Handling Missing Data in Time Series: Evaluating Kalman Smoothing and Other Imputation Methods

Francisca Adobea Ofori
University of Rhode Island

December 13, 2024

Abstract

Kalman smoothing was applied to impute missing data in a time-series dataset, with a particular focus on water quality monitoring. The method demonstrated strong performance, especially under the MAR (Missing at Random) mechanism, which yielded the most accurate results. Key metrics, including RMSE and MAE, were used to evaluate its effectiveness. These findings highlight Kalman smoothing's potential for addressing missing data challenges, ensuring reliable estimations critical for informed decision-making in environmental monitoring. Future work will explore its application in more complex scenarios and its integration with hybrid modeling approaches.

Keywords: Time Series Analysis, Missing Data, Kalman Smoothing, Environmental Monitoring, Exponential Smoothing

1 Introduction

Missing data presents a significant challenge in many fields, including environmental monitoring, finance, healthcare, and water quality assessments. It disrupts statistical analyses, reduces predictive accuracy, and introduces bias, particularly in time series data. Effective handling of missing data requires methods tailored to the underlying missingness mechanism—whether MCAR (Missing Completely at Random), MAR (Missing at Random), or MNAR (Missing Not at Random).

Common methods for dealing with missing data include Kalman smoothing, interpolation, exponential smoothing, and machine learning. Among these, the Kalman filter stands out for its versatility, particularly in time series analysis. Applied widely in systems like navigation and tracking, Kalman smoothing is known for its robustness in managing incomplete data, leveraging both past and future observations to make accurate estimates. In contrast, simpler methods like interpolation often fail to account for the data's temporal relationships.

This study focuses on Kalman smoothing for imputing missing values in water quality datasets, comparing its performance to exponential smoothing using metrics like RMSE, MAE, and bias. Addressing missing data in water quality is crucial for maintaining reliable environmental monitoring and informed decision-making.

2 Motivating Data

The dataset for this study, obtained from *data.world*, contains water quality measurements from drinking water treatment plants. The analysis focuses on four variables:

- **pH:** A measure of acidity or alkalinity (stationary).
- **Turbidity:** An indicator of water cloudiness (stationary).
- **Temperature:** Records water temperature (non-stationary).
- **Chemical cost per 1000 gallons:** Represents operational costs (non-stationary).

The combination of stationary and nonstationary variables makes this dataset suitable for evaluating the robustness of Kalman smoothing and other imputation methods.

3 Model and Method

In this study, we use two primary techniques for smoothing and prediction: Kalman smoothing and exponential smoothing. First, Kalman smoothing will be applied to impute missing data under MCAR, MAR, and MNAR conditions to determine the most effective method. Subsequently, we will compare the performance of Kalman smoothing with exponential smoothing.

3.1 Kalman Smoothing

Kalman smoothing operates by using a state-space model to estimate the underlying state of a time series. It involves two key steps:

- **Prediction Step:** Estimates the current state based on the previous state.
- **Update Step:** Adjusts the prediction using observed data.

When dealing with missing data, the Kalman smoother automatically accounts for gaps by relying on its recursive nature. Missing observations (z_t) are handled by skipping the update step for those time points. Instead, the algorithm propagates the predicted state estimate forward, leveraging information from prior and subsequent observations to refine estimates. This ensures that the imputation reflects both past trends and future patterns, making Kalman smoothing particularly effective for datasets with intermittent missing values.

Mathematically, the Kalman filter can be expressed as:

$$\hat{x}_{t|t-1} = A\hat{x}_{t-1|t-1} + Bu_t$$

where $\hat{x}_{t|t-1}$ is the predicted state at time t, A is the state transition matrix, $\hat{x}_{t-1|t-1}$ is the previous state estimate, B is the control input matrix, and u_t is the control vector at time t.

The update step incorporates the observed data z_t (when available) and updates the state estimate:

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(z_t - H\hat{x}_{t|t-1})$$

where $\hat{x}_{t|t}$ is the updated state estimate, K_t is the Kalman gain, H is the observation matrix, z_t is the actual observation at time t, and $(z_t - H\hat{x}_{t|t-1})$ is the innovation or measurement residual.

For missing z_t, the term $z_t - H\hat{x}_{t|t-1}$ is omitted, and the algorithm relies on prior predictions until new observations become available.This ability to combine past and future data makes Kalman smoothing a robust method for imputing missing values in time-series datasets.

