Application of Generalized Space-Time Autoregressive (GSTAR) Model on Distribution of Death Case in 8 District of Central Jakarta

Rahmat Febriyanto[1)], Yekti Widyaningsih[2)], Siti Nurrohmah[3)]

[1,2,3)]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universtity of Indonesia, West Java 16424, Indonesia*

[1)]*rahmatfebriyantoooo@gmail.com*
[2)]*yekti@sci.ui.ac.id*
[3)]*snurrohmah@sci.ui.ac.id*

Abstract. In this paper we apply Generalized Space-Time Model (GSTAR) model on Distribution of Death Case in 8 District of Central Jakarta. The GSTAR model is an extension of the STAR model. The main difference that exists between the GSTAR model and the STAR model is on assuming parameters. The parameters in the STAR model do not depend on the location, so this model is only suitable for locations with homogeneous characteristics. In the GSTAR model, the model parameters change for each location to form a diagonal matrix with parameters. The objective of this study is to model the data of death cases in Central Jakarta using GSTAR model. The data is provided by DINKES DKI (Public Health Office in Jakarta). The data describes the distribution of death cases in 8 districts of Central Jakarta in the period 2011-2017.

INTRODUCTION

Based on a report from the Global Burden of Disease Study, around the world, deaths are the majority due to infectious diseases, heart disease, conflict and terrorism. The study examined the state of health around the world by estimating the average life expectancy and the number of deaths. The study is coordinated by the Institute for Health Metrics and Evaluation (IHME) and involves more than 2,500 collaborators from 130 countries and regions of the world. The report found that for now, the average global life expectancy is 72.5 years (75.3 years for women and 69.8 years for men). Japan has the highest life expectancy in 2016 with an average life expectancy of 83.9 years. While the Central African Republic became the lowest life expectancy of 50.2 years. Overall, there are 54.7 million deaths worldwide by 2016. Nearly 72.3 percent of these deaths are caused by diseases such as heart disease, stroke and cancer. Approximately 19 percent of deaths in 2016 come from infectious diseases, maternal diseases, neonatal disease and nutritional deficiencies (CMNN). While 8 percent of deaths come from injuries.(livescience.com, 2017)

Jakarta as the capital city of Indonesia, has a high level of air pollution. The air quality monitoring conducted by Greenpeace since January 2017 at 21 locations in Jabodetabek (Jakarta Bogor Tangerang Depok Bekasi) shows similar results with the monitoring results of the US Embassy. Air quality in Jabodetabek for the past six months indicated to have entered unhealthy levels for humans and will cause more serious health impacts for sensitive groups, such as children, pregnant women, and elderly (senior citizen). The PM2.5 daily rate at that location is far beyond a tolerable standard, such as the WHO standard of 25μg / m3. PM2.5 can be inhaled and deposited in the respiratory organs. If exposed in the long run, PM2.5 can cause acute respiratory infections (especially for children) to lung cancer. In addition, PM2.5 can increase levels of toxins in the blood vessels that can stimulate stroke,

cardiovascular disease and other heart diseases, and can harm pregnant women because of the potential to attack the fetus.

World Health Organization statistics from 2008 estimated that 1.3 million urban dwellers die prematurely due to air pollution. If the WHO strict rules on clean air density levels are applied, nearly 1.1 million deaths can be avoided. Emissions from factories and power plants, gas pollution from cars, heated or coal-fired kitchens, mountainous debris is the largest source of dust and smoke particles. For several years WHO has issued a reference to air pollution limits for big cities. But these references are rarely met throughout the world. Namely the maximum limit of 20 micrograms per particle per cubic meter of air per year. In many cities the particle content even reaches more than 300 micrograms per cubic meter. (WHO, 2008)

In a study conducted by Michael Jerrett et al (2013), titled Spatio Temporal Analysis of Air Pollution and Mortality in California Based on the American Cancer Society Cohort shows that there is a consistent and strong effect on air pollution that spreads with death cases.

