A Generalized Space-Time Autoregressive Moving Average (GSTARMA) Model for Forecasting Air Pollutant in Surabaya

M S Akbar¹, Setiawan², Suhartono³, B N Ruchjana⁴, D D Prastyo⁵, A Muhaimin⁶, E Setyowati⁷

¹²³５６⁷ Department of Statistics, Institut Teknologi Sepuluh Nopember
Kampus ITS Sukolilo, Surabaya 60111, Indonesia
⁴Department of Mathematics, Universitas Padjadjaran
Jl. Raya Bandung Sumedang KM.21, Hegarmanah, Jatinangor, Sumedang, Jawa Barat 45363, Indonesia
m_syahid_a@statistika.its.ac.id¹, setiawan@statistika.its.ac.id², suhartono@statistika.its.ac.id³, budinr@unpad.ac.id⁴, dedy-dp@statistika.its.ac.id⁵, amrimuhaimin@gmail.com⁶, endahsetyowati3@gmail.com⁷

Abstract. The GSTAR model is a multivariate time series model that has time and location dependencies. The implementation of the GSTAR model was developed using the GSTARMA model by referring to the Autoregressive Moving Average (ARMA) model. This research provides the early warning stage of air pollution in Surabaya by forecasting the content of pollutant, especially the CO (Carbon monoxide). The data used in this research are CO data are taken from three monitoring stations in Surabaya with a period from January 1st to December 31st, 2018. There are two stages in this research, the first stage is time series regression with a dummy variable from the data pattern, and the second step is modeling residual time series regression with GSTAR and GSTARMA. This research is using two weights, uniform and inverse weight the distances, with two-parameter estimates, namely OLS and SUR. The results given the RMSE values tend to be small by using GSTARMAX(2₁[7]) model with an inverse weight the distance using SUR parameter estimation on SUF 1, GSTAR(2₁) model with an inverse weight the distance using SUR parameter estimation on SUF 6, and GSTARMA(2₁[7]) model using an inverse weight the distance and SUR parameter estimation on SUF 7. Based on the results of this research, the GSTARMA model can correct prediction errors from the GSTAR model on CO data.

Keywords: GSTARMA, GSTAR, Time Series Regression, Air Pollutant, Surabaya

1. Introduction

Surabaya is one of the metropolitan cities in Indonesia and is the center of the economy in eastern Indonesia [1]. As a city with a population density of 9.468 people per km², Surabaya can not separate from various population problems, one of which is air pollution [2]. The Air Pollution Standard Index describes the condition of air quality which consists of several parameters, namely particulate matter (PM₁₀), carbon monoxide (CO), sulfur dioxide (SO₂), nitrogen dioxide (NO₂), and ozone (O₃) [3]. In 2007, Surabaya had PM₁₀ and CO concentration values that passed the air quality standards. Carbon
monoxide (CO) from vehicles provides the largest contribution to air pollution. Thus, CO levels in the air become one of the main focuses in monitoring air quality in the city of Surabaya.

Air quality monitoring in Surabaya has been carried out by the Surabaya City Environment Agency, one of which is the Air Quality Monitoring System (AQMS). AQMS network is installing on seven SUF stations in Surabaya, but there are only 3 active SUF stations, namely SUF 1, SUF 6, and SUF 7. Collecting data using AQMS only records the air condition at that time, thus requiring the role of science statistics to estimate air conditions for the future. Previous research on air quality in Surabaya was conducted by Chrisdayanti and Suharsono with Double Seasonal ARIMA (DSARIMA) method. The data used in the research is PM10 data at two stations, namely SUF 6 and SUF 7 univariately. The air is fluid in an open system that allows linkages between SUF stations [4]. So, it can develop the modeling methods that can capture Spatio-temporal data patterns.

