JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 10, No. 1, January 2026, pp. 1368-1384

Dimensionality Reduction Evaluation of Multivariate Time Series of Consumer Price Index in Indonesia

Nina Valentika[1], I Made Sumertajaya[1*], Aji Hamim Wigena[1], Farit Mochamad Afendi[1]
[1]Department of Statistics and Data Science Study Program, IPB University, Indonesia
imsjaya@apps.ipb.ac.id

Article History:
Received : 20-08-2025
Revised : 13-11-2025
Accepted : 14-11-2025
Online : 01-01-2026

Keywords:
Principal Component Analysis;
Robust Principal Component Analysis;
Multivariate Time Series;
Dimensionality Reduction;
Consumer Price Index.

ABSTRACT

Multivariate time series (MTS) analysis of the Consumer Price Index (CPI) in Indonesia often encounters challenges such as outliers, missing data, and inter-variable correlations. Principal Component Analysis (PCA) is a practical approach for dimensionality reduction; however, its performance may vary depending on the data characteristics. This study is a quantitative comparative study that integrates empirical analysis and Monte Carlo simulation based on a first-order Vector Autoregressive (VAR(1)) model to evaluate three PCA approaches: Classical PCA, Robust PCA (RPCA), and PCA of MTS. These methods were applied to weekly price data of eight strategic food commodities across 70 districts and cities in Indonesia. The evaluation employed three criteria: (1) dimensionality reduction efficiency (empirical and simulation), (2) reconstruction accuracy measured using Root Mean Square Error (RMSE) (empirical), and (3) robustness to outliers and inter-variable correlations (simulation). Empirical results indicate that Classical PCA (lag 1) and RPCA (lag 1) are both efficient and effective in reducing dimensionality with minimal information loss. Using the first three principal components, all three methods were able to explain at least 85% of the total variance, with lag 1 identified as optimal. Simulation results reveal that RPCA (lag 1) provides the most stable and consistent performance in the presence of outliers, while Classical PCA (lag 2) performs better under conditions of high inter-variable correlation and a low proportion of outliers. These findings suggest that robust covariance estimation can improve the accuracy of dimensionality reduction and enhance the stability of multivariate time-series analysis for food price data in Indonesia.

https://doi.org/10.31764/jtam.v10i1.34151

A. INTRODUCTION

Various challenges in the analysis of multivariate time series, such as missing data, outliers, and inter-variable correlations, are frequently encountered in weekly Consumer Price Index (CPI) data across 70 districts/cities in Indonesia. Such complexities can hinder the estimation process and reduce the accuracy of predictive outcomes. Handling missing data is a common challenge in data analysis. In this regard, imputing missing values often becomes a crucial step in data analysis (Stekhoven & Bühlmann, 2012). This issue is especially relevant in the context of Indonesia's weekly CPI data, which often contain gaps due to reporting delays and irregularities across regions. The mechanism underlying missing data plays a critical role in determining the appropriate handling method. Three types of mechanisms are commonly distinguished: Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR) (Little & Rubin, 2002; Lotfipoor et al., 2023).

The choice of imputation method also depends on the type and structure of the data. Moreover, algorithms commonly used in the analysis of large-scale data often rely on a complete dataset (Stekhoven & Bühlmann, 2012). For univariate data, linear interpolation can be employed. This method imputes missing values by drawing a straight line between the last observed data point and the first subsequent observed data point following the missing value (Sumertajaya et al., 2023). Linear interpolation has been extensively studied and implemented in the *imputeTS* package by (Moritz & Bartz-Beielstein, 2017). In this study, imputation was applied to empirical CPI data to ensure that the three Principal Component Analysis (PCA) approaches, namely Classical PCA, Robust PCA (RPCA), and PCA of Multivariate Time Series (PCA of MTS), were evaluated under comparable conditions.

In addition to imputation, dimensionality reduction constitutes the next step in addressing multivariate datasets. PCA is a widely used and effective method for reducing dimensionality while retaining most of the data variability (Zamprogno et al., 2020). Several extensions of PCA have been developed to address challenges in the analysis of multivariate time series. Wei (2019) introduced the theoretical foundation of PCA for Multivariate Time Series (MTS), which in this study is referred to as PCA of MTS. Alshammri & Pan (2021) developed Moving Dynamic Principal Component Analysis (MDPCA) to address dynamic dependencies in non-stationary data. Zhao & Shang (2016) introduced Non-Stationary Principal Component Analysis (NSPCA), while Sundararajan (2021) proposed a frequency component–based PCA approach for the segmentation of multivariate time series. However, most of these developments focus on methodological extensions rather than systematic performance comparisons across PCA variants under realistic data conditions involving outliers, correlations, and autocovariance structures.

Since PCA is constructed based on the sample covariance or correlation matrix, the technique is highly sensitive to outliers, which may lead to misleading dimensionality reduction results (Cotta, 2019). An outlier is an observation that falls well above or well below the overall bulk of the data (Agresti et al., 2023). The presence of outliers may arise from errors in data recording, unusual but explainable events in observations, measurement errors, analytical mistakes, instrument failures, experimental errors, or substantial variability within the data (Montgomery et al., 2012; Gao & Fang, 2016). Outliers can substantially affect statistical estimation results, including the covariance structure of variables, which in turn may lead to inaccurate estimation of principal components (Reisen et al., 2018; Cotta et al., 2020). Therefore, in the context of PCA, robust estimators are required to mitigate the impact of outliers (Cotta, 2019).

Several studies have employed covariance or correlation matrices that are robust to outliers, including those by Cotta et al. (2020), Reisen et al. (2024), Cotta (2014), and Cotta et al. (2017). PCA methods incorporating robust covariance or correlation matrices have been specifically examined by Cotta et al. (2020) and Cotta (2014). For convenience, the PCA approach developed in these studies is referred to as Robust PCA (RPCA). In this study, Classical PCA refers to the approach of Reisen et al. (2024), which employs the Autocorrelation Function (ACF), but is implemented using the Autocovariance Function (ACOVF) via the *acf()* function in the stats package.

Reisen et al. (2024) demonstrated that robust covariance-based approaches are more stable to outliers in multivariate time series; however, their study was conducted using a factor-modeling approach rather than through a direct comparison among Classical PCA, RPCA, and PCA of MTS applied to economic time-series data. Cotta (2014) primarily analyzed RPCA and Classical PCA at lag 0, whereas the present study extends this analysis by constructing total autocovariance matrices for lags 1 to 7 in the Classical and Robust PCA methods, following the approach of Reisen et al. (2024), who developed a factor-modeling scheme based on the sum of autocovariances across multiple lags.

