Bayesian Inference for a New Negative Binomial-Samade Model for Time Series Data Counts with Its Properties and Applications

SIRINAPA ARYUYUEN, ISSARAPORN THAIMSORN, UNCHALEE TONGGUMNEAD*
Department of Mathematics and Computer Science,
Rajamangala University of Technology Thanyaburi,
39 Moo 1, Klong 6, Khlong Luang, Pathum Thani 12110,
THAILAND

**Corresponding Author*

Abstract: - A new distribution was developed that mixed the negative binomial (NB) and Samade distributions, called the negative binomial-Samade (NB-SA) distribution. The properties of this distribution were studied, and the newly created distribution was applied using the framework of generalized linear models to build a time series data count model. The characteristics of overdispersion and heavy-tailed distribution of the count response variables were applied in the actual dataset modeling. Distribution parameters and the regression coefficient were estimated using a Bayesian approach. Results showed that the NB-SA model had significantly the highest efficiency compared with the classical NB and Poisson models for analyzing factors influencing the daily number of COVID-19 deaths in Thailand.

Key-Words: Mixed NB model, Time series data, Bayesian inference, Count data, regression model, COVID-19.

Received: December 27, 2022. Revised: June 9, 2023. Accepted: June 28, 2023. Published: July 27, 2023.

1 Introduction

Time series data counts occur as daily occurrences whenever several events are observed over time. Increased understanding of such data and extraction of information requires statistical analysis or modeling. Different data counts may possess characteristics that cannot be used with particular models. During the past decades, researchers have developed models to derive better conclusions from the data. An important framework to handle time series data counts involves generalized linear models (GLM). When serial dependence is incorporated through the so-called link function, this leads to a more flexible class of models where covariates are easily included, enabling the models to better explain the dynamics of available information and provide trustworthy predictions. A famous special case of the GLM is integer-valued generalized autoregressive conditional heteroscedastic (INGARCH) models, also known as a linear order model, [1], [2], [3]. Later, researchers developed log-linear order models (p,q), [4]. In most traditions, the distribution assumption Y_t given $\mathbb{F}_{t-1}$ assumes a Poisson distribution where the mean is equal to the variance, called equidispersion. However, this distribution usually leads to the problem that the variance of random variables is greater than the mean, called overdispersion or underdispersion if the variance of the random variable is less than the mean. The researchers solved the problem using the negative binomial (NB) distribution for overdispersion and underdispersion, respectively, [5], [6]. For cases of overdispersion, the NB distribution may solve the problem but some cases have a high probability of no event of interest, resulting in a higher frequency of zero data values. As a result, excessive overdispersion is aggravated, making the NB distribution unsuitable for this data type. Various methods have been employed to develop a new class of discrete distributions, such as mixed Poisson, [7], generalized Poisson, [8], zero-inflated generalized Poisson, [9], mixed Poisson-inverse-Gaussian, [10], and mixed NB distributions. Many mixed NB distributions have also been proposed, such as the NB-Lindley (NB-L), [11], NB-generalized exponential, [12], NB-gamma, [13], NB-Sushila, [14], NB-generalized Lindley, [15], NB-Quasi Lindley, [16], and NB-modified Quasi Lindley, [17] distributions. Mixed NB distributions are applied to statistical model events for data counts in real life, such as actuarial and insurance models, [11], [13], [18], medical or industrial models, [16], [17], [18], and ecology and biodiversity, [19].

This paper developed a new mixed negative binomial distribution that was then applied as the GLM for time series data counts. When developing the GLM for a proposed distribution, one important aspect is the methods used for parameter estimation. One method for modeling the GLM is the maximum likelihood (ML) method which estimates an assumed probability distribution given observed data. The ML estimation is achieved by maximizing a likelihood function so that the observed data is most probable under the assumed statistical model. The GLM has been developed for mixed NB distribution using the ML method to estimate parameters, such as the NB-beta Weibull, [20], and NB-inverse Gaussian, [21], regression models. However, the ML method provides only a point estimate which may need to be more robust or may even fail to converge when the sample size is small or when the dispersion parameter is much larger than the mean.

The GLM does not consider prior information, which may be helpful in the case of missing observations. As an alternative, Bayesian inference can account for prior expert knowledge on variables of interest, especially in a small sample set, by providing a sample of estimators that may be helpful for uncertainty analysis, [22]. Practical advantages of the Bayesian approach are its flexibility and generality to better cope with complex problems, [23], [25]. Recently, some researchers have favored the Bayesian approach over the ML method, [22], [25], while the GLM for mixed NB distribution using the Bayesian method has been developed such as the NB, [24], NB-Quasi Lindley, [16], NB-modified Quasi Lindley, [17], and NB-Sushila, [25], regression models. In the past, the most commonly used GLM model in regression analysis was the Poisson and NB regression model. However, the Poisson regression model has a limitation: if the data is overdispersion, it will result in such a model being inappropriate. Although the NB regression model was developed to solve the problem of overdispersion, in some events, there may be a very high probability that the event of interest will not occur at all, and this makes the data value zero has a higher frequency. As a result, the problem of overdistribution is more serious. These situations make the Pois and NB distributions unsuitable for data of this nature. At the same time, if the numerical data is collected continuously over time in the form of a time series, there are better approaches than GLM modeling based on regression analysis. Because the limitations of regression analysis are that the observations of each dependent variable must be independent, this study extended the GLM model to the GLM for time series count data with a new mixed negative binomial. This study proposed a new mixed NB distribution for time series data counts as a flexible alternative for analyzing heavy-tailed data with overdispersion. The Bayesian approach was used to estimate model parameters with the GLM framework applied to build the time series data counts using both internal and external covariates. This proposed model was constructed for time series data counts of COVID-19 deaths in Thailand from 1 January 2021 to 30 September 2022, [26].

2 Materials and Methods

2.1 Preliminary about Some Distribution

The Poisson distribution: Let Y be a random variable with a Poisson distribution with a parameter μ, denoted by $Y \sim \text{Pois}(\mu)$. Its probability mass function (pmf) as follows:

$$f(y;\mu)=\frac{\exp(-\mu)\mu^{y}}{y!} \text{ or } y=0,1,2,... \text{ and } \mu>0. \quad (1)$$

Its mean and variance are, respectively

$$\text{E}(Y)=\mu \text{ and } \text{Var}(Y)=\mu. \quad (2)$$

The NB distribution: Let Y be a random variable as distributed the NB distribution with parameters r and m, denoted by $Y \sim \text{NB}(r,m)$. Its pmf is given by:

$$f(y;r,m)=\binom{y+r-1}{y}m^{r}(1-m)^{y} \text{ for } y=0,1,2,..., \quad (3)$$

where $r>0$ and $0<m<1$. Its mean and variance are respectively:

$$\text{E}(Y)=r\left(\frac{1-m}{m}\right) \text{ and } \text{Var}(Y)=r\left(\frac{1-m}{m^{2}}\right). \quad (4)$$

The Samade distribution: In 2021, the Samade (SA) distribution was proposed by [27], for analyzing lifetime data. The SA distribution is derived from a mixture of the gamma (Gam) and exponential (Exp) distributions, i.e., $\text{Gam}(a,b)$ and $\text{Exp}(b)$. The SA probability density function (pdf) is obtained as:

$$g(\lambda;a,b)=\left(\frac{b^{4}}{b^{4}+6a}\right)g_{\text{Exp}}(\lambda)+\left(\frac{6a}{b^{4}+6a}\right)g_{\text{Gam}}(\lambda) \quad (5)$$

where $\lambda > 0$, $g_{\text{Exp}}(\lambda)$ and $g_{\text{Gam}}(\lambda)$ are the pdf of the Exp and Gam distributions. Its pdf can be written as:

$$g(\lambda;a,b) = \frac{b^4}{b^4+6a}\left(b+a\lambda^3\right)e^{-b\lambda} \qquad (6)$$

where $\lambda > 0$, a and b are the shape and scale parameters, respectively. When $a=1$ and $a=0$, the SA distribution reduces to the one-parameter Pranav and Exp distributions, respectively, [27].

2.2 Notations and Model Specifications for Time Series Data Counts

Time series data counts occur whenever several events are observed over time. The GLM works out the appropriate linear predictor link functions (mean function) and inverse link functions, [28], to describe the covariate effects attached to the dependence observations. If Y_t is a time series data count, the GLM of the time series data will be composed of random components, system components, and link functions. A random component is the conditional distribution of the response given the past belongs to the exponential family of distributions having canonical form as follows:

$$f(y_t \mid \mathbb{F}_{t-1};\theta_t,\phi) = \exp\left\{\frac{y_t\theta_t - w(\theta_t)}{\alpha_t(\phi)}\right\} + c(y_t;\phi), \qquad (7)$$

where $w(\theta_t)$ denotes the function of θ_t, $\alpha_t(\phi) = \phi / w(\theta_t)$, ϕ is a dispersion parameter and $c(y_t;\phi)$ is a prior weight. The systematic component is a monotone function that can be represented as:

$$g(\mu_t) = \sum_{k=1}^{p}\beta_j Y_{(t-i_k)} \text{ for } j = 0,1,2,...,p, \qquad (8)$$

where i_k represents a positive integer, $0 < i_1 < i_2 < \cdots < i_p < \infty$, and $g(\mu_t)$ is a link function, which represents the linear predictor of the model. The GLM equation can be written as follows:

$$\int_{-\infty}^{\infty} f(y_t \mid \mathbb{F}_{t-1};\theta_t,\phi)dy_t = 1. \qquad (9)$$

The mean and variance of $Y_t \mid \mathbb{F}_{t-1}$ are as follows:

$\mu_t = \mathrm{E}\left[Y_t \mid \mathbb{F}_{t-1}\right] = w'(\theta_t)$

(10) $\mathrm{Var}\left[Y_t \mid \mathbb{F}_{t-1}\right] = \alpha_t(\phi)w''(\theta_t) = \alpha_t(\phi)\mathrm{Var}(\mu_t)$,

where $w'(\theta_t)$ and $w''(\theta_t)$ are the first and second derivative of $w(\theta_t)$ respectively. Because $\mathrm{Var}(Y_t \mid \mathbb{F}_{t-1}) > 0$, it follows that $w'(\theta_t)$ is monotone. Therefore, $\theta_t = w^{-1}(\mu_t)$ and $w(\theta_t)$ is the monotone function of μ_t. Thus, the link function can be defined as Eq. (8).

