputs bursts

For low values of burst length the model returns a consistent accuracy; however, for values above 5 the polynomial interpolation becomes unreliable.
As we can observe from Figure 3, when the gap between two values used for the interpolation becomes too large, the predicted values in the middle become higly inaccurate.

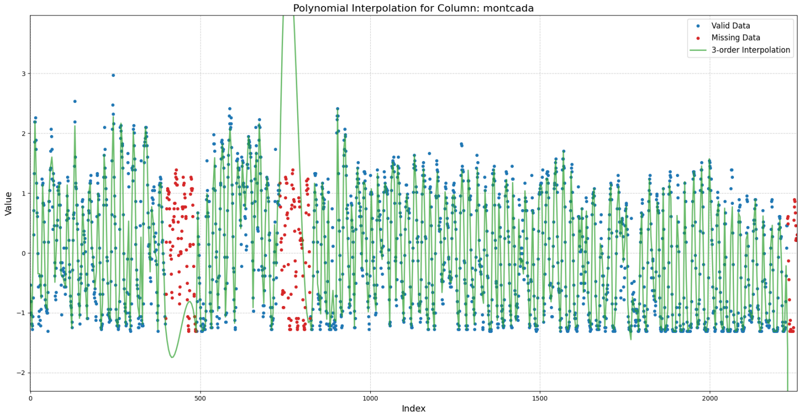

Figure 3: Interpolation function with real data (scaled)

2.2 Long Short-Term Memory (LSTM)

Neural Network Architecture

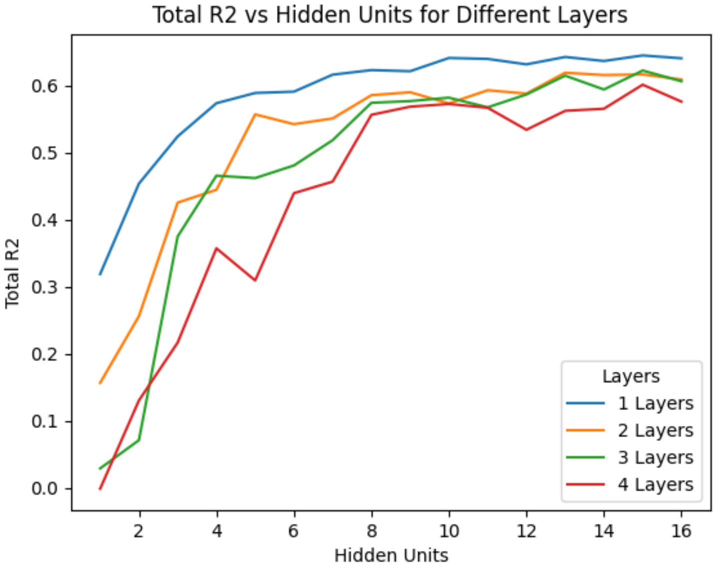

Figure 4: R-squared for number of neurons per layers, for different number of layers

Neurons Increasing the number of neurons per hidden layer significantly improves the accuracy of the model up to eight neurons, which is equivalent to the input layer dimension. Beyond this point, adding more neurons does not result in substantial performance gains; the model's accuracy stabilizes. This trend is consistent regardless of the number of hidden layers.

Hidden Layers For this particular dataset, a simpler model yields better results. Adding more hidden layers consistently reduces the model's accuracy, even if the decrease is minimal. This negative impact of additional layers is observed for nearly every number of neurons.

Epochs

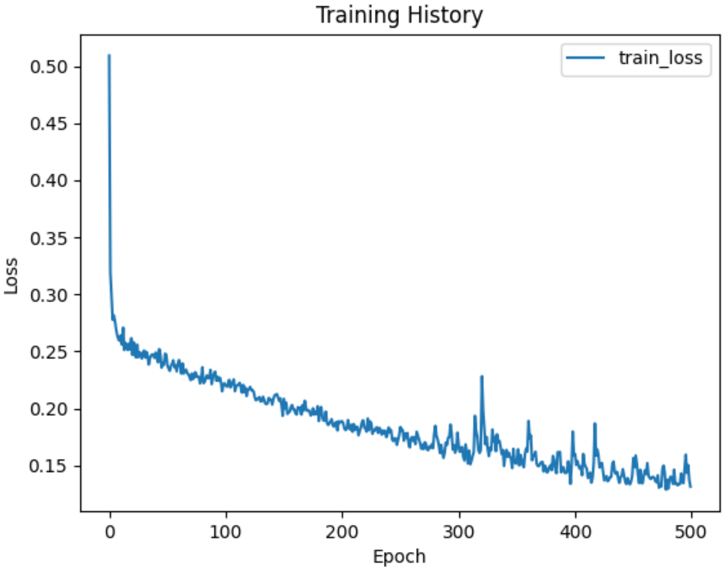

Figure 5: Loss for different numbers of Epochs

Except for a rapid decrease in loss during the initial epochs, the accuracy of the model improves slowly as the number of epochs increases, showing a progressive convergence. However, at higher epoch values, the model exhibits more noise in the loss curve.