By using spatio temporal data, the above study explains the linkage of location and time to mortality rate. To predict future mortality rates, we can use the right model of spatio temporal data, which is the data of each region as time series data. Time series is an observation value that is influenced by previous times to obtain a picture of the development of a future activity. One way to model the data is to use an autoregressive vector (VAR) model (Hannan, 1970). Although very flexible, the autoregressive vector model (VAR) has too many unknown parameters which must be estimated from the limited data. The Autoregressive Vector model only uses the observed value at a previous time to obtain a picture in the future.

By adding location weights into the autoregressive vector model introduced Space Time Autoregressive (STAR) model by Cliff and Ord (1973), Martin and Oepeen (1975) and developed by Pfeir and Deutsch (1980). As an autoregressive vector model, the STAR model has a linear dependency on the region and time. The main difference of the VAR model is that the spatial interconnection of the STAR model is built using the location weight matrix. By adding the location weight matrix into the VAR model, the distance between regions will be the weight of the model. But the STAR model has a weakness in the flexibility of parameters which assumes that the locations studied have a uniform characteristic (Homogeneous) where the autoregressive parameters in the model are assumed to be the same for each location, so that if faced with locations with heterogeneous characteristics the model is poor to use.

From these weaknesses Borovkova, Lopuhaa and Ruchjana (2002) developed the Generelized Space Time Autoregressive (GSTAR) model that can be used in locations with uniform (heterogeneous) characteristics where autoregressive parameters on different models for each location. Therefore, GSTAR model is more realistic, because in reality there are more models with different model parameters for different location (Wutsqa et al, 2010). The main problem in the GSTAR model lies in the determination and selection of location weights. Location weights in GSTAR are generally divided into three types: binary location weights, uniform location weights, and location weights. There have been many previous studies that have used the GSTAR model such as, Ruchjana (2002) conducted a study on the modeling of the petroleum production curve, Nunung et al. (2012) applied the GSTAR model to the GDP data of existing Western European countries. Susanti (2013) applies the GSTAR model to forecasting the number of tourist visits of four tourist sites in Batu with three weight locations and Herlin et al (2014) applying GSTAR model on the data of the number of West Java TKI.

Therefore, based on the description above, it will be used GSTAR model to model the spatio temporal data for the data of death cases in 8 districts in Central Jakarta by using three location weights i.e. binary location weight, uniform location weight, and weight of distance location. Central Jakarta consists of 8 districts namely Gambir, Sawah Besar, Kemayoran, Senen, Cempaka Putih, Johar Baru, Menteng and Tanah Abang. Spatio Temporal data on death cases to be used is monthly data of death cases in 2011-2017.

STAR MODEL

Let $\{\mathbf{Z}(t) : t = 0, \pm 1, , \pm 2, \dots\}$ be a multivariate time series of N components. STAR models require the definition of a hierarchical ordering of "neighbors" of each site. For instance, on a regular grid, one can define neighbors of a site "first order neighbors" of the first order neighbors- "second order neighbors" and so on. The next observation at each site is then modelled as a linear function of the previous observations at the same site and of the weighted previous observations at the neigbhoring sites of each order. The weights are incoporated in weight matrices $\boldsymbol{W}^{k}$ for spatial order k. Formally, the space-time autoregressive model of autoregressive order p and spatial orders $\lambda_1, \lambda_2, \dots, \lambda_p$ (STAR(p, $\lambda_1, \lambda_2, \dots, \lambda_p$)) is defined as:

$$\boldsymbol{Z}(t) = \sum_{s=1}^{p} \sum_{k=0}^{\lambda_s} \phi_{sk} \boldsymbol{W}^{(k)} \mathrm{Z(t-s)} + \mathbf{e(t)}, \quad \mathrm{t} = 0, \pm 1, \pm 2, \ \ldots \qquad (1)$$

Where λ_s is the spatial order of the s-*th* autoregressive term, ϕ_{sk} is the autoregressive parameter at time lag s and spatial lag k, $\boldsymbol{W}^{(k)}$ is the N x N weight matrix for spatial lag k = 0, 1, ..., λ_s and $\mathbf{e(t)}$ are random error disturbances with mean zero. Because there are no neighbors at spatial lag zero, $\boldsymbol{W}^{(0)} = \boldsymbol{I_N}$. For spatial order one or higher we only have first order or higher order neighbors, so that for k ≥ 1 the weight matrix should satisfy $\sum_{j=1}^{N} w_{ij}^{(k)} = 1$ for all i, k.

