MJFAS
Malaysian Journal of Fundamental and Applied Sciences

RESEARCH ARTICLE

Forecasting Model of Air Pollution Index using Generalized Autoregressive Conditional Heteroskedasticity Family (GARCH)

Nurul Asyikin Zamrus, Mohd Hirzie Mohd Rodzhan*, Nurul Najihah Mohamad

Department of Computational and Theoretical Sciences, Kulliyyah of Science, International Islamic University Malaysia, 25200 Kuantan, Pahang, Malaysia

Abstract The Air Pollution Index (API) of Malaysia has increased consistently in recent decades, becoming a serious environment issue concern. In this paper, we analyzed daily integer value time series data for API in Sarawak from January to June in 2019 using generalized autoregressive conditional heteroskedasticity (GARCH) family for discrete case namely Poisson integer value GARCH (INGARCH), negative binomial integer value GARCH (NBINGARCH) and integer value autoregressive conditional heteroskedasticity (INARCH) models. The parameters of the models will be estimated using quasi likelihood estimator (QLE) and we compare their Aiken information criterion (AIC) and Bayesian information criteria (BIC) to determine the best model fitted the data. Besides, the forecasting performance will be measured by using mean square error (MSE) and Pearson Standard Error (εt). The results showed that INGARCH (1,1) and INARCH (1,0) performed inconsistent results since the conventional methods of NBINGARCH (1,1) outperformed the performance of INGARCH (1,1) and INARCH (1,0). However, consistent results were achieved as the NBINGARCH (1,1) gave the smallest forecasting error compared to INGARCH (1,1) and INARCH (1,0). The findings are very important for controlling the API results in future and taking protection measure for conservation of the air.

Keywords: Forecast Time series, Generalized Autoregressive Conditional Heteroskedasticity (GARCH), Air Pollution Index, Integer-Value.

Introduction

Air quality prediction has become a crucial area of environmental science due to negative effects of high concentrations of different contaminants on human health. The air quality is defined by the API. Many researchers have used time series analysis to estimate the concentration of different pollutants and air quality in the literature [1]. Air pollution consists of a mixture of gases and particulate matter in harmful quantities released into the atmosphere by natural or human [2]. Carbon monoxide (CO), ozone (O_3), particulate matter (PM10), nitric oxide (NO), nitrogen dioxide (NO_2), nitrogen oxides (NOx), and Sulphur dioxide (SO_2) are the gases and particles involved. Air pollution is a fundamental problem in many parts of the world and it raises two important concerns. Firstly, the impact on human health and, secondly, on the environment [3]. It is extremely harmful and had been the causes of many deaths worldwide. It was estimated that 200,000 to 570,000 or about 0.4 to 1.1 per cent of global basis annual death recorded had been due to air pollution [4].

However, time series analysis has been used by many researchers in literature to predict the concentration of various pollutants and the air quality [5]. A hybrid model is proposed to deal with both

***For correspondence:**
mohdhirzie@iium.edu.my

Received: 21 June 2021
Accepted: 23 April 2022

linear and nonlinear data of a station in Delhi during 1999 to 2003 [6]. Kumar [7] in 2011 predicted the air quality index of Delhi based on three models namely Autoregressive Integrated Moving Average (ARIMA), principal component regression and hybrid of the first two. It was found that the model demonstrated the highest performance accuracy compared to other models. Further the importance of various meteorological parameters in model 3 which is integration between ARIMA and principle component regression was assessed based on principal component analysis. In the same year, the short term prediction of the concentration of ozone in Albany, New York was presented by Tsakiri [8] based on vector autoregressive model and the Kalman filter. The performance of linear, nonlinear and hybrid model was checked using mean absolute percentage error and relative error. It was found that the hybrid model outperformed both the linear and nonlinear models. In 2012, the data of eight stations in central Taiwan have been analyzed by using multivariate time series analysis models namely Autoregressive Conditional Heteroscedasticity (ARCH) and GARCH [9]. The models selected both the photochemical and fuel factors for evaluating the various time series patterns. In two years ahead, Kadiyala [10] developed a model to manage the indoor air quality using multivariate time series model to manage the concentrations of both carbon dioxide and carbon monoxide. This prediction was applied to design an optimal ventilation system for vehicles. Next, there is researcher analyzed the patterns of the relationship between various air pollutants of an Alpine Italian province [11]. The dynamic multiple time series analysis is carried out using common autoregressive stochastic model to find the improvement level in the pollution during the last decade. In 2016, Hoi and Mok [12] proposed a model named time-varying autoregressive model with linear exogenous input (TVAREX) for predicting the daily concentration of PM10 based on Kalman filter based autoregressive model. The results of TVAREX model were compared to artificial neural network (ANN) model and it was observed that TVAREX outperformed ANN. The prediction was found to be most accurate when the time series components of temperature and solar radiation were taken into consideration.