While these studies provided important insights into the robustness of covariance-based estimators, systematic comparative evaluations that jointly examine Classical PCA, RPCA, and PCA of MTS under varying correlation and outlier scenarios remain scarce. This study addresses that gap by comparing the three PCA approaches for dimensionality reduction of weekly multivariate time-series data on eight strategic food commodities across 70 districts and cities in Indonesia. It aims to evaluate Classical PCA, RPCA, and PCA of MTS through empirical analysis and Monte Carlo simulation based on a VAR(1) model to assess efficiency, effectiveness, reconstruction accuracy, and robustness to outliers and inter-variable correlations in Indonesia's weekly CPI data. The main contribution of this research lies in providing a PCA-based comparative framework distinct from factor modeling, extending the application of RPCA and Classical PCA from lag 1 to 7, and comparing them with PCA of MTS using evaluation criteria encompassing component efficiency, effectiveness, cumulative variance proportion, and reconstruction accuracy.

B. METHODS

This study is a quantitative comparative study that combines empirical analysis and Monte Carlo simulation to compare three PCA approaches for dimensionality reduction in multivariate time series data, namely PCA of MTS, Classical PCA, and RPCA. The empirical analysis aims to evaluate the performance of these methods on real CPI data, while the simulation analysis provides a controlled setting to assess their robustness against outliers and varying correlation structures. Both analyses form an integrated approach that connects empirical interpretability with simulation-based validation. The primary distinction among these approaches lies in the estimation of the covariance or autocovariance matrix employed in the eigen decomposition process.

1. Empirical Analysis Design

The empirical analysis was conceptually structured into five main components:

a. Data Preparation

Weekly price data for eight strategic food commodities, namely rice, chicken eggs, shallots, garlic, red chili, bird's eye chili, cooking oil, and sugar, were obtained from the National Strategic Food Price Information Center (Bank Indonesia, 2025) covering the period from July 21, 2021, to February 26, 2025. From the total of 74 districts/cities recorded, only 70 were included in the analysis, as four regions (Gorontalo and Banggai districts, Aceh Besar district, and Metro city) had insufficient data throughout the observation period. Two additional commodities, namely broiler chicken meat and beef, were excluded due to incomplete and inconsistent availability across locations.

Missing values were detected and, when present, imputed using linear interpolation. The Augmented Dickey-Fuller (ADF) test at the 10% significance level was used to assess stationarity. Non-stationary series were iteratively differenced until stationary, followed by mean-centering. The QS test (Molinaro & DeFalco, 2022) was applied to detect the presence of seasonal patterns, and its outcomes were later incorporated into simulation modeling when seasonal effects appeared in at least 20% of the series. For this purpose, the test was applied to weekly time series data that were converted into *ts()* objects with a frequency of 52. Although there are technically approximately 52.18 weeks in a year, most R functions that utilize *ts()* objects require the frequency parameter to be specified as an integer (Hyndman & Athanasopoulos, 2018).

b. PCA Modeling

The dimensionality reduction was conducted using three PCA approaches. In the PCA of MTS, once the multivariate time series was verified to be stationary through the ADF test and subsequently mean-centered, the covariance matrix $\hat{\Gamma}$ was constructed and subjected to eigendecomposition to obtain the sample principal components along with their associated eigenvalues and eigenvectors. The implementation was performed in R version 4.5.1 using the stats and tsqn packages for classical and robust autocovariance estimation, while PCA of MTS was implemented using the princomp() function following Wei (2019). For the Classical PCA and RPCA, eigendecomposition was applied to the total sample autocovariance matrix defined as

$$\widehat{\boldsymbol{M}} = \sum_{h=1}^{h_0} \hat{\boldsymbol{\Gamma}}^*(h)\hat{\boldsymbol{\Gamma}}^*(h)' \tag{1}$$

where $\hat{\boldsymbol{\Gamma}}^*(h)$ corresponds to the sample autocovariance matrix at positive lag $h = 1, 2, \dots, h_0$. This procedure was repeated iteratively for $h_0 = 1, 2, \dots, 7$. In the case of Classical PCA, $\hat{\boldsymbol{\Gamma}}^*(h)$ was taken as the sample autocovariance matrix $\hat{\boldsymbol{\Gamma}}(h)$, while for RPCA, $\hat{\boldsymbol{\Gamma}}^*(h)$ was defined as the robust autocovariance matrix $\hat{\boldsymbol{\Gamma}}^Q(h)$ estimated using the scale estimator $Q_n(.)$ (Reisen et al., 2024). The eigendecomposition of $\widehat{\boldsymbol{M}}$ then yielded the estimated eigenvalues $\hat{\lambda}_i$.

c. Evaluation Design

1) Efficiency and Effectiveness Evaluation

The evaluation of efficiency and effectiveness began by comparing Classical PCA and RPCA under identical lag settings. The optimal lag was defined as the smallest lag within the range of 1 to 7 that produced a discernible difference in evaluation outcomes between the two methods, specifically when one demonstrated superior performance over the other. The superior method at this optimal lag was subsequently compared with PCA of MTS, and the final method was selected based on the criterion of yielding the fewest principal components.

When the number of principal components was identical across methods, the method explaining the greater proportion of cumulative variance was considered more effective. The most efficient and effective approach was thus defined as the one

requiring the fewest components to achieve a minimum cumulative variance threshold of 85%, while retaining most of the data variability.

