TOML-MIRI Homework 3

Missing Value Imputation (MVI) in IoT Monitoring Networks

Ezgi Sena Karabacak
Spring 2024-2025

Introduction

Missing data is a common and critical problem in environmental time series, particularly in air quality monitoring systems where sensor faults, communication issues, or calibration gaps can lead to partial observations. Accurate imputation of these missing values is essential for downstream tasks such as forecasting, anomaly detection, and environmental policy evaluation.

In this study, we focus on the imputation of hourly ozone (O_3) concentration data collected from urban sensors across Barcelona. We evaluate a range of imputation strategies, spanning from classical statistical methods to modern machine learning approaches. Specifically, we compare polynomial interpolation, LSTM-based sequence modeling, Multivariate Imputation by Chained Equations (MICE) with both linear regression and k-nearest neighbors, and autoencoders.

To assess robustness and generalization, we systematically vary:

- the **percentage of missing values** (e.g., 5%, 10%, 20%),
- the **burst length** of missing intervals (e.g., 3, 5, 10 time steps),
- the **temporal window size** used in LSTM modeling,
- and the **number of sensors** used as input features.

Our analysis emphasizes the **Gràcia** sensor as a representative example while also benchmarking performance across multiple spatial locations. We evaluate the imputation quality using two key metrics: the root mean squared error (RMSE) and the coefficient of determination (R^2). The results offer insights into the trade-offs between model complexity, data availability, and imputation accuracy in practical urban sensing applications.

1 Data Analysis

1.1 Dataset Overview

We used hourly air quality data collected from three urban sensors in Barcelona: `gracia`, `eixample`, and `hebron`. The dataset spans several months and captures ozone (O_3) concentration levels. Each sensor provides a time series of measurements, and we treat this multivariate temporal data as the basis for imputation experiments.

Prior to any preprocessing, the data was cleaned to remove NaNs occurring at the start or end of the time series to ensure consistency across imputation models.

Figure 1 shows an example time series from the `gracia` sensor. It includes the original ozone data, as well as artificially introduced missing values (10% missing with bursts of 5 points). For illustration, we also show how a 3rd-degree polynomial fills in the missing parts. This example helps visualize the patterns in the data, the gaps caused by missing values, and the difficulty of recovering both sharp spikes and slow trends.

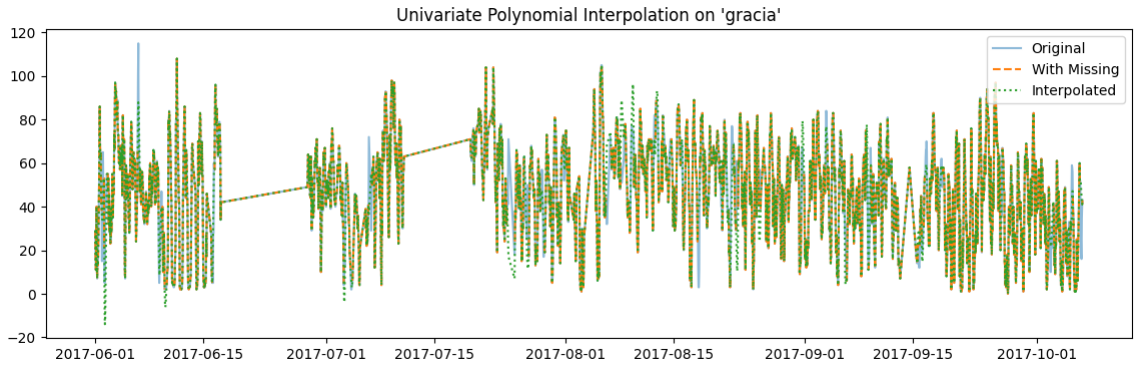

Figure 1: Ozone readings from the `gracia` sensor with missing values (10% bursty) and polynomial interpolation (degree 3) overlayed.

1.2 Sensor Selection and Spatial Considerations

To ensure that spatial correlations could be meaningfully leveraged during imputation, we selected a subset of three sensors located in close geographic proximity. Using the coordinates provided in the `Node-Location.csv` file, we computed pairwise Euclidean distances between all nodes and identified the tightest cluster. The sensors **Gràcia**, **Hebron**, and **Eixample** were selected as they form the most spatially cohesive group among the available nodes. Their proximity increases the likelihood of observing correlated pollution trends, which is particularly beneficial for methods that utilize spatiotemporal relationships, such as autoencoders and LSTMs.