Integer-valued time series models have been widely used, especially integer-valued autoregressive models and integer-valued generalized autoregressive conditional heteroscedastic (INGARCH) models [13]. Rodrigo and Wagner [14] propose a general class of Integer-valued Generalized Auto Regressive Conditional Heteroskedastic (INGARCH) models based on a flexible family of mixed Poisson (MP) distributions. Next, they proposed class of count time series models contains the negative binomial (NB) INGARCH process as particular case and open the possibility to introduce new models such as the Poisson-inverse Gaussian (PIG) and Poisson generalized hyperbolic secant processes. In particular, the PIG INGARCH model is an interesting and robust alternative to the NB model. The author illustrates the flexibility and robustness of the MPINGARCH models through two real-data applications about number of cases of Escherichia coli and Campylobacter infections. The Poisson integer-valued GARCH model is a popular tool in modeling time series of counts. However, Zhu [15] consider a class of flexible bivariate Poisson INGARCH(1,1) model whose dependence is established by a special multiplicative factor. Stationarity and ergodicity of the process are discussed. Poisson INGARCH model, which is more flexible and allows for positive or negative cross-correlation. In the integer-valued generalized autoregressive conditional heteroscedastic (INGARCH) models, parameter estimation is conventionally based on the conditional maximum likelihood estimator (CMLE). However, because the CMLE is sensitive to outliers, Lee and Kim [16] consider a robust estimation method for bivariate Poisson INGARCH models while using the minimum density power divergence estimator. They demonstrate the proposed estimator is consistent and asymptotically normal under certain regularity conditions.

Though majority of the research has been focused on the prediction of individual concentration of pollutants, there is a need to predict a single value that indicates the air quality. In this paper, the univariate time series analysis of the Kimanis, Limbang, ILP Miri, Kapit and Samarahan has been performed by using GARCH family namely INGARCH (1,1), NBINGARCH (1,1) and INARCH (1,0) models. INGARCH (1,1), NBINGARCH (1,1) and INARCH (1,0) models is used because the Air Pollution data is integer value which mean it is more accurate with the model. The detail about study area regarding the data description has been presented in the next section. Section 3 discusses the methodology for time series analysis of air index pollution dataset. A comparison between the performance evaluation results between INGARCH (1,1), NBINGARCH (1,1) and INARCH (1,0) has been discussed in Section 4. Hence, a conclusion has been presented in the last section.

Materials and Data

Description of sampling sites

Sarawak were selected for this study due to variety of locations it offered; urban areas that consists of both living and working areas and have high population and suburban areas that are mainly residential area with a larger population than rural areas. Moreover, these locations have often been affected by trans-boundary pollution from the neighbouring countries, which has been usually the main factor behind hazardous occurrences. This dangerous haze is caused by forest fires in Sumatra and Kalimantan. Due to the haze that hit the country, Malaysia ranks third in the world in the list of countries that record the highest API after Iran and Indonesia in 2019. In 2019, based on the API observations by the World Air Quality Index (WAQI) [17] Malaysia recorded an API reading of 271 while Iran and Indonesia are 385 and 303 respectively. In total, five areas in Sarawak still recorded very unhealthy API readings including Kuching, Samarahan, Sri Aman, Sibu and Sarikei. Sarawak give the big impact on this air pollution because the location is near to the Kalimantan (Figure 1) where the forest fire happens [4]. This make the spread of the haze is more impact in Sarawak due to the direction of the wind blow faster the process of air pollution to happen here. Hence, due to the haze, a total of 2,649 schools were closed including in Sarawak while the number of asthma and conjunctivitis cases was found to be increasing based on monitoring from 31 haze sentinel clinics. Hence, Kimanis, Limbang, ILP Miri, Kapit and Samarahan have been choosing in this research to measure the air quality using GARCH family model for integer value.