The general form for time series data counts follows a generalized linear model as:

$$g(\mu_t) = \beta_0 + \sum_{k=1}^{p}\beta_k \tilde{g}(Y_{t-ik}) + \sum_{l=1}^{q}\alpha_l g(\mu_{t-jl}) + \tilde{\eta}^T\mathbf{X}_t \qquad (11)$$

where the response variable Y_t represents a time series data count, μ_t represents the mean process, $\mathbb{F}_{t-1}$ represents history up to time t, and $\theta_t = \tilde{\theta}$ represents the regression coefficients:

$$\tilde{\theta} = \left(\beta_0,\beta_1,...,\beta_p,\alpha_1,...,\alpha_q,\tilde{\eta}^T\right)^T \in \Theta \subseteq \mathbb{R}^{1+p+q+s} \qquad (12)$$

where $g(\cdot)$ represents the link function, g : $\mathbb{R}^+ \to \mathbb{R}$, $\tilde{g}$ represents the transformation function, and $\tilde{\eta}$ represents the parameter vector, $\tilde{\eta} = (\eta_1,\ldots,\eta_s)$ corresponding to the effects of covariates. A popular special case from Eq. (11) is known as INGARCH models of order p and q, abbreviated as the INGARCH(p,q). In this model, the distribution assumption on Y_t given $\mathbb{F}_{t-1}$ is distributed as Poisson, NB, etc., Eq. (11) with the logarithmic link function $g(y_{t-1}) = \log(y_{t-1})$, and $\tilde{g}(y_{t-k}) = \log(y_{t-k}+1)$, can be written as [29], i.e., $g(\tilde{\mu}_t) = \log \mu_t$:

$$g(\tilde{\mu}_t) = \beta_0 + \sum_{k=1}^{p}\beta_k \log(Y_{t-k}+1) + \sum_{l=1}^{q}\alpha_l\mu_{t-l}. \qquad (13)$$

From the above equation, the internal and external regressor effects of covariates $\tilde{\eta}$ can be determined by adding internal regressors $\mathbf{X}_t^{(I)}$, such as intervention, cos(.) and sin(.) functions to represent seasonal cycles. The conditional mean of Eq. (13) can then be expressed as:

$$g(\tilde{\mu}_t) = \beta_0 + \sum_{k=1}^{p}\beta_k \log(Y_{t-k}+1) + \sum_{l=1}^{q}\alpha_l\mu_{t-l} + \tilde{\eta}^T\mathbf{X}_t^{(I)}. \qquad (14)$$

To further include the external regressor $\mathbf{X}_t^{(E)}$, the mean of Eq. (14) can be expressed as:

$$g(\tilde{\mu}_t) = \beta_0 + \sum_{k=1}^{p}\beta_k \log(Y_{t-k}+1) + \sum_{l=1}^{q}\alpha_l\nu_{t-l} + \tilde{\xi}^T\mathbf{X}_t^{(I)} + \tilde{\eta}^T\mathbf{X}_t^{(E)}. \qquad (15)$$

2.3 The GLM for Time Series Data Counts

Suppose there are discrete time series count data Y_t with t $t \in \mathbb{N}$. The conditional mean $\mathrm{E}\left[Y_t \mid \mathbb{F}_{t-1}\right]$ from time series counts data, for example, μ_t. Then the general GLM for time series count data is as follows:

$$g(\mu_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-ik}) + \sum_{l=1}^{q} \alpha_l g(\mu_{t-jl}) + \tilde{\eta}^T \mathbf{X}_t.$$

With: g: $\mathbb{R} \rightarrow \mathbb{R}^+$ is the link function, g : $\mathbb{N}_0 \rightarrow \mathbb{R}$ is a transformation function, and a vector parameter is $\tilde{\eta} = \left(\eta_1, \ldots, \eta_s\right)$. In GLM $g(\mu_t)$ called the linear predictor, The regression can be used for the past time response variable, defined as $P = \{i_1, i_2, ..., i_p\}$ and i is an integer where $0 < i_1 < i_2, ..., < i_p < \infty$. In the GLM for time series count data, it is possible to regressor observed lag $Y_{t-i_1}, Y_{t-i_2}, ..., Y_{t-i_p}$. The same analogy with lag in observation defined Q where $Q = \{j_1, j_2, ..., j_q\}$ and j is an integer where $0 < i_1 < i_2, ..., < i_p < \infty$. For the regressor variable on the lag for the conditional mean $\mu_{t-l_1}, \mu_{t-l_2}, ..., \mu_{t-l_q}$

Assume the time series data count Y_t to be a random variable with the Poisson distribution denoted by $Y_t \mid \mathbb{F}_{t-1} \sim \mathrm{Pois}(\mu_t)$. The Poisson model for the time series data count can then be written as:

$$f(y_t \mid \mathbb{F}_{t-1}; \mu_t) = \frac{\exp(-\mu_t)\mu_t^{y_t}}{y_t!}, \text{for } y_t = 0,1,2,... \tag{16}$$

where $\mu_t = g^{-1}(\tilde{\mu}_t)$. The mean and variance of $\{Y_t \mid \mathbb{F}_{t-1}; \mu_t\}$ are, respectively

$$\mathrm{E}[Y_t \mid \mathbb{F}_{t-1}; \mu_t] = \mu_t \text{ and } \mathrm{Var}[Y_t \mid \mathbb{F}_{t-1}; \mu_t] = \mu_t. \tag{17}$$

Hence in the case of a conditional Poisson response model, the conditional mean is identical to the conditional variance of the observed process. However, if the response variable occurs the overdispersion problem arises and the Poisson response is unsuitable for modeling.

The NB distribution is an alternative when dependent variable problems arise, such as overdispersion. Let Y be a random variable with the NB distribution and pmf as shown in Eq. (3). We can parameterize m in the term of r as $m = r/(\mu + r)$ for μ as the mean response variable Y and r as the reciprocal (or inverse of a dispersion parameter $\phi : \phi = 1/r$). Therefore, the pmf of the NB distribution can be rewritten as:

$$f(y; r, \mu) = \frac{\Gamma(r+y)}{\Gamma(r)\Gamma(y+1)} \left(\frac{r}{\mu + r}\right)^r \left(\frac{\mu}{\mu + r}\right)^y \tag{18}$$

where $\Gamma(\cdot)$ as a complete gamma function, denoted by $Y \sim \mathrm{NB}(\mu, r)$.

Let Y_t be a response random variable in the NB distribution, denoted as $Y_t \mid \mathbb{F}_{t-1} \sim \mathrm{NB}(\mu_t, r)$. The NB model for time series data counts then become

$$f(y_t \mid \mathbb{F}_{t-1}; r, \mu_t) = \frac{\Gamma(r+y_t)}{\Gamma(r)\Gamma(y_t+1)} \left(\frac{r}{\mu_t + r}\right)^r \left(\frac{\mu_t}{\mu_t + r}\right)^{y_t} \tag{19}$$

where $y_t = 0,1,2,...,$ and $\mu_t = g^{-1}(\tilde{\mu}_t)$. The mean and variance of $\{Y_t \mid \mathbb{F}_{t-1}; \mu_t, r\}$ are:

$$\mathrm{E}[Y_t \mid \mathbb{F}_{t-1}; \mu_t, r] = \mu_t \tag{20}$$

and $\mathrm{Var}[Y_t \mid \mathbb{F}_{t-1}; \mu_t, t] = \mu_t + \frac{\mu_t^2}{r}.$

However, even if the NB distribution can solve the problem of overdispersion, in some cases there are problems of heavy-tailed distribution and the NB response is not suitable for modeling.

2.4 Criteria for Model Evaluation

Three criteria were used to compare the performance of model suitability as the deviance, p_D, and the deviance information criterion (DIC). The DIC is a hierarchical modeling generalization of the Akaike information criterion, which is often and widely used as a goodness-of-fit measure using the Bayesian approach. The DIC is beneficial to Bayesian model comparison problems when posterior distributions have been obtained by the Markov chain Monte Carlo (MCMC) simulation. The model has the smallest value of DIC, and p_D, is the best model, [30], [31], [32]. Let $\mathrm{D}(\mathbf{\Omega}) = -2\log \mathrm{L}(y \mid \mathbf{\Omega})$ be the deviance, where $\mathrm{L}(y \mid \mathbf{\Omega})$ are the likelihood function and the conditional joint pdf of observations are given by unknown parameters. We have $\mathrm{DIC} = \bar{D}(\mathbf{\Omega}) + p_D$, $\bar{D}(\mathbf{\Omega}) = \mathrm{E}\left[-2\log \mathrm{L}(y \mid \mathbf{\Omega})\right]$ and $p_D = \mathrm{Var}\left[\mathrm{D}(\mathbf{\Omega})\right]/2$.

