SARIMA Modeling with a State Space Approach for Tax Revenue Forecasting

Ayu Indriani, I Gede Nyoman Mindra Jaya, Budi Nurani Ruchjana, Anna Chadidjah

Abstract— **Regional tax revenue is the main source of local income for Sumedang Regency, supporting economic and social development. Accurate forecasting of regional tax revenue is critical for effective fiscal planning and public service provision at the local level. This study evaluates the performance of classical SARIMA and SARIMA State Space models in forecasting monthly tax revenue collected by the Regency Primary Tax Service Office. The dataset exhibits seasonal effects, an upward trend, and irregular fluctuations, necessitating a model capable of capturing these characteristics robustly. The SARIMA(0,1,1)(0,1,1)^{12} model was chosen based on AIC values and analysis of the residuals, and subsequently reformulated using a State Space approach to assess its structural advantages. While both models show similar accuracy (MAPE of 27.27% for SARIMA State Space and 27.49% for ARIMA), the SARIMA State Space model yields lower RMSE and Theil's U, and offers a more interpretable and modular structure for future extensions. These results highlight the potential of state space modeling for complex fiscal time series, providing a replicable and theoretically grounded approach to enhance revenue forecasting strategies at the regional level.**

Index Terms— **Forecasting, SARIMA, State Space, Tax Revenue.**

I. Introduction

THE decentralization policy in Indonesia was officially implemented in 2001, marked by the enactment of Law No. 22 on Regional Autonomy and Law No. 25 concerning the distribution of fiscal authority between central and local governments [1]. These regulations positioned Locally Generated Revenue (Pendapatan Asli Daerah, PAD) as a critical component in supporting regional government financing. PAD encompasses all revenue collected from local sources in accordance with applicable regulations [2]. In Sumedang Regency, regional tax revenue constitutes one of the primary contributors to PAD. Revenue from local taxes can be obtained by local governments through three mechanisms: (1) directly collecting taxes, (2) imposing locally determined surcharges on taxes collected by higher levels of government, or (3) receiving fixed allocations from taxes collected by the central government [3]. Sumedang Regency generates tax revenue primarily by directly collecting local taxes through the Sumedang Regency Primary Tax Service Office.

Manuscript received January 30, 2025; revised September 12, 2025.

This work was supported by Universitas Padjadjaran through the RDDU Research Grant Number 4582/UN6.D/PT.00/2025.

*Ayu Indriani is a postgraduate student in the Master's Program in Applied Statistics, Department of Statistics, Padjadjaran University, Sumedang, Indonesia (e-mail: ayu20003@mail.unpad.ac.id).

I Gede Nyoman Mindra Jaya is the Head of the Master's Program in Applied Statistics, Department of Statistics, Universitas Padjadjaran, Sumedang, Indonesia (e-mail: mindra@unpad.ac.id).

Budi Nurani Ruchjana is a Statistics Professor and Head of the Indonesian Ethnoscience Studies Center, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Sumedang, Indonesia (e-mail: budi.nurani@unpad.ac.id).

Anna Chadidjah is a lecturer and Head of the Actuarial Science Undergraduate Program, Department of Statistics, Universitas Padjadjaran, Sumedang, Indonesia (e-mail: anna.chadidjah@unpad.ac.id).

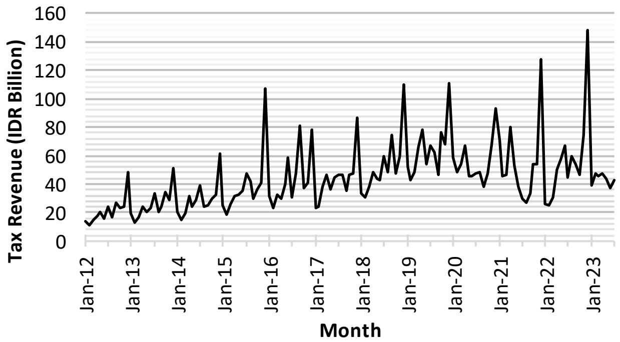


Fig. 1. Time series of tax revenue collected by the Primary Tax Office of Sumedang Regency.

Figure 1 presents the pattern of tax revenue collected by the Sumedang Regency Primary Tax Service Office, measured in billions of rupiah, from January 2012 to July 2023. The data exhibit seasonal fluctuations, with revenue typically peaking in December due to taxpayers' tendency to settle their obligations before the end of the fiscal year. In addition to the seasonal trend, a gradual increase in tax revenue is observed over time. However, a notable decline occurred in the latter half of 2019, likely resulting from the COVID-19 pandemic and the associated implementation of social distancing measures. This decline contributed to missed revenue targets in both 2019 and 2020, prompting the local government to revise its revenue targets downward in the following years.

In this context, forecasting tax revenue for the Sumedang Regency Primary Tax Service Office becomes essential for supporting policymakers in setting realistic and effective revenue targets. However, tax revenue forecasting poses considerable challenges due to its inherent uncertainty, non-linear dynamics, and sensitivity to a range of external factors, such as macroeconomic conditions, public awareness of taxation, the rise of the digital economy, changes in government policies, and taxpayers' financial capacity [4]. Furthermore, historical tax revenue data exhibit extreme and irregular fluctuations driven by random events, policy shocks, and other unobserved influences. Such complexities frequently challenge the predictive performance of traditional time series approaches, including ARIMA and SARIMA models.

To overcome these limitations, the state space model offers a flexible and robust framework for modeling and forecasting interrelated time series data with complex dynamics [5]. This approach is particularly effective for analyzing data with seasonal patterns, trends, and structural changes that arise at unknown points in time. By embedding the SARIMA model within a state space framework, it becomes possible to capture and accommodate non-linearities and abrupt fluctuations more effectively than using SARIMA alone.

Therefore, this study applies the SARIMA model within a state space context to address the complex characteristics of tax revenue data in Sumedang Regency. This study aims to enhance the precision of forecasts while offering data-driven insights to support policy formulation in the fiscal sector. Ultimately, this study contributes not only to the practical domain of regional tax management but also to the broader literature on forecasting complex time series data under structural uncertainty.