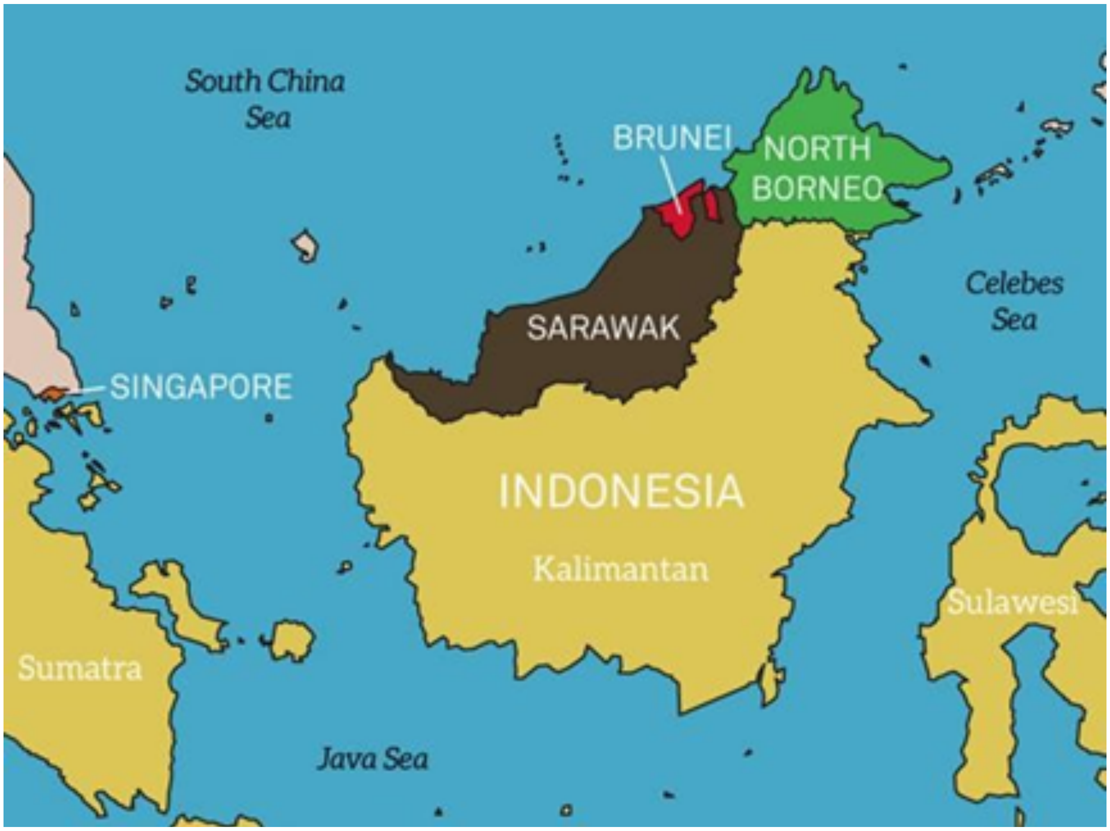


Figure 1. Sarawak-Kalimantan Map which is the location is side by side and cause the haze spread more easily to Sarawak from Kalimantan forest.

Air quality index data

The dateset use in this research is API. The frequency of the data collection is daily data within six (6) months from January 2019 until June 2019 (183 observations) to identify the API model. The data have been chosen from five (5) locations, Kimanis, Limbang, ILP Miri, Kapit and Samarahan which is located at Sarawak from the Department of Statistic Malaysia website [13]. Three statistical models; INGARCH (1,1), NBINGARCH (1,1) and INARCH (1,0) were used in forecasting the daily API data.

Methodology

Forecasting analyses

Time series forecasting analysis using INGARCH (p,q), NBINGARCH (p,q), and INARCH (p,q) model has been carried out for the prediction of air pollution index. The steps of the methodology of the time series analysis using INGARCH (p,q), NBINGARCH (p,q), and INARCH (p,q) model have been summarized in this section. In figure 2, show that the framework of the methodology which is divided into three phases. The first phase known as data collection, followed by second phase which is modelling assessment and lastly third phase, forecasting evaluation.