1.3 Experimental Setup

To evaluate imputation methods under controlled conditions, we simulated artificial missingness by randomly removing some percent of the observed values. This synthetic masking allows us to compute ground truth errors (RMSE, R^2) by comparing the imputed values with the originally available data.

To ensure a fair comparison across all imputation methods, we adopted a standardized procedure in which a fixed random seed was used to generate a single mask of missing entries. This mask was then reused consistently across all models—including MICE, autoencoder, and LSTM—thereby eliminating variability that could arise from differing patterns of missing data. This approach guarantees that performance differences reflect model capability rather than inconsistencies in the evaluation setup.

1.4 Evaluation Metrics

For each sensor, we compute the following metrics:

- **RMSE** (Root Mean Squared Error): Quantifies the average magnitude of error between imputed and true values.
- R^2 (Coefficient of Determination): Measures the proportion of variance in the ground truth explained by the imputed values.

Let y_i be the true value at timestamp i, and $\hat{y}_i$ the imputed value. Then:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

2 First Experimentation

In our default experimental setting, we simulate missing values in a controlled and reproducible manner. For each sensor, we randomly introduce missing data such that **10% of the values are missing**, applied in **bursts of length 5** (i.e., five consecutive time points per burst). For LSTM, we used a fixed-size window of **20 time steps** to frame input sequences for the default scenario.

The results of different imputation strategies are summarized below:

- **Polynomial Interpolation (degree 3)** on `gracia`:
  - RMSE: 15.82
  - R^2: 0.537
- **LSTM (Window-based)** on `gracia`:
  - RMSE: 18.79
  - R^2: 0.349

  LSTM performed relatively worse in this setting. This is likely due to the fixed-size window used to frame the input sequences, which may not be well-suited for recovering longer missing bursts or for capturing the slow dynamics in O_3 data without careful tuning.
- **MICE (Multivariate Imputation by Chained Equations)**:
  - Linear Regression:
    - `gracia`: RMSE = 7.56, R^2 = 0.894
    - `eixample`: RMSE = 8.00, R^2 = 0.860
    - `hebron`: RMSE = 13.88, R^2 = 0.773
  - K-Nearest Neighbors:
    - `gracia`: RMSE = 8.49, R^2 = 0.867
    - `eixample`: RMSE = 8.85, R^2 = 0.828
    - `hebron`: RMSE = 15.08, R^2 = 0.732

For multivariate models, we used 3 sensors for imputation for the default scenario.

- **Autoencoder**:
  - `gracia`: RMSE = 12.58, $R^2 = 0.707$
  - `eixample`: RMSE = 11.03, $R^2 = 0.733$
  - `hebron`: RMSE = 13.33, $R^2 = 0.790$

Visual comparisons for `gracia` sensor are shown in Figures 2, 3, 4, 5, and 6, providing qualitative support to the metrics above. MICE methods—particularly the linear regression variant—show superior accuracy in this setup, while deep models like the autoencoder and LSTM require more tuning to outperform classical methods.