Spatio-temporal cases are usually associated with multivariate data that correlates with previous time and other locations. Space-Time Autoregressive (STAR) model proposed by Pfeifer and Deutsch is a model that combines time and location dependencies in a multivariate time series model [5, 6]. Generalized Space-Time Autoregressive (GSTAR) model is an extension of the STAR model. Currently, GSTAR is more often used to model and forecast time series data in various locations [7]. Previous research of GSTAR has been applied to the forecasting of the Outflow of Currencies in Java, Indonesia, using GSTAR-SUR by involving calendar variations and intervention [8]. Whereas, the research about air pollutants using station differences has been carried out by Wutsqa, Suhartono, and Sutijo by applying the GSTAR model to air pollution data in Surabaya [9]. The data used in the research is PM10 data from three monitoring stations in Surabaya. The result shows a pattern refers to the Autoregressive Moving Average (ARMA) model, but MA pattern cannot accommodate only by the GSTAR model approach. Thus, it requires further research to get a better model with accommodate MA order.

Air quality modeling in Surabaya by using the GSTAR method tends to give less accurate because errors in the GSTAR model still form a pattern. Thus, this research developing Generalized Space-Time Autoregressive Moving Average (GSTARMA) model to increase the accuracy of prediction. The purpose of this research is to examine the best model from GSTAR and GSTARMA model based on RMSE criteria. The data used for modeling air pollution is CO data in Surabaya using three active SUF stations.

2. Methods

2.1. Intervention Analysis

Intervention analysis is a stochastic modeling technique for analyzing precise or precise data carried out by humans so that it can cause a significant difference in the average level of time series data [10]. Intervention analysis is using for unexpected events as well as internal events that are expected to influence predictive variables. General forms of intervention models such as equation (1) [11].

\[ Z_t = \frac{\omega_1(B)B^k}{\delta_1(B)} I_t + N_t, \]

where \( \omega_1(B) = (\omega_0 - \omega_1 B - \omega_2 B^2 - \ldots - \omega_j B^j), \delta_1(B) = (1 - \delta_1 B - \delta_2 B^2 - \ldots - \delta_j B^j), I_t = \) intervention variable, \( N_t = \) model error that following IIDN(0, \( \sigma_N^2 \)).

In general, there are two intervention variables, namely the step and pulse function. The step function is an intervention that occurs over a long period of time. While the pulse function is an intervention that affects time and is unsuccessful at other times. The step function and pulse function are denoted as equation (2) and (3).

\[ I_t : S_t = \begin{cases} 
0, & t < T. \\
1, & t \geq T.
\end{cases} \]
2.2. Generalized Space-Time Autoregressive Moving Average (GSTARMA)

The GSTARMA model is a development of the GSTAR model by re-modeling the GSTAR error model. In general, the GSTARMA model is like equation (5) [7].

\[
z_t = \sum_{i=1}^{p} \Phi_{io} + \sum_{k=1}^{q} \Phi_{ik} W^{(k)} \bigg| e_{t-s} - \sum_{i=1}^{q} \Theta_{io} + \sum_{k=1}^{q} \Theta_{ik} W^{(k)} \bigg| e_{t-s} + u_t,
\]

where \( \Phi_{io} = \text{diag}(\phi_{i0}^{(1)}, \ldots, \phi_{i0}^{(N)}) \), \( \Phi_{ik} = \text{diag}(\phi_{ik}^{(1)}, \ldots, \phi_{ik}^{(N)}) \), then \( \Phi_{io} \) and \( \Phi_{ik} \) are called autoregressive parameters with the \( i \)-th location, s-time sequence, and k-spatial sequence. \( W^{(k)} \) is the weight matrix \((N \times N)\) on the k-spatial location, s-time sequence, and \( k \)-spatial sequence. \( u_t \) is an error vector that is usually randomly distributed at time \( t \) with a mean of zero and covariance \( \Sigma \). If the GSTARMA error model correlates between locations, then estimating the model parameters uses GLS. The GSTARMA model involving exogenous variable is called GSTARMAX.