GSTAR MODEL

The main restriction on the STAR model defined in (1) is that the autoregressive parameters ϕ_{sk} are assumed to be the same for all locations. However, there is no a priori justification for this assumption; in fact these parameters are most likely to be different for different locations. A generalized STAR (GSTAR) model is a natural generalization of STAR models, allowing the autoregressive parameters to vary per location: ${\phi_{sk}}^{(i)}, \ i = 1,2. \ldots, N$. Note that STAR models are a subclass of GSTAR models. In this paper we shall consider GSTAR models; naturally all the results will continue to hold for STAR models. A GSTAR(p, $\lambda_1, \lambda_2, \ldots, \lambda_p$) model of autoregressive Order p and spatial order λ_s of the sth autoregressive term is given by

$$\boldsymbol{Z}(t) = \sum_{s=1}^{p} [\phi_{s0} \boldsymbol{Z}(\mathrm{t-s}) + \sum_{k=1}^{\lambda_s} \phi_{sk} \boldsymbol{W}^{(k)} \boldsymbol{Z}(\mathrm{t-s})] + \mathbf{e(t)}, \ \mathrm{t} = 0, \pm 1, \pm 2, \ \ldots \qquad (2)$$

Where ϕ_{sk} = diag(${\phi_{sk}}^{(1)}, {\phi_{sk}}^{(2)}, \ldots, {\phi_{sk}}^{(N)}$) $\boldsymbol{W}^{(k)}$ satisfies the same conditions as for model (1).

RESEARCH METODOLOGY

Generate Data

Data for simulation is generated from random and normal assumption with mean of data is zero (assumption of stationary), cosists of ..... areas and ..... time bservations.

Model Identification

The dependency of space and time needs to identify, to propose a suitable model for the data. In this study, model identification is used by analyzing the graph of Space Time Autocorrelation Function (STACF) and Space Time Partial Autocorrelation Function (STPACF)

Parameter Estimation

Parameter estimation is done to find out the significant of estimator and cheking stationary process regarding the estimator. The rules are as follow

Significance of Estimator

H_0: $\phi = 0$ (*the estimator of parameter is not significance*)
H_1: $\phi \neq 0$ (*the estimator of parameter is significance*)

Diagnostic Checking

The basic concept of diagnostic checking is checking the assumption that related with error and checking the residual (realization of error). In this study, normality test and ACF residual is analyzed. Finally, mean square error (MSE) is used to check the residual and determine the best model.

Forecasting

Forecasting is computed for January 2018 by using best model.

DATA

GSTAR model is applied to model the spatio temporal data of death cases data in 8 districts in Central Jakarta using three types of location weights i.e. binary location weight, uniform location weight, and weight of distance location. Central Jakarta consists of 8 sub-districts namely Gambir, Sawah Besar, Kemayoran, Senen, Cempaka Putih, Johar Baru, Menteng and Tanah Abang. Spatio temporal data on death cases is based on monthly observations of death cases in 2011-2017.

Table 1.The distances between two sub-districts in Central Jakarta (in kilometer)

	Gambir	Sawah Besar	Kemayoran	Senen	Cempaka Pth	Johar Baru	Menteng	Tanah Abang
Gambir	0	3.303	6.851	5.696	9.27	9.026	5.527	7.026
Sawah Besar	3.303	0	4.601	6.005	8.097	7.423	7.321	10.092
Kemayoran	6.851	4.601	0	4.718	4.04	3.974	7.486	11.876
Senen	5.696	6.005	4.718	0	4.342	3.019	2.966	7.794
Cempaka Pth	9.27	8.097	4.04	4.342	0	1.348	7.127	12.041
Johar Baru	9.026	7.423	3.974	3.019	1.348	0	5.78	10.694
Menteng	5.527	7.321	7.486	2.966	7.127	5.78	0	4.916
Tanah Abang	7.026	10.092	11.876	7.794	12.041	10.694	4.916	0

Table 1 shows the distances between two sub-districts in Central Jakarta (kilometers). Weight matrix is constructed based on this table. The distance between two sub-districts is gained by quadrating the differences of longtitude and latitude of two sub-districts, and then get the square root of it.

