Mathematical Foundations of Missing Value Imputation Methods

For Financial Time Series Analysis

FinTech Studio

MSc Banking and Finance

September 15, 2025

Outline

Missing Data in Financial Time Series

Types of Missingness

- **MCAR**: Missing Completely at Random
- **MAR**: Missing at Random
- **MNAR**: Missing Not at Random

Financial Data Characteristics

- High frequency observations
- Temporal dependencies
- Cross-sectional correlations
- Heteroscedasticity and fat tails
- Non-stationarity

Simple Mean Imputation

Mathematical Formula

Replace missing value $\hat{x}_{i,j}$ with column mean:

$$\hat{x}_{i,j} = \frac{1}{n - m_j} \sum_{k: x_{k,j} \text{ observed}} x_{k,j} \tag{1}$$

where:

- n = total observations
- m_j = missing values in column j

Advantages

- Simple: $O(n)$ complexity
- Fast computation
- Unbiased under MCAR

Disadvantages

- Reduces variance
- Ignores correlations
- Distorts distributions

Rolling Mean Imputation

Time-Localized Approach

Uses sliding window for temporal patterns:

$$\hat{x}_{i,j} = \frac{1}{|W_i|} \sum_{k \in W_i : x_{k,j} \text{ observed}} x_{k,j} \tag{2}$$

$$W_i = \{k : |k - i| \leq w\} \tag{3}$$

where W_i is the window of size w centered at time i.

Median Imputation

More robust to outliers:

$$\hat{x}_{i,j} = \text{median}\{x_{k,j} : x_{k,j} \text{ is observed}\} \tag{4}$$

Forward/Backward Fill

Forward Fill (LOCF)

Propagates most recent value forward in time:

$$\hat{x}_{i,j} = x_{k,j} \text{ where } k = \max\{t < i : x_{t,j} \text{ is observed}\} \tag{5}$$

Backward Fill (NOCB)

Next Observation Carried Backward:

$$\hat{x}_{i,j} = x_{k,j} \text{ where } k = \min\{t > i : x_{t,j} \text{ is observed}\} \tag{6}$$

Combined Strategy (Enhanced)

$$\hat{x}_{i,j} = \begin{cases} x_{k_{\text{prev}},j} & \text{if in the beginning of the series} \\ x_{k_{\text{next}},j} & \text{if after the middle of the series} \\ \frac{d_{\text{next}} \cdot x_{k_{\text{prev}},j} + d_{\text{prev}} \cdot x_{k_{\text{next}},j}}{d_{\text{prev}} + d_{\text{next}}} & \text{otherwise, interpolate} \end{cases} \tag{7}$$

Forward/Backward Fill Properties

Advantages

- Preserves temporal patterns
- Very fast: $O(n)$
- Realistic for financial data
- No distributional assumptions

Disadvantages

- Fails with long gaps
- Creates artificial plateaus
- Forward fill unavailable initially
- Doesn't model volatility

Financial Application

Ideal for short gaps in high-frequency financial data where temporal continuity is strong.

K-NN Distance Calculation

Euclidean Distance

Distance between observations, excluding missing components:

$$d(x_i, x_l) = \sqrt{\sum_{j \in O_{i,l}} (x_{i,j} - x_{l,j})^2} \tag{8}$$

where $O_{i,l} = \{j : x_{i,j} \text{ and } x_{l,j} \text{ both observed}\}$

Simple K-NN Imputation

Average of K nearest neighbors:

$$\hat{x}_{i,j} = \frac{1}{K} \sum_{l \in N_K(i)} x_{l,j} \tag{9}$$

where $N_K(i)$ represents the K nearest neighbors of observation i.

Weighted and Temporal K-NN

Weighted K-NN

Weights neighbors by inverse distance:

$$\hat{x}_{i,j} = \frac{\sum_{l \in N_K(i)} w_{i,l} \cdot x_{l,j}}{\sum_{l \in N_K(i)} w_{i,l}} \tag{10}$$

$$w_{i,l} = \frac{1}{d(x_i, x_l) + \epsilon} \tag{11}$$

Temporal K-NN for Financial Data

Incorporates temporal distance:

$$d_{\text{combined}}(x_i, x_l) = d_{\text{feature}}(x_i, x_l) + \lambda \cdot d_{\text{temporal}}(i, l) \tag{12}$$

$$d_{\text{temporal}}(i, l) = \frac{|i - l|}{n} \tag{13}$$

where λ controls temporal vs feature similarity trade-off.

K-NN Properties and Complexity

Advantages

- Captures local patterns
- Non-parametric
- Adapts to data structure
- Temporal version preserves time series properties

Disadvantages

- Expensive: $O(n^2)$
- Curse of dimensionality
- Choice of K critical
- May need feature scaling

Complexity Analysis

- **Time:** $O(n^2 \cdot p)$ for distance calculations
- **Space:** $O(n \cdot p)$ for storing distances

MICE: Core Algorithm

Multiple Imputation by Chained Equations

Models each variable as function of others through iterative regression:

$$X_j^{(t+1)} | X_{-j}^{(t)} \sim f_j(X_{-j}^{(t)}, \theta_j^{(t)}) \tag{14}$$

where:

- $X_j^{(t)}$ = values of variable j at iteration t
- $X_{-j}^{(t)}$ = all variables except j at iteration t
- f_j = conditional distribution model for variable j
- $\theta_j^{(t)}$ = model parameters at iteration t