The probability integral transform (PIT) is a tool for assessing the probability calibration of the predictive distribution. This follows a uniform distribution if the predictive distribution is correct.

The shape of the PIT histograms suggests the calibration accuracy of the predictive distribution. A convex shape indicates an underdispersed predictive distribution, whereas concave histograms refer to overdispersed predictive distributions, [14].

3 Results and Discussion

3.1 A New Mixed NB Distribution

Definition 1: Let Y be a random variable distributed as the NB distribution with parameters $r>0$ and $m=\exp(-\lambda)$ where λ is distributed as the SA distribution with parameters $a>0$ and $b>0,$ i.e., $Y\sim \text{NB}(r,m=\exp(-\lambda))$ and $\lambda\sim \text{SA}(a,b).$ Then a random variable $Y\,|\,\lambda$ follows a negative binomial-Samade (NB-SA) distribution with parameters r, a and $b,$ denoted by $Y\sim \text{NB-SA}(r,a,b).$

Theorem 1: Let $Y\sim \text{NB-SA}(r,a,b),$ then its pmf is

$$f(y;r,a,b)=\binom{y+r-1}{y}\sum_{j=0}^{y}\binom{y}{j}(-1)^j \times\frac{b^5(b+r+j)^3+ab^4}{(b^4+ab)(b+r+j)^4}, \tag{21}$$

where $y=0,1,2,...,\ r>0,\ a>0$ and $b>0.$

Proof: Let $Y\sim \text{NB}(r,m=e^{-\lambda})$ and $\lambda\sim \text{SA}(a,b),$ then the marginal pmf of Y can be obtained using the expression:

$$f(y;r,a,b)=\int_0^\infty f(y\,|\,r,\lambda)g(\lambda;a,b)d\lambda,$$

where $g(\lambda;a,b)$ is the pdf of the SA distribution as Eq. (5), and $f(y\,|\,r,\lambda)$ is the NB's pmf as Eq. (3),

$$f(y\,|\,r,\lambda)=\binom{y+r-1}{y}e^{-r\lambda}\left(1-e^{-\lambda}\right)^y,$$

where $\left(1-e^{-\lambda}\right)^y=\sum_{j=0}^{y}\binom{y}{j}(-1)^j e^{-\lambda j},$ and

$$\begin{aligned} f(y;r,a,b) &=\binom{y+r-1}{y}\sum_{j=0}^{y}\binom{y}{j}(-1)^j\int_0^\infty e^{-\lambda(r+j)}g(\lambda;a,b)d\lambda \\ &=\binom{y+r-1}{y}\sum_{j=0}^{y}\binom{y}{j}(-1)^j \text{M}_\lambda\left[-(r+j)\right], \end{aligned}$$

where $\text{M}_\lambda\left[-(r+j)\right]$ is the moment generating function (mgf) of the SA distribution, which is shown in Lemma 1.

Lemma 1: Let $\lambda\sim \text{SA}(a,b),$ then its mgf is:

$$\text{M}_\lambda(s;a,b)=\frac{b^5(b-s)^3+ab^4}{(b^4+6a)(b-s)^4}\ \text{for}\ s=0,1,2,..., \tag{22}$$

$a>0$ and $b>0.$

Proof: If $\lambda\sim \text{SA}(a,b)$, then its mgf is

$$\begin{aligned} \text{M}_\lambda\left[-(r+j)\right] &=\frac{b^4}{b^4+6a}\int_0^\infty e^{-(b-s)\lambda}(b+a\lambda^3)d\lambda \\ &=\frac{b^4}{b^4+6a}\left[b\int_0^\infty e^{-(b-s)\lambda}d\lambda+a\int_0^\infty \lambda^3 e^{-(b-s)\lambda}d\lambda\right] \\ &=\frac{b^4}{b^4+6a}\left[\frac{b}{b-s}+\frac{a}{(b-s)^4}\right]=\frac{b^4\left[b(b-s)^3+a\right]}{(b^4+6a)(b-s)^4}. \end{aligned}$$

Thus, the pmf of the NB-SA distribution can be written as Eq. (21). The NB-SA distribution sub-models follow the NB-Pranav (for $a=1$) and the NB-Exp (for $a=0$) distributions. The pmf's shapes for the NB-SA distributions are provided in Figure 1. Some basic properties of the proposed distribution are obtained from the factorial moment as follows.

Theorem 2: Let $Y\sim \text{NB-SA}(r,a,b),$ then its k^{th} factorial moment is

$$\mu'_{[k]}=\frac{\Gamma(r+k)}{\Gamma(r)}\sum_{j=0}^{k}\binom{k}{j}(-1)^j\frac{b^5(b-k+j)^3+ab^4}{(b^4+ab)(b-k+j)^4}, \tag{23}$$

where $k=1,2,3,...$ and $r,a,b>0.$

Proof: Let $Y\sim \text{NB}(r,m),$ then its factorial moment is

$$\mu'_{[k]}=\text{E}\left[Y(Y-1)\cdots(Y-k+1)\right]=\frac{(1-m)^k\Gamma(r+k)}{m^k\Gamma(r)},$$

where $k=1,2,3,....$ For $Y\,|\,\lambda\sim \text{NB}(r,m=e^{-\lambda})$ and $\lambda\sim \text{SA}(a,b)$, we have

$$\mu'_{[k]}=\text{E}_\lambda\left[\frac{\Gamma(r+k)}{\Gamma(r)}\frac{(1-e^{-\lambda})^k}{e^{-\lambda k}}\right]=\frac{\Gamma(r+k)}{\Gamma(r)}\text{E}_\lambda\left[(e^\lambda-1)^k\right],$$

Using a binomial expansion for the term $(e^{-\lambda}-1)^k,$ the $\mu'_{[k]}$ can be written as:

$$\begin{aligned} \mu'_{[k]} &=\frac{\Gamma(r+k)}{\Gamma(r)}\sum_{j=0}^{k}\binom{k}{j}(-1)^j\text{E}_\lambda\left[e^{\lambda(k-j)}\right] \\ &=\frac{\Gamma(r+k)}{\Gamma(r)}\sum_{j=0}^{k}\binom{k}{j}(-1)^j\int_0^\infty e^{\lambda(k-j)}g(\lambda;a,b)d\lambda \\ &=\frac{\Gamma(r+k)}{\Gamma(r)}\sum_{j=0}^{k}\binom{k}{j}(-1)^j\text{M}_\lambda\left[k-j\right]. \end{aligned}$$

From $\text{M}_\lambda\left[k-j\right]$ is the mgf of the SA distribution in (22) for $s=k-j$, we have the k^{th} factorial moment as follows

$$\mu'_{[k]} = \frac{\Gamma(r+k)}{\Gamma(r)}\sum_{j=0}^{k}\binom{k}{j}(-1)^j \frac{b^5(b-k+j)^3+ab^4}{(b^4+ab)(b-k+j)^4}.$$

Based on the k^{th} factorial moment in Theorem 2, the first two factorial moments can be written as:

$$\mu'_{[1]} = \mathrm{E}(Y) = r\left(\delta_1 - \delta_0\right),$$

$$\mu'_{[2]} = \mathrm{E}(Y^2) - \mu'_{[1]} = r(r+1)\left(\delta_2 - 2\delta_1 + \delta_0\right),$$

where $\delta_j = \dfrac{b^5(b-j)^3+ab^4}{(b^4+ab)(b-j)^4}$ for $j = 0,1,2.$

Its mean and variance are: $\mathrm{E}(Y) = r\left(\delta_1 - \delta_0\right)$ and $\mathrm{Var}(Y) = r^2\left[\delta_2 - (2+\delta_1)\delta_1 + (1+2\delta_1-\delta_0)\delta_0\right] + r(\delta_2 - \delta_1).$

The proposed distribution can fit overdispersed and underdispersed data or both and the index of dispersion (ID) needs to be found. The ID is

$$\mathrm{ID} = \frac{r\left[\delta_2 - (2+\delta_1)\delta_1 + (1+2\delta_1-\delta_0)\delta_0\right] + \left(\delta_2 - \delta_1\right)}{\delta_1 - \delta_0}.$$

3.2 A New Mixed NB Distribution

Let $Y_t \sim \text{NB-SA}(r,a,b)$ then the GLM for the time series data count is developed with $\{Y_t \mid \mathbb{F}_{t-1};\mathbf{\Omega}\}$ where $\mathbf{\Omega} = (r,a,b,\tilde{\theta})$ as:

$$\mathrm{P}\left(Y_t \mid \mathbb{F}_{t-1};\mathbf{\Omega}\right) = \int_0^\infty f(y_t;\lambda\mu_t,r)g(\lambda;a,b)d\lambda = \frac{\Gamma(y_t+r)}{\Gamma(y_t+1)\Gamma(r)} \times \int_0^\infty \left(\frac{r}{\lambda\mu_t+r}\right)^r \left(\frac{\mu_t}{\lambda\mu_t+r}\right)^{y_t} \left(\frac{b^4(b+a\lambda^3)}{b^4+6}\right) e^{-b\lambda}d\lambda,$$

where $r>0$, $\mu_t = g^{-1}(\tilde{\mu}_t) > 0$ and $\tilde{\theta}$ defines as Eq. (12). Its mean and variance are

$$\mathrm{E}\left(Y_t \mid \mathbb{F}_{t-1};\mathbf{\Omega}\right) = \mu_t \mathrm{E}(\lambda), \tag{24}$$

and $\mathrm{Var}\left(Y_t \mid \mathbb{F}_{t-1};\mathbf{\Omega}\right) = \mu_t\mathrm{E}(\lambda) + \mu_t^2\left((1+r)/r\right)\mathrm{E}(\lambda^2) - \left[\mu_t\mathrm{E}(\lambda)\right]^2$, where $\mathrm{E}(\lambda)$ and $\mathrm{E}(\lambda^2)$ are the first and second moments of the original SA random variable, [27],

$$\mathrm{E}(\lambda) = \frac{24a+b^4}{6ab+b^5} \text{ and } \mathrm{E}(\lambda^2) = \frac{120a+2b^4}{6ab^3+b^6}. \tag{25}$$

A flow diagram of the steps creating the time series model for the NB-SA distribution is shown in Figure 2 (Appendix). The vector of unknown parameters $\mathbf{\Omega}$ is customarily estimated using the Bayesian approach, which allows consideration of previous information for parameter estimation.

