Poisson GSTAR model: spatial temporal modeling count data follow generalized linear model and count time series models

Wardhani L P 1  Setiawan 2  Suhartono 3  Kuswanto H 4
1  Department of Mathematics  2,3,4  Department of Statistics
Faculty of Mathematics, Computational and Data Science, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.

Email : laksmiprita61@gmail.com

Abstract. This paper discusses the formation of one temporal spatial model, the Generalized Space Time Autoregressive (GSTAR) model, if the data of model is count data. The development of the GSTAR model is an update or problem completion of the count data which tends to be stationary and non-normal / Gaussian data, because GSTAR model is assumed normal distributed and stationarity.

GSTAR modeling for count data refers to the Time series model for count data, which are the Generalized Autoregressive Moving Average (GARMA) model and modeling Count Time Series which has the Generalized Linear Model (GLM) concept. The model formed is called the Poisson GSTAR model

1. Introduction
In daily life, often found a data that not only includes the location dependence in the previous time but also has a connection to the location or other place called spatial data. The statistical model that accommodates data that depends on time and location is called the Space Time Model or the spatial Temporal Model. The existing Space Time model is a development of the time series model itself. The development of a time series model that considers spatial and time aspects is expected to increase the accuracy of forecasting. One of the Space Time models is the GSTAR model.

The Generalized Space Time Autoregressive (GSTAR) model developed by Ruchjana (2002) does not require that the parameter values be the same for all locations, which means it is assumed that the location characteristics are not the same (heterogeneous). So that the GSTAR model is more realistic because more models are found with different parameters for different locations. The GSTAR model has two types of parameters namely its time series parameters and spatial parameters. The GSTAR model is a temporal Spatio model in which time series data and the errors are assumed to have Normal distribution. In fact the real problem is that not all multivariate data are matched or precisely modeled in the Normal model, one of which is the count data (non-negative integer data) which is data from variables that state the number of events such as data about the number of accidents in a given time, data on the number of dengue fever sufferers, data on the number of tourist visits to a certain area, data on the number of luxury vehicles sold in an area at a certain time interval, the number of flights in an airport. For count time series data, the Normal distribution is suitable for stationary data, but for non-stationary data is not proper, even though the count data tends to be non-stationary w.r.t to time.

The standard distribution that is often used in data counts is the Poisson Distribution. Time series data count model with the assumption that the Poisson distribution response variable is mostly done, one of them is the formulation of the model that follows the rules in the Generalized Linear Model (GLM), such as the Poisson regression model. The time series regression model for data count differs in terms of covariates and their error structure. Other Time series data count models that follow the GLM include the GARMA model, the Poisson GARMA model from Benjamin, et al (2003), the Zeger-Qaqish model (Kendem, 2002) and the Modeling Count Time series, (Liboschick, et al, 2016), the Time Model data count series whose formulation follows GLM is easier and more flexible in applying it to model conditional observations of information that has already occurred (past
This methodology implements the selection of a distribution for data count and also the proper link function (Liboschik, 2016). In this paper the GSTAR model was developed to count data that has never been done in the development of the GSTAR model. Implementation of development is done by making the structure of the GSTAR model such as the GARMA model which is a combination of the Poisson Regression model and the ARMA model, with the reason for seeing Ruchyana (2002) making the GSTAR model can be formed into a linear regression model. The development of the GSTAR model is called the Poisson GSTAR model because the standard distribution is Poisson with the link function a logarithmic function. This paper only discusses the GSTAR model, GARMA model, GLM and about the steps of forming the model and its simulations, while estimate the model parameters will be done in another paper.

2. Literature Review

The formation of space time models for data counts, namely GSTAR Poisson is done through the thought of several time series models for data count such as the GARMA model, the multivariate count autoregression model and certainly the GSTAR model itself.