Missing input values: Percentage of Missings

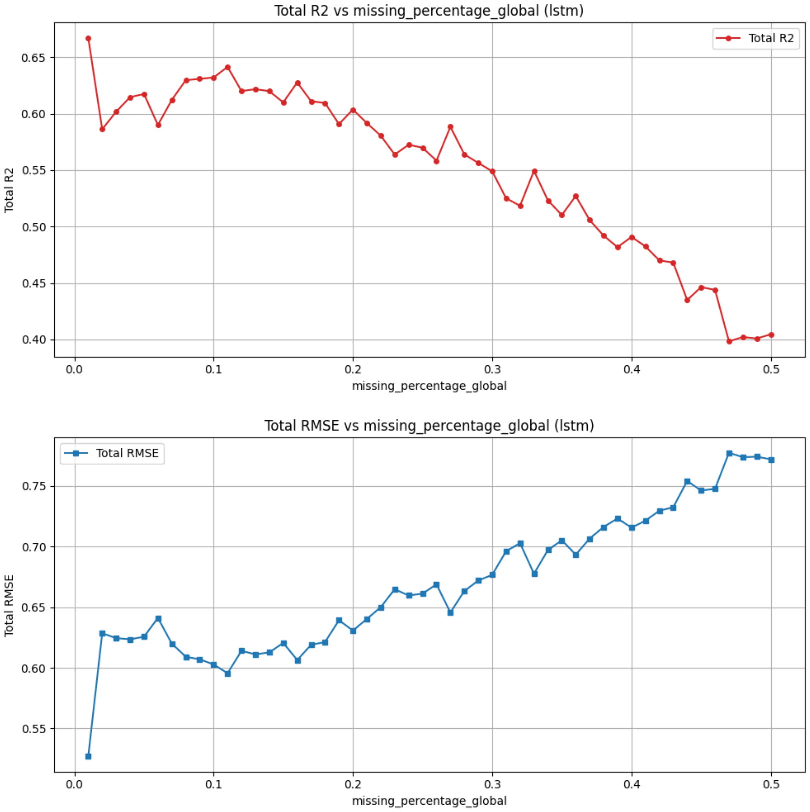

Figure 6: Performance of the model for different values of missing inputs

The accuracy of the LSTM model gradually decreases as the percentage of missing values increases. While a significant loss in performance is expected when a large portion of the dataset is missing, the LSTM model loses approximately 60% of its accuracy when the percentage of missing values increases from 0.01 to 0.5. This significant decline indicates a weakness in the model's robustness against missing values.

Missing input values: Bursts Length

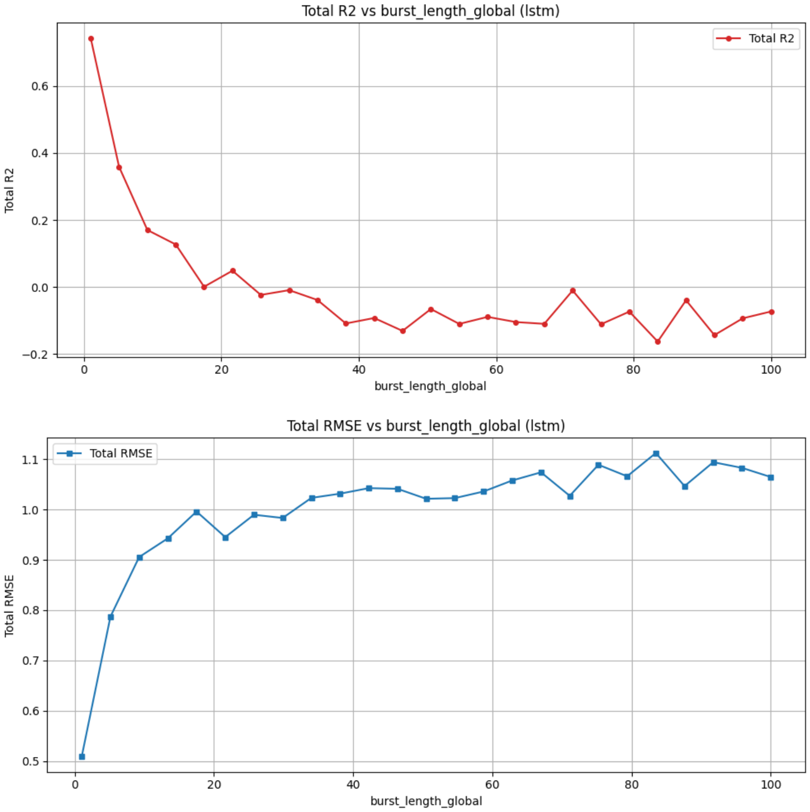

Figure 7: Performance of the model for different lengths of missing inputs bursts

When increasing the length of the bursts, the LSTM model rapidly becomes completely inaccurate. This is because LSTMs heavily rely on recent past values to make predictions. Consequently, having multiple consecutive missing inputs disrupts the continuity of the data, leading to unreliable predictions from the model. This dependency on sequential data makes LSTMs particularly vulnerable to long bursts of missing values, significantly degrading their performance.