II. Data

This study utilized secondary data obtained from the Regency Primary Tax Service Office. The variable analyzed was the tax revenue of the Sumedang Regency Primary Tax Service Office, measured in billions of rupiah. The dataset is a univariate time series consisting ofa total of 139 monthly data points, spanning from January 2012 through July 2023.

III. Methods

A. Seasonal Autoregressive Integrated Moving Average (SARIMA)

SARIMA is a type of time series model that builds upon the ARIMA framework by integrating seasonal components. This model is well-suited for time series that exhibit periodic patterns or recurring fluctuations over a specified seasonality cycle.

The SARIMA formulation can be represented as:

$$\mathrm{ARIMA}(p,d,q)(P,D,Q)^{s} \quad (1)$$

where p, d, and q refer to the non-seasonal AR, differencing, and MA orders; P, D, and Q indicate their seasonal counterparts; s corresponds to the number of periods per season. The general equations of the SARIMA model are as follows:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_Q(B^s)\alpha_t \quad (2)$$

where Y_t denotes the time series value at time t; $\Phi_p(B)$ and $\theta_q(B)$ denote the AR and MA operators for the non-seasonal componen; $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ refer to the corresponding seasonal operators; $(1-B)^d$ and $(1-B^s)^D$ are the differencing operators. The mean of the series, denoted as μ, is subtracted if d=0 or D=0.

B. State Space Model

The state space model represents an innovative approach for modeling and forecasting interrelated time series data, in which variables interact dynamically[5]. It is applicable to both forecasting and parameter estimation tasks [6]. The model comprises two primary equations: the measurement (or observation) equation and the state transition equation, which together describe the relationships among the system's input, output, and internal state variables [7]. The representation of the state space model is formulated as:

Transition Equation:

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1}+\boldsymbol{v}_t \quad (3)$$

Measurement Equation:

$$\boldsymbol{Z}_t = \boldsymbol{H}\boldsymbol{X}_t+ \boldsymbol{w}_t \quad (4)$$

$\boldsymbol{Z}_t$ denotes the observed measurement vector at time t, $\boldsymbol{X}_t$ represents the unobserved state vector, $\boldsymbol{H}$ is the measurement matrix, and $\boldsymbol{A}$ is the transition matrix. The stochastic disturbances $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ are modeled as white noise processes, each having zero mean and associated covariance matrices $\boldsymbol{Q}_t$ and $\boldsymbol{R}_t$, respectively. Moreover, $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are assumed to be uncorrelated to ensure identifiability of the model structure.

C. Stationary Test

Stationarity is a fundamental assumption for numerous time series models. The Augmented Dickey–Fuller (ADF) test is applied to identify the presence of a unit root, indicating non-stationarity in the series. The hypotheses are formulated as:

H_0: δ=0 (data contain a unit root/non-stationary)
H_1: $\delta\neq$0 (data are stationary)

The test statistic is calculated as:

$$t_\delta = \frac{\delta}{SE(\delta)} \quad (5)$$

where δ is the estimated coefficient and SE(δ) is its standard error, approximated by $\sqrt{1-\delta^2/n}$ with n being the sample size. The null hypothesis is rejected when the absolute test statistic surpasses the threshold value $t_{\alpha,n}$ or the p-value falls below the pre-specified significance level α.

D. Box-Cox Transformation

When the time series exhibits non-constant variance (heteroskedasticity), the Box–Cox transformation can be employed as a variance-stabilizing technique. The transformation is given by:

$$T(Z_t) = \begin{cases} (Z_t^\lambda - 1)/\lambda, \lambda = 0 \\ \ln Z_t \quad , \lambda \neq 0 \end{cases} \quad (6)$$

where λ serves as the transformation coefficient, determined by optimizing the log-likelihood function. A time series is considered variance-stationary if the optimal λ is close to 1 and $\boldsymbol{Z}_t$ values are positive.

E. Differencing

Differencing is applied when the time series demonstrates non-stationarity in its mean. The process stabilizes the mean by subtracting the observation of the previous period from the current period. Differencing of order d is defined as:

$$Z_t^{(d)}=(1-B)^d Z_t \quad (7)$$

Where B is the backshift operator and d denotes the differencing order.

F. ACF and PACF

In SARIMA modeling, the autocorrelation function (ACF) helps identify the moving average (MA) components, while the partial autocorrelation function (PACF) assists in detecting the autoregressive (AR) structure. The sample ACF at lag k is computed as [5]:

$$\rho_k = \frac{\sum_{t=1}^{n-k}\left(Z_t-\overline{Z}\right)\left(Z_{t+k}-\overline{Z}\right)}{\sum_{t=1}^{n-k}\left(Z_t-\overline{Z}\right)^2} \tag{9}$$

where Z_t and Z_{t+k} are the observed values at times t and $t+k$, respectively, and $\overline{Z}$ is the sample mean.

The PACF at lag k is calculated using [5]:

$$\phi_{kk} = \frac{\operatorname{cov}\left[\left(Z_t-\overline{Z}_t\right)\left(Z_{t+k}-\overline{Z}_{t+k}\right)\right]}{\sqrt{\operatorname{var}\left(Z_t-\overline{Z}_t\right)\operatorname{var}\left(Z_{t+k}-\overline{Z}_{t+k}\right)}} \tag{10}$$

G. SARIMA Model Parameter Estimation and Testing

The SARIMA model parameters ϕ, θ, Φ, and Θ are estimated using the Maximum Likelihood Method (MLM). The residuals a_t assumed to follow a white noise process characterized by a distribution (0, σ_a^2), the log-likelihood function is expressed as:

$$\left(a\,|\,\phi,\theta,\Phi,\Theta,\sigma_a^2\right) = -\frac{n}{2}\ln\left(2\pi\sigma_a^2\right) - \frac{S\cdot\phi,\theta,\Phi,\Theta,\sigma_a^2}{2\sigma_a^2} \tag{11}$$

The parameters that have been estimated, then tested for significance to find out whether these parameters can be included in the model or not.

The hypotheses are:

H_0: The parameter has no effect (value equals zero).

H_1: The parameter is significant (value differs from zero).

The following test statistic is applied:

$$t_p = \frac{\beta}{SE\left(\beta\right)} \tag{12}$$

where β is the estimated parameter value and $SE(\beta)$ is its standard error. The null hypothesis H_0 is rejected if $|t_p| \geq t_{\alpha/2,df}$ or if the p-value is less than the significance level α, with degrees of freedom $df=n-k$.