2) Reconstruction Performance Evaluation

The reconstruction performance of each method was evaluated by assessing its ability to reconstruct the preprocessed data, which included missing value imputation using linear interpolation, stationarity testing and differencing when necessary, and mean-centering, using three principal components. The preprocessed data were denoted as a matrix $\boldsymbol{X} \in \mathbb{R}^{nxp}$, where n represents the number of time periods in the preprocessed dataset and p corresponds to the number of commodities (eight in total). PCA analysis was conducted through the eigendecomposition of the covariance matrix of $\boldsymbol{X}$, yielding eigenvectors and eigenvalues, with the matrix of eigenvectors denoted as $\boldsymbol{V}$. The number of principal components (PCs) used in the reconstruction process was determined as the maximum number of PCs across the three PCA approaches, each of which satisfied the minimum cumulative variance threshold of 85%. The orthonormal submatrix corresponding to the first three principal components was denoted as $\boldsymbol{V}_{reduced} \in \mathbb{R}^{px3}$. The reduced principal component scores were computed as

$$\boldsymbol{U}_{reduced} = \boldsymbol{X} \cdot \boldsymbol{V}_{reduced}, \tag{2}$$

and the data were subsequently reconstructed as

$$\widehat{\boldsymbol{X}} = \boldsymbol{U}_{reduced} \cdot \boldsymbol{V}_{reduced}^{T}. \tag{3}$$

Reconstruction accuracy was assessed using the Root Mean Square Error (RMSE), defined as

$$RMSE = \sqrt{\frac{1}{n \text{ x } p}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(X_{ij} - \hat{X}_{ij}\right)^2} \tag{4}$$

where X_{ij} denotes the preprocessed data value at the i-th row and j-th column, and $\hat{X}_{ij}$ represents the reconstructed value obtained from the projection onto the first three principal components. A model was considered to demonstrate satisfactory performance when the RMSE was smaller than the standard deviation of both the observations and the reconstructed model (Liemohn et al., 2021), where the standard deviations were computed from the preprocessed data matrix $\boldsymbol{X}$ and the reconstructed matrix $\widehat{\boldsymbol{X}}$, respectively.

The identification of the best method for reconstruction performance was carried out by comparing the RMSE values across the three principal components obtained from each approach. The method that yielded the lowest RMSE was considered superior,

with good performance defined by the criterion that the RMSE must be smaller than both the standard deviation of the observed data and that of the reconstructed data.

d. Outlier Detection
Following the confirmation of stationarity and the application of mean-centering, additive outliers were examined through the analysis of the autocovariance function (ACOVF) at the lag identified as optimal based on the PCA evaluation. The classical ACOVF was employed for Classical PCA, while the robust ACOVF, based on robust scale estimation, was utilized for RPCA.

e. Visualization
The results of the analysis were further complemented with visualization to support the evaluation of reconstruction performance. The projected PCA scores ($\boldsymbol{U}_{reduced}$) were plotted in a two-dimensional space defined by the first two principal components (PC1 and PC2). To preserve directional and scaling proportionality in the graphical representation, $\boldsymbol{U}_{reduced}$ was rescaled relative to $\boldsymbol{V}_{reduced}$, thereby producing a visualization analogous to a biplot. This visualization illustrates both the temporal relationships across periods (rows of $\boldsymbol{U}_{reduced}$) and the directional contributions of each commodity (columns of $\boldsymbol{V}_{reduced}$) to the first two principal components.

2. Simulation Analysis Design

The simulation analysis in this study was conducted to replicate the key stages of the empirical analysis under controlled conditions, allowing for the evaluation of the three PCA approaches when the correlation structure and the presence of additive outliers were known in advance. The procedures followed the same conceptual structure as the empirical analysis, consisting of data preparation, PCA modeling, and evaluation of efficiency and effectiveness, but were applied to simulated datasets generated from a parametric VAR(1) process. The simulation analysis in this study was conducted through the following steps:

a. Pre-Simulation Data
The pre-simulation stage followed the same procedures as the empirical analysis, which included the collection of empirical data, detection and treatment of missing values, stationarity testing, and seasonal testing.

b. Simulation Modelling
The simulated data were generated based on a first-order Vector Autoregressive model without intercept (VAR(1)), with the process mean set to zero, formulated as:

$$\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_t \sim N_p(\mathbf{0}, \boldsymbol{\Gamma}_\varepsilon) \quad (5)$$

where $\boldsymbol{Y}_t \in \mathbb{R}^p$ denotes the observation vector at time t, $\boldsymbol{\Phi}$ is a pxp autoregressive coefficient matrix, and $\boldsymbol{\Gamma}_\varepsilon$ is the covariance matrix of the white noise process $\boldsymbol{\varepsilon}_t$. The number of variables was fixed at $p = 8$, while the number of time periods matched the length of the pre-simulation data. Parameters $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}_\varepsilon$ were estimated from empirical data of a selected district/city using Maximum Likelihood estimation of the VAR(1)

model without intercept, with the residuals employed to compute $\boldsymbol{\Gamma}_{\varepsilon}$. To ensure stationarity, $\boldsymbol{\Phi}$ was normalized according to:

$$\boldsymbol{\Phi}^{*} = \frac{0.5}{\max|\lambda_i|} \cdot \boldsymbol{\Phi}, \tag{6}$$

where λ_i are the eigenvalues of the estimated $\boldsymbol{\Phi}$.

c. Correlation Structures
To incorporate varying levels of cross-variable dependence, three correlation structures were considered. For the low correlation case, the covariance matrix was specified as a diagonal form,

$$\boldsymbol{\Gamma}_{low} = diag\left(\sigma_1^2, \dots, \sigma_p^2\right) \tag{7}$$

where σ_i denotes the empirical residual standard deviation of variable i. For the moderate and high correlation cases, the covariance structure was instead constructed by transforming a target correlation matrix $\boldsymbol{R}$ into a covariance matrix using

$$\boldsymbol{\Gamma} = \boldsymbol{D}\,\boldsymbol{R}\,\boldsymbol{D}, \tag{8}$$

where $\boldsymbol{R}$ is a correlation matrix with constant off-diagonal entries of 0.4 (for moderate correlation) or 0.8 (for high correlation), and diagonal entries equal to 1. The diagonal scaling matrix is defined as

$$\boldsymbol{D} = diag\left(\sigma_1, \dots, \sigma_p\right), \tag{9}$$

which contains the empirical residual standard deviations along its diagonal.

d. Data Generation
The data were generated over a predetermined number of time steps, with the initial 50 observations discarded as a burn-in period to eliminate the influence of initial conditions. Only the remaining observations beyond the burn-in phase were retained and utilized as the primary simulated dataset for subsequent analysis.

e. Formation of Simulation Datasets: With and Without Outliers
Two types of simulated datasets were constructed, namely baseline datasets without outliers and datasets containing outliers. The baseline datasets were generated from the VAR(1) model for each category of inter-variable correlation. The datasets with outliers were obtained by injecting anomalies into the baseline datasets according to a combination of factors: (a) inter-variable correlation categories (low, moderate, and high), (b) outlier proportions (1%, 5%, and 15% of the total observations), (c) outlier magnitudes (4, 10, and 15 standard deviations of the residuals for each variable), and (d) outlier locations (restricted to the first variable or simultaneously across all variables). This design resulted in 18 outlier scenarios plus one baseline scenario for each correlation category, yielding a total of 19 scenarios per category. With 100 replications for each scenario, a total of 1,900 datasets were produced per correlation category,

amounting to 5,700 simulated datasets overall. The datasets containing outliers were generated through an injection process until the resulting series satisfied the stationarity criterion based on the ADF test at the 10% significance level.