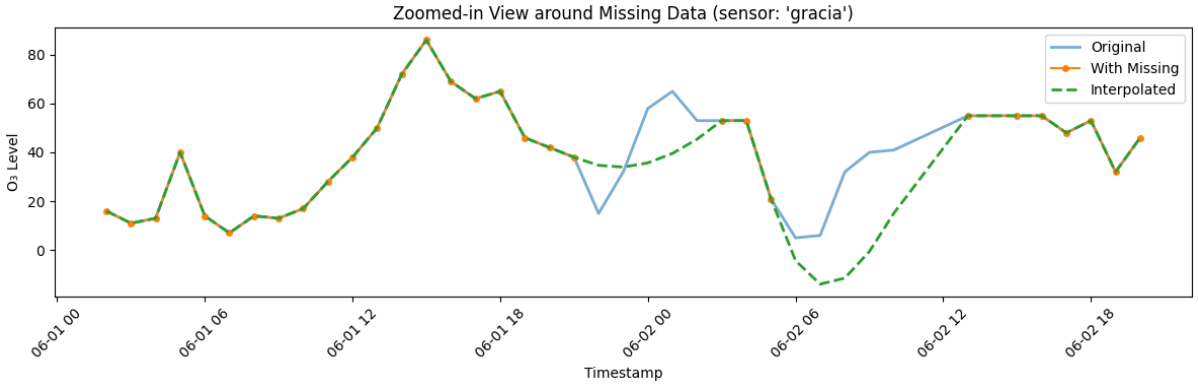

Figure 2: Univariate Polynomial Interpolation (degree 3) on `gracia`

We can observe how interpolation extends the existing trend by smoothly connecting known values. While this method can be effective for filling small gaps, it struggles to handle complex patterns in the data. As a result, interpolation may not perform well when the underlying signal contains sudden changes or nonlinear behavior.

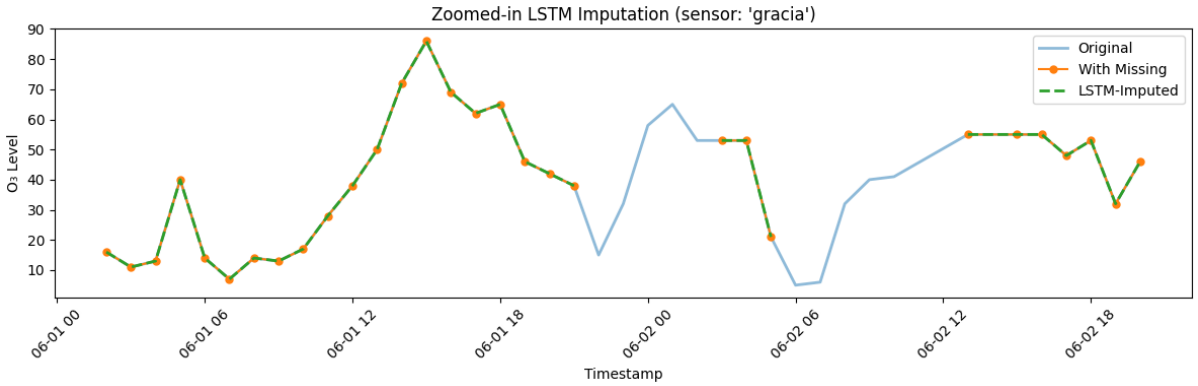

Figure 3: Zoomed-in LSTM Imputation on `gracia`

For LSTM, we observe that no meaningful predictions were made for this seed due to limitations caused by the fixed input window size. Since the missing values were not preceded by enough valid data, the model could not generate outputs, leading to a low imputation score.

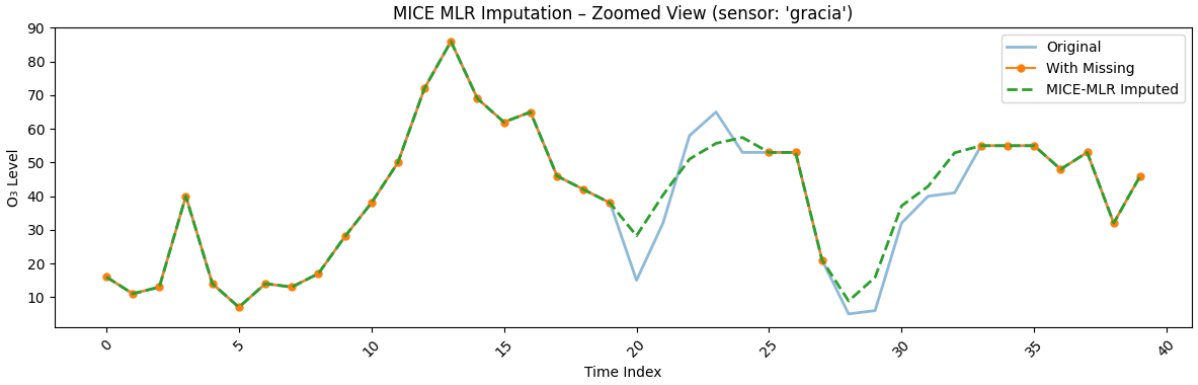

Figure 4: MICE MLR Imputation – Zoomed View (`gracia`)