2.1. GSTAR Model

GSTAR \((p_{i}, \lambda_{s})\) model, is an autoregressive model of order \(p\), and spatial order \(\lambda_{s}\) from autoregressive shape to \(s\) for the number of locations \(m\) is defined as equation 1 below, (Borovkova, Lopuhaa, & Ruchjana, 2005)

\[
Z(t) = \sum_{k=1}^{p} \Phi_{0k} Z(t-k) + \sum_{k=1}^{\lambda_{s}} \Phi_{sk} W^{(k)} Z(t-k) + e(t)
\]

where \(\Phi_{0k} = \text{diag}(\phi_{01}^{(1)}, \phi_{02}^{(2)}, \ldots, \phi_{0m}^{(m)})\), is the parameter autoregressive matrix on time lag -s sized \(m \times m\); \(\Phi_{sk} = \text{diag}(\phi_{sk}^{(1)}, \ldots, \phi_{sk}^{(m)})\), is the spatial parameter matrix on time lag -s and spatial lag -k, sized \(m \times m\) ; \(W^{(k)}\): matrix of weight sized \(m \times m\) on spatial lag \(k\). \(Z(t)\): random vector sized \((m \times 1)\) on time \(t\), which are \(Z(t) = (z_{1}(t), \ldots, z_{m}(t))'\) and \(e(t)\) : White Noise vector sized \((m \times 1)\) distributed Multivariate Normal with mean 0 and variance covariance matrix \(\sigma^{2}I_{m}\).

The GSTAR \((1,1)\) model in equation (1) with \(\phi_{0} = \phi_{i}^{(j)}, k = 0,1,\ldots\) can be expressed as

\[
Z(t) = \Phi_{0} Z(t-1) + \Phi_{1} WZ(t-1) + e(t)
\]

or for location \(i\),

\[
Z_{i}(t) = \phi_{0i} Z_{i}(t-1) + \phi_{i0} \sum_{j=1}^{m} w_{ij} Z_{j}(t-1) + e_{i}(t), \quad i = 1,2,\ldots, m
\]

GSTAR model can be presented as a linear model so that the autoregressive parameter can be estimated by Least Square method. The model equation for location \(i\) can be written in linear regression as follows \(Z_{i} = \beta_{i} + u_{i}\).

The GSTAR model can be written in the form of the vector Autoregressive (VAR (1)) model as follows

\[
Z(t) = \Phi Z(t-1) + e(t)
\]

with \(\Phi = [\Phi_{10} + \Phi_{11} W]\) and \(\Phi_{10} = \begin{pmatrix} \phi_{00} & 0 & 0 \\ 0 & \phi_{20} & 0 \\ 0 & 0 & \phi_{30} \end{pmatrix}\), \(\Phi_{11} = \begin{pmatrix} \phi_{10} & 0 & 0 \\ 0 & \phi_{21} & 0 \\ 0 & 0 & \phi_{31} \end{pmatrix}\). As with the time series model, the first step is to identify the model, which is characterized by time and location or spatial
orders. Spatial orders are currently limited to order 1 because higher orders are difficult to interpret. As for temporal orders in accordance with the VAR or VARMA model by looking at the Matrix Autocorrelation Function (MACF) and Matrix Partial Autocorrelation Function (MPACF), then the appointment of orders using the Akaike Information Criteria (AIC) (Wutsqa and Suhartono, 2010),

$$AIC(i) = \ln \left( \frac{\hat{\sigma}_i^2}{T} \right) + \frac{2k_i^2}{T},$$

where $$k_i$$ is the number of parameters in the model and $$T$$ is the number of observations. Estimation of the GSTAR Model is done by presenting the model as a linear model with its autoregression parameters can be estimated by the Least Square method.

2.2. Generalized Linear Model

T2e Generalized Linear Model or GLM broadly analyzes data that has a non-normal distribution and is an exponential family. GLM is an extension of the classical linear model so that the shape of the linear model becomes the starting point for discussion of GLM (McCullagh and Nelder, 1989). McCullagh and Nelder (1989) approach the linear regression model in discussing GLM, making it easier to transition to GLM. Kendem and Fokianos (2002) define time series following GLM and divide GLM into 2 sections i.e random and systematic component.