H. SARIMA Model Diagnostic Checking

The diagnostic checking of the SARIMA model involves two essential conditions: the residuals must behave as white noise and follow a normal distribution [5]. White noise implies that the residuals are uncorrelated (i.e., non-autocorrelated) and exhibit constant variance (homoskedasticity).

1) Residual Normality Test

In this study, the Kolmogorov–Smirnov (K–S) test was utilized to assess whether the model residuals follow a normal distribution. This method is appropriate for large samples, specifically those with more than 30 observations.

The hypotheses are:

H_0: The residuals follow a normal distribution.

H_1: The residuals deviate from normality.

The test statistic is:

$$D = \max\left|F_0\left(\alpha_t\right) - S_n\left(\alpha_t\right)\right| \tag{13}$$

where $F_0\,(\alpha_t)$ is the cumulative distribution function under H_0 and $S_n\,(\alpha_t)$ is the empirical cumulative distribution function. The null hypothesis is rejected if $D \geq D_{(1-\alpha;n)}$ or if the p-value is less than α.

2) Non-Autocorrelation Test

The Q-Ljung Box test is used to verify that residuals are not autocorrelated. The hypotheses are:

H_0: Residuals are not autocorrelated.

H_1: Residuals are autocorrelated.

The test statistic is:

$$Q = n\left(n+2\right)\sum_{k=1}^{K}\frac{\rho_k^2}{\left(n-k\right)} \tag{14}$$

where K is the maximum lag and ρ_k^2 is the sample autocorrelation at lag k. The null hypothesis is rejected if $Q \geq \chi^2_{\alpha,K-p-q}$ or if the p-value is less than α.

3) Homoskedasticity Test

Homokedasticity means the residuals have constant variance. The squared residuals are tested using a similar Q-Ljung Box statistic applied to α_t^2. The hypotheses are:

H_0: Residual variance is homogeneous.

H_1: Residual variance is heterogeneous.

The test statistic is identical to (14), but applied to squared residuals. The null hypothesis is rejected if $Q \geq \chi^2_{\alpha,K-p-q}$ or if the p-value is less than α.

I. Selection of SARIMA Models

The best model based on the in-sample is selected based on the Akaike Information Criterion (AIC), computed as:

$$AIC(M) = n\ln\left(\sigma_\alpha^2\right) + 2M \tag{15}$$

where σ_α^2 denotes the residual variance estimated through maximum likelihood, M indicates the number of model parameters, and n represents the total number of observations [5]. The model with the smallest AIC value is considered the optimal model.

J. ARIMA Representation in State Space Form

The conventional ARIMA model can be reformulated into a state space representation by expressing the autoregressive process through state and observation equations. This representation provides a structured framework for modeling dynamic systems [8]. Based on Equations (3) and (4), the ARIMA model in state space form, with $r=\max(p,q+1)$, is defined as follows [8]:

State transition equation:

$$\mathbf{X}_t = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{r-1} & \phi_r \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}\mathbf{X}_{t-1} + \begin{bmatrix} \alpha_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{16}$$

Measurement/observation equation:

$$\mathbf{Z}_t = \begin{bmatrix} 1 & \theta_1 & \theta_2 & \cdots & \theta_{r-1} \end{bmatrix}\mathbf{X}_t + \mathbf{w}_t \tag{17}$$

K. Kalman Filter

The Kalman Filter is applied in the SARIMA model to allow time-varying parameters, thereby enhancing the model's adaptability to changes in the system's dynamics and improving the forecasting accuracy. The filter algorithm comprises the following steps:

Initialization stage

The process begins by setting the initial estimate of the state vector and the initial error covariance matrix:

$$P_0 = P_{x_0}, x_0 = x_0 \tag{18}$$

Time Update (Prediction Step)

This stage estimates the current state using the previous time step's information. The predicted state and its error covariance are calculated as follows:

$$\bar{\mathbf{x}}_t = \mathbf{A}\mathbf{x}_{t-1} \tag{19}$$

$$\mathbf{P}_t^- = \mathbf{A}\mathbf{P}_{t-1}\mathbf{A}^T + \mathbf{Q} \tag{20}$$

In Equation (19), $\bar{\mathbf{x}}_t$ is the predicted (a priori) state vector at time t, $\boldsymbol{A}$ is the state transition matrix, $\mathbf{x}_{t-1}$ is the estimated (a posteriori) state vector from the previous time step. Equation (20) provides the predicted error covariance matrix $\mathbf{P}_t^-$, where $\mathbf{P}_{t-1}$ is the error covariance at the previous time, $\mathbf{A}^{\mathrm{T}}$ denotes the transpose of matrix $\mathbf{A}$, and $\mathbf{Q}$ is the covariance matrix of the process noise.

Measurement Update (Correction Stage)

The correction stage improves the prediction by incorporating the current measurement through a measurement model. The estimation error is minimized using the Kalman Gain matrix [9].

$$\mathbf{K}_t = (\mathbf{P}_t^- \mathbf{H}^t (\mathbf{H}\mathbf{P}_t^- \mathbf{H}^t + \mathbf{R}))^{-1} \tag{21}$$

$$\mathbf{x}_t = \bar{\mathbf{x}}_t + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}\bar{\mathbf{x}}_t) \tag{22}$$

$$\mathbf{P}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{H})\mathbf{P}_t^- \tag{23}$$

In Equation (21), $\mathbf{K}_t$ refers to the Kalman Gain at time t, $\mathbf{H}$ represents the observation matrix, and $\mathbf{R}$ denotes the covariance matrix of the measurement noise. Equation (22) computes the updated state estimate $\mathbf{x}_t$, where $\mathbf{Z}_t$ corresponds to the observed value at time t, and $\bar{\mathbf{x}}_t$ is the predicted state. Finally, Equation (23) updates the error covariance matrix $\mathbf{P}_t$, where $\mathbf{I}$ denotes the identity matrix.