Matrix of Spatial Weight

Binary Weight

$$w_{ij}^{(l)} = \begin{cases} 1, & if\ d_{ij} = \min\{d_{ij}, d_{ik}\}\ for\ j \neq k \\ 0, & others \end{cases}$$

Uniform Weight

$$w_{ij}{}^{(l)} = \begin{cases} \frac{1}{n_i{}^{(l)}}, & \text{j is the neighbor i on the spatial lag l} \\ 0, & others \end{cases}$$

Distance Weight

$$w_{ij}{}^{(l)} = \begin{cases} \frac{\frac{1}{1+d_{ij}{}^{(l)}}}{\sum_{j=1}^{N}\frac{1}{1+d_{ij}{}^{(l)}}}, & \text{j is the neighbor i on the spatial lag } l \\ 0, & others \end{cases}$$

RESEARCH RESULT AND DISCUSSION

The result of this study consists of model identification, parameter estimation, diagnostic model (assumption checking) and forecasting Weight Matrix.

BINARY WEIGHT MATRIX LAG 1

0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00

UNIFORM WEIGHT MATRIX LAG 1

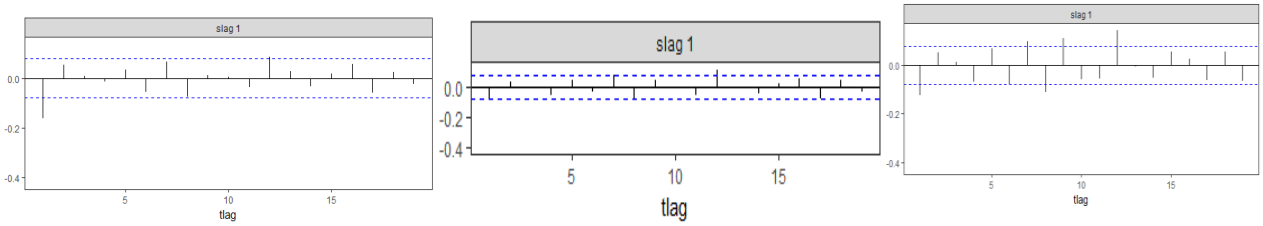
0.0	1.00	0.0000000	0.0000000	0.0000000	0.0000000	0.00	0.0
0.5	0.00	0.5000000	0.0000000	0.0000000	0.0000000	0.00	0.0
0.0	0.25	0.0000000	0.2500000	0.2500000	0.2500000	0.00	0.0
0.0	0.00	0.2500000	0.0000000	0.2500000	0.2500000	0.25	0.0
0.0	0.00	0.3333333	0.3333333	0.0000000	0.3333333	0.00	0.0
0.0	0.00	0.3333333	0.3333333	0.3333333	0.0000000	0.00	0.0
0.0	0.00	0.0000000	0.5000000	0.0000000	0.0000000	0.00	0.5
0.0	0.00	0.0000000	0.0000000	0.0000000	0.0000000	1.00	0.0

DISTANCE WEIGHT MATRIX LAG 1

0.00	0.248/1.01	0.127/1.01	0.149/1.01	0.097/1.01	0.111/1.01	0.153/1.01	0.125/1.01
0.248/1.009	0.00	0.179/1.009	0.`143/1.009	0.11/1.009	0.119/1.009	0.12/1.009	0.09/1.009
0.127/1.076	0.179/1.076	0.00	0.175/1.076	0.198/1.076	0.201/1.076	0.118/1.076	0.078/1.076
0.149/1.269	0.143/1.269	0.175/1.269	0.00	0.187/1.269	0.249/1.269	0.252/1.269	0.114/1.269
0.097/1.221	0.110/1.221	0.198/1.221	0.187/1.221	0.00	0.429/1.221	0.123/1.221	0.077/1.221
0.111/1.343	0.119/1.343	0.201/1.343	0.249/1.343	0.429/1.343	0.00	0.148/1.343	0.086/1.343
0.153/1.11	0.12/1.11	0.118/1.11	0.252/1.11	0.123/1.11	0.148/1.11	0.00	0.196/1.11
0.125/0.766	0.09/0.766	0.078/0.766	0.114/0.766	0.077/0.766	0.086/0.766	0.196/0.766	0.00