3.3 Bayesian Inference of the NB-SA Model for Time Series Data Counts

The Bayesian framework for the NB-SA model was proposed for time series data counts, and created based on the likelihood function, prior distribution, and posterior distribution, denoted respectively by $\mathrm{L}(Y_t \mid \mathbb{F}_{t-1};\mathbf{\Omega})$, $\pi(\mathbf{\Omega})$, and $p(\mathbf{\Omega} \mid y)$. The Bayesian NB-SA model is $y_t \mid \mu_t \sim \text{NB-SA}(r,a,b,\tilde{\theta})$, where $\mu_t = g^{-1}(\tilde{\mu}_t)$ and $g(\tilde{\mu}_t)$ denoted in Figure 2 (Appendix).

The likelihood function of $\{Y_t \mid \mathbb{F}_{t-1};\mathbf{\Omega}\}$ is:

$$\mathrm{L}\left(Y_t \mid \mathbb{F}_{t-1};\mathbf{\Omega}\right) = \prod_{t=1}^{n} \frac{\Gamma(y_t+r)}{\Gamma(y_t+1)\Gamma(r)} \int_0^\infty \left(\frac{r}{\gamma\mu_t+r}\right)^r \times \left(\frac{\mu_t}{\gamma\mu_t+r}\right)^{y_t} \left[\frac{b^4\left(b+a\lambda^3\right)}{b^4+6a}\right] e^{-b\lambda}d\lambda. \tag{26}$$

This function can be executed using the representation of the hierarchical model implicit in the integral and the definition of the SA distribution. The SA distribution is a mixture between the $\mathrm{Exp}(b)$ and $\mathrm{Gam}(a,b)$ distributions with the pdf as Eq. (5), while the NB-SA distribution is conditional on the unobserved site-specific frailty term λ, which describes the additional heterogeneity, [33].

Consequently, the hierarchical framework can be represented as:

$$\mathrm{P}\left(Y_t \mid \mathbb{F}_{t-1};\mu_t,r \mid \lambda\right) = f_{\mathrm{NB}}(y_t \mid \lambda\mu_t,r),$$

$$\lambda \sim \left(\frac{b^4}{b^4+6a}\right)g_{\mathrm{Exp}}(\lambda) + \left(\frac{6a}{b^4+6a}\right)g_{\mathrm{Gam}}(\lambda).$$

In Bayesian inference, the prior distribution plays a defining role in estimating the unknown parameters in any distribution. This model contains all unknown parameters. Accordingly, under the squared error loss function, the Bayesian estimator of $\mathbf{\Omega}$ will be $\mathrm{E}\left(\mathbf{\Omega} \mid y_t\right)$. In the GLM context, the most frequently used informative prior distribution is the normal distribution, [34]. We define the prior distribution of parameters r, a and b as the gamma distribution,

$r \sim \mathrm{Gam}(\tau_r,\gamma_r)$, $a \sim \mathrm{Gam}(\tau_a,\gamma_a)$, and
$b \sim \mathrm{Gam}(\tau_b,\gamma_b)$.

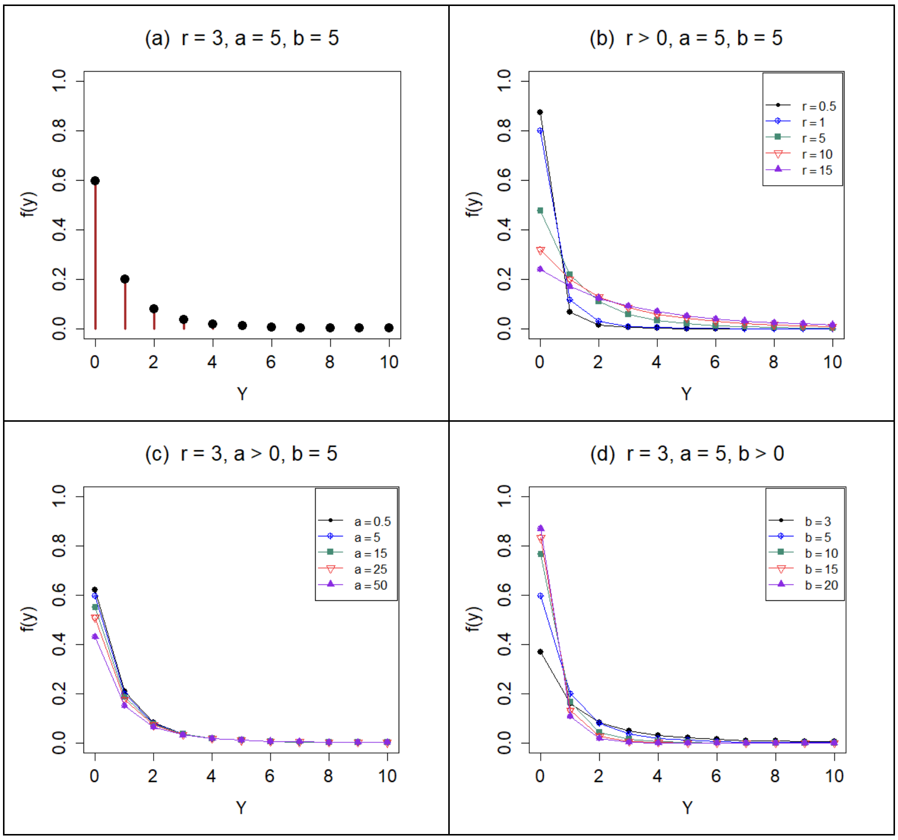


Fig. 1: The pmf plots of the NB-SA distribution with specified parameter values

Let prior distribution for $\tilde{\theta}$ be the normal (N) distribution, denoted by $\tilde{\theta} \sim \mathrm{N}(\mu_{\tilde{\theta}}, \sigma^2_{\tilde{\theta}})$ for the positive real values of τ_r, γ_r, τ_a, γ_a, τ_b, γ_b, $\mu_{\tilde{\theta}}$, and $\sigma^2_{\tilde{\theta}}$ are known or fixed. Suppose that $\mu_{\tilde{\theta}}$ is a $(k \times 1)$ hyper-parameter vector and $\sigma^2_{\tilde{\theta}}$ is a $(k \times k)$ known non-negative specific matrix, while k is the number of the regression coefficient. Each parameter is supposed to be independently distributed, and the joint prior distribution of all unknown parameters can be written as:

$$\pi(\mathbf{\Omega}) = \pi(r)\pi(a)\pi(b)\pi(\tilde{\theta}). \qquad (27)$$

The posterior distribution will integrate the sample information from the likelihood function as Eq. (26), with accessible parameter information from the prior distribution as Eq. (27). The posterior distribution can then be derived as follows:

$$p(\mathbf{\Omega} \,|\, \mathbf{y}) \propto L\left(Y_t \,|\, \mathbb{F}_{t-1}; \mathbf{\Omega}\right)\pi(r)\pi(a)\pi(b)\pi(\tilde{\theta}). \qquad (28)$$

Since the posterior distribution does not have an explicit form, a computational method called the Gibbs sampler is used in this study. The best known MCMC sampling algorithm was applied to find $\mathrm{E}\left(\mathbf{\Omega} \,|\, \mathbf{y}\right)$. By setting some initial value, the Gibbs sampler algorithm will randomly walk through parameter space. The basic scheme Gibbs sampler is given as follows, [35], [36].

- **Step 0.** Choose an arbitrary starting point $\mathbf{\Omega}^{(0)}$.
- **Step 1.** Generate $\mathbf{\Omega}^{(j+1)}$ as follows:

Generate $r^{(j+1)} \sim p(r \,|\, a^{(j)}, b^{(j)}, \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_k^{(j)}, \mathbf{y});$
Generate $a^{(j+1)} \sim p(a \,|\, r^{(j)}, b^{(j)}, \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_k^{(j)}, \mathbf{y});$
Generate $b^{(j+1)} \sim p(b \,|\, r^{(j)}, a^{(j)}, \theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_k^{(j)}, \mathbf{y});$
Generate $\theta_1^{(j+1)} \sim$
$p(\theta_1 \,|\, r^{(j)}, a^{(j)}, b^{(j)}, \theta_2^{(j)}, \ldots, \theta_k^{(j)}, \mathbf{y}); \vdots$
Generate $\theta_k^{(j+1)} \sim p(\theta_k \,|\, r^{(j)}, a^{(j)}, b^{(j)}, \theta_1^{(j)}, \ldots, \theta_{k-1}^{(j)}, \mathbf{y});$

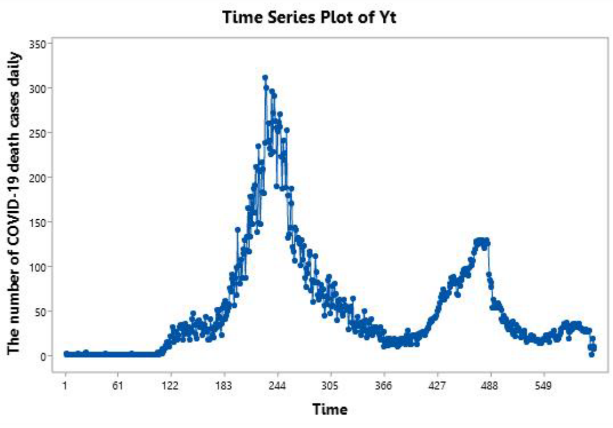
- **Step 2.** Set $j = j + 1$ and go to step 1.