f. Preparation before PCA
   The simulated time series were first tested for stationarity using the ADF test at the 10% significance level and subsequently mean-centered prior to conducting the dimensionality reduction analysis.
g. PCA Modeling
   Three dimensionality reduction approaches were applied, consistent with those employed in the empirical analysis.
h. Evaluation of Efficiency and Effectiveness
   The evaluation of efficiency and effectiveness in the simulation data followed the same principles as in the empirical analysis, namely by assessing the dimensionality reduction performance based on the minimum number of principal components (PCs) required and the cumulative proportion of explained variance. For each scenario and method (Classical PCA, RPCA, and PCA of MTS), the average covariance matrix was computed across 100 simulation replications. Classical PCA and RPCA were applied over lags ranging from 1 to 7, whereas PCA of MTS was applied only once. All covariance matrices were subsequently subjected to eigendecomposition to obtain the proportion of variance explained. The identification of the best method was carried out based on efficiency and effectiveness criteria, selecting the approach that demonstrated the most favorable balance between the minimum number of principal components required and the cumulative proportion of explained variance.

C. RESULT AND DISCUSSION

The Results and Discussion section is organized into two main parts, namely the empirical data analysis and the simulation-based modelling analysis. The empirical data analysis encompasses the distribution and imputation of missing values, stationarity testing, seasonality testing, evaluation of the efficiency and effectiveness of dimensionality reduction, assessment of data reconstruction performance, detection of outliers, and visualization. Meanwhile, the simulation modelling analysis focuses on the process of data generation and the evaluation of PCA performance across various correlation scenarios, both in the absence and presence of outliers.

1. Distribution and Imputation of Missing Data

Prior to the main analysis, an initial exploration was conducted to assess the completeness of the dataset. The results indicated the presence of missing values in several commodities across a number of districts/cities. The proportion of missing data for the eight commodities in districts/cities with incomplete records is presented in Figure 1.

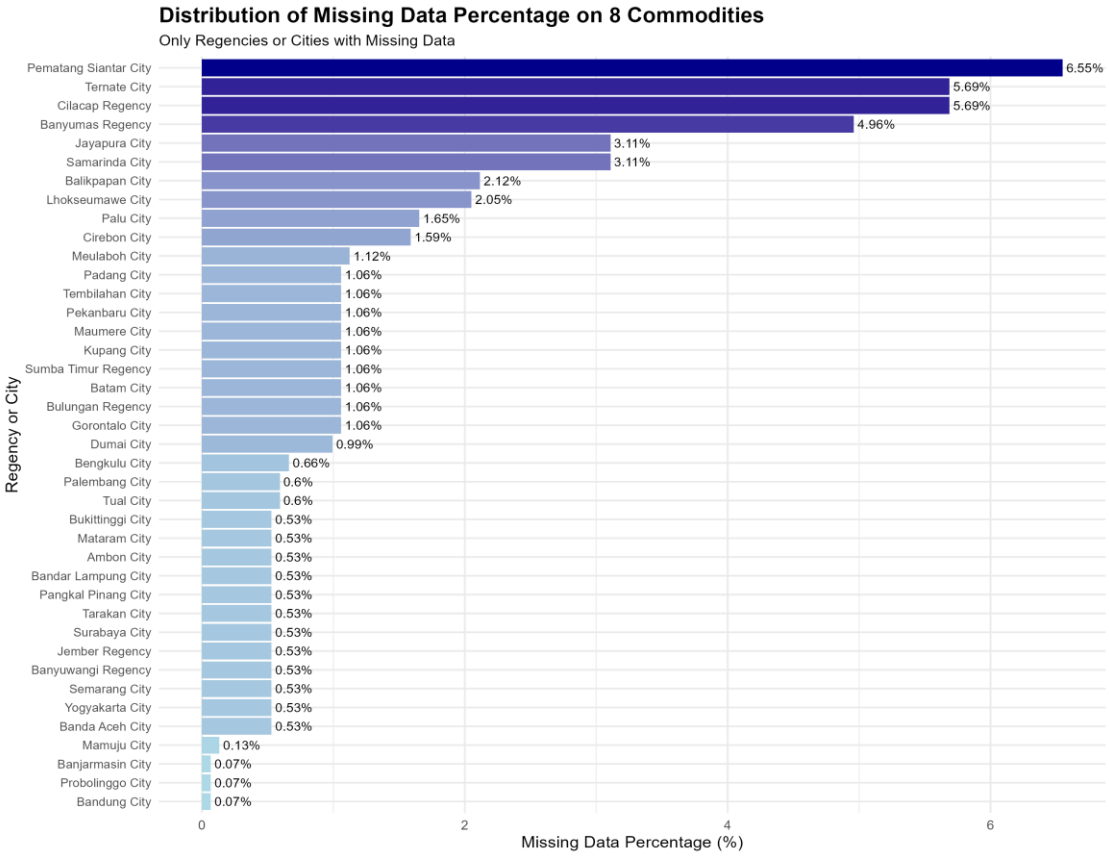
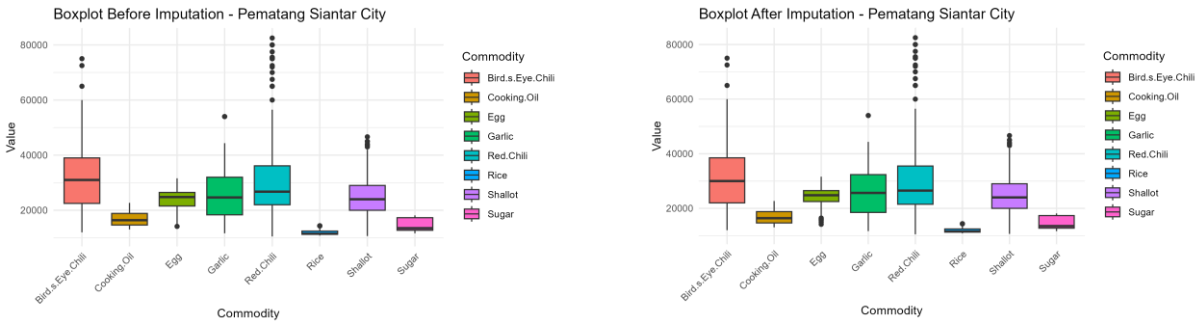


Figure 1. Percentage of Missing Data Across Eight Commodities for Districts/Cities With Incomplete Records

Based on Figure 1, Pematang Siantar City exhibited the highest proportion of missing data at 6.55%. In contrast, 30 other districts/cities had no missing data and were therefore not included in the graph. These findings indicate the presence of missing values in certain regions, necessitating imputation prior to further analysis. To address this issue, a linear interpolation method was employed. As an illustration, Figure 2 presents a comparison of the data distribution before (a) and after (b) the imputation process in Pematang Siantar City, which represents the region with the highest proportion of missing data.