3 Multivariate Models

3.1 MICE with Multiple Linear Regression

Missing input values: Percentage of Missings

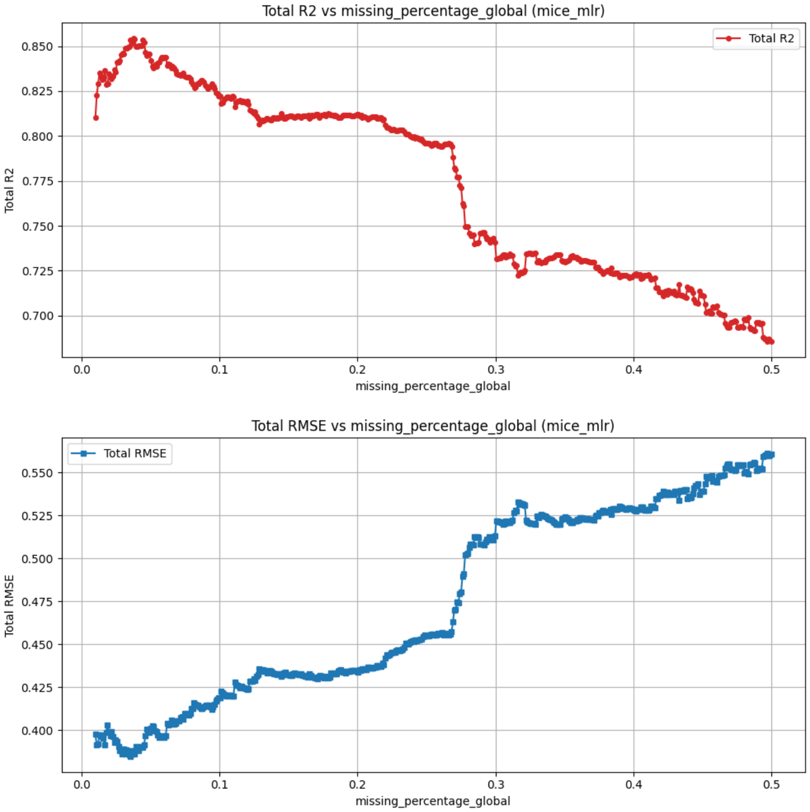

Figure 8: Performance of the model for different values of missing inputs

The performance of the model worsens as the percentage of missing values increases, with the R-squared score decreasing from greater than 0.8 when almost no values are missing to less than 0.7 when half of the dataset is missing. Although this performance loss is significant, it is expected given the reduction to half of the dataset. R-squared values remain above 0.65 for every percentage of missing data, demonstrating the model's robustness to high portions of missing data.

Missing input values: Bursts Length

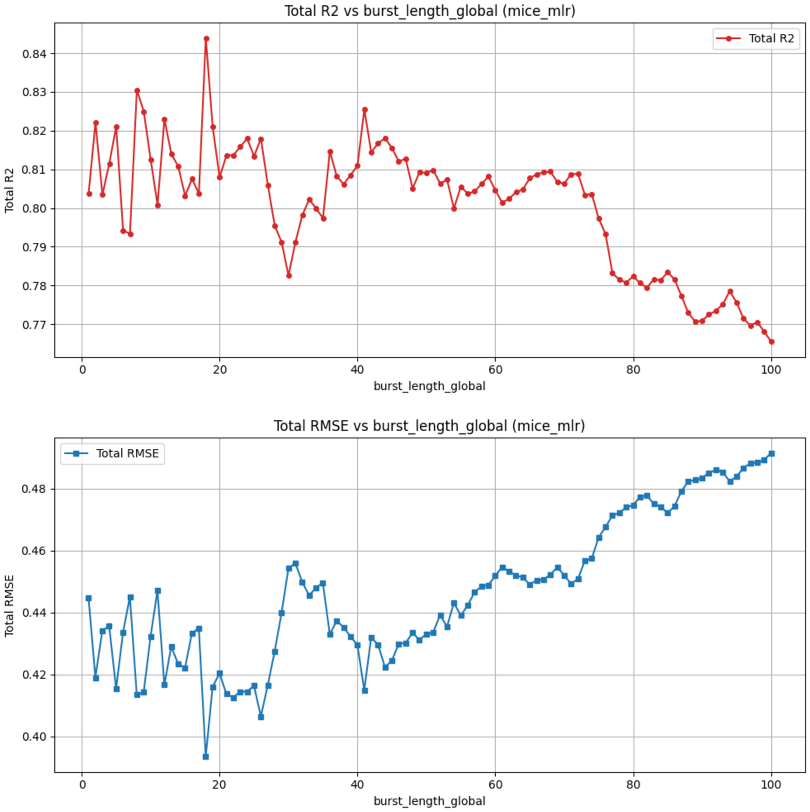

Figure 9: Performance of the model for different lengths of missing inputs bursts

MICE with Multiple Linear Regression (MLR) exhibits only a slight loss in accuracy as the length of the missing input bursts increases, albeit with significant variance. Even with bursts of 100 missing values, the model achieves R-squared scores above 0.75.

Incorporating values from other sensors in predicting the missing inputs enhances the model's robustness, ensuring good performance even with longer bursts of missing data. This indicates that leveraging additional sensor information effectively mitigates the impact of extended periods of missing inputs, maintaining the model's accuracy.