L. Model Evaluation

In this study, MAPE (Mean Absolute Percentage Error) is employed to evaluate the forecasting performance of the model. It measures the average magnitude of prediction errors in percentage terms and is defined as [10]:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100\% \tag{24}$$

A lower MAPE value indicates better forecasting accuracy. Based on the classification proposed in the literature [11], forecasting accuracy can be categorized as follows:

TABLE I
FORECASTING ACCURACY CATEGORIES BASED ON MAPE

MAPE Values	Accuracy of Forecasting
<10%	High accuracy
10% - 20%	Good accuracy
20% - 50%	Fair accuracy
0%	Poor accuracy

IV. RESULTS

A. Data Exploration

The monthly tax revenue data collected by the Sumedang Regency Primary Tax Service Office from January 2012 to July 2023 exhibit both seasonal and trending behavior. Recurrent peaks are typically observed in December, likely due to the tendency of taxpayers to settle their obligations before the fiscal year ends. In contrast, lower revenues tend to occur during the early months of the year. A temporary decline is also noticeable during the second half of 2019, which can be attributed to the economic impact of the COVID-19 pandemic and the accompanying mobility restrictions.

Descriptive statistics of the observed tax revenue are summarized in Table 1. The average monthly revenue was approximately IDR 44.33 billion, with a median of IDR 42.44 billion. The values range from a minimum of IDR 11.29 billion to a third quartile of IDR 52.25 billion, indicating a slightly right-skewed distribution. This pattern reflects the presence of seasonal peaks, particularly during year-end periods, which justify the application of variance-stabilizing transformations during model development.

TABLE II
SUMMARY STATISTICS OF MONTHLY TAX REVENUE (IN IDR BILLION)

Min	Median	Max	Mean	1st Quartile	3rd Quartile
11.29	42.44	147.93	44.33	28.74	52.25

Figure 1 displays the original time series of monthly tax revenue, while Figure 2 illustrates its decomposition into trend, seasonal, and residual components. The decomposition reveals strong seasonal effects and a clear upward trend, reinforcing the need for a modeling framework that can adequately capture such dynamics.

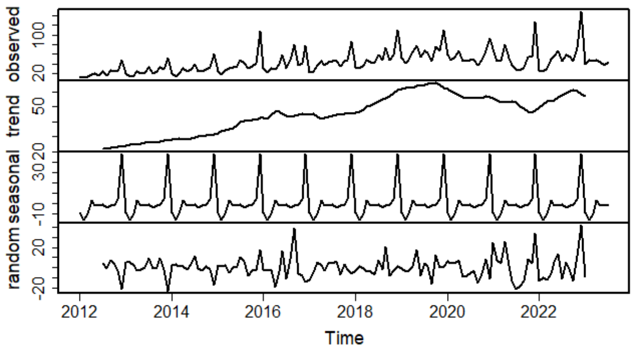


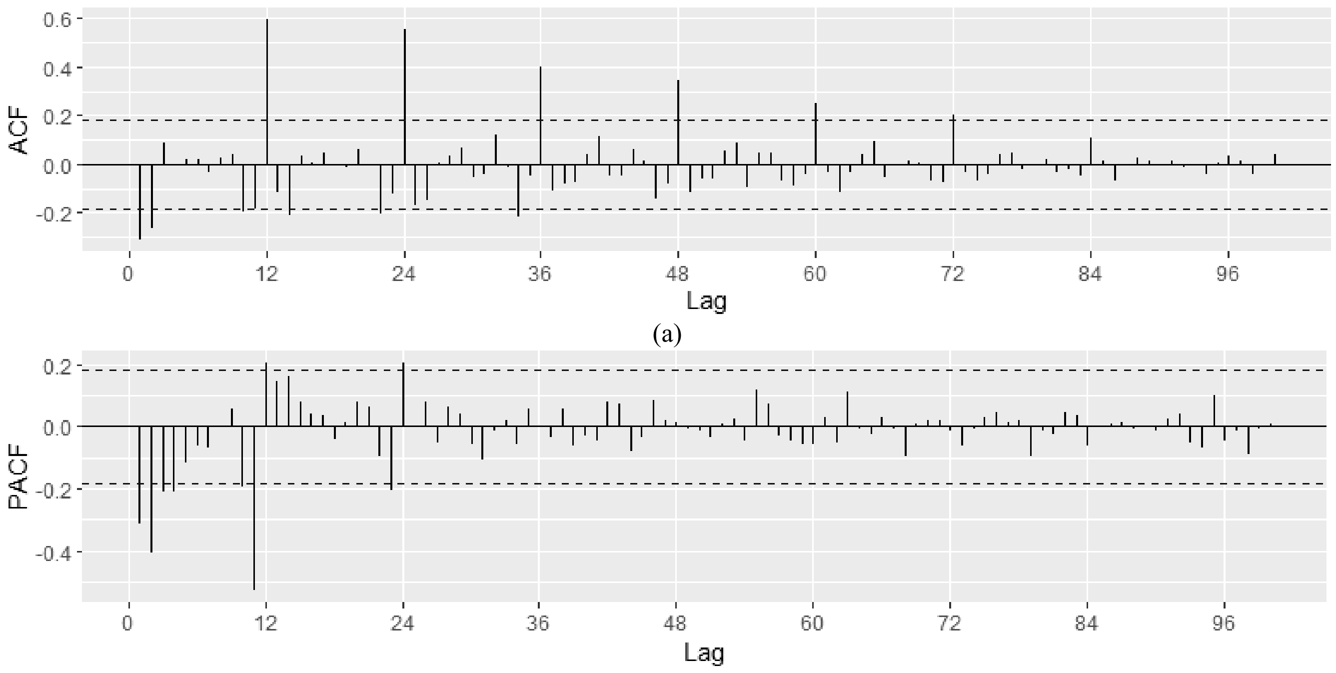
Fig. 2. Seasonal-trend decomposition of monthly tax revenue collected by the Sumedang Regency Primary Tax Service Office.

B. SARIMA Modeling

This section outlines the modeling process using the Seasonal Autoregressive Integrated Moving Average (SARIMA) framework to forecast monthly tax revenue. The procedure includes testing for stationarity, identifying model structure, estimating parameters, and conducting residual diagnostics to ensure the model's validity and forecasting suitability.

1) Stationarity Checking

To assess variance stationarity, the Box–Cox transformation was applied. The initial lambda (λ) value was



(a)

(b)

Fig. 3. Autocorrelation and partial autocorrelation plots of the stationary tax revenue series: (a) ACF (b) PACF

0.1977, indicating non-stationary variance. After transformation, the λ value increased to 0.9937, suggesting that the data became variance-stationary.