2.3. GSTARMA Model Identification

One important step in building the GSTARMA model is determining the order of \( p \) in the autoregressive model and the order of \( q \) in the moving average model. Determination of the order \( p \) and \( q \) on the GSTARMA model using the Cross-Correlation Function (CCF) and Partial Cross-Correlation Function (PCCF). CCF serves to measure the magnitude and direction of the correlation between two random variables [11]. CCF is used to identify the order of the MA model. PCCF is used to identify the order of the AR model. The cross-correlation function between \( x_t \) and \( y_t \) is follow equation (6).

\[
\hat{\rho}_{xy}(k) = \frac{1}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})(y_i + k - \bar{y})}}
\]

PCCF matrix in lag \( s \) as in equation (7) [10].

\[
P(s) = \left[D_v(s)\right]^{-1} V_{uv}(s) \left[D_u(s)\right]^{-1}
\]

where \( D_v(s) \) is a diagonal matrix in the i-th diagonal element is the root of the i-diagonal element of \( V_v(s) \) and \( D_u(s) \) is defined as \( V_u(s) \).

\[
U_{s-1,s+k} = \begin{cases} Z_{t+s} - \sum_{k=1}^{s-1} \alpha_{i,k} Z_{t+s-k}, & s \geq 2 \\ Z_{t+1}, & s = 1 \end{cases} \quad \text{and} \quad V_{s-1,s} = \begin{cases} Z_{t} - \sum_{k=1}^{s-1} \beta_{i-1,k} Z_{t+k}, & s \geq 2 \\ Z_{t}, & s = 1 \end{cases}
\]

So \( V_{u(s)} \) is \( \text{var}(U_{s-1,s}) \), \( V_{v(s)} \) is \( \text{var}(V_{s-1,s}) \), and \( V_{uv(s)} \) is \( \text{cov}(V_{s-1,s}, U_{s-1,s}) \).
2.4. Model Evaluation Criteria
The selection of the best model is based on the Root Mean Square Error (RMSE) criteria which can be calculated by the following equation [12], where L is the length of the forecast.

\[
RMSE = \sqrt{\frac{1}{L} \sum_{i=1}^{L} (Z_{n+1} - \hat{Z}_n(i))^2}
\]  

(9)

3. Data
The data used to GSTARMAX model are the daily data of air pollution in Surabaya at three locations, namely SUF 1 (Taman Prestasi), SUF 6 (Wonorejo), and SUF 7 (Kebonsari). Air pollution data were obtained from the Air Quality Monitoring System (AQMS). The data used is from January, 1st until December 2018, 31st. The data divide into two-part, January 1st until December, 24th as training data and December 25th – 31st as testing data.

4. Results
The time series plot of CO data on the three SUFs in Figure 1 shows values that tend to be high at several points. There are 3 data patterns at CO data on SUF 1, the first part from the 1st to 154th data, the second part from the 155th to 297th data, and the third part from the 298th to 358th data. The CO data on SUF 6 shows that the data pattern consists of several patterns, based on Table 1. Then, the CO data on SUF 7 divides into several patterns which are grouping in the dummy variable according to Table 1. Furthermore, the dummy variable is used as a predictor variable in the GSTARMAX model through the time series regression stage. In the time series regression, each CO data in the three SUFs regressed with the dummy variables according to Table 1 and parameters were significant at the 5% significance level.

![Figure 1. Time Series Plot of CO Data on (a) SUF 1, (b) SUF 6, and (c) SUF 7](image-url)
The time series regression model of CO data on SUF 1, 6, and 7 as follows:

\[ Y_{1,t} = 0.5148D_{11,t} + 0.433D_{12,t} - 0.72D_{13,t} + 0.00621D_{14,t} + N_{1,t} \]

\[ Y_{2,t} = 0.9257D_{21,t} + 0.6616D_{22,t} + 0.304D_{23,t} + 0.19D_{24,t} + 0.4963D_{25,t} + 0.7543D_{26,t} + 2.48D_{28,t} + N_{2,t} \]

\[ Y_{3,t} = -0.038D_{31,t} + 1.588D_{32,t} + 0.046D_{33,t} + 0.016D_{34,t} + 15.603D_{35,t} + 0.390D_{36,t} + 0.594D_{37,t} + 0.0026D_{38,t} + N_{3,t} \]