In this study, the model parameters of $\mathbf{\Omega}$ can be estimated from the Bayesian method using the JAGS (Just Another Gibbs Sampler) as an implementation of an MCMC algorithm called Gibbs sampling to sample the posterior distribution of a Bayesian model. In this paper, the expected posterior of parameters is calculated using the JAGS function in the R2jags package of the R language, [32], [37]. Factors influencing the number of COVID-19 cases resulting in death included the daily number of COVID-19 cases in Thailand from

1 January 2021 to 30 September 2022, comprising 607 days, [26]. Figure 3 shows the number of daily COVID-19 deaths in Thailand.

In Figure 3(a) the time series motion is in the form of sine and cosine waves. When the time series reaches its peak, the value of each cycle decreases in the form of a transient shift. Therefore, this study considered internal covariates as sine and cosine functions and transient shift interventions. The external covariate was the number of new COVID-19 infection cases. Random variables used in this study were as follows:

- Y_t is time series data of the number of daily COVID-19 deaths (unit: people) where $t = 1,...,607$.
- X_t is the number of new COVID-19 infection cases (unit: people), with the mean and standard deviation as 7638.90 and 7548.74, respectively (minimum, median, and maximum as 26, 4563, and 28379). From the actual data, the mean and variance of the number of COVID-19 deaths (unit: people) were 52.88 and 3,646.61, respectively (minimum, median, maximum, and standard deviation as 0, 31, 312, and 60.39, respectively). Since the variance of COVID-19 deaths was greater than the mean, this dataset had an overdispersion problem. The histogram (Figure 3 (b)) and normal Q-Q plot (Figure 3(c)) also showed that the data had a heavy-tailed distribution. Therefore, the new NB-SA distribution was developed to solve these problems and a GLM was created for the NB-SA distribution. The procedure for creating the model is shown in Figure 2.



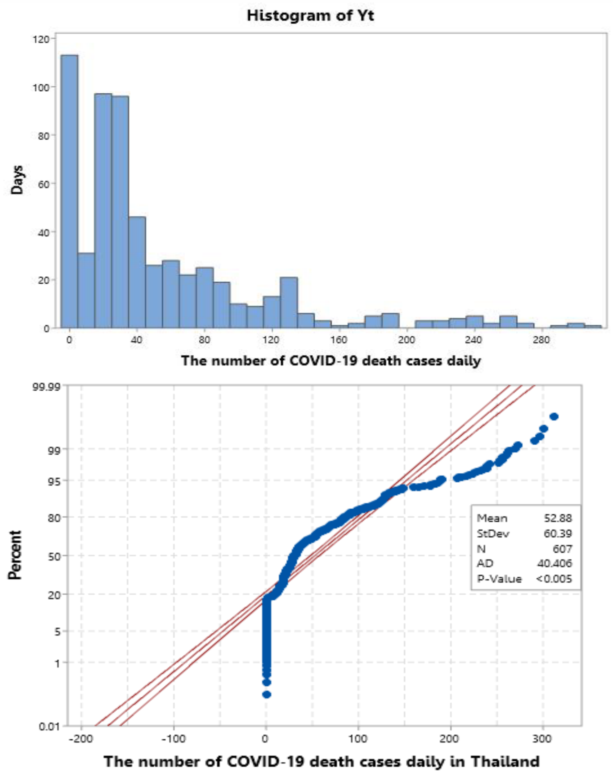


Fig. 3: Empirical data: (a) Plot of the daily number of COVID-19 deaths in Thailand, (b) Bar chart of the daily number of COVID-19 deaths in Thailand, and (c) The normal Q-Q plot of the number of COVID-19 deaths in Thailand.

3.4 Data Analysis Results

From the time series data of the number of daily COVID-19 deaths, P and Q initial values were defined by forming an ACF and PACF plot specifying the optimal combination based on the smallest AIC value. The results showed that NB-SA, NB, and Poisson gave the minimum AIC when P=1 and Q=1. Figure 4(c) shows the optimal combination of P and Q of the NA-SA. The result shows that when P=1, Q=1 yields an AIC of 322.30. These P and Q values were used for modeling, using the NB-SA, Poisson, and NB distributions by adding internal and external covariate effects and estimating model parameters with the Bayesian approach by defining the model as:

$$\hat{Y}_t = \exp\left\{\hat{\beta}_0 + \hat{\beta}_1 \log\left(Y_{t-1}+1\right) + \hat{\alpha}_1\upsilon_{t-1} + \hat{\xi}_1 X_t + \hat{\eta}_1 I_1 + \hat{\eta}_2 I_2 + \hat{\eta}_3 F_1 + \hat{\eta}_4 F_2\right\} \tag{29}$$

where

- I_1 is the intervention with transient shift for

the time $t \geq 230$ for $I_1 = 0.8^{t-230} I_{230}$, for $I_{230} = 1$ if $(\mathrm{t} \geq 230)$, otherwise $I_{230} = 0$.

- I_2 is the intervention with transient shift for

the time $t \geq 483$ where $I_2 = 0.8^{t-483} I_{483}$, for $I_{483} = 1$ if $(\mathrm{t} \geq 483)$, otherwise $I_{483} = 0$.

- F_1 is the internal covariate $\sin(2\pi t/365)$.
- F_2 is the internal covariate $\cos(2\pi t/365)$.

After developing a new distribution, the NB-SA distribution was studied. Its GLM framework was applied with time series data counts of the daily number of COVID-19 deaths to build the GLM derived from the NB-SA distribution. The time series data count of Y_t was provided in the NB-SA model. The Bayesian approach estimated the model parameters as the posterior means (estimates), standard error (s.e.), 95% credible intervals (Cr.I.) of each parameter, and statistics for comparing model performance, such as the deviance, DIC, and p_D of the Poisson, NB, and NB-SA models, as shown in Table 1. The study variable X_t was standardized to a standard score, i.e., $Z_t = (X_t - 7638.90)/7548.74$, represents the external covariate influenced by the daily number of COVID-19 deaths. Transient intervention shifts for time $t \geq 230$ and $t \geq 483$. The F_1 and F_2 are internal covariate effects due to the data because there was a surge in the number of daily COVID-19 deaths in Thailand (Y_t).

Based on these prior distributions, three parallel independent MCMC chains were generated with 300,000 iterations for each parameter, discarding the first 150,000 iterations as a burn-in for computation. Results indicated that the deviance, DIC and p_D values of the NB-SA model were smallest compared with the Poisson and NB models. The density plots and trace plots of the three MCMC chains with the MCMC plots package in R, [37], from the NB-SA model are shown in Figure 5 and Figure 6, respectively. Results showed that the density plots of all parameters in three parallel chains overlapped well after the burn-in period, while the trace plots showed that graphs of the values of simulated parameters against the drawn lines were almost vertical and dense. The motion of the trace plots revealed characteristics of a converged manner, and the sequence was stable. Therefore, the NB-SA model could be fitted for this dataset. Results of the GLM for time series data count models are shown in Table 1, with the estimated parameters r, a and b for the NB-SA model $\hat{r} = 31.846 (\text{s.e.} = 0.149)$, $\hat{a} = 1.678$ $(\text{s.e.} = 0.2626)$ and $\hat{b} = 11.183 (\text{s.e.} = 0.802)$. From Eq. (25), we have $\mathrm{E}(\lambda) = 0.0896$. The daily number of COVID-19 deaths in Thailand with NB-SA distribution can be represented as:

$$\begin{aligned}\hat{Y}_{t,\text{NB-SA}} &= 0.0896 \exp\{1.619 + 0.940 \log(Y_{t-1} + 1) \\ &\quad +0.001\upsilon_{t-1} + 0.076 Z_t - 0.010 I_1 + 0.008 I_2 \\ &\quad -0.059 F_1 - 0.074 F_2\}.\end{aligned}$$