Mean stationarity was evaluated using the ADF test. The non-seasonal component returned a p-value of 0.001473, which is below the 0.05 threshold, confirming that the mean is stationary. However, the seasonal component produced a p-value of 0.99, indicating the need for seasonal differencing.

Following the application of seasonal differencing (D=1), the ADF test conducted at lag 12 yielded a p-value of 0.0363, suggesting that the seasonal component had become stationary. To account for remaining trend elements, a first-order non-seasonal differencing (d=1) was also applied. The final ADF test, evaluated at both lag 1 and lag 12, returned a p-value of 0.01, confirming that the time series is fully stationary in both the seasonal and non-seasonal dimensions.

1) Model Identification

To determine the appropriate SARIMA model order, the ACF and PACF plots of the stationary series were analyzed. The PACF, shown in Figure 3(b), displays a cut-off at lag 4, indicating the presence of non-seasonal autoregressive components up to AR(4). Spikes at lags 12 and 24 also suggest possible seasonal AR terms, including SAR(1) and SAR(2).

Meanwhile, the ACF plot in Figure 3(a) exhibits a cut-off at lag 2, which implies the inclusion of non-seasonal moving average components such as MA(1) or MA(2). A distinct spike at lag 12 indicates the need for a seasonal MA term, potentially SMA(1).

2) Estimation and Diagnostics

Based on the identified differencing orders (d=1 dan D=1) and a seasonal period of 12, a total of 62 SARIMA model specifications were estimated. From these, six models were found to have statistically significant parameters based on standard t-tests. These candidate models include various combinations of non-seasonal AR and MA terms, as well as seasonal AR and MA components.

To determine the adequacy of each candidate model, three diagnostic tests were conducted: the Kolmogorov–Smirnov test for residual normality, the Q–Ljung Box test for autocorrelation, and the Breusch–Pagan test for homoskedasticity. A model was considered statistically acceptable if it passed all three diagnostic checks. The complete parameter estimates and their significance levels are summarized in Table III, while the diagnostic test results are reported in Table IV.

TABLE III
PARAMETER ESTIMATES AND SIGNIFICANCE TESTS FOR SELECTED SARIMA MODELS

Model	Parameter	Estimate	p-value
ARIMA(1,1,0)(0,1,1)^{12}	ϕ_1	-0.4375	1×10^{-6}
	Θ_1	-0.7418	2.25×10^{-8}
ARIMA(2,1,0)(0,1,1)^{12}	ϕ_1	-0.6482	7.34×10^{-13}
	ϕ_2	-0.4488	6.43×10^{-7}
	Θ_1	-0.6735	7.78×10^{-10}
ARIMA(1,1,0)(1,1,0)^{12}	ϕ_1	-0.4373	$9.57e^{-7}$
	Φ_1	-0.4940	1.74×10^{-9}
ARIMA(2,1,0)(1,1,0)^{12}	ϕ_1	-0.6481	2.08×10^{-13}
	ϕ_2	-0.4701	8.6×10^{-8}
	Φ_1	-0.5185	2.03×10^{-10}
ARIMA(0,1,1)(0,1,1)^{12}	θ_1	-0.8118	2.08×10^{-36}
	Θ_1	-0.5944	3.24×10^{-7}
ARIMA(0,1,1)(1,1,0)^{12}	θ_1	-0.7999	9.95×10^{-38}
	Φ_1	-0.4552	1.28×10^{-7}

TABLE IV
SUMMARY OF DIAGNOSTIC TESTS FOR SARIMA CANDIDATE MODELS

Model	Residual Normality (p-value)	Residual Non-Autocorrelation (p-value)	Residual Homoskedasticity (p-value)
ARIMA(1,1,0)(0,1,1)^{12}	0.7773	0.0365	1.006×10^{-5}
ARIMA(2,1,0)(0,1,1)^{12}	0.0297	0.8601	0.109
ARIMA(1,1,0)(1,1,0)^{12}	0.3616	0.0243	0.0004
ARIMA(2,1,0)(1,1,0)^{12}	0.5584	0.7046	0.5675
ARIMA(0,1,1)(0,1,1)^{12}	0.5584	0.9853	0.1789
ARIMA(0,1,1)(1,1,0)^{12}	0.162	0.8597	0.6706

The Kolmogorov–Smirnov test indicated that all models satisfied the residual normality assumption, with the exception of ARIMA(2,1,0)(1,1,0)^{12}. The Q–Ljung Box test revealed residual independence for all models except ARIMA(0,1,1)(0,1,1)^{12} and ARIMA(1,1,0)(1,1,0)^{12}. Furthermore, the homoskedasticity test indicated that only ARIMA(1,1,0)(0,1,1)^{12} and ARIMA(1,1,0)(1,1,0)^{12} failed to meet the constant variance assumption.

3) SARIMA Model Selection

Based on the diagnostic tests above, three models were deemed statistically appropriate for forecasting: ARIMA(2,1,0)(1,1,0)^{12}, ARIMA(0,1,1)(0,1,1)^{12}, and ARIMA(0,1,1)(1,1,0)^{12}. To determine the best-fitting model, Akaike Information Criterion (AIC) values were compared, as presented in Table V.

TABLE V
AIC VALUES OF SELECTED SARIMA MODELS

Model	AIC Value
ARIMA(2,1,0)(1,1,0)^{12}	160.7793
ARIMA(0,1,1)(0,1,1)^{12}	154.6360
ARIMA(0,1,1)(1,1,0)^{12}	157.4833

Among the three models, ARIMA(0,1,1)(0,1,1)^{12} yielded the lowest AIC value and was therefore selected as the most suitable model for forecasting monthly tax revenue.