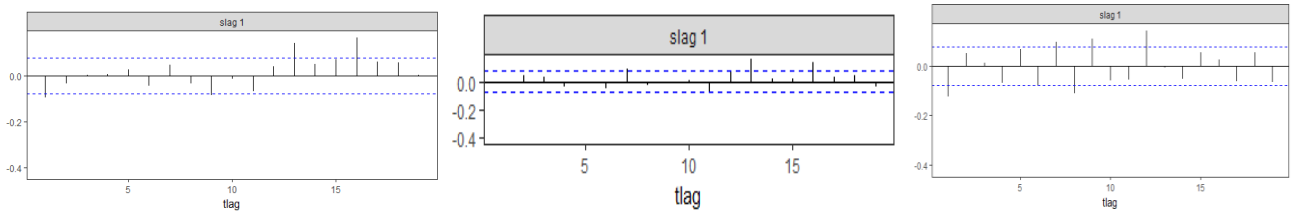
Model Identification

Model identification is determined based on the graph pattern of Space Time Autocorrelation Function (STACF) and Space Time Partial Autocorrelation Function (STACF). By using starma package in R program, the results of STACF and STPACF are showed by figure 2 and 3 as the following.



(a) (b) (c)

Figure 2. Graph of STACF Uniform (a), Binary (b) and Distance (c)



(a) (b) (c)

Figure 3. Graph of STPACF, Uniform (a), Binary (b) and Distance (c)

The STACF values in the GSTAR model is decreasing and the STPACF values cut off after time lag-p[4] , we choose GSTAR(1, 13) for binary weight matrix, GSTAR(1,7) for uniform weight matrix and GSTAR(1,3) for distance weight matrix.

PARAMETER ESTIMATION

We use least square methods to find the estimators, and find the estimators with R program [1] and choose the best model with the lowest MSE.

GSTAR(1,13) uniform	GSTAR(1,7) biner	GSTAR(1,3) distance
MSE =93.27382	MSE = 96.11254	MSE = 93.16643

GSTAR (1,3) with distance weight matrix is choosen as the best model since it has smallest MSE. We use the least square method to obtain parameters in the GSTAR (1.3). With p-value on F test smaller than alpha then model already explain data.As for the insignificant parameters that should be omitted, consider the weight of each location, the parameter elimination is not performed. According to Kestenko (2008) and Armstrong (2006) in Ike Fitrianingsih (2012: 45), states that insignificant variables can still be used to make the forecasting process.

```
       Estimate Std. Error t value Pr(>|t|)
X1    -0.5379158  0.2270493  -2.369 0.018149
X2    -0.4348300  0.1505340  -2.889 0.004011
X3    -0.4625361  0.0723060  -6.397 3.22e-10
X4    -0.5082148  0.1336545  -3.802 0.000158
X5    -0.3594602  0.1512020  -2.377 0.017754
X6    -0.6771433  0.1039330  -6.515 1.55e-10
X7    -0.4267103  0.1950222  -2.188 0.029059
X8    -0.5707598  0.1129528  -5.053 5.80e-07
X9    -0.2012057  0.2522748  -0.798 0.425442
X10   -0.3176873  0.1561748  -2.034 0.042379
X11   -0.3016088  0.0783426  -3.850 0.000131
X12   -0.0669728  0.1589666  -0.421 0.673688
X13   -0.2073409  0.1550400  -1.337 0.181626
X14   -0.3463498  0.1195226  -2.898 0.003897
X15    0.0397564  0.2188292   0.182 0.855898
X16   -0.1605892  0.1282299  -1.252 0.210936
X17   -0.1477078  0.2331261  -0.634 0.526589
X18   -0.1956795  0.1466486  -1.334 0.182604
X19    0.0003475  0.0728326   0.005 0.996195
X20   -0.0830500  0.1428650  -0.581 0.561247
X21   -0.1702465  0.1599183  -1.065 0.287498
X22   -0.1253769  0.1055625  -1.188 0.235427
X23   -0.0456552  0.1956569  -0.233 0.815576
X24   -0.0307082  0.1161299  -0.264 0.791541
Residual standard error: 9.652 on 592 degrees of freedom
Multiple R-squared:  0.2693,    Adjusted R-squared:  0.2101
F-statistic: 4.546 on 48 and 592 DF,  p-value: < 2.2e-16
```