(a) (b)

Figure 2. Boxplot Visualization Results Before (a) and After (b) the Imputation Process for Commodity Data in Pematang Siantar City

Based on the visualization in Figure 2, the distributional patterns of the data did not exhibit any significant changes following the imputation process. The median, interquartile range (IQR), and overall distributional shape remained consistent between the pre- and post-imputation datasets. This indicates that the application of linear interpolation did not substantially alter the statistical characteristics of the time series data. Consequently, the imputed data are deemed suitable for subsequent stages of analysis. All subsequent tests and analyses in this study were therefore conducted using data imputed through the linear interpolation method. Conceptually, the imputation step ensured that all three PCA approaches, namely Classical PCA, RPCA, and PCA of MTS, were evaluated under identical data conditions, thereby maintaining fairness and methodological consistency across both the empirical and simulation analyses.

2. Stationarity Testing

Stationarity is a fundamental assumption in the application of PCA to multivariate time series. To verify this assumption, the Augmented Dickey–Fuller (ADF) test was applied to eight commodities across 70 districts/cities. Three forms of data were examined: (1) data at level, denoted as I(0); (2) first-order differenced data, denoted as I(1); and (3) mean-centered I(1) data. Table 1 presents the percentage of districts/cities classified as stationary (S) and non-stationary (NS) based on the results of the ADF test applied to 560 weekly time series (8 commodities across 70 districts/cities).

Table 1. Percentage of districts/cities classified as stationary (S) and non-stationary (NS) based on the results of the ADF test applied to 560 weekly time series (8 commodities across 70 districts/cities)

Data Type	S (%)	NS (%)
I(0)	43.21	56.79
I(1) before MC	100	0
I(1) after MC	100	0

Notes: I(0) = data at level; I(1) = first-order differenced data; MC = mean-centering; S = stationary at the 10% significance level; NS = non-stationary at the 10% significance level.

Based on Table 1, the majority of the level data were found to be non-stationary at the 10% significance level. However, after applying first-order differencing, all time series became stationary, both before and after mean-centering. Consequently, all subsequent PCA analyses, reconstruction procedures, and visualizations in this study were conducted using first-order differenced data that had been mean-centered. Conceptually, this preprocessing step ensured that the multivariate time series fulfilled the stationarity assumption required for PCA and established methodological consistency with the simulation design, where all generated series were designed to be stationary.

3. Seasonality Testing

The next step involved testing for the presence of seasonal patterns in the data through autocorrelation analysis at seasonal lags using the QS test. Three types of data were examined using the QS method: (1) level data (I(0)), (2) first-order differenced data (I(1)) before mean-centering, and (3) I(1) data after mean-centering. Table 2 presents the percentage of time series

exhibiting seasonal and non-seasonal patterns based on the QS test applied to 560 weekly time series (covering eight commodities across 70 districts/cities).

Table 2. Percentage of Time Series Indicating Seasonal and Non-Seasonal Patterns Based on the QS Test Applied to 560 Weekly Time Series (Eight Commodities Across 70 Districts/Cities)

Data Type	M (%)	TM (%)
I(0)	5.4	94.6
I(1) before MC	6.1	93.9
I(1) after MC	6.3	93.8

Note: MC = Mean-centering; M = Time series indicating seasonal effects at the 10% significance level; TM = Time series not indicating seasonal effects at the 10% significance level.

Table 2 indicates that, overall, the proportion of time series exhibiting evidence of seasonal effects at the 10% significance level is less than 20%. Consequently, seasonal effects were not further addressed in the subsequent analysis. Conceptually, this finding supports the short-memory assumption adopted in the simulation model, where the absence of strong seasonal patterns justifies modeling the data using a VAR(1) process.

4. Evaluation of Dimensionality Reduction Efficiency and Effectiveness

The evaluation of the efficiency and effectiveness of dimensionality reduction considers the fact that the strongest correlations in multivariate time series generally occur at small lag values (Reisen et al., 2024). Accordingly, the selection of lags 1 through 7 for Classical PCA and RPCA follows the simulation procedure of Cotta et al. (2017) as well as the established practice of employing fixed positive lags in the study by Reisen et al. (2024). In contrast, the PCA of MTS approach is applied directly to the multivariate time series data without explicitly accounting for lag variation. The optimal lag selected in this study is lag 1, as it represents the smallest lag that consistently reveals differences in evaluation outcomes between Classical PCA and RPCA, both in terms of the minimum number of principal components required and the cumulative proportion of explained variance. Consequently, lag 1 was adopted as the basis for subsequent analyses involving these two methods. A summary of the comparative results between Classical PCA (lag 1), RPCA (lag 1), and PCA of MTS with respect to the minimum number of principal components needed to achieve a cumulative explained variance of at least 85% is presented in Table 3.

Table 3. Summary of the optimal number of principal components from the comparative results between Classical PCA (lag 1), RPCA (lag 1), and PCA of MTS.

Best Method	Number of PCs
Classical PCA (lag 1)	1 or 2
RPCA (lag 1)	1 or 2
PCA of MTS	2 or 3

Based on Table 3, it can be observed that both Classical PCA and RPCA generally require only one to two principal components to achieve the 85% variance threshold. In contrast, PCA of MTS tends to require a larger number of principal components (two to three) to reach the

same level of explained variance. This finding indicates that, within the context of the present dataset, PCA of MTS is relatively less efficient and effective compared to the other two methods.