The model of using INGARCH (p,q), NBINGARCH (p,q), and INARCH (p,q) are fitted in the form of (1,1) by QLE. The Poisson assumption is right to get a standard maximum likelihood estimator (MLE). However, if we assume a mixed Poisson distribution, we get a MLE. The vector of regression parameters is denoted by the symbol θ=(β_0,β_1,...,β_p,α_1,...,α_q). The parameter space for the INGARCH (1,1) model with covariates is given by regardless of the distributional assumption [18].

$$\Theta = \left\{\theta \in R^{p+q+r+1};\ \ \beta_0 > 0, \beta_1, \dots, \beta_p, \alpha_1, \dots, \alpha_q, \geq 0, \sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l < 1 \right\} \tag{1}$$

Next, the efficacy of the ARCH is investigated. Until running simulations with the time series combination of ARCH and GARCH models, the model calibration phase must be completed first to ensure that the residual series is not connected to the first order series, often known as white noise, and that the model is acceptable. In testing the presence of ARCH effect, the generalised autoregressive representation of the squared residuals ($\hat{u}^2$) with the the error (e_t) is given as:

$$\hat{u}_t^2 = b_0 + b_1 u_{t-1}^2 + b_2 u_{t-2}^2 + b_3 u_{t-3}^2 + \dots + b_q u_{t-q}^2 + e_t \tag{2}$$

After that, the residual square test is used to see if the model has the ARCH effect. The significance of the parameters b_i indicates the presence of conditional volatility (ARCH effect) under the null hypothesis of no ARCH effect:

$$b_1 = b_2 = b_3 =. . . = b_q = 0 \tag{3}$$

Therefore, before running the model, the ARCH (1) effect is tested to clarify that the API data is significant with the GARCH family model:

$$\hat{u}_t^2 = b_0 + b_1 u_{t-1}^2\ + e_t \tag{4}$$

The Lagrange multiplier (LM) test is being used to test for the arch effect before running the GARCH family model. Iterative nonlinear calculations for estimating model parameters can only be done with the model that has ARCH effectiveness [9].

After that, the data were tested for stationary using the Augmented Dickey–Fuller (ADF) test, with the null hypothesis (H_0) being that the time series is non-stationary. The ADF test revealed that the time-series data were non-stationary (p > 0.05), indicating that they were non-stationary. The INGARCH (1,1), NBINGARCH (1,1), and INARCH (1,0) model's value were chosen based on the (AIC) value, which is given as follow.

$$\mathrm{AIC}_{p,q} = \frac{-2ln(\text{maximized likelihood})+2r}{n} \approx ln(\sigma_i^2) + r\frac{2}{n} + \text{constant} \tag{5}$$

where n is the number of data observations, r=p+q+1 and σ_2 is the maximum likelihood prediction.

We have tested different values of α and β parameters ranging from 0 to 5, while d was chosen to be 1 based on ADF test. We found that the best INGARCH (1,1) model that gives the lowest AIC values is (1,1).

Hence, three model under the GARCH family have been chosen which are called as INGARCH (1,1), NBINGARCH (1,1), and INARCH (1,0) model.

a. INGARCH (p,q)

In this subsection, we focus on INGARCH (p, q) model given by equation (6) [18].