Convergence Criterion

Algorithm converges when parameter estimates stabilize:

$$|\theta_j^{(t+1)} - \theta_j^{(t)}| < \epsilon \text{ for all } j \tag{15}$$

MICE: Bayesian Linear Model

Bayesian Ridge Regression

For continuous variables:

$$X_j | X_{-j}, \beta_j, \sigma_j^2 \sim N(X_{-j}\beta_j, \sigma_j^2 I) \tag{16}$$

$$\beta_j | \sigma_j^2 \sim N(\mu_0, \sigma_j^2 \Sigma_0^{-1}) \tag{17}$$

$$\sigma_j^2 \sim \text{InvGamma}(\alpha_0, \beta_0) \tag{18}$$

Random Forest MICE

For non-linear relationships:

$$\hat{X}_j = \frac{1}{B} \sum_{b=1}^{B} T_b(X_{-j}) \tag{19}$$

where T_b is the b-th decision tree trained on bootstrap sample.

MICE: Uncertainty Quantification

Multiple Imputation Variance

MICE provides uncertainty estimates:

$$\mathrm{Var}(\hat{X}_j) = \underbrace{\frac{1}{M}\sum_{m=1}^{M}\mathrm{Var}(\hat{X}_j^{(m)})}_{\text{Within}} + \underbrace{\frac{M+1}{M}\cdot\frac{1}{M-1}\sum_{m=1}^{M}(\hat{X}_j^{(m)} - \bar{X}_j)^2}_{\text{Between}} \tag{20}$$

Advantages

- Models multivariate relationships
- Preserves distributions
- Provides uncertainty
- Handles mixed data types

Disadvantages

- Computationally intensive
- Convergence not guaranteed
- May amplify biases
- Requires model selection

Performance Evaluation

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{|M|} \sum_{(i,j) \in M} |x_{i,j} - \hat{x}_{i,j}| \tag{21}$$

Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{|M|} \sum_{(i,j) \in M} (x_{i,j} - \hat{x}_{i,j})^2} \tag{22}$$

Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{100}{|M|} \sum_{(i,j) \in M} \left| \frac{x_{i,j} - \hat{x}_{i,j}}{x_{i,j}} \right| \tag{23}$$

where M represents the set of originally missing positions (i, j).

Correlation and Variance Metrics

Correlation Coefficient

$$r = \frac{\sum_{(i,j)\in M}(x_{i,j} - \bar{x})(\hat{x}_{i,j} - \bar{\hat{x}})}{\sqrt{\sum_{(i,j)\in M}(x_{i,j} - \bar{x})^2 \sum_{(i,j)\in M}(\hat{x}_{i,j} - \bar{\hat{x}})^2}} \tag{24}$$

Variance Preservation

Critical for financial modeling:

$$\text{Variance Ratio} = \frac{\text{Var}(\hat{X})}{\text{Var}(X)} \tag{25}$$

Ideal ratio $= 1.0$ (perfect variance preservation)

Financial Interpretation

- MAE/RMSE in dollar terms for direct interpretation
- MAPE for scale-independent comparison
- Correlation measures pattern preservation

Computational Complexity Comparison

Method	Time	Space	Assumptions	Best Use
Mean/Median	$O(n \cdot p)$	$O(1)$	MCAR	Fast baseline
Forward/Back Fill	$O(n \cdot p)$	$O(1)$	Temporal continuity	Time series gaps
K-NN	$O(n^2 \cdot p)$	$O(n \cdot p)$	Local similarity	Non-linear patterns
MICE	$O(T \cdot p^2 \cdot n)$	$O(n \cdot p)$	MAR	Multivariate relationships

Notation

- n = number of observations
- p = number of features
- T = MICE iterations
- MCAR = Missing Completely at Random
- MAR = Missing at Random

Recommendations for Financial Time Series

Gap Length Guidelines

1. **Short gaps (<5 periods):** Forward fill or linear interpolation
2. **Medium gaps ($5-20$ periods):** Temporal K-NN or scaled MICE
3. **Long gaps (>20 periods):** Model-based approaches
4. **High dimensionality:** MICE with feature selection
5. **Real-time:** Forward fill only (no look-ahead bias)

Financial Data Properties

- **Heteroscedasticity:** Variance changes over time
- **Fat tails:** Extreme values more common
- **Temporal dependence:** Values correlated across time
- **Cross-sectional correlation:** Assets move together
- **Non-stationarity:** Statistical properties evolve

Implementation Strategy

Agile Development Approach

1. Start with simple methods (Mean, Forward fill)
2. Implement evaluation framework
3. Add sophisticated methods (K-NN, MICE)
4. Compare performance on your specific data
5. Optimize best-performing method

Quality Assurance

- Cross-validation on held-out missing data
- Sensitivity analysis on hyperparameters
- Stress testing with different missing patterns
- Documentation of assumptions and limitations

Key Takeaways

Method Selection Criteria

- **Data size:** Large datasets favor simple methods
- **Missing pattern:** MCAR vs MAR vs MNAR
- **Gap length:** Short gaps favor fill methods
- **Relationships:** Complex correlations favor MICE
- **Real-time:** Only forward-looking methods acceptable

Implementation Success Factors

- Understand your data's missing mechanism
- Start simple, add complexity gradually
- Always evaluate on held-out data
- Consider computational constraints
- Document assumptions and validate results

Final Recommendation

No single method dominates - choose based on your specific data characteristics and use case requirements.

Thank you for your attention!

Questions and Discussion