As a result of this selection process, Classical PCA emerged as the best method for 33 districts/cities, while RPCA was identified as the best method for 37 districts/cities. These findings indicate that RPCA demonstrates a slight advantage over Classical PCA in the context of the analyzed data, particularly in terms of the minimum number of principal components required and the cumulative proportion of variance explained at lag 1. Padang Sidempuan City is the only region that achieved a cumulative proportion of variance of 100%, exhibited no missing data, and attained the best results using the RPCA method at lag 1 with a single principal component. This result is consistent with Cotta (2014) and complements the insights of Reisen et al. (2024), indicating the advantage of RPCA in achieving stable and efficient dimensionality reduction.

5. Reconstruction Performance Evaluation

The evaluation of reconstruction performance was conducted by calculating the Root Mean Square Error (RMSE) between the pre-processed data and the reconstructed data obtained from three principal components, which were consistently applied across all methods. A model was categorized as having good performance if the RMSE value was smaller than both the standard deviation of the observations and the standard deviation of the reconstructed data, in accordance with the criteria established by Liemohn et al. (2021). The summary of the number and percentage of districts/cities based on the reconstruction performance criteria is presented in Table 4.

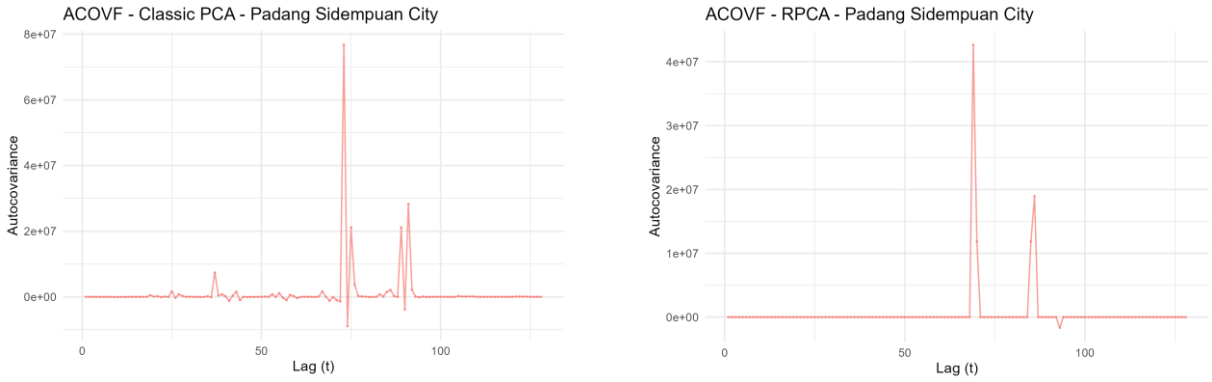
Table 4. Summary of the Number and Percentage of Districts/Cities Based on Model Performance Criteria

Method	Good Model Performance		Poor Model Performance	
	Number	Percentage (%)	Number	Percentage (%)
Classical PCA (lag 1)	70	100	0	0
RPCA (lag 1)	70	100	0	0
PCA of MTS	70	100	0	0

Based on Table 4, when using the first three principal components, Classical PCA (lag 1), PCA of MTS, and RPCA (lag 1) consistently achieved the best reconstruction performance across all districts/cities. To complement this evaluation, data reconstruction was also performed using all principal components without dimensionality reduction for the 70 districts/cities. The results revealed that the RMSE values were extremely small, ranging from 1.91×10^{-12} to 8.67×10^{-12} for Classical PCA, 0 to 1×10^{-11} for RPCA, and 1.05×10^{-12} to 1.14×10^{-11} for PCA of MTS. These near-zero values indicate that the full variability of the data can be almost perfectly reproduced when all principal components are employed. This suggests that the use of three principal components is sufficient to accurately represent the main structure of the data across all regions. Furthermore, the consistent reconstruction accuracy suggests that the empirical and simulation results are in good agreement, implying that dimensionality reduction using PCA tends to preserve the main data structure, in line with its theoretical purpose of retaining data variability.

6. Outlier Detection

An outlier detection assessment was conducted by applying both classical and robust ACOVF analyses to the I(1) data after mean-centering at the optimal lag (lag 1). Consistent with Reisen et al. (2024), the classical and robust ACF exhibited similar patterns in the absence of additive outliers; however, the classical ACF was substantially distorted in the presence of additive outliers, whereas the robust ACF remained stable. Padang Sidempuan City was selected as a representative case study, as it contained no missing data and achieved a cumulative variance proportion of 100% using the RPCA method at lag 1 with a single principal component. Figure 3 presents a comparison of the ACOVF derived from Classical PCA and RPCA applied to the I(1) data after mean-centering in Padang Sidempuan City.



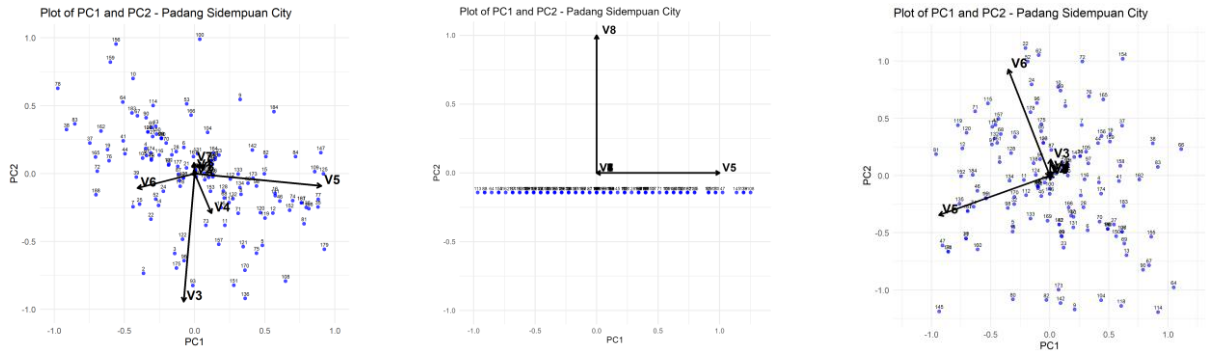
(a) (b)

Figure 3. Comparison of Classical (a) and Robust (b) ACOVF on I(1) Data After Mean-Centering in Padang Sidempuan City

The visualization results presented in Figures 3(a) and 3(b) reveal that the classical ACOVF in Padang Sidempuan City produces elevated autocovariance values, indicating the presence of additive outliers. In contrast, the robust ACOVF demonstrates lower and more stable autocovariance values, reflecting its resilience against outliers. These findings provide clear evidence of additive outliers within the data. Conceptually, this result supports the simulation findings, indicating that RPCA provides more stable covariance estimation under additive outlier conditions.