1. Random component

The conditional distribution of response variables for the past is an exponential family of natural or canonical distributions. For $$t=1,\ldots,N$$

$$f(y_t; \delta_t, \phi|F_{t-1}) = \exp \left\{ \frac{y_t \delta_t - b(\delta_t)}{k_*(\phi)} + c(y_t; \phi) \right\}$$

(4)

With $$k_*(\phi)$$ is a parameter function which is a form $$\frac{\phi}{\sigma_t}$$, $$\phi$$ is a dispersion parameter and $$\sigma_t$$ is a parameter called weight or prior weight. The parameters $$\delta_t$$ are the parameters is called natural parameters of the distribution. $$F_{t-1}$$ is $$\sigma^{-}$$-field built from with is $$Y_{t-1}, Y_{t-2}, \ldots$$, and $$z_{t-1}, z_{t-2}, \ldots$$, a covariate process.

2. Systematic Components.

There is a monotone function $$g(.)$$ such that

$$g(\mu_t) = \eta_t = \sum_{j=1}^{r} \beta_j z_{(t-1)j} = Z_t^\top \beta$$

(5)

with $$\mu_t = E(Y_t|F_{t-1}), t=1,2,\ldots,T$$ . The function $$g(.)$$ is called the link function while $$\eta_t$$ is referred to as the linear predictor of the model.

2.3. Time series model for count data

Let a count time series by $$\{Y_t: t \in \mathbb{N}\}$$ and $$\{X_t: t \in \mathbb{N}\}$$ a time-varying r-dimensional covariate vector, so $$X_t = (X_{t,1}, \ldots, X_{t,r})^\top$$. Model of the conditional mean $$E(Y_t|F_{t-1})$$ of the count time series is called $$\lambda_t: t \in \mathbb{N}\}$$, such that $$E(Y_t|F_{t-1}) = \lambda_t$$. The count time series model say ARMA($$p,q$$) count data which is following GLM has a general form is [Liboschick, T., 2016]

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{r} \phi_k \tilde{g}(Y_{t-k}) + \sum_{l=1}^{q} \psi_l g(\lambda_{t-l}) + \eta^\top X_t$$

(6)

where $$g: \mathbb{R} \rightarrow \mathbb{R}$$ is a link function, $$\tilde{g}: \mathbb{N}_0 \rightarrow \mathbb{R}$$ is a transformation function and $$\eta=(\eta_1, \ldots, \eta_r)^\top$$ is the parameter vector is corresponded on effects of covariates. In concept of GLM, $$g(\lambda_t)$$ is called the linear predictor. Consider model ARMA (1,0) count data, with the logarithmic link function.
\(g(x) = \log(x), \tilde{g}(x) = \log(x + 1)\) so a log linear model of order 1 is obtained. If set \(v_t = \log\left(\lambda_t\right)\) then equation (5) to be equation (7) as follows,

\[v_t = \beta_0 + \phi_1 \log(Y_{t-1} + 1)\]

(7)

The function of transformation \(\tilde{g}(Y_{t+1}) = \log(Y_{t+1} + 1)\) follow by Fokianos and Tjoostheim (2019), because they are on the same scale a the linear predictor \(v_t\). So far, we have only specified the mean of but not its distribution.

2.4. GARMA model

The special model of the count time series model is proposed by Benyamin et al (2003) is the GARMA model. GARMA model develops the count time series model by requiring that the distribution follow an exponential family and adding a regression covariate according to GLM rules. This model can accommodate non-stationary behavior that might arise due to the influence of exogenous variables and also through the use of mixed models, enabling easier parameterization than the pure autoregressive process or pure moving average used in the previous observation approach (Benyamin, et al, 2003). The GARMA model is generally defined as follows

\[g(\mu_t) = \eta_t = x_t^T \beta + \tau_t\]

(8)

where \(\tau_t = \sum_{j=1}^{p} \phi_j A(y_{t-j}, x_{t-j}, \beta) + \sum_{j=1}^{q} \theta_j M(y_{t-j}, \mu_{t-j})\), \(A\) and \(M\) are a functions that represents a form of autoregressive and moving average with \(\phi' = (\phi_1, \phi_2, \ldots, \phi_p)\) and \(\theta' = (\theta_1, \theta_2, \ldots, \theta_q)\), each of them is the autoregressive and moving average parameters. Equation (8) is too general to be applied, so that a parsimony and flexible sub-model is formed, then equation (9) is obtained as follows, (Benjamin, et al, 2003)

\[g(\mu_t) = \eta_t = X_t^T \beta + \sum_{j=1}^{p} \phi_j \left\{ g(y_{t-j}) - X_{t-j}^T \beta \right\} + \sum_{j=1}^{q} \theta_j \left\{ g(y_{t-j}) - \mu_{t-j} - \beta \right\}\]