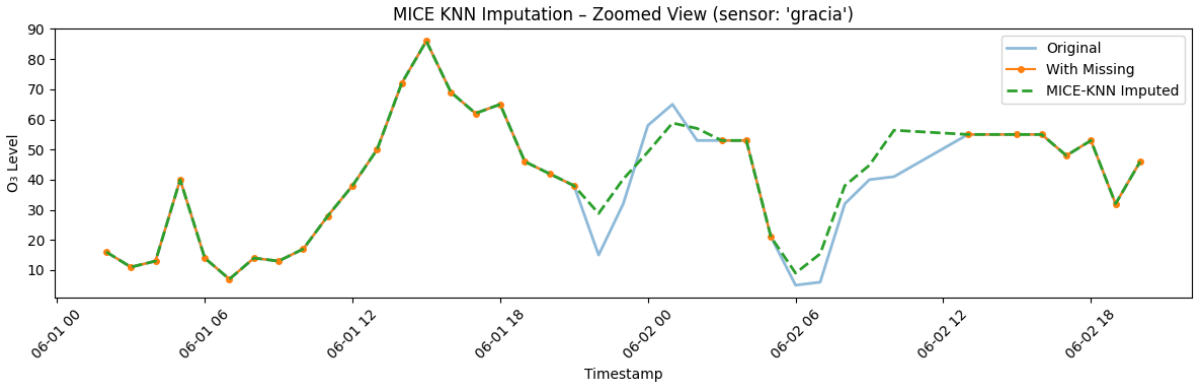

Figure 5: MICE KNN Imputation – Zoomed View (`gracia`)

MICE performs really well for this data.

For autoencoder, it makes complex predictions but they are not closely aligned for this specific case.

3 Effect of Missing Burst Length

Table 1 summarizes the imputation performance across three different burst lengths: 3, 5, and 10.

As shown above, shorter bursts (length 3) consistently yield better reconstruction across all methods, with lower RMSE and higher R^2 scores. The default case (burst 5) exhibits moderate degradation, while the longest burst length (10) significantly hampers performance—especially for simpler methods like polynomial interpolation and for LSTM, which struggles to reconstruct long sequences with missing windows.

Interestingly, while LSTM benefits from time-aware modeling, its performance sharply declines as the missing window increases. This is likely due to its fixed window size during training, which limits the model's capacity to infer long-range dependencies in the presence of consecutive gaps. In contrast, MICE and autoencoders show more gradual performance decline, making them more robust to bursty missingness.

Figure 6: Autoencoder Imputation – Zoomed View (`gracia`)

Table 1: Imputation performance on the Gràcia region for different burst lengths.

Method	**Burst**	**RMSE**	**R^2**
Polynomial (deg 3)	3	10.54	0.787
Polynomial (deg 3)	5	15.82	0.537
Polynomial (deg 3)	10	21.93	0.185
LSTM	3	16.68	0.475
LSTM	5	18.79	0.349
LSTM	10	25.57	−0.108
MICE Linear	3	7.90	0.880
MICE Linear	5	7.56	0.894
MICE Linear	10	6.18	0.935
MICE KNN	3	8.60	0.858
MICE KNN	5	8.49	0.867
MICE KNN	10	6.93	0.919
Autoencoder	3	11.95	0.727
Autoencoder	5	12.58	0.707
Autoencoder	10	7.92	0.894

4 Effect of Missing Percentage

In this section, we explore how varying the overall proportion of missing data affects the performance of different imputation methods. We fix the burst length to 5 and compare three different levels of missing data: 5%, 10% (default), and 20%. Table 2 summarizes the Root Mean Squared Error (RMSE) and coefficient of determination (R^2) for each method and sensor combination.

As expected, increasing the missing percentage leads to a general decline in performance across all methods. Polynomial interpolation suffers the most as missing data becomes more prevalent, likely due to its reliance on local continuity. MICE methods, particularly with linear regression, remain robust even at 20% missingness. Autoencoder performance remains stable but slightly deteriorates, particularly on the hebron sensor. LSTM shows high variability and lower performance, which may be attributed to limited training data or unsuitable hyperparameters for higher missing rates.