7. Visualization of Principal Components

The visualization of the projection of the first two principal components (PC1 and PC2) obtained from the three PCA approaches, namely Classical PCA (lag 1), RPCA (lag 1), and PCA of MTS, was conducted for a representative region, namely Padang Sidempuan City, to illustrate the temporal relationships across observation periods as well as the contribution of each commodity to the overall data variability. For ease of visual interpretation, the eight commodities were denoted as follows: V1 for Rice, V2 for Egg, V3 for Shallot, V4 for Garlic, V5 for Red Chili, V6 for Bird's Eye Chili, V7 for Cooking Oil, and V8 for Sugar. The visualization of the projection of the first two principal components for Padang Sidempuan City using the three PCA methods is presented in Figure 4, corresponding to Classical PCA (a), RPCA (b), and PCA of MTS (c).



(a) (b) (c)

Figure 4. Visualization of the Projection of the First Two Principal Components in Padang Sidempuan City: (a) Classical PCA (Lag 1), (b) RPCA (Lag 1), and (c) PCA of MTS

Based on Figure 4, Red Chili is identified as the commodity with a consistently dominant contribution across all PCA approaches in Padang Sidempuan City. Conceptually, this pattern indicates that the main sources of variability captured by Classical PCA, RPCA, and PCA of MTS are consistent, suggesting that the empirical and simulation analyses are in agreement in identifying dominant variables driving joint price dynamics.

8. Simulation-Based Modeling

The simulation analysis indicates that the data generation process satisfies the assumptions of a VAR(1) model, with a stationary coefficient matrix $\mathbf{\Phi}$ and a disturbance covariance matrix $\mathbf{\Gamma}_{\varepsilon}$ that is symmetric and positive definite. Based on simulations without outliers, three categories of inter-variable correlation levels, namely low, moderate, and high, were identified, as visualized in Figure 5.

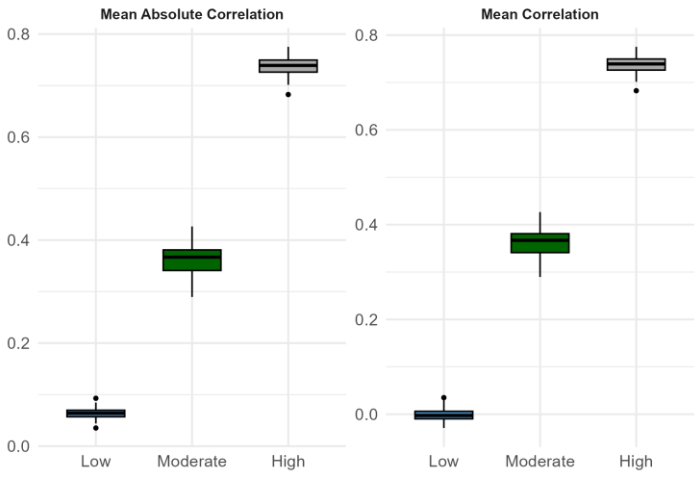


Figure 5. Boxplot of Mean Absolute Correlation and Mean Correlation from Simulation Results Without Outliers for Low, Moderate, and High Correlation Categories

Based on Figure 5, the mean absolute correlation values among variables range from 0.0352 to 0.0929 for the low correlation category, 0.2896 to 0.4265 for the moderate category, and 0.6826 to 0.7752 for the high category. Meanwhile, the mean correlation values (accounting for the sign) vary between -0.0291 and 0.035 for the low correlation category,

0.2896 to 0.4265 for the moderate category, and 0.6826 to 0.7752 for the high category. Furthermore, the ADF test confirmed that all simulated datasets, both the baseline and those containing outliers, were stationary at the 10% significance level. Conceptually, these simulation results are consistent with the empirical findings, reinforcing the integration between empirical analysis and simulation-based validation while successfully reflecting short-memory dynamics and inter-variable relationships observed in the data.

9. Evaluation of PCA Methods on Simulated Data

The evaluation of PCA methods on simulated data was conducted with respect to cumulative variance proportion, the number of principal components (PCs), and optimal lag across various simulation scenarios. The results indicate that only under the scenario characterized by high correlation, outliers present in all variables, large magnitude (15), and a low proportion of outliers (1%), Classical PCA achieved the best performance, with a single principal component explaining 90.68% of the variance at lag 2. In all other scenarios, the method demonstrating superior performance was RPCA, as summarized in Table 5.

Table 5. Average Variance Proportion, Number of Principal Components, and Optimal Lag by RPCA

Correlation	Outlier Condition	Proportion (%)	Average PC	Optimal Lag
Low	Without Outliers	93.46	2	1
	All Variables	93.51	2	1
	First Variable	93.46	2	1
Moderate	Without Outliers	95.11	2	1
	All Variables	94.83	2	1
	First Variable	95.11	2	1
High	Without Outliers	90.68	1	1
	All Variables	90.01	1	1
	First Variable	90.68	1	1
Overall Scenarios		92.99	1.68	1

Table 5 indicates that RPCA consistently emerges as the most efficient and effective method, achieving an average variance proportion of 92.99%, with a relatively low average number of principal components (1.68), an optimal lag consistently identified at lag 1, and stable performance even in the presence of outliers. To further support this result, the autocovariance behavior was visually examined to illustrate how outliers influence the stability of classical and robust estimators. Visual inspection of the autocovariance behavior (Figure 3) shows that classical ACOVF produces elevated autocovariance values due to outliers, whereas the robust ACOVF remains more stable, consistent with the robust ACF behavior reported by Reisen et al. (2024). The simulation results confirm that the strongest correlations occur at small lags, while autocovariance estimation becomes less accurate at larger lags (Reisen et al., 2024), with lag 1 identified as the optimal lag. Unlike Reisen et al. (2024), who implemented a factor modeling approach based on the summation of autocovariances, and Cotta (2014), who compared RPCA and Classical PCA only at lag 0, this study also includes PCA of MTS within a short-memory VAR(1) setting under three levels of inter-variable correlation. Within this structure, both Classical PCA and RPCA were constructed from the summation of autocovariances across

several lags, while RPCA achieved the highest efficiency and variance proportion, with Classical PCA outperforming only under high-correlation, low-outlier conditions.