$$\lambda_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j} \tag{6}$$

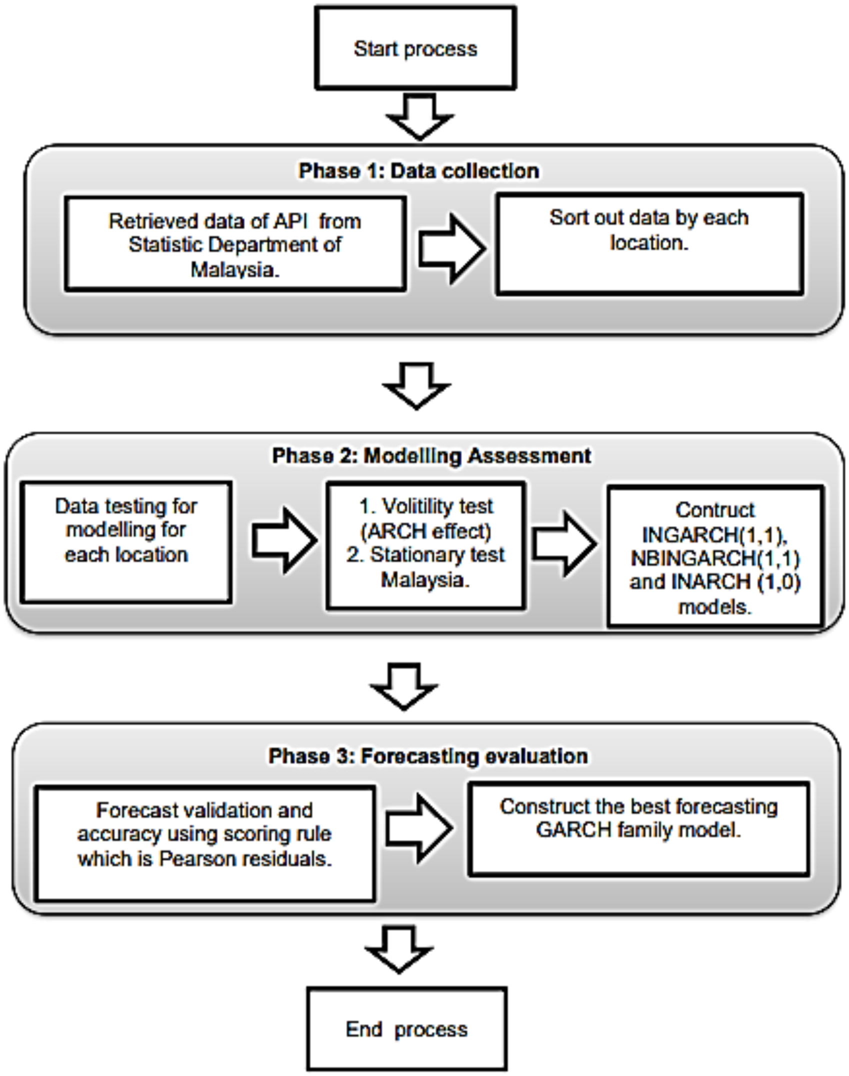

Start process
Phase 1: Data collection
Retrieved data of API from Statistic Department of Malaysia.
Sort out data by each location.
Phase 2: Modelling Assessment
Data testing for modelling for each location
1. Volitility test (ARCH effect)
2. Stationary test Malaysia.
Contruct INGARCH(1,1), NBINGARCH(1,1) and INARCH (1,0) models.
Phase 3: Forecasting evaluation
Forecast validation and accuracy using scoring rule which is Pearson residuals.
Construct the best forecasting GARCH family model.
End process

Figure 2. Framework of the methodology

where $\alpha_0 > 0$, $\alpha_i \geq 0$, $1 \leq i \leq p$, $\beta_j \geq 0$, $1 \leq j \leq q$.

b. NBINGARCH (p,q)

Let $\{X_t\}$ be a time series of counts for Air Pollution Index. We assume that conditional on F_{t-1}, the random variables $X_1, \ldots, X_n$ are independent and the conditional distribution of X_t is specified by a normal binomial distribution. To be specific, we consider the following model:

$$X_t|F_{t-1} : NB(r, p_t) \tag{7}$$

where F_{t-1} is the r-field generated by $\{X_t , X_{t-2}, . . . \}$, r is a positive number and p_t satisfies the model.

$$\frac{1-p_t}{p_t} = \lambda_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j} \tag{8}$$

where $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, i=1,...,p, j=1,...,q ,$p \geq 1$, $q \geq 0$.

c. INARCH (p,q)

The purely autoregressive INARCH (p, 0) model is also called an (p, 0) model by [19]. The (p, 0) model is defined as equation (9).

$$\lambda_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{t-i} \tag{9}$$

where $t \in Z$, $\beta_0 > 0$, $\beta_i \geq 0$, , i=1,...,p.

Evaluation performance

For forecasting, the model with the lowest AIC and BIC value was chosen. In terms of the mean square error, the optimal 1-step-ahead predictor $\hat{Y}_{n+1}$ for Y_{n+h} , given F_n. The past of the process up to time n and potential covariates at time $n+1$, is the conditional expectation λ_{n+1}. By construction of the model the conditional distribution of $\hat{Y}_{n+1}$ is a Poisson (equation 6) respectively Negative Binomial (equation 7) distribution with mean $\lambda_{n+1.}$ An h-step-ahead prediction $\hat{Y}_{n+h}$ for Y_{n+h} is obtained by recursive 1- step-ahead predictions, where unobserved values $Y_{n+1},\ldots,$ Y_{n+h-1} are replaced by their respective 1- step-ahead prediction, $h \in N$.

Pearson residual is the statistical tests that is used to measure the model validation. The model is adequate if the Pearson residuals is close to one [20]. The Pearson residual is given by:

$$\varepsilon_t = \frac{X_1 - \widehat{\hat{\lambda}_t}}{\sqrt{\widehat{\hat{\lambda}_t} + \hat{\lambda}_t^2\, \hat{\sigma}_2}} \tag{10}$$

Where, $X_1 \ldots X_{183}$ is the observation for API, ε_t is Pearson residuals for the GARCH family model. Hence, the fitted values are denoted by $\hat{\lambda}_t$.