(9)

Equations (8) and (9) are called the GARMA model \((p, q)\). The parameters \(\beta, \phi, \theta\) are estimated by Maximum Likelihood Estimation (MLE) method and completion of the optimization using the Iterative Reweighted Least Square (IRLS) approach.

If \(y_t\) has a Poisson distribution, \(y_t = \lambda_t, \mu_t\), then model (9) becomes to Poisson GARMA \((p,q)\) model and the model is defined as follow,

\[\ln(\mu_t) = \eta_t = X_t^T \beta + \sum_{j=1}^{p} \phi_j \left\{ \ln(y_{t-j}) - X_{t-j}^T \beta \right\} + \sum_{j=1}^{q} \theta_j \left\{ \ln(y_{t-j}) - \mu_{t-j} - \beta \right\}\]

with \(y_{t-j}^* = \max\{y_{t-j}, c\}, 0 < c < 1\). \(c\) as a threshold parameter in the equation (10). Link function for Poisson distribution is logarithmic function, so \(g(\mu_t) = \ln \mu_t\). Example of the Poisson GARMA model (1.0) based on equation (10) where is

\[\ln(\mu_t) = X_t^T \beta + \phi_1 \ln(y_{t-1}^*) - X_{t-1}^T \beta\]

(11)

The equation (11) is Poisson GARMA model is if \(p = 1\) and \(q = 0\)

3. Research Methodology

The formation of the GSTAR model with the Poisson distribution response variable is carried out in a manner similar to the formation of the GARMA model (1.0) as equation (11). To build the Poisson GSTAR model, the following steps are taken,
1. Create a GARMA (1.0) Poisson model as in equation (11) without predictors.

2. Create the GSTAR (1, 1) Poisson model by following the equation from step 1 and trying to adopt the Vector Auto Regressive or VAR (1) model.

4. GSTAR Poisson modelling

The Poisson GSTAR model that will be formed is the development of the GSTAR (1, 1) model for the response variable in the form of Poisson distributed data count. The GSTAR (1, 1) model as in equation (2) is a multivariate time series model, with multivariate here expressed for one variable in many locations and the interrelationships between locations expressed in the weight matrix. Consider to the Box Jenskin Autoregressive AR (1) model (1), **Z_t = \phi Z_{t-1} + e_t** , with **e_t ~ N(0, \sigma^2)** then the developing model for this model if the response variable is the count data and has the Poisson distribution then the Autoregressive model for the data count obtained by Kendem and Fokianos , 2002, Liboschick (2015) as equation (7). Meanwhile Zeger and Qaqish model in Kendem and Fokianos (2002), Benjamin et al (2003) and Liboschick et al. (2015) stated that the GARMA model (1.0) for data count is as follows

\[
\begin{align*}
  g(\mu_t) &= X^T \beta + \phi \left( g(Z_{t-1}) - X^T \beta \right) \\
  \text{or} \\
  g(\mu_t) &= \beta_0 + \phi g(y_{t-1})
\end{align*}
\]

where \( \beta_0 = c_1 - \phi c_1 \) with \( c_1 \) is a constant and \( g(\cdot) \) is a link function of the count data distribution. So, if \( Z_t \sim P(\theta) \) and link function of Poisson distribution is a natural logarithmic function, then equation (14) can be written as

\[
\ln(\mu_t) = \beta_0 + \phi \ln(Z_{t-1})
\]

