Comparison of ARIMA, SSA, and ARIMA – SSA Hybrid Model Performance in Indonesian Economic Growth Forecasting

Action Area C. Integrated statistics for integrated analysis (SC1)

Methodological approaches to integrated analysis:
Use of sound methodologies

Presenter:
Muhammad Fajar
Statistics Indonesia
BACKGROUND

- The development of forecasting methods is increasingly rapid and complex as advances in the development of computing technology

- Using Hybrid Forecasting
METODOLOGY

• Data Source

The data used in this research is economic growth (quarter to quarter, q to q) 1983 Q2 (quarter 2) – 2018 Q2 taken from Badan Pusat Statistik-Statistics Indonesia (BPS). The data for testing is divided into 20% observations (28 forecast ahead), 10% observations (14 forecast ahead), 5% observations (7 forecast ahead), and 3% observations (4 forecast ahead).
• ARIMA-SSA Hybrid

ARIMA – SSA hybrid method is a combination of ARIMA and Singular Spectrum Analysis (SSA) method. Time series data is assumed to consist of linear and nonlinear components, thus could be represented as:

\[ x_t = P_t + N_t \]

with \( P_t \) is a linear component and \( N_t \) is a nonlinear component. ARIMA is used to forecast on linear component, then the residual from the linear component is the nonlinear component. Then, SSA is used to forecast the nonlinear component.

\[ \hat{x}_{T+h} = \hat{P}_{T+h} + \hat{N}_{T+h} \]

with \( \hat{x}_{T+h} \) is the \( x \) forecasting result on the \( T + h \) period, \( \hat{P}_{T+h} \) is the \( P \) forecasting result on the \( T + h \) period, \( \hat{N}_{T+h} \) \( N \) forecasting result on the \( T + h \) period, and \( h \) is the ahead period.
METODOLOGY

• **ARIMA (Autoregressive-Moving Average)**

In general, ARIMA \((p, d, q)(P, D, Q)^S\) model for \(x_t\) time series is:

\[
\Phi_P B^S \phi_p(B) (1 - B)^d (1 - B^S)^D x_t = \theta_q(B) \Theta_Q(B^S) \epsilon_t
\]

- \(B\): lag operator.
- \(p, q\): nonseasonal autoregressive order and nonseasonal moving average order.
- \(P, Q\): seasonal autoregressive order and seasonal moving average order.
- \(d\): nonseasonal differencing order.
- \(D\): seasonal differencing order.
- \(S\): seasonal period, for monthly data \((S = 12)\), quarter data \((S = 4)\).
- \(\phi_p(B)\): nonseasonal autoregressive component.
- \(\Phi_P B^S\): seasonal autoregressive component.
- \(\theta_q(B)\): nonseasonal moving average component.
- \(\Theta_Q(B^S)\): seasonal moving average component.
- \((1 - B)^d\): nonseasonal differencing.
- \((1 - B^S)^D\): seasonal differencing.
- \(\epsilon_t\): error term.
METODOLOGY

• Singular Spectrum Analysis (SSA)

Step 1. Embedding

Given a $x_1, x_2, \ldots, x_T$ time series, choose an even number $L$, where $L$ parameter is the window length defined as $2 < L < T/2$, and $K = T - L + 1$.

The cross matrix is:

$$ X = (X_1, \ldots, X_T) = \begin{pmatrix} x_1 & x_2 & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & \cdots & x_T \end{pmatrix} $$

The cross matrix proves to be a Hankel matrix, which means every element in the main anti diagonal has the same value. Thus, $X$ could be assumed as multivariate data with $L$ characteristic and $K$ observations so that the covariance matrix is $S = XX'$ with dimension of $L \times L$. 
METODOLOGY

Step 2. Singular Value Decomposition (SVD)

Suppose that $S$ has eigen value and eigen vector $\lambda_i$ and $U_i$, respectively. Where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_L$ and $U_1, \ldots, U_L$. Thus, obtained SVD from $X$ as follows:

$$X = E_1 + E_2 + \cdots + E_d$$

where $E_i = \sqrt{\lambda_i} U_i V_i'$, $i = 1,2,\ldots,d$, $E_i$ is the main component, $d$ is the number of eigen value $\lambda_i$, and $V_i = X' U_i / \sqrt{\lambda_i}$. 