| Data     | Dummy Variable | Point to-Information |
|----------|----------------|----------------------|
| CO SUF 1 |                |                      |
| D_{11,t} |                | 1-154                |
| D_{12,t} |                | 155-297              |
| D_{13,t} |                | 298-358              |
| D_{14,t} |                | 298-358              |
|          |                | Binary dummy         |
| CO SUF 6 |                |                      |
| D_{21,t} |                | 1-102                |
| D_{22,t} |                | 103-176              |
| D_{23,t} |                | 17-185               |
|          |                | Binary dummy         |
|          |                | Trend dummy          |
| CO SUF 7 |                |                      |
| D_{31,t} |                | 1-36                 |
| D_{32,t} |                | 1-36                 |
| D_{33,t} |                | 37-60                |
| D_{34,t} |                | 61-100               |
| D_{35,t} |                | 101-164, 184-219, 247-275, 285-297, 355-358 |
|          |                | 276-284              |
| D_{37,t} |                | 298-310              |
| D_{38,t} |                | 311-329              |

The residual time series regression model is using as the data to GSTAR and GSTARMA model. The MCCF and MPCCF plots of residual time series regression given in Figure 2 and Figure 3. GSTAR modeling is using two weights namely uniform and inverse weights the distance. Uniform weights assume that the CO data between locations is homogeneous and inverse weights the distance shows the actual distance from three SUF locations. Besides, this research is using two-parameter estimates, namely Ordinary Least Square (OLS) and Seemingly Unrelated Regression (SUR). Based on the MPCCF scheme in Figure 3, it shows that the autoregressive order used is 2, so the GSTAR model uses the GSTAR(2,1) model. Furthermore, the residuals of the GSTAR(2) model are modeled by determining the moving average order of the GSTAR(2) model residuals. The results of modeling with the GSTAR(2) model using uniform weights it is known that MCCF residual model is
significant in lag 7, so the GSTARMA model used is GSTARMA(2,[7]). Then, the MCCF of residual GSTAR(2) model using inverse weights the distance are significant at lag 7 too, so the GSTARMA model used is GSTARMA(2,[7]). Thus, the GSTARMA model using uniform weights and namely inverse weights the distance is GSTARMA(2,[7]).

Based on Figure 4, the distance is too near, so the GSTARMA model used is GSTARMA(2,[7]). Then, the MCCF of residual GSTAR(2) model using inverse weights the distance are significant at lag 7 too, so the GSTARMA model used is GSTARMA(2,[7]). Thus, the GSTARMA model using uniform weights and namely inverse weights the distance is GSTARMA(2,[7]).

| Variable/ | Lag 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 |
|----------|-----------------------------------------------|
| SUF1     | +++ +.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.   |
| SUF6     | ++.++.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.   |
| SUF7     | ++.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.    |

**Figure 2.** Schematic of MCCF Residual CO Data from SUF 1, SUF 6, and SUF 7

| Variable/ | Lag 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 |
|----------|-------------------------------------------------------|
| SUF1     | ... .+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.     |
| SUF6     | ... .+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.     |
| SUF7     | ... .+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.+.     |

**Figure 3.** Schematic of MCCF Residual CO Data from SUF 1, SUF 6, and SUF 7

To determine the best method, the criteria for evaluating the model is RMSE. Figure 4 shows the comparison of the RMSE values of the GSTAR and GSTARMA models using uniform and inverse weights the distance. Based on Figure 4 it can be seen that CO SUF 1 data tends to be well modeled using the GSTARMA(2,[7]) model using inverse weights the distance and parameter estimation using OLS. Whereas in CO SUF 6 data it is known that the GSTAR(2) model tends to be better than
the GSTARMA(2, [7]) model. However, the RMSE value in CO SUF 6 tends to be small in the GSTAR(2) model with the inverse weight the distance and using SUR parameter estimation. In the SUF 7 CO data, the GSTARMA(2, [7]) model by using inverse weight the distance and parameter estimation using SUR tends to produce a smaller RMSE value. Based on the comparison of the RMSE values, the GSTARMA model can correct prediction errors from the GSTAR model.