Table 2: Imputation performance for the Gràcia region across different missing data percentages.

Method	Missing %	RMSE	R^2
Polynomial (deg 3)	5%	13.83	0.662
	10%	15.82	0.537
	20%	16.27	0.506
LSTM	5%	19.02	0.372
	10%	18.79	0.349
	20%	17.23	0.444
MICE Linear	5%	7.67	0.896
	10%	7.56	0.894
	20%	10.01	0.813
MICE KNN	5%	8.60	0.869
	10%	8.49	0.867
	20%	11.67	0.746
Autoencoder	5%	11.61	0.762
	10%	12.58	0.707
	20%	11.74	0.743

5 Effect of LSTM Window Size

To investigate how the LSTM input sequence length affects imputation quality, we varied the window size parameter T while keeping all other settings fixed. The goal was to determine whether using shorter or longer temporal context helps the model recover missing ozone measurements in the Gràcia region.

Figure 7 presents zoomed-in imputation results for four different window sizes: $T = 6, 12, 24, 48$. The predicted values (dashed green) are compared against both the original (clean) signal and the masked signal with missing values.

Table 3: Imputation performance of LSTM across different window sizes T on the Gràcia sensor.

Window Size T	RMSE	R^2
6	17.88	0.403
12	17.27	0.443
24	21.33	0.148
48	22.11	0.049

As shown in Table 3, the LSTM achieves its best performance at $T = 12$, balancing between too little context (e.g., $T = 6$) and overly long sequences (e.g., $T = 48$) which may introduce noise or over-smoothing. When T becomes large, the model's ability to adapt to rapid changes in ozone levels deteriorates, possibly due to the challenge of learning long-range dependencies.

6 Effect of Sensor Count on Gràcia Imputation Performance

To assess the impact of sensor count on imputation quality for a single location, we focus on the Gràcia sensor and evaluate the performance of three methods—MICE Linear Regression, MICE KNN, and Autoencoder—under varying numbers of input sensors. Specifically, we compare three configurations: using the default 3 sensors, an extended set of 5 sensors, and a larger set of 8 sensors.

Table 4: Imputation performance for the **Gràcia** sensor across varying numbers of input sensors.

Method	Sensor Count	RMSE	R^2
MICE Linear	3 sensors	6.18	0.935
	5 sensors	10.06	0.811
	8 sensors	9.78	0.822
MICE KNN	3 sensors	6.93	0.919
	5 sensors	8.60	0.869
	8 sensors	10.86	0.780
Autoencoder	3 sensors	7.92	0.894
	5 sensors	15.62	0.544
	8 sensors	12.88	0.690

From Table 4, we observe that the best imputation results are obtained with the default 3-sensor setup for all three methods, especially for the Autoencoder, where increasing the number of sensors significantly degrades performance. This suggests that including more sensors may introduce noise or misaligned correlations, especially for neural architectures not optimized for multi-sensor fusion. Conversely, while MICE methods show degradation beyond 3 sensors, their performance remains more robust, especially for linear regression.

Conclusion

In this study, we evaluated multiple imputation methods—including polynomial interpolation, LSTM, MICE (Linear and KNN), and Autoencoders—on ozone time series data from multiple urban sensors. We systematically varied the percentage and structure of missing data, LSTM window size, and the number of sensors used as input to assess the robustness and generalization of each method.

Overall, we find that:

- **MICE with Linear Regression** consistently performs best across sensors and missing data scenarios, achieving the lowest RMSE and highest R^2 values, especially when the number of sensors is small.
- **LSTM** performance is sensitive to window size; smaller window sizes (e.g., $T = 12$) yield better results than longer histories. This highlights the importance of tuning temporal context for sequence models.

- **Autoencoders**, while competitive in some settings, degrade in performance when more sensors are added, suggesting that their capacity to generalize across multivariate spatial inputs may be limited without specialized architectures.
- **Polynomial interpolation**, although simple and fast, fails to capture nonlinear patterns or sharp transitions, particularly as the missing rate increases or burst length becomes longer.