Back to the GSTAR model (1, 1) a spatial time series model, where the GSTAR model in equation (2) can be formed as a structural model such as the VAR (1) model (equation (3)) and can also be structurally formed as a univariate model for each location \( i, i = 1, 2, \ldots, m \) as follows

\[
\begin{align*}
  Z_1(1) & = \begin{bmatrix} Z_t(0) & V_t(0) & \ldots & 0 & 0 \\
  Z_t(1) & V_t(1) & \ldots & 0 & 0 \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  Z_t(T) & V_t(T-1) & \ldots & 0 & 0 \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  Z_N(1) & 0 & 0 & \ldots & Z_N(0) \\
  Z_N(2) & 0 & 0 & \ldots & Z_N(1) \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  Z_N(T) & 0 & 0 & \ldots & Z_N(T-1)
\end{bmatrix} & = \begin{bmatrix} \phi_{01} \\
  \phi_{11} \\
  \vdots \\
  \phi_{0N} \\
  \phi_{1N} \\
  \vdots \\
  \phi_{TN} \\
  e_1(T) \\
  e_2(T) \\
  \vdots \\
  e_N(T)
\end{bmatrix}
\end{align*}
\]

with \( V_t = \sum_{j=1}^{N} W_{ij} Z_j(t) \).

Observing the development of the time series count data model and the structural form of the GSTAR model, the Poisson GSTAR model was formed for 3 locations.
In what follows we assume \( \{ Y_i = \{ y_{i,t} \}, i = 1,2,3, t = 1,2,\ldots,T \} \), denotes a 3-dimensional count time series. Let \( \{ \theta_i = (\theta_{i,t}), i = 1,2,3, t = 1,2,\ldots,T \} \) be the corresponding 3-dimensional intensity process and \( \theta_{i,t} = E (y_{i,t} | F_{i,t-1}) \), we assume \( y_{i,t} \sim P(\theta_{i,t}) \) and the log linear model for Poisson GSTAR (1,1) model for each location \( i, i = 1,2,3 \) as follow

For location 1:
\[
v_{1,t} = \beta_{01} + \phi_{01} \ln y_{1,t-1} + \phi_{11} (w_{12} \ln y_{2,t-1} + w_{13} \ln y_{3,t-1}),
\]
For location 2:
\[
v_{2,t} = \beta_{02} + \phi_{02} \ln y_{2,t-1} + \phi_{21} (w_{21} \ln y_{1,t-1} + w_{23} \ln y_{3,t-1}),
\]
For location 3:
\[
v_{3,t} = \beta_{03} + \phi_{03} \ln y_{3,t-1} + \phi_{31} (w_{31} \ln y_{1,t-1} + w_{32} \ln y_{2,t-1}).
\]

where \( v_{i,t} = \log (\theta_{i,t}); i = 1,2,3 \). So that Poisson GSTAR (1,1) model can be defined as follow,

\[
\begin{pmatrix}
  v_{1,t} \\
v_{2,t} \\
v_{3,t}
\end{pmatrix} = \begin{pmatrix}
  \beta_{01} & 0 & 0 \\
  \phi_{02} & 0 & 0 \\
  \phi_{03} & 0 & 0
\end{pmatrix} \ln \begin{pmatrix}
  y_{1,t} \\
y_{2,t} \\
y_{3,t}
\end{pmatrix} + \begin{pmatrix}
  \phi_{11} & 0 & 0 \\
  \phi_{21} & 0 & 0 \\
  \phi_{31} & 0 & 0
\end{pmatrix} \begin{pmatrix}
  w_{12} & w_{13} \\
  w_{21} & w_{23} \\
  w_{31} & w_{32}
\end{pmatrix} \ln \begin{pmatrix}
  y_{1,t} \\
y_{2,t} \\
y_{3,t}
\end{pmatrix}
\]

Generally, in matrix model (16) with \( m \) location can defined

\[
v = \beta_0 + [\Phi_0 + \Phi_1 W] \tilde{Y}^*,
\]

where
\[
v = \begin{pmatrix}
  v_{1,t} \\
v_{2,t} \\
v_{m,t}
\end{pmatrix}, v_{i,t} = \ln \theta_{i,t}, \tilde{Y}^* = \begin{pmatrix}
  \tilde{y}_{1,t-1} \\
  \tilde{y}_{2,t-1} \\
  \vdots \\
  \tilde{y}_{m,t-1}
\end{pmatrix}, \tilde{y}_{i,t-1} = \ln y_{i,t-1}, \text{ and } y_{i,t} = \max (y_{i,t}, \zeta), 0 < \zeta < 1.
\]

This equation (17) we called the Poisson GSTAR (1,1) model.