D. CONCLUSION AND SUGGESTIONS

This study synthesizes empirical and simulation-based analyses to evaluate three PCA approaches, namely Classical PCA, RPCA, and PCA of MTS, for dimensionality reduction of strategic food price data in Indonesia. Both Classical PCA (lag 1) and RPCA (lag 1) proved to be efficient and effective, requiring only one to two principal components to explain at least 85% of the total variance. At the optimal lag (lag 1), RPCA demonstrated the most stable and consistent performance, particularly in the presence of outliers, whereas Classical PCA (lag 2) was superior only in scenarios with high inter-variable correlation and a low proportion of outliers. On average, RPCA achieved 92.99% of explained variance with a relatively small number of components (1.68). Scientifically, these findings emphasize that robust covariance estimation plays an essential role and is indicated to enhance dimensionality-reduction accuracy while maintaining the stability of multivariate analysis under imperfect data conditions. Practically, the application of RPCA is indicated to improve the accuracy of price monitoring and policy evaluation in the food sector. Future research is encouraged to develop high-dimensional or long-memory robust models to strengthen the resilience and flexibility of analysis when dealing with more complex data characteristics.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Indonesian Education Scholarship (BPI), the Center for Higher Education Funding and Assessment (PPAPT), Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, and the Indonesian Endowment Fund for Education (LPDP), Ministry of Finance of the Republic of Indonesia, for their financial support in the completion of this research.

REFERENCES

Agresti, A., Franklin, C. A., & Klingenberg, B. (2023). *Statistics: The Art and Science of Learning from Data* (5th ed.). Pearson Education Limited. https://www.pearson.com/en-gb/subject-catalog/p/statistics-the-art-and-science-of-learning-from-data-global-edition/P200000008773/9781292444796

Alshammri, F., & Pan, J. (2021). Moving dynamic principal component analysis for non-stationary multivariate time series. *Computational Statistics*, *36*(3), 2247–2287. https://doi.org/10.1007/s00180-021-01081-8

Bank Indonesia. (2025). *Tabel Harga Pedagang Besar Berdasarkan Daerah*. Pusat Informasi Harga Pangan Strategis (PIHPS) Nasional. https://www.bi.go.id/hargapangan/TabelHarga/PedagangBesarDaerah

Cotta, H. H. A. (2014). *Análise de componentes principais robusta em dados de poluição do ar: aplicação à otimização de uma rede de monitoramento* [Master's thesis, Universidade Federal do Espírito Santo]. https://dspace5.ufes.br/items/72c34454-d695-48f9-9eec-3297f7c6a0d5

Cotta, H. H. A. (2019). *Robust Methods in Multivariate Time Series* [Doctoral Thesis, Universidade Federal do Espírito Santo]. https://sappg.ufes.br/tese_drupal/tese_14040_UFESFINALTheseHigor0809%20%282%29.pdf

Cotta, H. H. A., Reisen, V. A., Bondon, P., & Filho, P. R. P. (2020). Identification of Redundant Air Quality Monitoring Stations using Robust Principal Component Analysis. *Environmental Modeling and Assessment*, *25*(4), 521–530. https://doi.org/10.1007/s10666-020-09717-7

Cotta, H. H. A., Reisen, V. A., Bondon, P., & Stummer, W. (2017). Robust estimation of covariance and correlation functions of a stationary multivariate process. *International Work-Conference on Time Series*, 47–58. https://centralesupelec.hal.science/hal-01578459

Gao, X., & Fang, Y. (2016). *Penalized Weighted Least Squares for Outlier Detection and Robust Regression*. http://arxiv.org/abs/1603.07427

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts. https://otexts.com/fpp2/

Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M., & Mukhopadhyay, A. (2021). RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial Physics*, *218*, 105624. https://doi.org/10.1016/j.jastp.2021.105624

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119013563

Lotfipoor, A., Patidar, S., & Jenkins, D. P. (2023). Transformer network for data imputation in electricity demand data. *Energy and Buildings*, *300*, 113675. https://doi.org/10.1016/j.enbuild.2023.113675

Molinaro, A., & DeFalco, F. (2022). Empirical assessment of alternative methods for identifying seasonality in observational healthcare data. *BMC Medical Research Methodology*, *22*(1), 182. https://doi.org/10.1186/s12874-022-01652-3

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). John Wiley & Sons. https://books.google.co.id/books?id=0yR4KUL4VDkC

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, *9*(1), 207–218. https://doi.org/10.32614/RJ-2017-009

Reisen, V. A., Lévy-Leduc, C., Monte, E. Z., & Bondon, P. (2024). A dimension reduction factor approach for multivariate time series with long-memory: a robust alternative method. *Statistical Papers*, *65*(5), 2865–2886. https://doi.org/10.1007/s00362-023-01504-2

Reisen, V. A., Monte, E. Z., da Conceição Franco, G., Sgrancio, A. M., Molinares, F. A. F., Bondon, P., Ziegelmann, F. A., & Abraham, B. (2018). Robust estimation of fractional seasonal processes: Modeling and forecasting daily average SO_2 concentrations. *Mathematics and Computers in Simulation*, *146*, 27–43. https://doi.org/10.1016/j.matcom.2017.10.004

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Sumertajaya, I. M., Rohaeti, E., Wigena, A. H., & Sadik, K. (2023). Vector Autoregressive-Moving Average Imputation Algorithm for Handling Missing Data in Multivariate Time Series. *IAENG International Journal of Computer Science*, *50*(2), 727–735. https://www.iaeng.org/IJCS/issues_v50/issue_2/IJCS_50_2_42.pdf

Sundararajan, R. R. (2021). Principal component analysis using frequency components of multivariate time series. *Computational Statistics and Data Analysis*, *157*, 107164. https://doi.org/10.1016/j.csda.2020.107164

Wei, W. W. S. (2019). Principal Component Analysis of Multivariate Time Series. In *Multivariate Time Series Analysis and Applications* (1st ed., pp. 139–161). John Wiley & Sons Ltd. https://doi.org/10.1002/9781119502951.ch4

Zamprogno, B., Reisen, V. A., Bondon, P., Aranda Cotta, H. H., & Reis Jr, N. C. (2020). Principal component analysis with autocorrelated data. *Journal of Statistical Computation and Simulation*, *90*(12), 2117–2135. https://doi.org/10.1080/00949655.2020.1764556

Zhao, X., & Shang, P. (2016). Principal component analysis for non-stationary time series based on detrended cross-correlation analysis. *Nonlinear Dynamics*, *84*(2), 1033–1044. https://doi.org/10.1007/s11071-015-2547-6