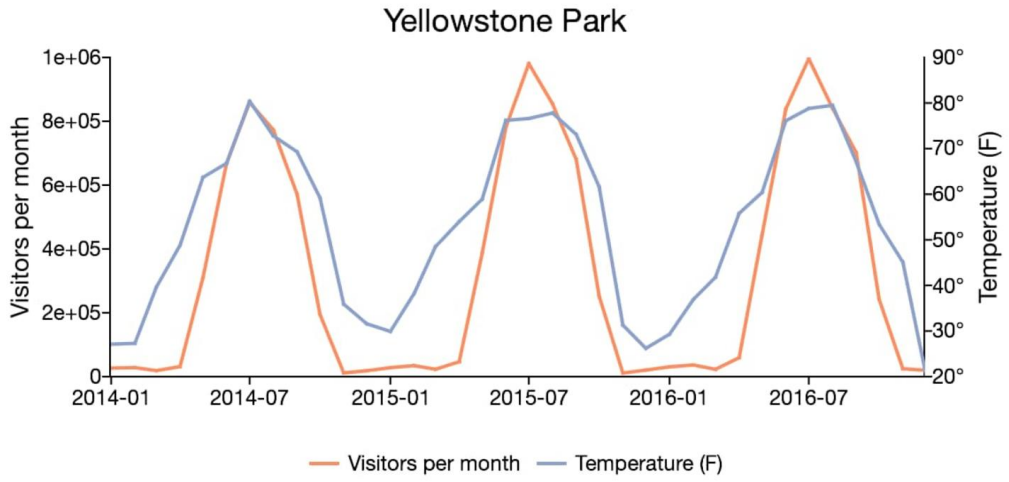
APPLICATION OF MACHINE LEARNING MODELS TO SOLVE THE PROBLEM OF UNIVARIATE TIME SERIES DATA IMPUTATION

Author: Nguyen Thanh Luan

1. Overview

In daily life, the benefits of time series data are undeniable. This type of data is recorded at fixed time intervals, such as years, quarters, months, hours, or minutes. For instance, stock price data in finance or revenue data in banking are common examples. However, real-time data recording is prone to risks of data loss due to factors such as storage limitations, hardware failures, or software issues. Such data loss significantly impacts the performance of statistical, prediction, and forecasting models.

Therefore, addressing and imputing missing data is a crucial task in time series analysis. Successfully handling this issue not only improves analytical results but also ensures high reliability in implementing statistical models. This article proposes and presents a method for addressing this problem in univariate time series data. It is essential to ensure that the correlation and relationships between observations are preserved during experimentation.



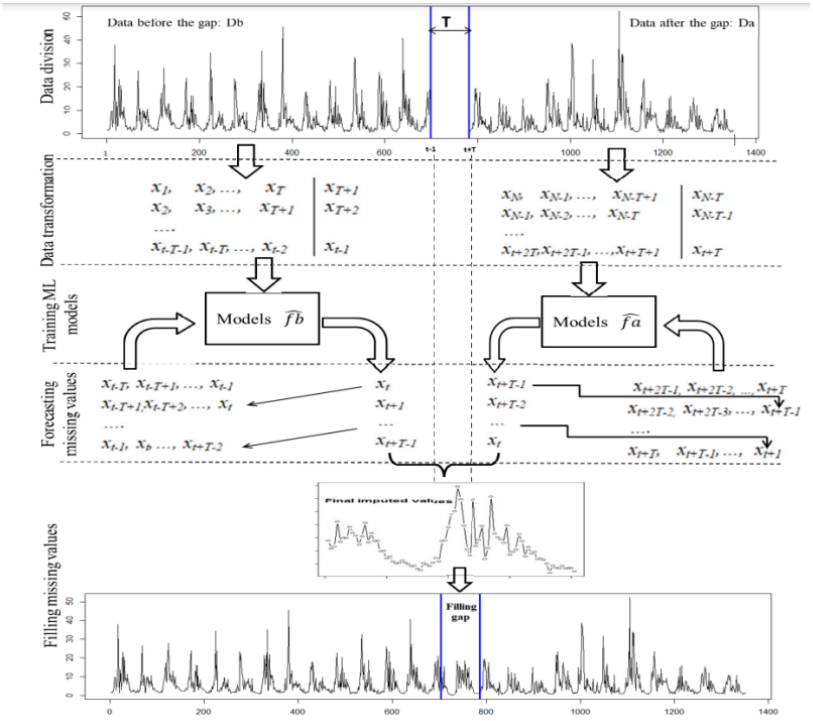
2. Methodology

Case Definition: To experimentally address the problem of missing data, I divided it into three main cases and developed an effective handling workflow. First, if gaps occur at the beginning or end of the time series, they are defined as such when the gap duration (T) is within 3*T at the start or end of the series. In this case, the model operates on the adjacent data. If the gaps do not fall into these two cases, they are classified as "middle gaps," where the series is divided into two parts: before and after the gap. This flexible method ensures accuracy and reliability for the computational process.

Data Preparation: Transform the univariate data into (T+1)-dims data. This transformation enables the efficient utilization of predictions from earlier time points as inputs for subsequent predictions.

Algorithms: ARIMA and regression-based models such as SVM and decision tree algorithms were selected for experimentation in this study.

Evaluation Metrics: The metrics used include Similarity, MAE, RMSE, FB (Forecast Bias), and FSD (Fraction of Standard Deviation). Among these, FB and FSD are overall evaluation metrics, while the remaining three assess individual data samples.



3. Experiment

The experimental results conducted over 20 folds for each case are as follows: code

	Sim	MAE	RMSE	FB	FSD
ML	0.5988	2.8077	3.1903	0.0014	0.4233
Arima	0.6760	3.0518	3.4493	0.001	0.2318
Arima + ML	0.6825	3.2987	3.6861	0.0004	0.1034

a. Beginning cases

	Sim	MAE	RMSE	FB	FSD
ML	0.8153	1.1432	1.5654	0.0026	0.4774

Arima	0.7589	1.1908	1.3922	0.0002	0.7392
Arima + ML	0.7923	1.1440	1.3806	0.0001	0.6649

b. Ending cases

	Sim	**MAE**	**RMSE**	**FB**	**FSD**
ML	0.3767	2.2302	2.5084	0.0045	1.5522
Arima	0.7356	1.3732	1.7089	0.0017	0.7706
Arima + ML	0.7658	1.3274	1.7027	0.0016	0.6590

c. Middle case

In this paper, I do not prioritize any specific evaluation metric, as it depends on the range and value of the data under study. However, an overall evaluation across the three cases shows that ARIMA yields better results compared to independent ML models. Nevertheless, combining both methods proves to be more effective, with ARIMA as the first layer and SVM as the second layer (leveraging residuals from the first-layer model).

4. Repo: https://github.com/thanhluan7702/Imputation-1D-timeseries-data