3.2 MICE with K-Nearest Neighbors

Missing input values: Percentage of Missings

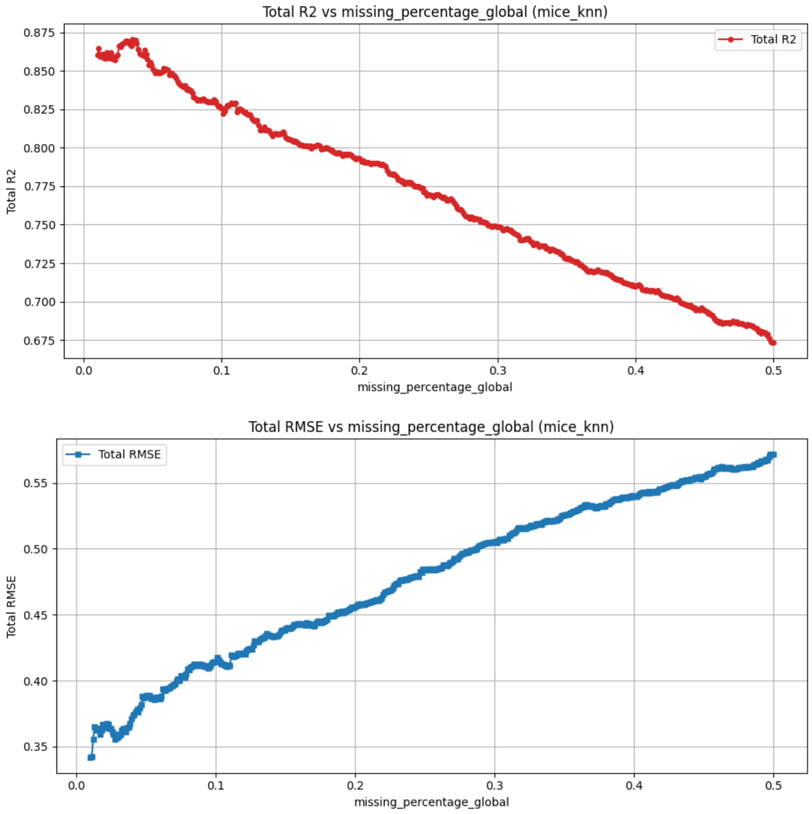

Figure 10: Performance of the model for different values of missing inputs

The accuracy of the model smoothly decays as the percentage of missing values increases, with the R-squared score decreasing from greater than 0.85 when almost no values are missing to less than 0.7 when half of the dataset is missing. MICE with KNN performs better than MICE with MLR for low number of values, but exhibits less robustness towards high number of missing inputs. The high number of missing values leads to difficulties in accurately imputing data using the KNN method, as KNN relies heavily on the availability of suitable neighboring data points to make predictions.

Missing input values: Bursts Length

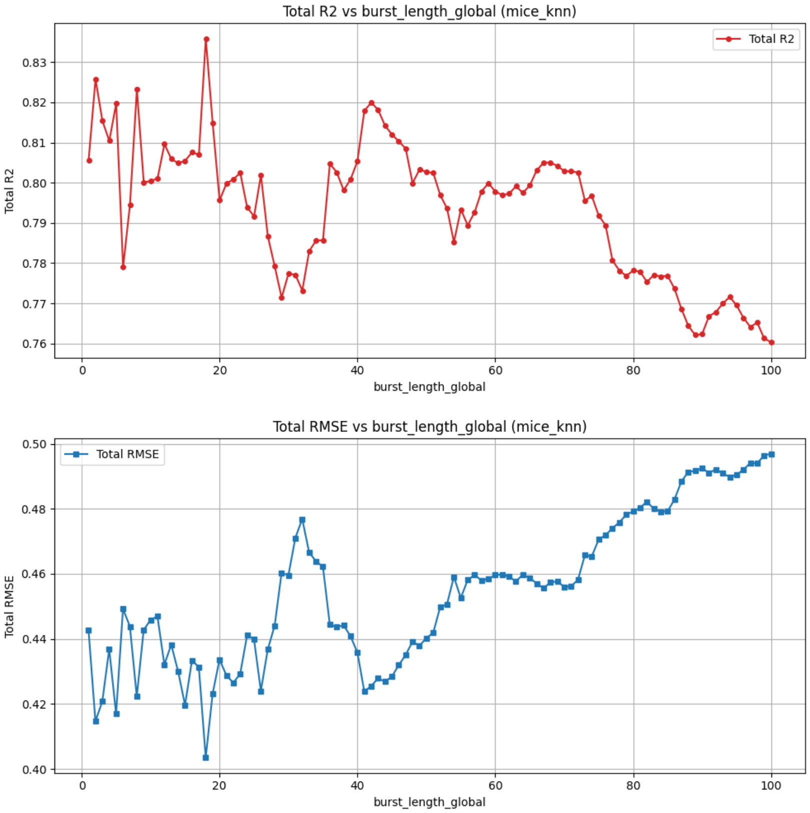

Figure 11: Performance of the model for different lengths of missing inputs bursts

MICE with KNN exhibits only a slight loss in accuracy as the length of the missing input bursts increases, with significant variance. Even with bursts of 100 missing values, the model achieves R-squared scores above 0.75.
Incorporating values from other sensors in predicting the missing inputs guarantees good robustness against longer bursts of missing data.