3.2 Exponential Smoothing

Exponential smoothing is a forecasting method that applies exponentially decreasing weights to past observations. It is defined by the formula:

$$S_t = \alpha \cdot Y_t + (1 - \alpha) \cdot S_{t-1}$$

where S_t is the smoothed value at time t, Y_t is the observed value at time t, and α is the smoothing parameter $(0 < \alpha < 1)$.

For missing data, exponential smoothing handles gaps by treating missing observations as unobserved Y_t values. The algorithm skips updating the smoothed value S_t for those time points, effectively propagating the last available S_{t-1} forward. While this approach is computationally efficient, it lacks the ability to incorporate future observations into the imputation process, making it less accurate than Kalman smoothing for datasets with extensive missingness or complex patterns.

3.3 Simulating Missing Data Mechanisms

Three missing data mechanisms were simulated on the water quality dataset:

- **MCAR:** Missing values were randomly assigned using a 10% missing rate.
- **MAR:** Missing values depended on relationships between variables (e.g., pH missing when turbidity is high).
- **MNAR:** Missingness was influenced by the variable's own values (e.g., higher pH values were more likely to be missing).

Simulations were performed using R, and missing data patterns were visualized with the `VIM` package. The simulation involved applying MCAR, MAR, and MNAR mechanisms to the dataset, performing 50 simulations for each. Kalman and exponential smoothing methods were used, and their performance was evaluated using MSE and MAE, averaged across simulations.

4 Real Data Analysis and Results

4.1 Missing Data Pattern

The combined missingness patterns for MCAR, MAR, and MNAR are shown in Figure 1.This shows a general missing pattern These patterns highlight the different types of missing data used in the simulations for imputation analysis.

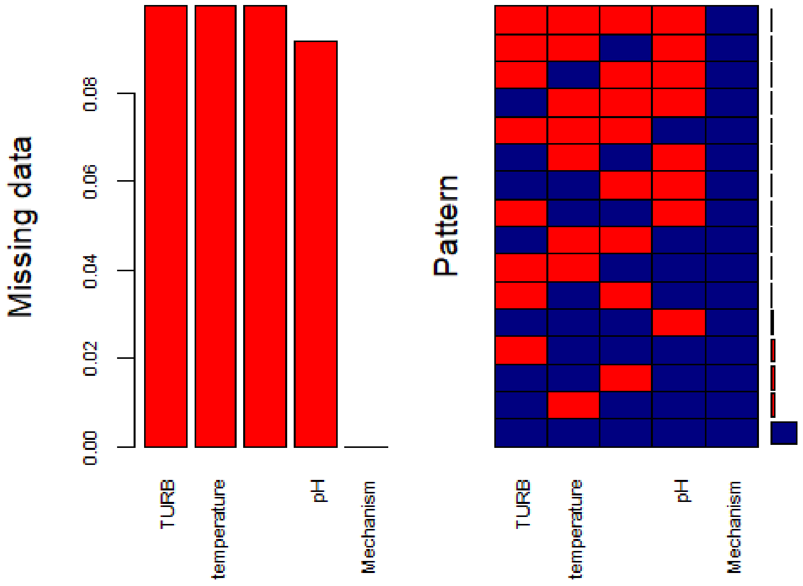

Figure 1: Combined Missingness Patterns for MCAR, MAR, and MNAR

4.2 Kalman Smoothing Results for RMSE and MAE

The performance of Kalman smoothing was evaluated using RMSE and MAE across different missing data mechanisms (MAR, MCAR, and MNAR). Both metrics showed that Kalman smoothing performed best under the MAR (Missing at Random) mechanism, achieving the lowest mean RMSE and MAE. This highlights its ability to effectively leverage relationships between observed data for accurate imputation.

In contrast, MCAR (Missing Completely at Random) exhibited higher RMSE and MAE due to the randomness of missingness, which reduces predictive context. MNAR (Missing Not at Random), where missingness depends on unobserved variables, presented additional challenges, leading to intermediate error values between MAR and MCAR.

Table 1 and Figures 2 and 3 summarize the results, demonstrating that Kalman smoothing's performance varies depending on the missingness mechanism, with MAR providing the most favorable conditions for imputation.

Mechanism	Mean RMSE	Mean MAE
MAR	0.383	0.0569
MCAR	0.410	0.0584
MNAR	0.425	0.0581

Table 1: Mean RMSE and MAE across different missing data mechanisms.