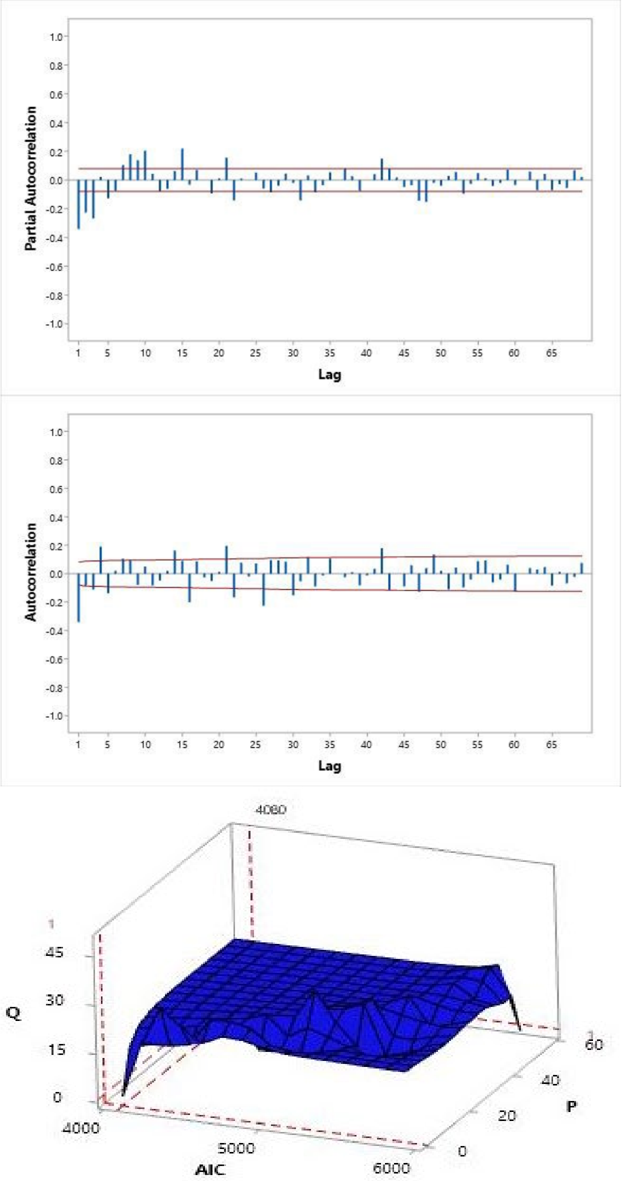


Fig. 4: (a) Partial Autocorrelation (b) Autocorrelation (c)The optimal combination based on the smallest AIC value for the NB-SA model.

Table 1. Parameter estimates and important statistics of the Poisson, NB and NB-SA models

Models		Estimate (s.e.)	95% Cr.I.
Poisson	β_0	0.450 (0.0022)	(0.344, 0.556)
	β_1	0.852 (0.0006)	(0.823, 0.882)
	α_1	0.001 (0.0001)	(0.000, 0.002)
	ξ_1	0.120 (0.0005)	(0.097, 0.142)
	η_1	0.047 (0.0017)	(-0.032, 0.130)
	η_2	0.123 (0.0026)	(-0.003, 0.241)
	η_3	-0.096 (0.0005)	(-0.120, -0.072)
	η_4	-0.100 (0.0006)	(-0.129, -0.071)
			Deviance = 4,566.2, DIC = 9,054.7, $p_D = 4{,}488.6$
NB	β_0	0.219 (0.0038)	(-0.047, 0.322)
	β_1	0.940 (0.0010)	(0.887, 0.987)
	α_1	0.001 (0.00004)	(0.000, 0.002)
	ξ_1	0.075 (0.0009)	(0.035, 0.116)
	η_1	-0.013 (0.0049)	(-0.247, 0.230)
	η_2	0.011 (0.0054)	(-0.244, 0.282)
	η_3	-0.059 (0.0009)	(-0.105, -0.016)
	η_4	-0.074 (0.0009)	(-0.118, -0.032)
			Deviance = 4,127.7, DIC = 4,137.4, $p_D = 9.7$
NB-SA	β_0	1.619 (0.0562)	(-0.868, 4.403)
	β_1	0.940 (0.0010)	(0.893, 0.988)
	α_1	0.001 (0.00004)	(0.000, 0.002)
	ξ_1	0.076 (0.0008)	(0.036, 0.115)
	η_1	-0.010 (0.0049)	(-0.243, 0.231)
	η_2	0.008 (0.0054)	(-0.257, 0.264)
	η_3	-0.059 (0.0009)	(-0.104, -0.014
	η_4	-0.074 (0.0009)	(-0.116, -0.029)
			Deviance = 4,127.6, DIC = 4,136.4, $p_D = 8.7$

The deviance, DIC, and p_D of the NB-SA model were 4127.6, 4136.4, and 8.7, respectively (Table 1). Results indicated that the average of Y_t was influenced by the number of Y_t on the previous day. The average daily number of COVID-19 deaths in Thailand was also influenced by the number of υ_{t-1} on the previous day. To consider the traditional NB distribution, we have $\hat{r} = 31.909 (\text{s.e.} = 0.152)$ and $\mathrm{E}(Y_i \mid \mathbb{F}_{t-1}) = \mu_t$. Consequently, the GLM for the time series data count approach with the NB distribution can be represented as:

$$\hat{Y}_{t,\mathrm{NB}} = \exp\{0.129 + 0.940\log(1+Y_{t-1}) + 0.001\upsilon_{t-1}$$
$$+0.075Z_t - 0.013I_1 + 0.011I_2$$
$$-0.059F_1 - 0.074F_2\}.$$

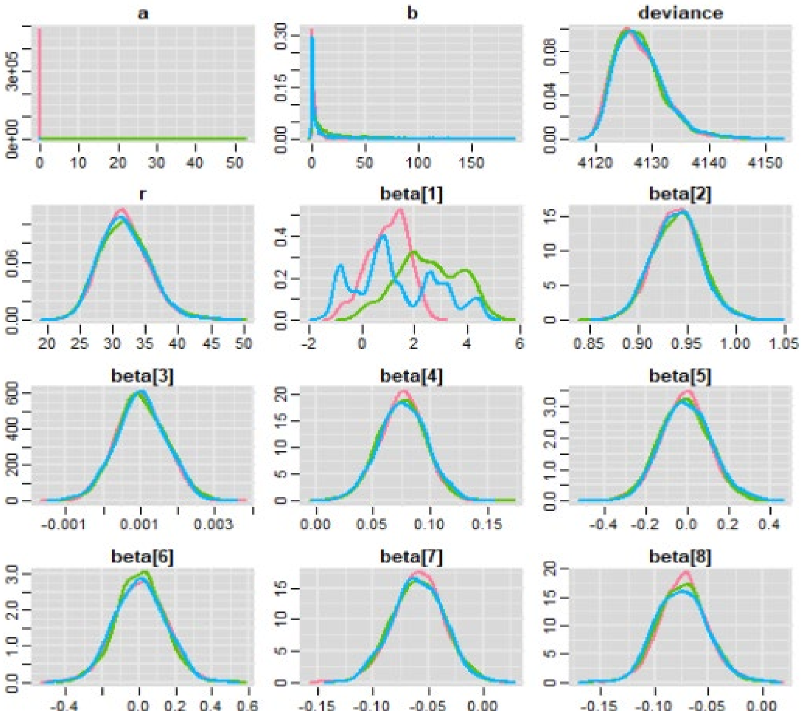


Fig. 5: Density plots of three MCMC chains for values of $r, a, b,$ deviance, β_0 = beta[1], β_1 = beta[2], α_1 = beta[3], ξ_1 = beta[4], η_1 = beta[5], η_2 = beta[6], η_3 = beta[7], and η_4 = beta[8] from the NB-SA model.

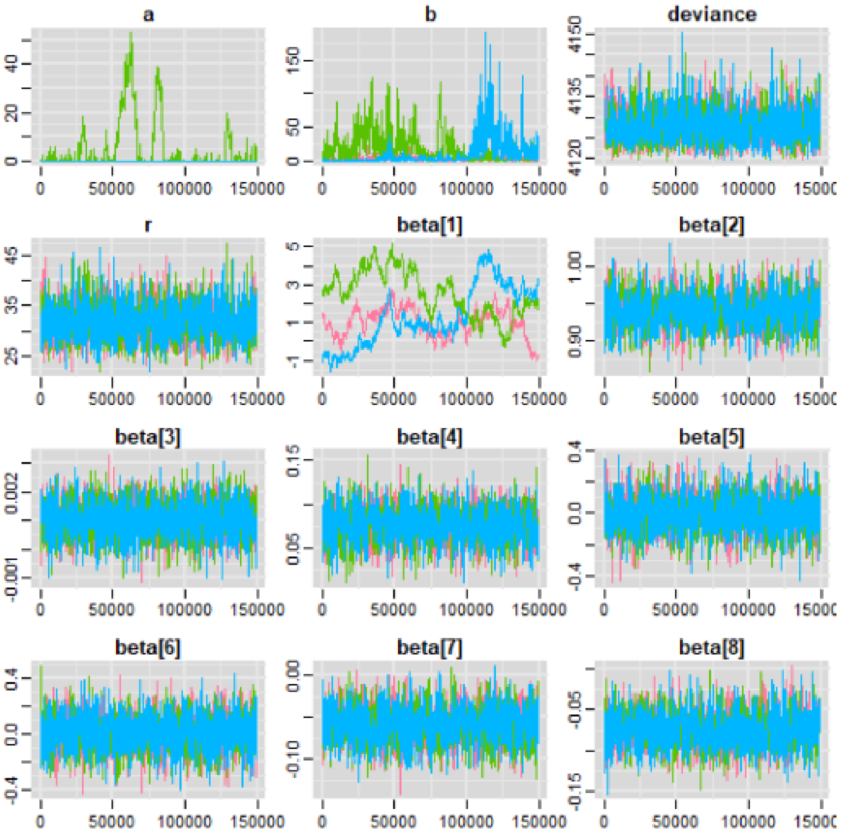


Fig. 6: Trace plots of three MCMC chains for values of $r, a, b,$ deviance, β_0 = beta[1], β_1 = beta[2], α_1 = beta[3], ξ_1 = beta[4], η_1 = beta[5], η_2 = beta[6], η_3 = beta[7], and η_4 = beta[8] from the NB-SA model.