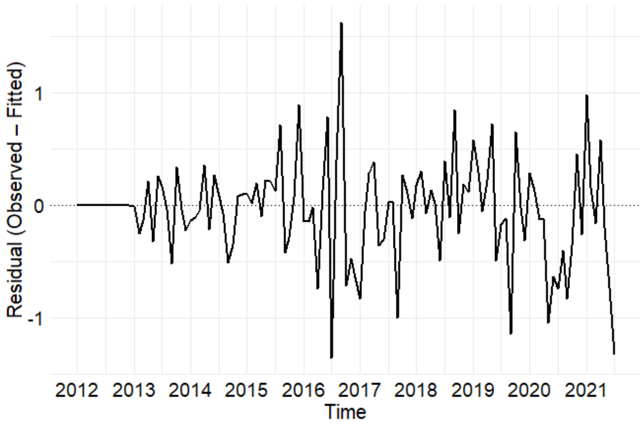


Fig. 4. Residual series of the selected SARIMA(0,1,1)(0,1,1)^{12} model from January 2012 to July 2023.

To further validate the adequacy of this model, the residual series was plotted over the training period (see Figure 4). The residuals oscillate around zero without any discernible trend or seasonal structure, suggesting that the model effectively captures both short- and long-term dynamics in the data. Although minor fluctuations occur, particularly during mid-2016 and late 2019, the variance remains relatively stable throughout the series. These visual patterns are consistent with the diagnostic test results reported earlier, supporting the suitability of the selected SARIMA model for forecasting purposes.

The selected SARIMA model can be expressed by the following equations:

$$(1-B)(1-B^{12})Y_t = \theta_1(B)\Theta_1(B^{12})a_t$$

$$(1-B-B^{12}+B^{13})\, Y_t = (1-\theta_1 B)(1-\Theta_1 B^{12})a_t$$

$$Y_t - Y_{t-1} - Y_{t-12} - Y_{t-13} = a_t - \theta_1 a_{t-1} - \Theta_1 a_{t-12} - \Theta_1 a_{t-13}$$

$$Y_t = Y_{t-1} + Y_{t-12} + Y_{t-13} + a_t - \theta_1 a_{t-1} - \Theta_1 a_{t-12} - \Theta_1 a_{t-13} \quad (25)$$

Based on the parameter estimates presented in Table III, the final forecasting equation becomes:

$$Y_t = Y_{t-1} + Y_{t-12} + Y_{t-13} + a_t - 0.8181a_{t-1} - 0.5944a_{t-12} - 0.4825a_{t-13} \quad (26)$$

C. SARIMA State Space

The optimal SARIMA model selected in the previous section is reformulated into its state space representation based on Equations (17) and (18). This formulation consists of a system of equations describing the transition of latent states over time and their relation to the observed data. The state space representation comprises the following equations:

Transition Equation:

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{v}_t$$

$$\boldsymbol{X}_t = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \boldsymbol{X}_{t-1} + \begin{bmatrix} -a_t \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -a_t \\ a_t \end{bmatrix} \quad (27)$$

Observation Equation:

$$\boldsymbol{Z}_t = \boldsymbol{H}\boldsymbol{X}_t + \boldsymbol{w}_t$$

$$\boldsymbol{Z}_t = [\,1\ \theta_1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ \Theta_1\,]\ \boldsymbol{X}_t + \boldsymbol{w}_t \quad (28)$$

This state space structure forms the basis for recursive forecasting using the Kalman filter. Before parameter estimation, the model requires initialization, including the specification of the initial state vector and the initial error covariance matrix. In this study, the estimation process focuses on the moving average parameters from the ARIMA(0,1,1)(0,1,1)^{12} model:

MA(1) lag 1 parameter : $\theta_1 = -0.812$
SMA(1) lag 12 parameter : $\Theta_1 = -0.737$

D. Model Performance Evaluation and Comparative Analysis

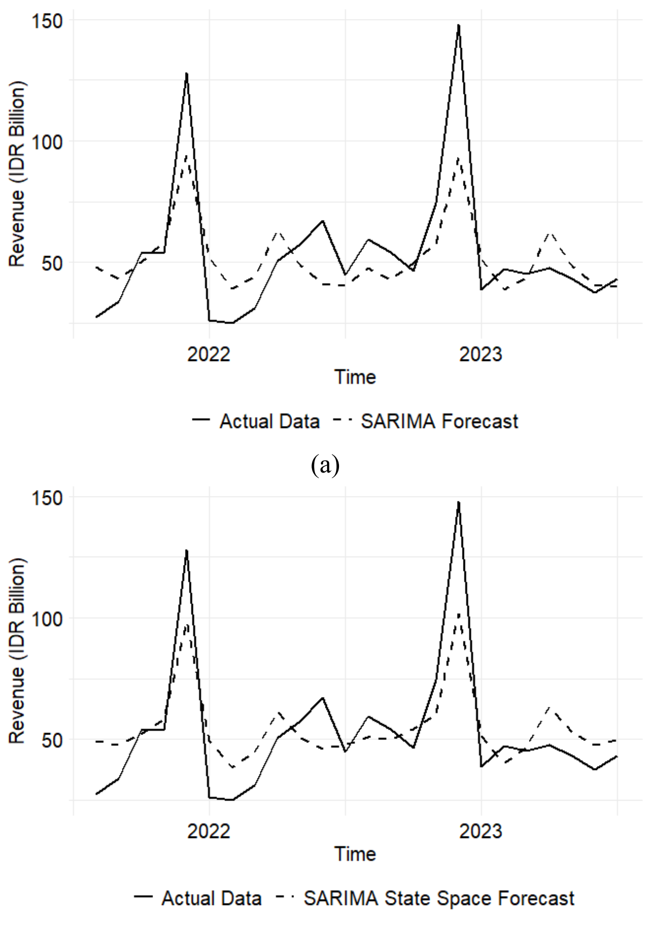
The predictive performance of both the classical SARIMA model and its state space reformulation was evaluated on the test dataset. Several commonly used accuracy metrics were calculated and are summarized in Table VI.

TABLE VI
FORECAST ACCURACY METRICS ON TEST DATA

Metric	SARIMA	SARIMA State Space
MAPE	27.49%	27.27%
RMSE	18.020	16.216
MAE	13.599	12.853
Theil's U	0.6697	0.6368

The SARIMA State Space model demonstrates superior performance across most error metrics. In particular, it yields lower RMSE, MAE, and Theil's U values, indicating improved accuracy and a better ability to track data fluctuations. Meanwhile, the MAPE values for both models are nearly identical, suggesting that both models exhibit comparable relative forecasting error in percentage terms.

A visual comparison of the predicted values and actual observations is provided in Figure 5.