3.3 Autoencoder

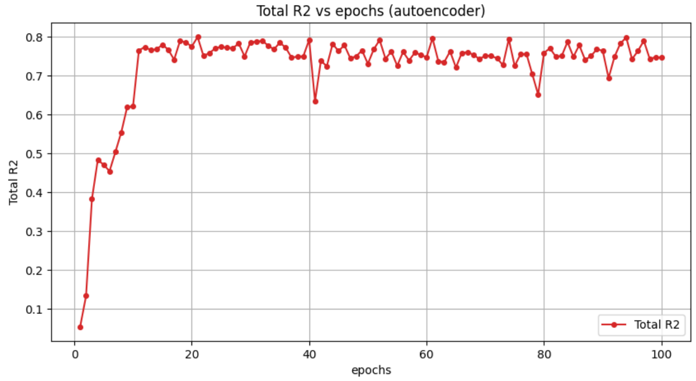

Figure 12: R-squared vs number of Epochs

Epochs

For low values, increasing the number of epochs immediately results in significant improvements in model performance. However, for values above 10, the accuracy of the model seems to converge, making it unnecessary to use a higher number of epochs. Beyond this point, additional epochs only increase the computational burden without any performance gain.

Latent Space Dimension

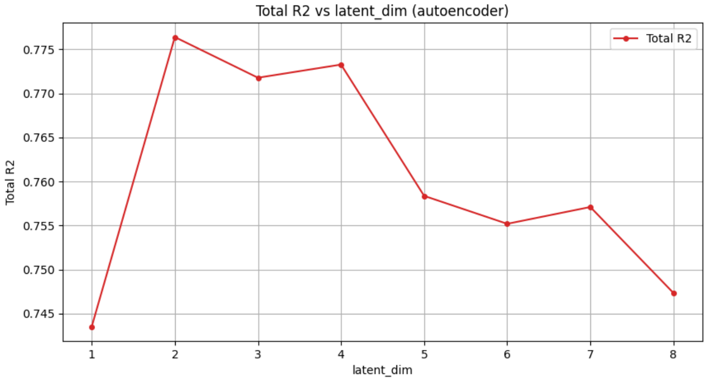

Figure 13: R-squared vs dimension of Latent Space

Aside from a latent dimension of 1, where a single neuron exhibits limited capability in capturing the dataset's patterns, the autoencoder demonstrates higher accuracy for lower dimensions of the latent space, specifically for dimensions of 2, 3, or 4.

For latent dimensions beyond 5, the model experiences a decrease in performance. This decline is likely due to overfitting or the model capturing noise, as the increased capacity of the latent space allows the model to memorize the training data rather than generalizing from it.

Missing input values: Percentage of Missings

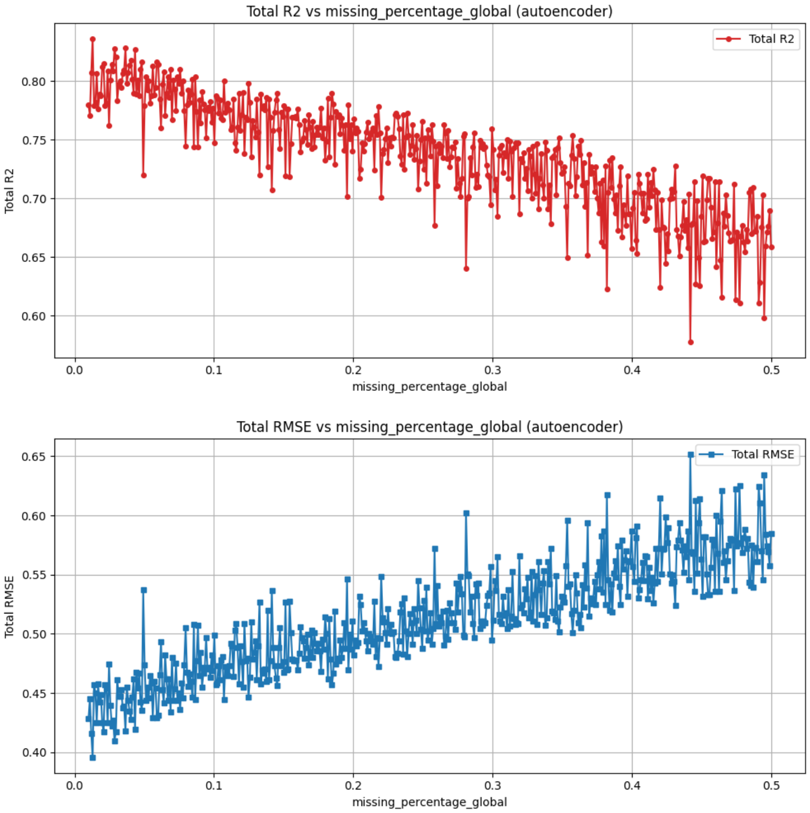

Figure 14: Performance of the model for different values of missing inputs

The accuracy of the model decreases as the percentage of missing data increases, exhibiting significant variance. Initially, R-squared values are above 0.8 for low percentages of missing data, ending below 0.7 with high percentage values. Although this performance loss is significant, it is expected given the reduction to half of the dataset. Also, after removing outliers, R-squared values remain above 0.65 for every percentage of missing data, demonstrating the model's robustness to high portions of missing data.