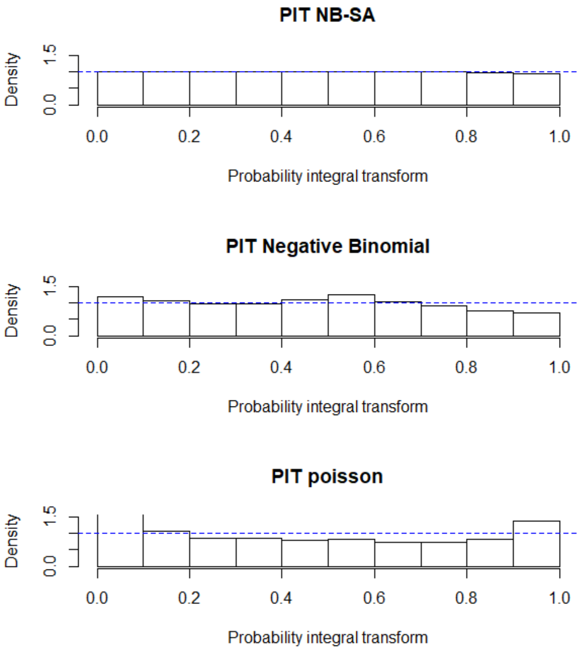


Fig. 7: PIT histograms for the NB-SA, NB, and Poisson models

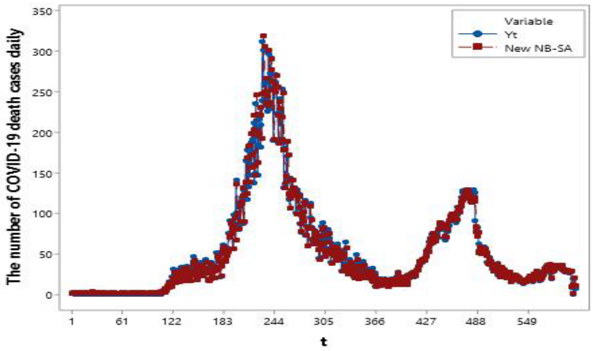


Fig. 8: Comparison of actual data and the NB-SA model

The deviance, DIC, and p_D of the generalized linear models of the NB model were 4127.7, 4137.4, and 9.7, respectively (Table 1). In the same way, the GLM for the time series data count approach with the Poisson distribution can be represented as:

$$\hat{Y}_{t,\text{Pois}} = \exp\{0.450 + 0.852\log(1+Y_{t-1}) + 0.001\upsilon_{t-1} + 0.120Z_t + 0.047I_1 + 1.123I_2 - 0.096F_1 - 0.100F_2\}.$$

The deviance, DIC, and p_D of the generalized linear models of the Poisson model were 4566.2, 9054.7, and 4488.6, respectively (Table 1).

3.5 Comparison of Model Performance

When comparing model performance based on deviance, DIC, and p_D the NB-SA model had the highest efficiency compared with the NB and Poisson models. PIT histograms for the NB-SA, Poisson, and NB distributions are shown in Figure 7. The Poisson model did not have a uniform distribution, while the NB-SA and NB models approached uniformly, and the NB-SA model was significantly more uniform than the NB model. Therefore, the NB-SA model performed better compared to the Poisson and NB models. Figure 8 compares the data plot and the NB-SA model. Results showed that the proposed model had reasonably good values that followed the actual data pattern. The NB-SA model had a value closer to the actual data plot, with the smallest BIC, DIC, and p_D values, while the PIT histogram approached uniform distribution.

4 Conclusion

A new mixed NB distribution was proposed called the negative binomial-Samade (NB-SA) distribution. Its properties were studied using a GLM framework to build the regression model. Data used in modeling comprised actual datasets of daily numbers of COVID-19 deaths in Thailand from 1 January 2021 to 30 September 2022 covering 607 days with data overdispersion and heavy-tailed. Comparing the accuracy of the proposed distribution with the Poisson and an NB model showed that the NB-SA model had the highest efficiency. One reason why the NB-SA model was more effective than NB was that the NB-SA distribution was developed from the NB distribution. The NB-SA model described the data better than the NB model for time series data counts with overdispersed or heavy-tailed distribution since the NB-SA model was more flexible with two shape parameters. By contrast, the NB model had one shape parameter. The deviance, DIC, and p_D of the new model were lower than the Poisson model at 10.6240, 118.9029, and 51,493.1035%, respectively. When comparing the new model and the NB model, the deviance, DIC, and p_D values of the NB-SA model were lower than the NB model at 0.0010, 0.0242, and 11.49%, respectively. When considering model performance using the PIT histogram, results showed that the NB-SA model approached uniform distribution. Based on deviance, DIC, p_D and the PIT histogram the proposed model was also suitable for forecasting the number of daily COVID-19 deaths in Thailand. Results indicated that the NB-SA model for time series data count was an efficient alternative to overcome overdispersion and heavy-tailed problems. Therefore, the NB-SA model was a suitable alternative to create or develop a model related to overdispersion and heavy-tailed distribution of time series data counts. An advantage of constructing the GLM for time series count data with a new mixed NB is that the developed distribution can fit with the actual data. It is robust to overdispersion problems and data with heavy-tailed distributions. As a result, the GLM for time series count data with a new mixed NB model is efficient and accurate in forecasting. But the challenge that may become a limitation of the development of new mixed NB distributions is the complexity of forming the moment-generating function of the distribution combined with the NB distribution into close form. When considering an application, the developed model can be applied to the actual count data in many fields such as medical, insurance and financial, industrial ecology and biodiversity, etc. Especially when the data have overdispersion and are heavy-tailed, the GLM for time series count data with a new mixed NB model is more efficient and accurate in forecasting. Further future work worth considering refers to a new mixed NB distribution for the multivariate time series analysis with more than one time-dependent variable. Each variable depends not only on its past values but also has some dependency on other variables. A study on Vector Auto Regression (VAR) describes the relationships between variables based on their past values.

Acknowledgment:

This research was supported by The Science, Research, and Innovation Promotion Funding (TSRI) (Grant no. FRB660012/0168). This research block grant was managed under Rajamangala University of Technology Thanyaburi (FRB66E0641N.2). The authors gratefully acknowledge this funding. We are also thankful to those who could not be mentioned here for their kindness and encouragement. And finally, the authors would like to thank the anonymous reviewers for their comments and suggestions.

References:

[1] Heinen, A. Modelling time series count data: an autoregressive conditional Poisson model. Available at SSRN 1117187, 2003.

[2] Ferland, R., Latour, A., and Oraichi, D. Integer-valued GARCH process. *Journal of time series analysis*, Vol.27, No.6, 2006, pp. 923-942.

[3] Fokianos, K., Rahbek, A., and Tjøstheim, D. Poisson autoregression. *Journal of the American Statistical Association*, Vol. 104, No.488, 2009, pp.1430-1439.

[4] Fokianos, K., and Tjøstheim, D. Log-Linear Poisson autoregression. *Journal of Multivariate Analysis*, Vol.102, No.3,2011, pp. 563-578.

[5] Genest, C., amd Nešlehová, J. A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA*, Vol.37, No.2, 2007, pp. 475-515.

[6] Lawless, J. F. Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 1987, pp. 209-225.

[7] Karlis, D., and Xekalaki, E. Mixed poisson distributions. *International Statistical Review/Revue Internationale de Statistique*, 2005, pp. 35-58.

[8] Joe, H., and Zhu, R. Generalized Poisson distribution: the property of mixture of Poisson and comparison with negative binomial distribution. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, Vol.47, No.2, 2005, pp. 219-229.

[9] Wagh, Y. S., and Kamalja, K. K. Zero-inflated models and estimation in zero-inflated Poisson distribution. *Communications in Statistics- Simulation and Computation*, Vol.47, No.8, 2018, pp. 2248-2265.

[10] Dean, C., Lawless, J. F., and Willmot, G. E. A mixed poisson–inverse-gaussian regression model. *Canadian Journal of Statistics*, Vol.17, No.2, 1989, pp. 171-181.

[11] Zamani, H., and Ismail, N. Negative binomial- Lindley distribution and its application. *Journal of mathematics and statistics*, Vol.6, No.1, 2010, pp. 4-9.

[12] Aryuyuen, S., and Bodhisuwan, W. The negative binomial-generalized exponential (NB-GE) distribution. *Applied Mathematical Sciences*, Vol.7, No.22, 2013, pp. 1093-1105.

[13] Gençtürk, Y., and Yiğiter, A. Modelling claim number using a new mixture model:negative binomial gamma distribution. *Journal of Statistical Computation and Simulation*, Vol.86, No.10, 2016, pp. 1829-1839.

[14] Yamrubboon, D., Bodhisuwan, W., Pudprommarat, C., and Saothayanun, L. The negative binomial-Sushila distribution with application in count data analysis.

Thailand Statistician, Vol.15, No.1, 2017, pp. 69-77.

[15] Aryuyuen, S. Bayesian inference for the negative binomial-generalized Lindley regression model: properties and applications. *Communications in Statistics-Theory and Methods*, 2012, pp.1-19.

[16] Aryuyuen, S., and Tonggumnead, U. Bayesian Inference for the Negative Binomial- Quasi Lindley Model for Time Series Count Data on the COVID-19 Pandemic. *Trends in Sciences*, Vol.19, No.21, 2022, pp.3171-3171.