(a)

(b)

Fig. 5. Forecast Comparison on the Test Dataset: (a) SARIMA model (b) SARIMA State Space model.

Figure 5 shows that both models successfully replicate the overall patterns in the observed data. While the differences between the two forecasts are not pronounced, the SARIMA State Space model exhibits a marginally better fit, particularly in capturing short-term variations. This result aligns with the quantitative evaluation, confirming the advantage of incorporating the state space approach in forecasting.

In addition to point forecast comparisons, the absolute forecast errors of both models were plotted to assess their temporal performance more closely. Figure 6 displays the evolution of one-step-ahead absolute forecast errors for each model during the test period from July 2021 to July 2023.

The SARIMA State Space model generally exhibits smaller forecast errors across most periods, particularly during high-variance months such as December 2022 and January 2023. This indicates better adaptability in capturing sudden fluctuations in tax revenue, likely due to its recursive filtering structure.

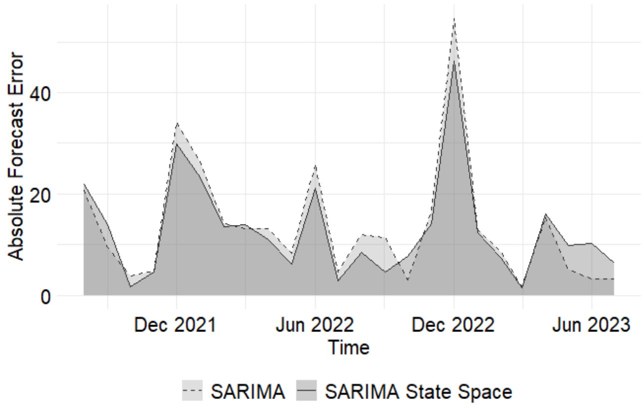


Fig. 6. Absolute forecast errors of the SARIMA and SARIMA State Space models over the test period (July 2021–July 2023).

To statistically assess the difference in forecast accuracy, a Diebold–Mariano (DM) test was conducted using the squared error loss function. The test yielded a statistic of 1.5783 with a p-value of 0.1282, indicating that the difference in predictive performance between the two models is not statistically significant at the 5% level. Nonetheless, the visual patterns and consistent reduction in error magnitude observed in Figure 6 suggest practical advantages offered by the state space formulation.

E. Forecasting

Based on the model evaluation in the previous section, the SARIMA State Space model was selected as the final forecasting model for monthly tax revenue at the Sumedang Regency Primary Tax Service Office. The forecasts cover the period from August 2023 to December 2025. The results are summarized in Table VII, with values expressed in billion Rupiah.

TABLE VII
FORECASTED MONTHLY TAX REVENUE FOR THE SUMEDANG REGENCY PRIMARY TAX SERVICE OFFICE (IN IDR BILLION)

Month	2023	2024	2025
January	-	40.36461	42.28570
February	-	35.02723	36.74261
March	-	40.92048	42.86267
April	-	53.48575	55.89110
May	-	49.11951	51.36668
June	-	47.03937	49.21021
July	-	48.80438	51.04003
August	44.47180	44.47180	46.54746
September	46.48206	46.48206	48.63233
October	50.48502	50.48502	52.78196
November	57.17818	59.71529	62.34193
December	106.91922	111.10273	115.41656

Figure 7 presents a time series plot of the historical and forecasted monthly revenue, including the 95% confidence intervals. The projected pattern indicates recurring seasonal peaks in December, in line with previous observations. Additionally, the overall trajectory reflects a consistent increasing trend throughout the forecast horizon.

This forecasted growth is in line with the recent initiatives introduced by the Sumedang Regency Primary Tax Service Office, such as taxpayer education, compliance monitoring, and administrative restructuring. The anticipated operational improvements, including the construction of a new office facility, are expected to contribute positively to tax revenue performance in the upcoming years.

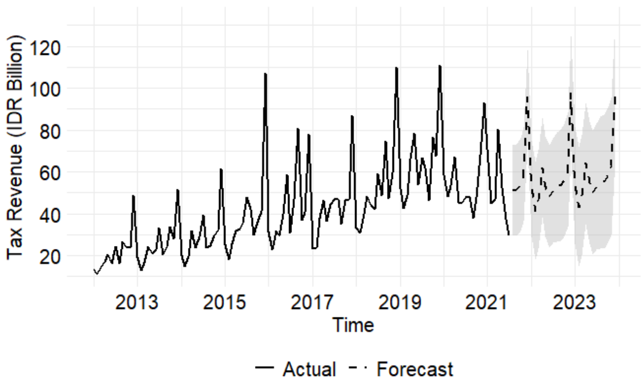


Fig. 7. Forecasted Tax Revenue for for the period August 2023 to December 2025.

V. Discussion

The empirical findings of this study offer important implications for understanding and forecasting regional tax revenue, particularly within the context of the Regency Primary Tax Service Office. The identification of SARIMA(0,1,1)(0,1,1)^{12} as the optimal baseline model confirms the strong presence of seasonal patterns and non-stationary trends, which are typical characteristics in fiscal time series data. This finding is consistent with previous studies that observed seasonal peaks in tax revenue during the third and fourth quarters, coinciding with periods of heightened economic activity such as increased consumption and investment [12].

Transforming the SARIMA model into a state space framework not only preserves its temporal structure but also enhances its flexibility for handling latent components and enabling recursive estimation. While the reduction in MAPE from 27.49% to 27.27% may appear limited, the improvements in RMSE, MAE, and Theil's U demonstrate the practical advantage of the state space model in generating more stable forecasts. These gains are especially relevant for fiscal data, which often exhibit reporting lags, administrative noise, and seasonal disbursement schedules. State space models have been effectively utilized in macroeconomic forecasting to handle variables affected by policy changes and administrative irregularities [13].

In contrast to traditional models used in earlier literature, such as exponential smoothing or basic ARIMA models, the SARIMA State Space model allows for a more robust decomposition of signal and noise components. It also supports extensions such as intervention modeling, time-varying parameters, or multivariate specifications. Such adaptability is crucial for fiscal forecasting under macroeconomic uncertainty, as multivariate state space models have been shown to better capture complex interactions among economic variables [14].