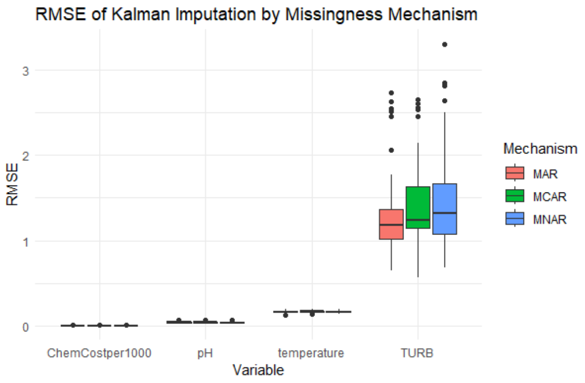

Figure 2: RMSE for Kalman Smoothing Imputation across Missingness Mechanisms

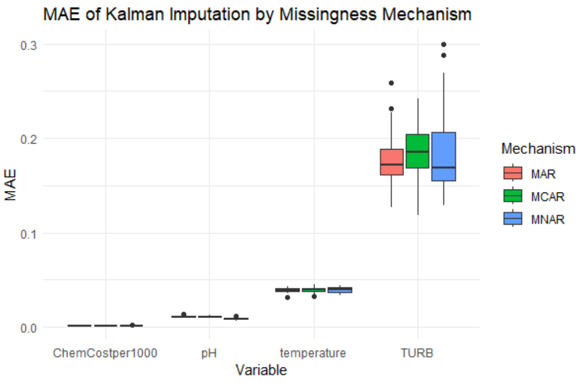

Figure 3: MAE for Kalman Smoothing Imputation across Missingness Mechanisms

4.3 Results of Bias, Variance, and Confidence Interval

Bias: The smallest bias was observed under MAR (0.000504), indicating minimal systematic error and highlighting the method's effectiveness when missingness is related to observed variables. In contrast, a slight negative bias was detected under MCAR and MNAR, reflecting the challenges posed by purely random or data-dependent missingness mechanisms.

Variance: The variance of the imputed values remained consistent across all missingness mechanisms, ranging from 29.4 to 29.5. This suggests a stable spread of predictions, regardless of the missing data pattern, further reinforcing the robustness of the Kalman smoothing approach.

Confidence Interval Width: The mean confidence interval width was identical across all mechanisms at 0.345, indicating consistent uncertainty bounds in the imputed values. This uniformity suggests that the Kalman smoothing approach maintains stable estimation precision regardless of the missing data mechanism (MAR, MCAR, or MNAR).

Mechanism	**Mean Bias**	**Mean Variance**
MAR	0.000504	29.5
MCAR	-0.00162	29.4
MNAR	-0.000793	29.4

Table 2: Comparison of Bias, Variance, and Confidence Interval Width across mechanisms.