Missing input values: Bursts Length

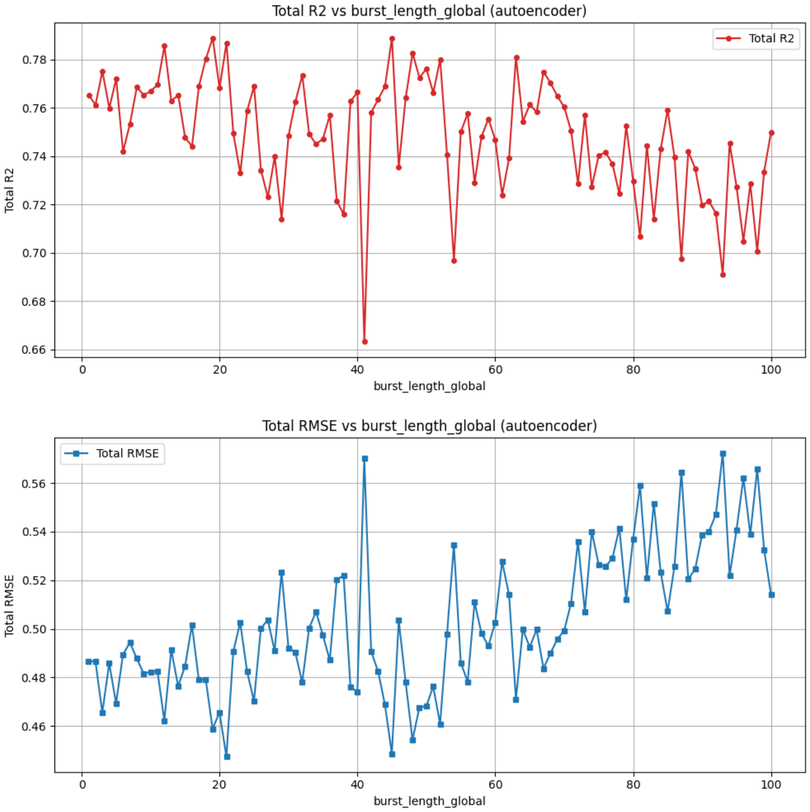

Figure 15: Performance of the model for different lengths of missing inputs bursts

The autoencoder exhibits only a slight loss in accuracy as the length of the missing input bursts increases. Even with bursts of 100 missing values, the model achieves R-squared scores above 0.7.
Incorporating values from other sensors in predicting the missing inputs enhances the model's robustness, allowing it to handle longer bursts of missing data more effectively.

3.4 Dual Linear Regression

As observed in previous assignments, O_3 concentrations exhibit both short-term and long-term patterns. Short-term patterns include higher concentrations during the day and lower concentrations at night, while long-term patterns show higher levels during the summer and lower levels during the winter. Additionally, despite variations in ozone levels across different locations, significant correlations exist between O_3 measurements in different areas of Barcelona.

The idea behind this model is to combine both short-term and long-term effects, along with correlations between sensors. The short-term effects are captured using a local estimator (linear regression with recent sensor values), and the long-term effects and inter-sensor correlations are addressed using a global estimator (MICE with MLR).

Also, sensors exhibit varying degrees of correlation with each other. For sensors with weak correlations to others, the local estimator should be given more weight. Conversely, for sensors with strong correlations, the global estimator should be prioritized.

To implement this, a new linear regression model is fitted for each sensor, using the local and global estimations as independent variables.

The final estimation of a value is, therefore, a combination of the local estimator and the global estimator, weighted by the coefficients of the trained linear regression model.

Missing input values: Percentage of Missings

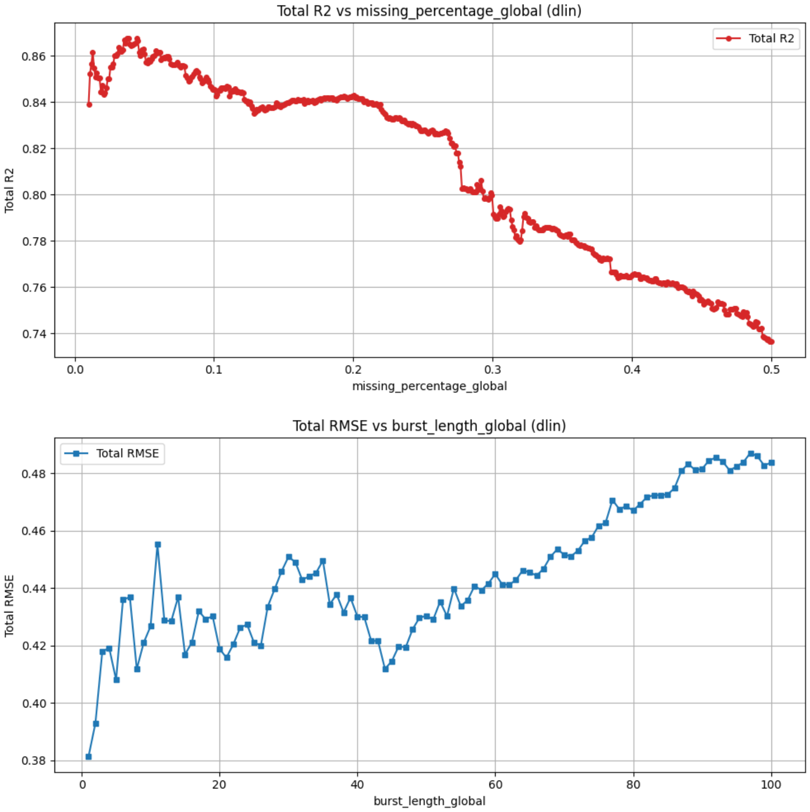

Figure 16: Performance of the model for different values of missing inputs

The performance of the Dual Linear Regression model worsens as the percentage of missing values increases, with the R-squared score decreasing from greater than 0.85 when almost no values are missing to less than 0.75 when half of the values are missing.