#apstatsweek2020
**METODOLOGY**

- **Step 3. Grouping**

In this step, \( X \) is additively grouped into subgroups based on patterns that form a time series, they are trend, periodic, quasi-periodic, and noise component. Partition the index set \( \{1, 2, \ldots, d\} \) into several groups \( I_1, I_2, \ldots, I_n \), then correspond \( X_I \) matrix into group \( I = \{i_1, i_2, \ldots, i_b\} \) which is defined as:

\[
X_I = E_{i_1} + E_{i_2} + \cdots + E_{i_b}
\]  

Thus, the decomposition represents as:

\[
X = X_{I_1} + X_{I_2} + \cdots + X_{I_n}
\]  

with \( X_{I_j} (j = 1, 2, \ldots, n) \) is reconstructed component (RC). \( X_I \) component contribution measured with corresponding eigen value contribution: \( \sum_{i \in I} \lambda_i / \sum_{i=1}^{d} \lambda_i \). Using the close frequency range from the main components is based on the study of grouping process using auto grouping (Alexandrov & Golyandina, 2005). Main components with relatively close frequency ranges are grouped into one reconstructed component. So on, until several reconstructed components are formed.
Step 4. Reconstruction

In this last step, $X_{Ij}$ is transformed into a new time series with $T$ observations obtained from diagonal averaging or Hankelization. Suppose that $Y$ is a matrix with $L \times K$ dimensions and has $y_{ij}, 1 \leq i \leq L, 1 \leq j \leq K$ elements. Then, $L^* = \min(L, K), K^* = \max(L, K)$, and $T = L + K - 1$. Then, $y_{ij}^* = y_{ij}$ if $L < K$ and $y_{ij}^* = y_{ji}$ if $L > K$. $Y$ matrix transferred into $y_1, y_2, ..., y_T$ series with using the following formula:

$$y_k = \begin{cases} 
\frac{1}{k} \sum_{m=1}^{k} y_{m,k-m+1}^*, & 1 \leq k \leq L^* \\
\frac{1}{L} \sum_{m=1}^{L} y_{m,k-m+1}^*, & L^* \leq k \leq K^* \\
\frac{1}{T-K^*+1} \sum_{m=k-K^*+1}^{T} y_{m,k-m+1}^*, & K^* \leq k \leq T 
\end{cases}$$  

(4)

Diagonal averaging on equation (4) is applied to every matrix component $X_{Ij}$ on equation (3) resulting a $\mathbf{X}^{(k)} = (\tilde{x}_1^{(k)}, \tilde{x}_2^{(k)}, ..., \tilde{x}_T^{(k)})$ series. Thus, $x_1, x_2, ..., x_T$ series is decomposed into an addition of reconstructed $m$ series:

$$x_t = \sum_{k=1}^{m} \tilde{x}_t^{(k)}, t = 1, 2, ..., T$$  

(5)
Step 4. Reconstruction

In this last step, \( X_{ij} \) is transformed into a new time series with \( T \) observations obtained from diagonal averaging or Hankelization. Suppose that \( Y \) is a matrix with \( L \times K \) dimensions and has \( y_{ij}, 1 \leq i \leq L, 1 \leq j \leq K \) elements. Then, \( L^* = \min(L, K), K^* = \max(L, K) \), and \( T = L + K - 1 \). Then, 
\[ y_{ij}^* = y_{ij} \text{ if } L < K \text{ and } y_{ij}^* = y_{ji} \text{ if } L > K. \]
\( Y \) matrix transferred into \( y_1, y_2, \ldots, y_T \) series with using the following formula:

\[
y_k = \begin{cases} 
\frac{1}{k} \sum_{m=1}^{k} y_{m,k-m+1}^*, & 1 \leq k \leq L^* \\
\frac{1}{L^*} \sum_{m=1}^{L^*} y_{m,k-m+1}^*, & L^* \leq k \leq K^* \\
\frac{1}{T-K^*+1} \sum_{m=K^*-1}^{T-K^*} y_{m,k-m+1}^*, & K^* \leq k \leq T
\end{cases}
\]  

(4)

Diagonal averaging on equation (4) is applied to every matrix component \( X_{ij} \) on equation (3) resulting a \( \tilde{X}^{(k)} = (\tilde{x}_1^{(k)}, \tilde{x}_2^{(k)}, \ldots, \tilde{x}_T^{(k)}) \) series. Thus, \( x_1, x_2, \ldots, x_T \) series is decomposed into an addition of reconstructed \( m \) series:

\[
x_t = \sum_{k=1}^{m} \tilde{x}_t^{(k)}, \ t = 1,2, \ldots, T
\]  

(5)
SSA Forecasting

SSA forecasting used in this research is SSA recurrent, with estimating min-norm LRR (Linear Recurrence Relationship) coefficient. The LRR coefficient is calculated with the following algorithm:

1. Input: $\mathbf{P} = [P_1: \ldots : P_r]$ matrix, $\mathbf{P}$ is a matrix composed of $U_i$ eigen vector from SVD step. Suppose that $\mathbf{P}$ is a $\mathbf{P}$ that the last row is removed, and $\mathbf{\bar{P}}$ is a $\mathbf{P}$ that the first row is removed.

2. For every $P_i$ vector column from $\mathbf{P}$, calculate $\pi_i$, where $\pi_i$ is a the last component from $P_i$ , and $\mathbf{\bar{P_i}}$ is a $P_i$ that the last component is removed.

3. Calculate: $\nu^2 = \sum_{i=1}^r \pi_i^2$. If $\nu^2 = 1$, then STOP with a warning message “Verticality coefficient equals 1.”
4. Calculate the min-norm LRR coefficient ($\mathcal{R}$):

$$\mathcal{R} = \frac{1}{1 - v^2} \sum_{i=1}^{r} \pi_i P_i$$

5. From point (4) obtained: $\mathcal{R} = (\alpha_{L-1} \ldots \alpha_1)'$.

6. Then, calculate the forecasting value with:

$$\hat{x}_n = \sum_{i=1}^{L-1} \alpha_i \hat{x}_{n-1}, \quad n = T + 1, \ldots, T + h$$
RESULTS

Applied in Indonesian Economic Growth Forecasting (Quarterly)

Table 2.1 RMSE of ARIMA, SSA, and ARIMA-SSA Hybrid Method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Forecast Ahead</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>28</td>
</tr>
<tr>
<td>ARIMA (0,0,0) (1,0,1)^4</td>
<td>1.764</td>
</tr>
<tr>
<td>SSA</td>
<td>2.207</td>
</tr>
<tr>
<td>ARIMA (0,0,0) (1,0,1)^4-SSA hybrid</td>
<td>1.861</td>
</tr>
</tbody>
</table>

Table presents RMSE according to the number of test data used from the observed method. In general, when the test data is smaller, the RMSE from ARIMA (0,0,0) (1,0,1)^4 and ARIMA (0,0,0) (1,0,1)^4 – SSA hybrid is decreasing, whereas the RMSE result of SSA is unstable. ARIMA-SSA hybrid method gives a minimum RMSE compared to the other two methods. This shows that forecasting performance of ARIMA-SSA hybrid method is better than ARIMA and SSA.
THANK YOU

Questions, please send to:
mfajar@bps.go.id
REFERENCES