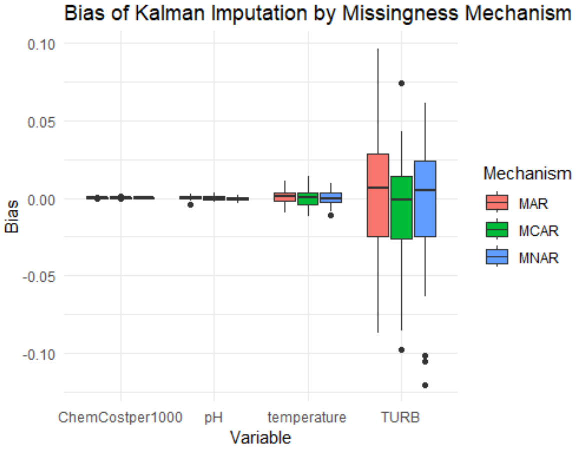

Figure 4: Bias Plot

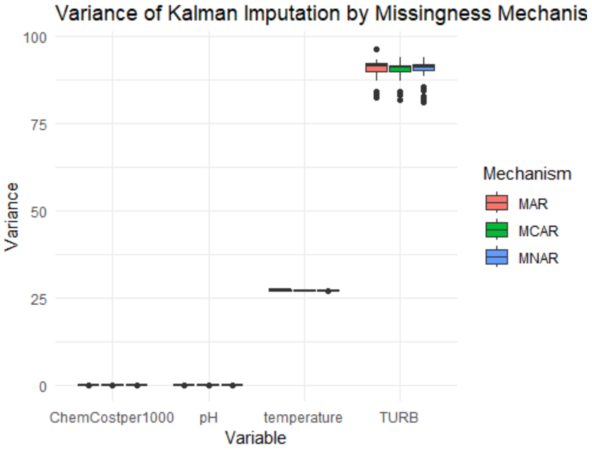

Figure 5: Variance Plot

4.4 Performance Comparison between Kalman and Exponential Smoothing

Kalman Smoothing outperforms Exponential Smoothing in RMSE, MAE, and Bias, offering better accuracy and consistency, though with slightly higher variance. Its higher variability makes it ideal for dynamic data, such as environmental monitoring, while Exponential Smoothing is more suitable for stationary data with predictable trends. The comparison is summarized in the table and figures below:

Metric	**Exponential Smoothing**	**Kalman Smoothing**
RMSE	1.4754	0.3828
MAE	0.6477	0.0569
Bias	0.0187	0.0005
Variance	27.23	29.54

Table 3: Comparison of performance metrics for Exponential Smoothing and Kalman Smoothing.

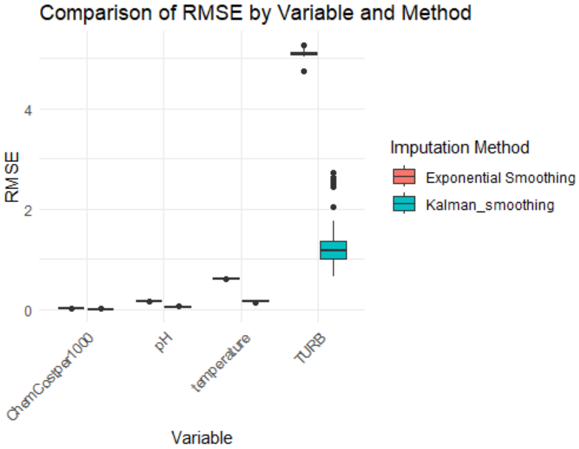

Figure 6: RMSE Plot

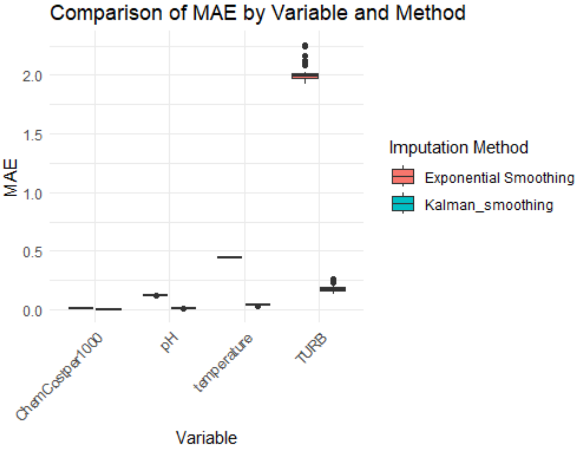

Figure 7: MAE Plot

5 Conclusions

Kalman smoothing provides a robust framework for imputing missing values in time-series datasets, demonstrating superior performance under the MAR mechanism, with a mean RMSE of 0.383 and MAE of 0.0569. Its ability to leverage both past and future data points enhances imputation accuracy, making it particularly suitable for applications like water quality monitoring.

Beyond imputation, Kalman smoothing's dynamic adjustment capabilities make it invaluable for real-time anomaly detection—such as identifying sudden pollution events—and for analyzing seasonal or long-term trends. In environmental management, it supports critical decision-making in resource allocation, pollution control, and policy evaluation.

However, the method's computational complexity and sensitivity to model assumptions warrant careful consideration, especially for large-scale or real-time datasets. Future research should explore hybrid approaches that integrate Kalman smoothing with machine learning techniques to improve accuracy and scalability, as well as extend its application to other missingness mechanisms and domains such as healthcare and finance.

References

[1] Aravkin, A., Burke, J. V., Ljung, L., Lozano, A., & Pillonetto, G. (2017). Generalized Kalman smoothing: Modeling and algorithms. *Automatica*, 86, 63-86. doi:10.1016/j.automatica.2017.08.011

[2] Junger, W. L., & Ponce de Leon, A. (2015). *Imputation of Missing Data in Time Series for Air Pollutants.* Atmospheric Environment, 102, 96–104. Elsevier. doi:10.1016/j.atmosenv.2014.11.049

[3] Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.