Since all estimated parameter is not equal to zero , then all values are significants

Diagnostic Checking

Normality of Residual

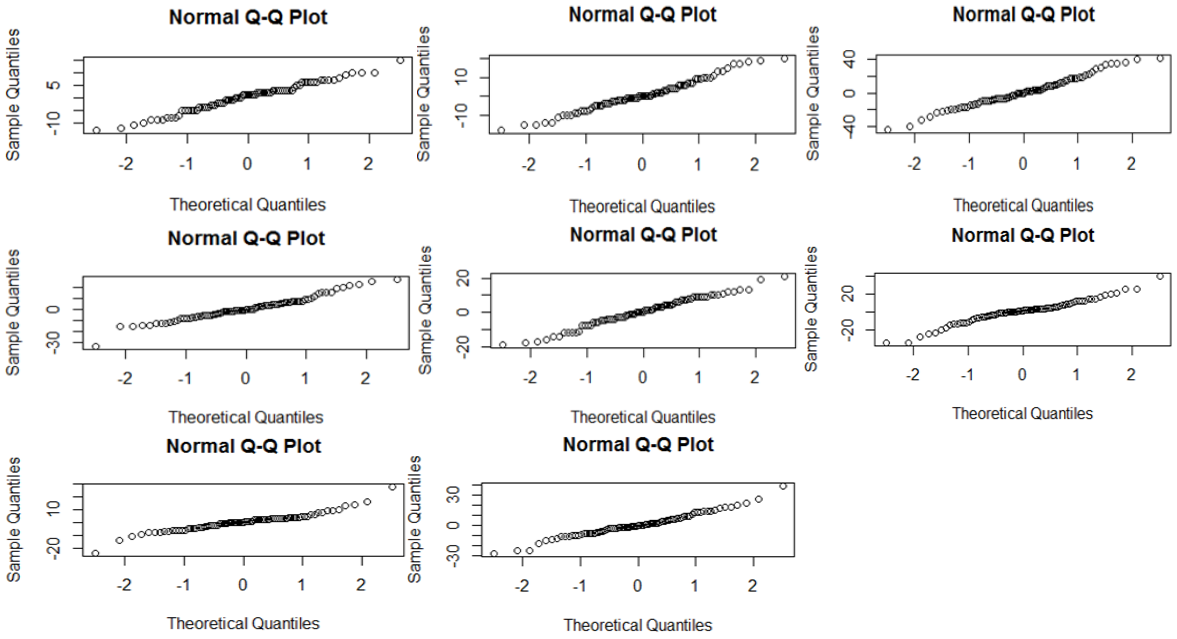


Figure 4. Q-Q Plot of Residual Data

Figure 4 show us that 8 disctrict in central jakarta has spreaded point. But, since the majority locations shows that point is grouped in the equation of x = y , then residual of data is assumed normal.

