IRREGULAR MULTIVARIATE TIME SERIES MISSING VALUE IMPUTATION

Asal Roudbari,Alireza Moradi
Georgia Institute of Technology
alirezamoradi@gatech.edu,asal@gatech.edu

1 Problem Statement and Formulation

Missing data are a pervasive challenge in intensive care unit (ICU) settings. ICU patient data such as vital signs and laboratory results are often recorded at irregular intervals due to clinical workflows, leading to gaps in the time series records. For example, a stable patient may have vital signs charted only every few hours, whereas during critical events measurements become more frequent, resulting in unevenly spaced observations. Practical factors like patient transport, sensor malfunction, or manual recording delays further contribute to non-random missingness in the data.

The presence of these missing values poses a significant obstacle to developing reliable predictive and diagnostic models. Many machine learning algorithms assume regularly sampled, complete datasets; hence, missing data can introduce bias, reduce statistical power, and distort learned patterns. Imputation—the process of estimating and filling in missing observations—is therefore a critical preprocessing step for any downstream analysis of ICU time series data.

The *PhysioNet 2012 Challenge* dataset provides a realistic and widely studied context for this problem. It contains multivariate time series from the first 48 hours of 12,000 ICU stays, with up to 37 physiological variables per patient. Measurements in this dataset occur at irregular intervals ranging from hourly to once every few days, and many entries are explicitly marked as missing (recorded as `-1` for "unknown"). These characteristics make it an ideal testbed for studying imputation methods in irregularly sampled clinical time series.

1.1 Mathematical Formulation

Let the dataset consist of N patient records. For each patient $i \in 1, \ldots, N$, we observe a multivariate time series:

$$X_i = (t_{i1}, \mathbf{x}_{i1}), (t_{i2}, \mathbf{x}_{i2}), \ldots, (t_{iT_i}, \mathbf{x}_{iT_i}), \tag{1}$$

where t_{ij} denotes the timestamp (in hours since ICU admission), and $\mathbf{x}_{ij} \in \mathbb{R}^D$ represents the D physiological variables measured at that time. Due to missing observations, some entries in $\mathbf{x}_{ij}$ are unobserved.

We can separate each patient's data into observed and missing subsets:

$$X_i^{\text{obs}} = x_{ij,d} \mid (i,j,d) \in \Omega, \qquad X_i^{\text{miss}} = x_{ij,d} \mid (i,j,d) \notin \Omega, \tag{2}$$

where Ω denotes the set of observed indices.

1.2 Objective

The goal of this project is to learn an imputation function

$$\hat{f} : (t, X_i^{\text{obs}}) \to \hat{X}_i^{\text{miss}}, \tag{3}$$

that estimates the missing entries X_i^{miss} and reconstructs a complete, physiologically consistent time series

$$\hat{X}_i = X_i^{\text{obs}} \cup \hat{X}_i^{\text{miss}}. \tag{4}$$

2 Dataset Description

The dataset used in this study is the *PhysioNet/Computing in Cardiology Challenge 2012* database [1], which contains records from 12,000 intensive care unit (ICU) stays. All patients were adults admitted to cardiac, medical, surgical, or trauma ICUs for diverse conditions. ICU stays shorter than 48 hours were excluded. The timestamps indicate elapsed time since ICU admission in hours and minutes. Missing or unknown values are encoded as `-1`.

2.1 Variables

Each record contains up to 42 variables, including six general descriptors collected at admission and 37 physiological time-series variables:

- **General descriptors:** RecordID, Age, Gender, Height, Weight, and ICUType (1: CCU, 2: Cardiac Surgery Recovery, 3: Medical ICU, 4: Surgical ICU).
- **Time-series variables:** vital signs and lab measurements such as heart rate (HR), invasive and non-invasive blood pressures (SysABP, DiasABP, NISysABP, etc.), respiratory rate, temperature, oxygen saturation (SaO_2), urine output, and biochemical indicators (e.g., Creatinine, BUN, Glucose, Na, K, Platelets, etc.).

2.2 Sampling and Characteristics

Measurements are irregularly spaced, reflecting real ICU workflows—ranging from frequent (hourly) to sparse (once or twice daily). Different measurement methods (e.g., invasive vs. non-invasive blood pressure) may yield multiple readings at nearly identical timestamps. Occasional outliers and gaps are present, making this dataset a realistic and challenging benchmark for imputation and time-series modeling in clinical data.

3 Methodology

We compare three statistical approaches for imputing missing ICU time-series data from the *PhysioNet 2012* dataset.

Smoothing Splines: Each variable is modeled as a smooth function of time by fitting cubic splines through observed points. Missing values are estimated by evaluating the fitted curve at unobserved timestamps [2].

Gaussian Processes: A Gaussian Process regression uses time as input and learns temporal correlations via a kernel (e.g., RBF). The posterior mean provides imputations with uncertainty estimates, making it suitable for irregular sampling [3].

Bayesian Imputation: Using Multiple Imputation by Chained Equations (MICE), missing entries are sampled iteratively from conditional posterior distributions. This approach captures uncertainty and inter-variable dependencies [4].

3.1 Evaluation Metrics

To assess model performance, a subset of observed data will be artificially masked to simulate missingness. The MAE, MSE, RMSE, and R^2 metrics will be computed between the true and imputed values. The comparison across methods will highlight trade-offs between accuracy, smoothness, and computational cost.

4 Expected Results

This project will compare three imputation methods smoothing splines, Gaussian Processes (GPs), and Bayesian multiple imputation on ICU time series data. We expect spline based methods to perform competitively on frequently sampled vital signs due to their simplicity and robustness, while GPs may excel for nonlinear temporal patterns and provide valuable uncertainty estimates. The Bayesian approach is anticipated to offer probabilistic insights through multiple imputations but may be computationally intensive. Beyond model performance, the project will enhance our understanding of nonparametric regression, kernel methods, and Bayesian inference, producing both a practical imputation framework and a comparative analysis of each method's strengths and limitations in clinical time-series data.

References

[1] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[2] Yuedong Wang. *Smoothing splines: methods and applications*. CRC press, 2011.

[3] Eric Schulz, Maarten Speekenbrink, and Andreas Krause. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of mathematical psychology*, 85:1–16, 2018.

[4] Ian R White, Patrick Royston, and Angela M Wood. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399, 2011.