However, several limitations should be acknowledged. The dataset used spans only eleven years and pertains to a single local tax office, limiting the generalizability of the findings. Moreover, the assumption of normally distributed and linearly related residuals may not fully reflect real-world tax collection dynamics, especially during periods of disruption or regulatory transition. In such cases, the application of non-Gaussian models has been suggested to better handle volatility and irregular patterns in fiscal data [15].

Despite these limitations, the results illustrate the value of SARIMA State Space models as analytical tools for regional fiscal management. Future research could explore spatial hierarchies, cross-sectional dependencies, or the integration of real-time compliance indicators to extend the model's applicability.

VI. Conclusion

This study confirms that incorporating a SARIMA model within a state space framework enhances the accuracy and robustness of regional tax revenue forecasting. The model effectively captures the observed seasonality, trend, and irregular variation in the monthly tax revenue data, offering a statistically coherent and operationally feasible approach for local fiscal planning.

Although the improvement in MAPE is relatively minor, the consistent reductions in RMSE, MAE, and Theil's U indicate that the state space specification provides more stable and reliable forecasts. This is particularly important for local governments that rely on timely and accurate revenue projections to design effective budgetary policies.

In addition to its forecasting capability, the state space approach provides a flexible foundation for incorporating external covariates and policy interventions, making it a promising avenue for future research and a valuable tool for public financial management. Therefore, this study not only contributes to the growing literature on time series modeling in the public sector but also underscores the importance of methodological innovation in achieving more effective revenue governance.

Acknowledgment

The authors gratefully acknowledge the Primary Tax Office (Kantor Pelayanan Pajak Pratama) Sumedang for providing the data used in this study. The views expressed in this paper are solely those of the authors and do not necessarily reflect the position or policies of the Primary Tax Office Sumedang.

References

[1] S. M. Fatem, S. A. Awang, S. Pudyatmoko, M. A. K. Sahide, A. A. Pratama, and A. Maryudi, "Camouflaging economic development agendas with forest conservation narratives: A strategy of lower governments for gaining authority in the re-centralising Indonesia," Land use policy, vol. 78, pp. 699–710, Nov. 2018, doi: 10.1016/J.LANDUSEPOL.2018.07.018.

[2] A. Halim and M. S. Kusufi, Teori, Konsep, dan Aplikasi Akuntansi Sektor Publik. Jakarta: Salemba Empat, 2014.

[3] A. Sipos, "Determining Factors of Tax-morale with Special Emphasis on the Tax Revenues of Local Self-governments," Procedia Econ. Financ., vol. 30, pp. 758–767, Jan. 2015, doi: 10.1016/S2212-5671(15)01325-8.

[4] L. Li-Xia, Z. Yi-Qi, and X. Y. Liu, "Tax forecasting theory and model based on SVM optimized by PSO," Expert Syst. Appl., vol. 38, no. 1, pp. 116–120, Jan. 2011, doi: 10.1016/J.ESWA.2010.06.022.

[5] W. W. S. Wei, “Time Series Analysis Univariate and Multivariate Methods (2nd Edition),” New introduction to Multiple Time Series Analysis. 2006.
[6] J. Durbin and S. J. Koopman, “Time Series Analysis by State Space Methods: Second Edition,” Oxford Stat. Sci. Ser., 2012.
[7] A. V Balakrishnan, Kalman Filtering Theory. in Series in communication and control systems. Optimization Software, Publications Division, 1987. [Online]. Available: https://books.google.co.id/books?id=jFsnAQAAIAAJ
[8] J. D. Hamilton, Time Series Analysis. in Book collections on Project MUSE. Princeton University Press, 1994. [Online]. Available: https://books.google.co.id/books?id=B8_1UBmqVUoC
[9] G. Welch and G. Bishop, “An Introduction to the Kalman Filter,” Proc. Siggraph Course, vol. 8, 2006.
[10] D. C. Montgomery, C. L. Jennings, and M. Kulahci, Introduction to Time Series Analysis and Forecasting. in Wiley Series in Probability and Statistics. Wiley, 2015. [Online]. Available: https://books.google.co.id/books?id=Xeh8CAAAQBAJ
[11] Colin D. Lewis, “Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting,” Butterworth Sci., no. June 1981, 1982.
[12] R. Handoko, "Seasonal Patterns in Indonesian Income Tax Receipts and Their Economic Linkages," Kajian Ekonomi Keuangan, vol. 3, no. 2, pp. 147-165, 2019.
[13] C. Allen and R. Pasupathy, "State Space Models for Macroeconomic Forecasting," Federal Reserve Bank of St. Louis Working Paper Series, 1997.
[14] M. West, "Bayesian Multivariate Dynamic Models," arXiv preprint arXiv:1911.09656, 2019.
[15] L. Elabed and M. Masmoudi, "Modeling Financial Time Series with Non-Gaussian State Space Models," Open Journal of Statistics, vol. 4, pp. 759-766, 2014.

Ayu Indriani received her B.S. degree in Statistics from the Department of Statistics, Universitas Padjadjaran, and is currently pursuing her Master’s degree in Applied Statistics at the same department.

I Gede Nyoman Mindra Jaya received his B.S. degree (S.Si) in Statistics from the Department of Statistics, Universitas Padjadjaran, in 2003, his Master of Science degree from Institut Pertanian Bogor in 2009, and his Doctor of Philosophy degree from the Faculty of Spatial Sciences, University of Groningen, Netherlands, in 2022. His research focuses on spatial and Bayesian methodologies. He is currently the Head of the Master’s Program in Applied Statistics at Universitas Padjadjaran.

Budi Nurani Ruchjana received her Doctoranda degree in Mathematics from Universitas Padjadjaran in 1987, her Master of Science degree from Institut Pertanian Bogor in 1992, and her Doctor of Philosophy degree from Institut Teknologi Bandung in 2002. She is a Professor of Statistics at the Department of Mathematics, Universitas Padjadjaran, and the Head of the Indonesian Ethnoscience Studies Center at the Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran.

Anna Chadidjah received her Doctoranda degree in Actuarial Science from Universitas Padjadjaran in 1985 and her Master of Engineering degree from Institut Teknologi Bandung in 2002. She is currently the Head of the Actuarial Science Undergraduate Program at Universitas Padjadjaran.