These results suggest that classical statistical models remain highly effective for structured time-series imputation under realistic missingness conditions. Deep learning methods require careful tuning and may benefit from additional context-aware design to outperform traditional approaches.

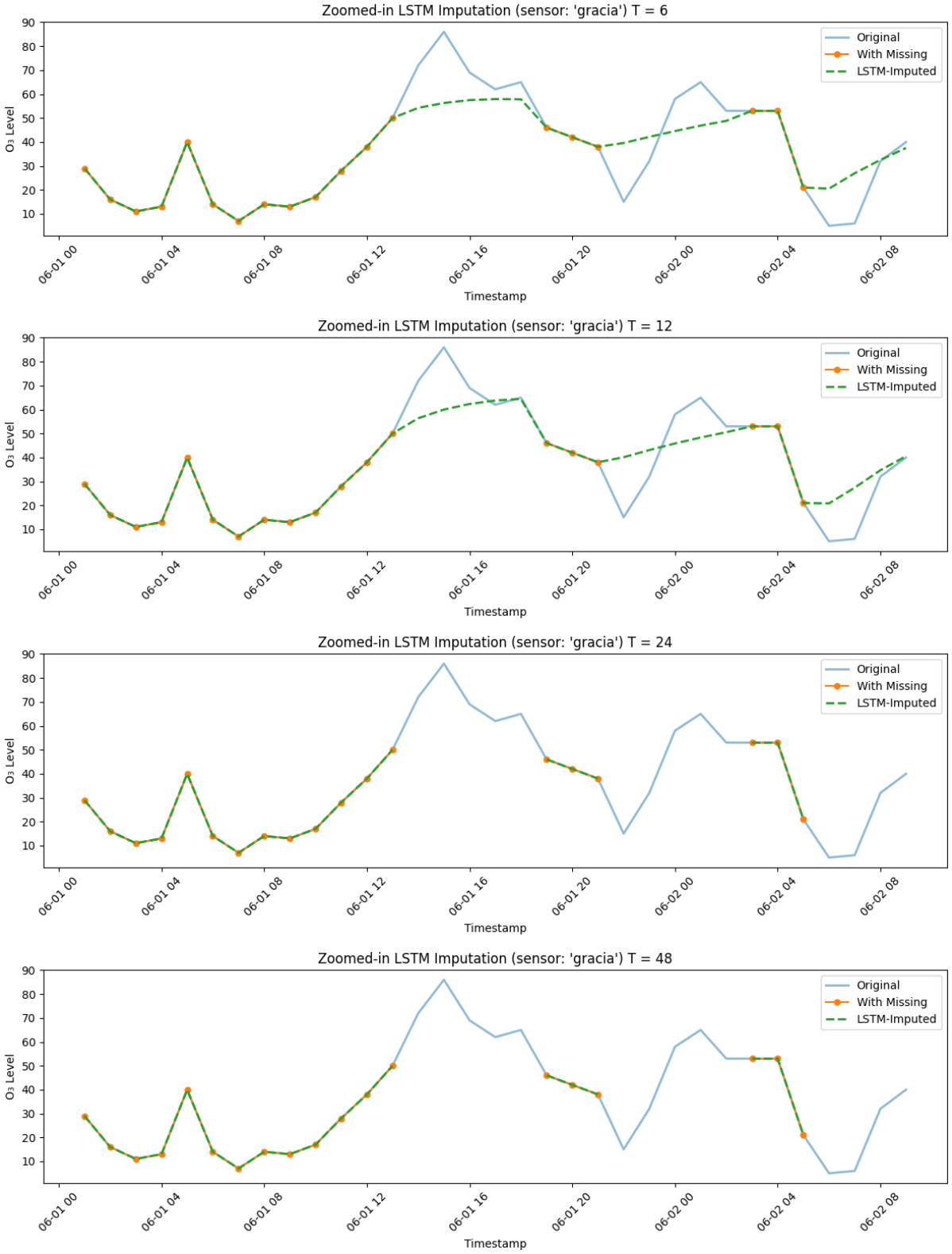

Figure 7: Zoomed-in LSTM imputations at different sequence lengths T for the Gràcia sensor.