While a decrease in accuracy is expected when a large portion of data is missing, the Dual Linear Regression model exhibits strong robustness towards missing inputs. This robustness is attributed to the mitigation effect provided by the combination of the global predictor and the local predictor. The dual approach ensures that even with significant amounts of missing data, the model maintains relatively high performance, leveraging the strengths of both predictors to compensate for the gaps in the data.

Missing input values: Bursts Length

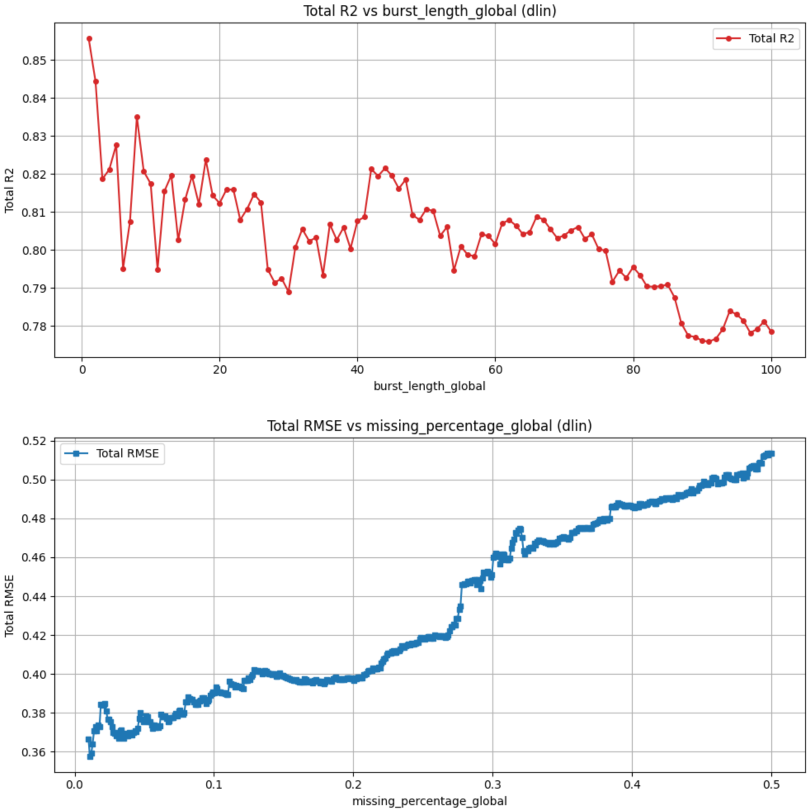

Figure 17: Performance of the model for different lengths of missing inputs bursts

The Dual Linear Regression exhibits only a slight loss in accuracy as the length of the missing input bursts increases. Even with bursts of 100 missing values, the model achieves R-squared scores of approximately 0.78.
When consecutive values are missing, the local estimator becomes weaker and the model relies more on the global estimator. The flexibility and adaptivity of the model enhance its robustness toward dirty datasets, as it could be one with many consecutive missing inputs.

4 Model Comparison

Missing input values: Percentage of Missings

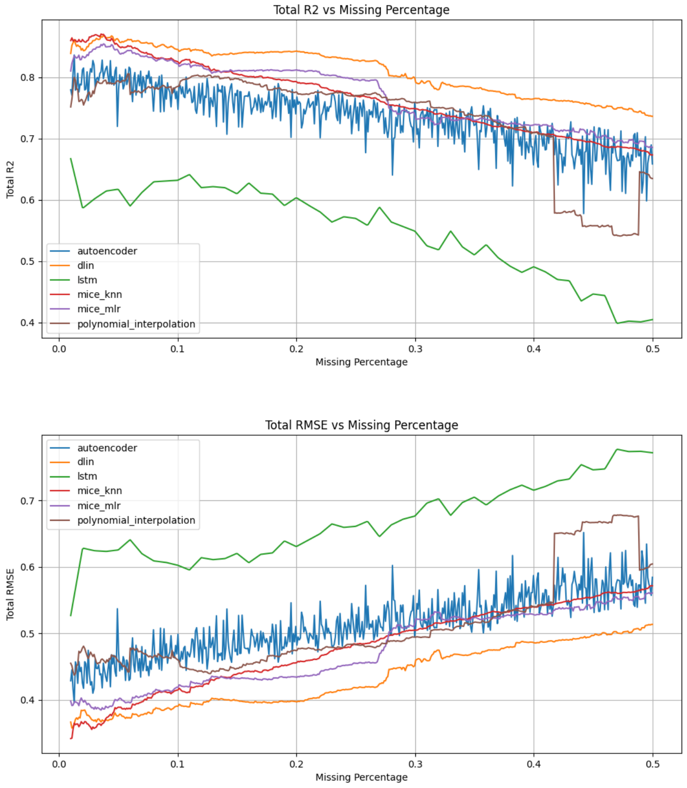

Figure 18: Comparison between models for values of input missing percentage

Except for very low values, the Dual Linear Regression model exhibits the best performance across the entire range of missing percentages. Following it, the two MICE models perform better than the other approaches; in particular, MICE with KNN shows high performance with low percentages of missing inputs, while MICE with MLR achieves better accuracy for higher percentages of missing inputs.