4.1. Simulation for the Poisson GSTAR (1,1) model.

The GSTAR () Poisson model simulation is carried out with the following steps, first determine the number of locations i.e \( m = 3 \) and the number of observations of \( T \) that change to \( T = 10, 50, 100, 300 \) and second, determine the parameters

\[
\Phi_0 = \begin{pmatrix}
  0.35 & 0 & 0 \\
  0 & 0.45 & 0 \\
  0 & 0 & 0.40
\end{pmatrix}, \Phi_1 = \begin{pmatrix}
  0.30 & 0 & 0 \\
  0 & 0.25 & 0 \\
  0 & 0 & 0.2
\end{pmatrix}, W = \begin{pmatrix}
  0.5 & 0 & 0.5 \\
  0.5 & 0.5 & 0
\end{pmatrix} \text{ and } c=0,5.
\]

\[
\beta_0 = (1 \ 1 \ 1)^T, \text{third determine the initial value of } \theta, \text{ i.e. } \theta^{(0)}, \theta^{(0)} = (10 \ 25 \ 23)^T. \text{Fourth Generate for } t = 1 \ i = 1,2,3, y_{i,t} \sim P(\theta^{(0)}_{i,t}), \text{ then calculate } \theta^{(i)} \text{ into the model (17) and so on until } t = T. \text{ Result of the simulation can see in Figure 1.}
\]
Figure 1 shows the same data patterns at each location and at the same number of observations, only a slight difference at the end of the observation. For $T = 10$ in Figure 1 (a) there are different patterns for the 3 locations, but $T$ the greater the data pattern the more the same. Stationarity of simulation increasingly seen as the greater $T$.

5. Conclusion
This Poisson GSTAR model is one of the alternative models, proposals for the space time model for data count which is a development of the Count data model that follows the GLM concept and time series count data model. This model is still in the stage of forming the model, while for other stages such as parameter estimation, test parameters and other model characteristics will be written in another paper.

References
[1] Andrade et al 2015 Bayesian GARMA models for count data. Commun Stat Case Study, data analysis, and applications, 1(4) pp192
[2] Benjamin M A et al 2003 Generalized Autoregressive Moving Average Models J Am Stat Assoc 98:461 214-223, DOI: 10.1198/016214503388619238.
[3] Cameron A C and Trivedi P K 1998 Regression analysis of count data Cambridge University
[4] Craigmile P F 2012 Spatio Temporal Modelling, School of Mathematics & Statistics, University of Glasgow, http://www.stat.osu.edu/~pfc/, accessed date 22 January 2014.

[5] De Jong P and Heller G Z 2008 Generalized Linear Models for Insurance Data, Cambridge University Press.

[6] Fokianos K et al 2019 Multivariat count Autoregression, Submitted to Bernoulli, accessed data 19 May 2019.

[7] Jun, R C et al 2006 Time series of count data: Modelling, Estimation and Diagnostic, *Journal Computational statistics & data analysis, Science Direct*, DOI: 10.1016/j.csda.2006.08.001.

[8] Kendem B and Folkianos K 2002 Regression Model for Time series Analysis John Willey & Sons,Inc, New Jersey.

[9] Liboschick, T. et al 2016 tscout: An R Package for Analysis of count Time series Following Generalized Linear Models vignette of *R package Tscout* DOI:10.17877/DE 290R-7239

[10] McCullagh P and Nelder J A 1989 Generalized Linear Models Second Edition, Chapman and Hall, London.

[11] Ruchjana B N 2002 Suatu model generalisasi Space Time autoregresi dan penerapannya pada produksi minyak bumi Disertasi, Institut Teknologi Bandung.

[12] Wutsqa D U et al 2010 Generalized Space-Time Autoregressive Modeling. Proceedings of the 6th IMT-GT Conference on Mathematics, Statistics and its Application (ICMSA 2010). Universiti Tunku Abdul Rahman, Kuala Lumpur, Malaysia.