Correlation of Residual

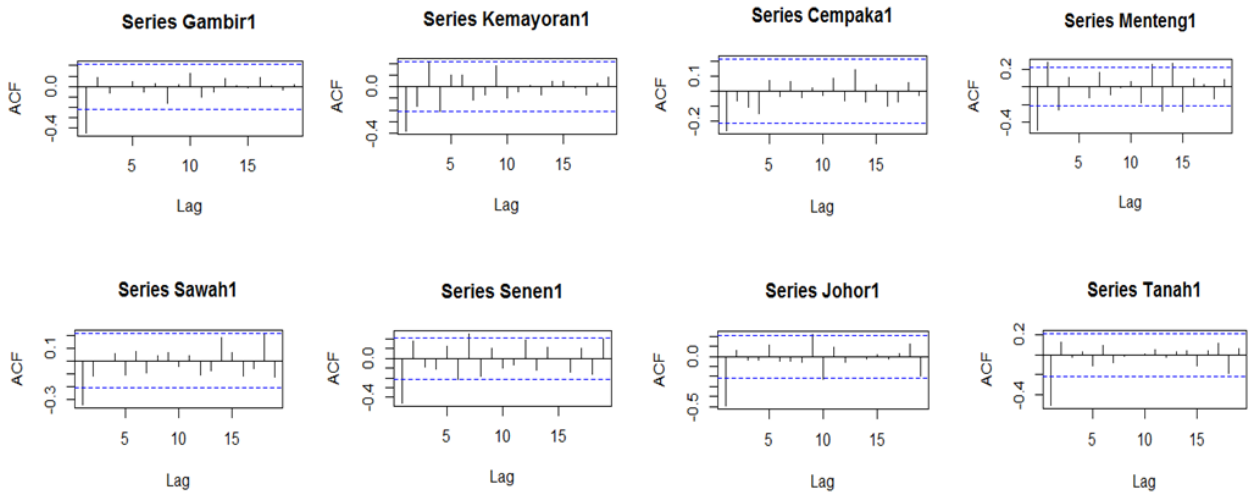


Figure 5. Plot of ACF of Residual Data

Figure 5 provides us ACF plot of redisual. Since ost market did not pass the significance level , it means the resdiuals is uncorralated.

Forecasting

Location Month	Gambir	Sawah Bsr	Kemayoran	Senen	Cempaka Pth	Johor Baru	Menteng	Tanah Abang
Januari 2018	28	55	109	36	30	55	29	60

Figure 6. Predicted number of deaths in January 2018

Location Month	Gambir	Sawah Bsr	Kemayoran	Senen	Cempaka Pth	Johor Baru	Menteng	Tanah Abang
Januari 2018	53	38	112	41	36	63	38	61

Figure 7. Number of deaths in January 2018

Summary

The deaths occurring in eight central Jakarta sub-districts can be modeled with GSTAR (1,3) using a matrix of distance weights. By comparing the predicted results with the actual number of deaths, only Gambir and Sawah Besar have considerable differences.

ACKNOWLEDGMENTS

We wish to thank DRPM Universitas Indonesia through Hibah Publikasi Internasional Terindeks untuk Tugas Akhir Mahasiswa UI (PITTA) 2018

REFERENCES

1. Ruchjana, Budi Nurani, et al. "R Software for Parameter Estimation of Spatio Temporal Model." World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering 3.12 (2016).
2. Fitriyaningsih, Ike. Perbandingan Bobot Lokasi Seragam, Invers Jarak, Dan Normalisasi Korelasi Silang Model Generalized Space Time Autoregressive (Gstar)(Studi Kasus: Harga Sayuran Pada Lima Pas. Diss. Universitas Brawijaya, 2012.
3. Ruchjana, Budi Nurani, Svetlana A. Borovkova, and H. P. Lopuhaa. "Least squares estimation of Generalized Space Time AutoRegressive (GSTAR) model and its properties." *AIP Conference Proceedings*. Vol. 1450. No. 1. AIP, 2012.
4. Karlina, Herlin Dewi, Rini Cahyandari, and Asep Solih Awalluddin. "Aplikasi Model Generalized Space Time Autoregressive (GSTAR) pada Data Jumlah TKI Jawa Barat dengan Pemilihan Lokasi Berdasarkan Klaster DBSCAN." *Jurnal Matematika Integratif* 10.1 (2014): 37-48.
5. Nurhayati, Nunung, Udjianna S. Pasaribu, and Oki Neswan. "Application of Generalized Space-Time Autoregressive Model on GDP Data in West European Countries." *Journal of Probability and Statistics* 2012 (2012).
6. Jerrett, Michael, et al. "Spatial analysis of air pollution and mortality in California." *American journal of respiratory and critical care medicine* 188.5 (2013): 593-599.

7. Escaramís, Geòrgia, et al. "Spatio-temporal analysis of mortality among children under the age of five in Manhiça (Mozambique) during the period 1997-2005." *International journal of health geographics* 10.1 (2011): 14.
8. Khagayi, Sammy, et al. "Bayesian spatio-temporal modeling of mortality in relation to malaria incidence in Western Kenya." *PloS one* 12.7 (2017): e0180516.