Polynomial Interpolation performs slightly better than the Autoencoder, which also displays significantly higher variance compared to the other models. LSTM performs significantly worse than the other models across all ranges of missing input percentages.

Missing input values: Bursts Length

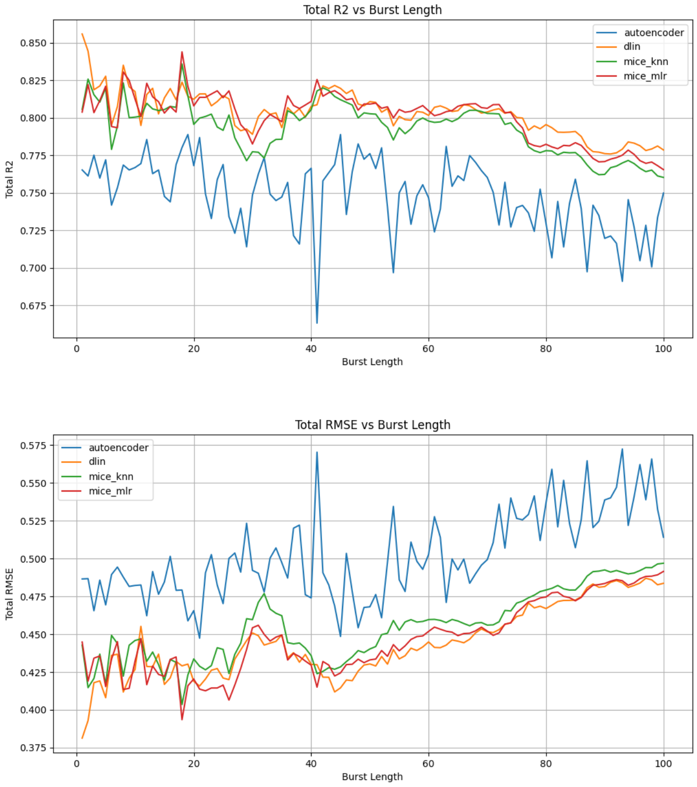

Figure 19: Performance of the model for different lengths of missing inputs bursts

Except for very low values where Dual Linear Regression has the highest accuracy, the two MICE models and the Dual Linear Regression model performs similarly. In fact, when the length of bursts is increased the Dual Linear Regression heavily relies on the global estimator, which is a MICE with MLR.
The Autoencoder has significantly lower accuracy and higher variance compared to the other models. The performances of the univariate models (Polynomial Interpolation and LSTM) for long bursts were very poor and the results were unreliable, so they are not included in this analysis.

5 Conclusions

The two univariate models performed consistently well with short lengths of missing bursts. However, due to their predictive nature, which heavily relies on recent past values, they become unreliable when multiple consecutive inputs are missing.

Conversely, the multivariate models exhibited robustness towards long-length bursts, thanks to their ability to consider measurements from other sensors when estimating missing inputs. This capability allows them to maintain higher accuracy and reliability even when faced with extended periods of missing data.

In general, the neural network methods (LSTM and Autoencoder) performed worse than the traditional methods (Polynomial Interpolation and MICE). For a limited dataset (2258 elements from 8 sensors), the neural networks were not able to show their potential, and the traditional statistical methods proved to be more suitable for accurately predicting missing values.
Within the class of traditional methods, the MICE models outperformed Polynomial Interpolation, highlighting the importance of including correlated sensors in the prediction process. This result underscores the advantage of leveraging multivariate data to improve the accuracy and robustness of missing value imputation.

Among all the models, the Dual Linear Regression model performed the best. By combining a local estimator (linear regression with recent sensor values) and a global estimator (MICE with MLR), this method captures both short-term and long-term effects, along with inter-sensor correlations. The two estimators are aggregated using a linear regression model tailored to each sensor. This flexible approach enhances the performance and robustness of the model for two main reasons:

- **Inter-sensor Correlations:** Some sensors have strong correlations, while others have less. More importance is given to the global estimator for strongly correlated sensors, while the local estimator is weighted more for weakly correlated sensors.
- **Handling Missing Bursts:** For short-length bursts of missing values, the local estimator significantly increases the performance of the model but becomes unreliable for long missing bursts. The Dual Linear Regression ensures consistent results even in the case of long missing bursts. However, it does not perform significantly better than a traditional MICE model.

Overall, this work demonstrated the possibility of reconstructing missing inputs using machine learning methods. Specifically, for models such as Dual Linear Regression and MICE methods, it is possible to estimate missing values with significant accuracy, even when a large part of the dataset is missing or long bursts of missing values are present.