Figure 4. Comparison of RMSE values from CO Testing Data on (a) SUF 1, (b) SUF 6, and (c) SUF 7

5. Conclusion

Modeling of air pollution, especially CO data in Surabaya, is carried out using the GSTARMAX method. This modeling consists of two stages, the first is time series regression, and the second is the GSTAR model. The time series regression model from CO data on SUF 1, SUF 6, and SUF 7 are

\[ Y_{1,j} = 0.5148D_{1,1,j} + 0.433D_{1,12,j} - 0.72D_{1,3,j} + 0.00621D_{1,14,j} + N_{1,j} \]

\[ Y_{2,j} = 0.9257D_{2,1,j} + 0.6616D_{2,12,j} + 0.304D_{2,3,j} + 0.19D_{2,14,j} + 0.4963D_{2,25,j} + 0.7543D_{2,26,j} + 2.48D_{2,28,j} + N_{2,j} \]

\[ Y_{3,j} = -0.038D_{3,1,j} + 1.588D_{3,2,j} + 0.046D_{3,3,j} + 0.016D_{3,14,j} + 15.603D_{3,25,j} + 0.390D_{3,26,j} + 0.594D_{3,28,j} + 0.0026D_{3,38,j} + N_{3,j} \]

There are several GSTAR and GSTARMA model from residual time series regression, namely GSTARX (2, 1) and GSTARMAX(2, [7]), either by using OLS or SUR parameter estimates. Determination of the best method is based on the smallest RMSE value. The results show that CO data tend to have small RMSE values using GSTARMAX(2, [7]) model using inverse weight the distance and parameter estimation using OLS on SUF 1, GSTAR(2) with inverse weight the distance using SUR parameter estimation on SUF 6, and GSTARMAX(2, [7]) model by using inverse weight the
distance and parameter estimation using SUR on SUF 7. Hence, the GSTARMA model can decrease the prediction errors from the GSTAR model on CO data in Surabaya.

References
[1] Badan Pusat Statistik (Statistics Indonesia) 2018 Retrieved from Badan Pusat Statistik (Statistics Indonesia) Surabaya City: https://surabayakota.bps.go.id/publication/2018/08/21/35de76f19338e3ec5d45b838/kota-surabaya-dalam-angka-2018
[2] Department of Population and Civil Registration 2016 Department of Population and Civil Registration Surabaya: Population Statistics http://dispendukcapil.surabaya.go.id
[3] Ministerial Decree of Environment and Forestry No. 45 1997 The Air Pollution Standard Index (Jakarta: Ministry of Environment and Forestry)
[4] Chrisdayanti B and Suharsono A 2015 Forecasting the content of Particular Matter (PM10) in Ambient Air Surabaya City Using Double Seasonal ARIMA (DSARIMA) Jurnal Sains dan Seni ITS 4(2)
[5] Pfeifer P E and Deutsch S J 1980 A Three Stage Iterative Procedure for Space-Time Modeling Technometrics 22(1): 35-47
[6] Pfeifer P E and Deutsch S J 1980 Identification and Interpretation of First Order Space-Time ARMA Models Technometrics 22(1): 397-408
[7] Ruchjana B N 2002 Modeling of Oil Production Curve Using Generalization S-TAR Model Statistics and Computing Forum (Bandung: Institut Teknologi Bandung)
[8] Akbar M, Setiawan, Suhartono, Ruchjana B N and Riyadi M A A 2018).STAR-SUR Modeling With Calendar Variations and Intervention to Forecast Outflow of Currencies In Java Indonesia Journal of Physics: Conference Series 974 012060
[9] Wutsqa D U, Suhartono and Ulama B S S 2012 Application of Generalized SpaceTime Autoregressive Model on Air Pollution Data in Surabaya City Pythagoras 7(2): 17-30
[10] Hipel K W, Lennox W C, Unny T E and Mcelough A I 1975 Intervention Analysis in Water Resources Water Resources research 11(6): 855-861
[11] Wei W W S 2006 Time Series Analysis: Univariate and Multivariate Methods Pearson Education, Inc.
[12] Ahmad I S, Setiawan, Suhartono and Masun N H 2015 Forecasting of Monthly Inflow and Outflow Currency Using Time Series Regression and ARIMAX: The Idul Fitri Effect AIP Conference Proceedings 1691 050002