[17] Aryuyuen, S., and Tonggumnead, U. A new mixed negative binomial regression model to analyze factors influencing the number of patients with respiratory disease and long-term effects of lung cancer. *Communications in Mathematical Biology and Neuroscience.,* 2022, Article-ID.

[18] Aryuyuen, S. The negative binomial-new generalized Lindley distribution for count data: properties and application. *Pakistan Journal of Statistics and Operation Research*, 2022, pp.167-177.

[19] Stoklosa, J., Blakey, R. V., and Hui, F. K. An overview of modern applications of negative binomial modelling in ecology and biodiversity. *Diversity*, Vol.14, No. 5, 2022, pp.320.

[20] Ortega, E. M., Cordeiro, G. M., and Kattan, M. W. The negative binomial–beta weibull regression model to predict the cure of prostate cancer. *Journal of Applied Statistics*, Vol.39 No.6, 2012, pp.1191-1210.

[21] Tzougas, G., Hoon, W. L., and Lim, J. M. The negative binomial-inverse Gaussian regression model with an application to insurance ratemaking. *European Actuarial Journal,* NO.9, 2019, pp. 323-344.

[22] Fu, S. A hierarchical Bayesian approach to negative binomial regression. *Methods and Applications of Analysis*, Vol.22, No.4, 2015, pp. 409-428.

[23] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian data analysis*. CRC press, 2013.

[24] Fu, S. Hierarchical Bayesian LASSO for a negative binomial regression. *Journal of Statistical Computation and Simulation*, Vol. 86, No,11, 2016, pp.2182-2203.

[25] Yamrubboon, D., Thongteeraparp, A., Bodhisuwan, W., Jampachaisri, K., and Volodin, A. Bayesian inference for the negative binomial-Sushila linear model. *Lobachevskii Journal of Mathematics*, No. 40, 2019, pp. 42-54.

[26] Department of Disease Control. Daily covid-19 report, Thailand information. Daily COVID-19 Report, Available at: https://data.go.th/dataset/covid-19-daily. [accessed January 2023].

[27] Aderoju, S. Samade probability distribution: its properties and application to real lifetime data. *Asian Journal of Probability and Statistics*, Vol.14, No.1, 2021, pp. 1-11.

[28] Cameron, A. C., and Trivedi, P. K. *Regression analysis of count data (Vol. 53)*. Cambridge university press, 2013.

[29] Liboschik, T., Fokianos, K., and Fried, R. tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software*, No.82, 2017, pp. 1-51.

[30] Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. *The BUGS book: A practical introduction to Bayesian analysis*. CRC press, 2013.

[31] Spiegelhalter, D. J., Best, N. G., and Carlin, B. P. Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol.64, No.4, 2002, pp.583-639.

[32] Su, Y. S., and Yajima, M. R2jags: Using R to run 'JAGS'. *R package version 0.5-7,* 2015.

[33] Geedipally, S. R., Lord, D., and Dhavala, S. S. The negative binomial-Lindley generalized linear model: *Characteristics and application using crash data. Accident Analysis & Prevention*, No.45, 2012, pp. 258-265.

[34] Dey, D. K., Ghosh, S. K., and Mallick, B. K. *Generalized linear models: A Bayesian perspective*. CRC Press, 2000.

[35] Wongrin, W., Srianomai, S., and Klomwises, Y. Bayesian Unit-Lindley Model: Applications to Gasoline Yield and Risk Assessment Data. *Naresuan University Journal: Science and Technology (NUJST)*, Vol.28, No.2, 2000, pp. 41-51.

[36] Bar-Joseph, Z., Gifford, D. K., and Jaakkola, T. S. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics,* 17(suppl_1), 2001, S22-S2.

[37] R Core Team. *R: a language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing, 2021.

Appendix

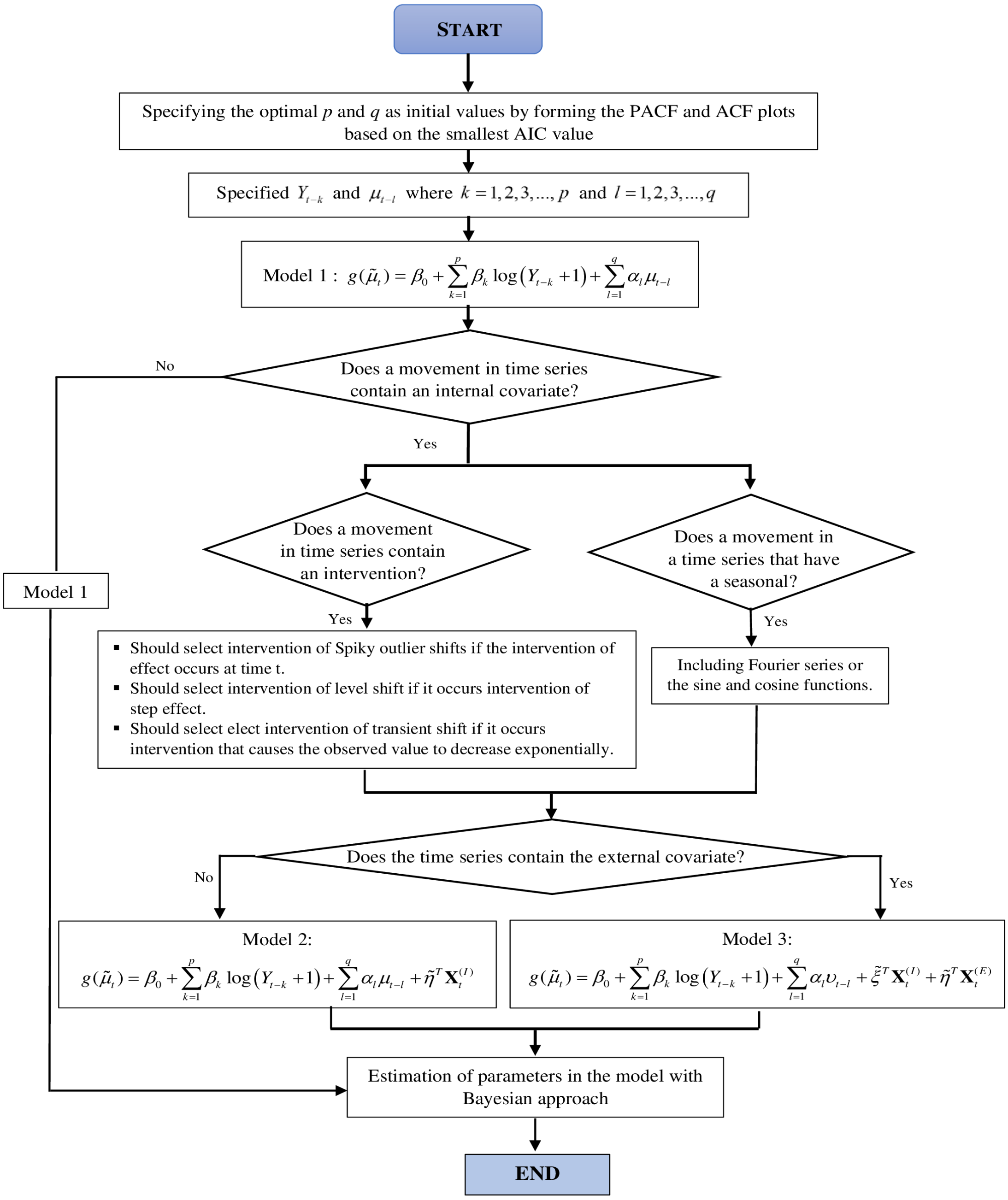

START
Specifying the optimal p and q as initial values by forming the PACF and ACF plots based on the smallest AIC value
Specified Y_{t-k} and μ_{t-l} where $k=1,2,3,\dots,p$ and $l=1,2,3,\dots,q$
Model 1 : $g(\tilde{\mu}_t)=\beta_0+\sum_{k=1}^{p}\beta_k\log(Y_{t-k}+1)+\sum_{l=1}^{q}\alpha_l\mu_{t-l}$
Does a movement in time series contain an internal covariate?
No
Yes
Model 1
Does a movement in time series contain an intervention?
Yes
Does a movement in a time series that have a seasonal?
Yes
Should select intervention of Spiky outlier shifts if the intervention of effect occurs at time t.
Should select intervention of level shift if it occurs intervention of step effect.
Should select elect intervention of transient shift if it occurs intervention that causes the observed value to decrease exponentially.
Including Fourier series or the sine and cosine functions.
Does the time series contain the external covariate?
No
Yes
Model 2:
$g(\tilde{\mu}_t)=\beta_0+\sum_{k=1}^{p}\beta_k\log(Y_{t-k}+1)+\sum_{l=1}^{q}\alpha_l\mu_{t-l}+\tilde{\eta}^T\mathbf{X}_t^{(I)}$
Model 3:
$g(\tilde{\mu}_t)=\beta_0+\sum_{k=1}^{p}\beta_k\log(Y_{t-k}+1)+\sum_{l=1}^{q}\alpha_l\upsilon_{t-l}+\tilde{\xi}^T\mathbf{X}_t^{(I)}+\tilde{\eta}^T\mathbf{X}_t^{(E)}$
Estimation of parameters in the model with Bayesian approach
END

Fig. 2: Flow diagram creating the time series model for $\{Y_t\mid\mathbb{F}_{t-1};\mathbf{\Omega}\}$

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors declare that there is no conflict of interests.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0 https://creativecommons.org/licenses/by/4.0/deed.en_US