Results and Discussion

The time series graph shown in Figure 3 explained that the pattern for the graph is irregular because there is a huge different for the highest and lowest values. The comparison for the graph pattern indicates that there is a minor change between Kimanis, Limbang, ILP Miri, Kapit and Samarahan. There is high volatility in the graph throughout the dateset from January 2019 until June 2019. The volatility also indicates there is ARCH effect in our result. The highest mean is ILP Miri with the value 48.53 and the lowest mean which is Samarahan with the value 33.30.

The skewness result of Air Index Pollution for Limbang and ILP Miri show that the data is substantially skewed distribution because the number of skewness is less than -1. Meanwhile, the skewness result for Bintulu, Kimanis, Kapit and Samarahan are between -1 and +1 indicates non substantially skewed distribution. The kurtosis for all the dateset is not approached to 3, so the dateset has no "heavy tails" and no “light-tailed”. Therefore, the distributions exhibiting skewness and/or kurtosis that exceed these guidelines are considered nonnormal [21]. From the result we can conclude that, the distribution in nonnormal behaviour because all the result are non- zero for both skewness and kurtosis.

The autocorrelation functions (ACF) of time series can be used to infer their stability or instability, as well as memory characteristics: Short-memory processes with non-zero autocorrelations at only a few lags are stable, whereas long-memory processes with major autocorrelations on many lags are unstable. Therefore, in this research lag (1,1) is using for INGARCH and NBINGARCH whereas the INARCH model is using lag (1,0). Stationary means that the statistical characteristics of a process under study do not change over time [22]. The Air Pollution Index is a stationary which mean the p-value is less than 0.05. When the result is less than 0.05, the H_0 is rejected and H_1 is accepted. Hence, the result is significant for stationary test using the ADF test.

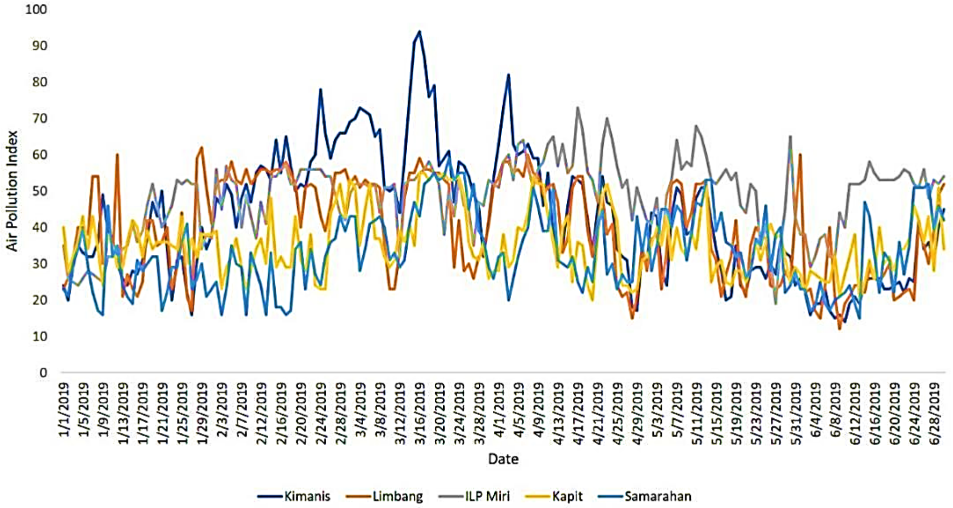


Figure 3. Time Series Graph for Kimanis, Limbang, ILP Miri , Kapit and Samarahan

Then, the volatility of the data is being evaluated to make sure there is ARCH effect in API database. The ARCH effect is being evaluated by using Langrage multiplier (LM) test. The result shows, that there is ARCH effect in the model because the p-value is less than 0.05. In univariate time series models, the LM test for ARCH is widely used as a specification test [23].

Table 1. Statistical summary

Location	Mean	Skewness	Kurtosis	Unit Root test	Arch Effect
Kimanis	42.47	0.51	-0.35	-3.6245**	75.3820***
Limbang	39.77	-0.15	-1.41	-3.7054**	48.9190**
ILP Miri	48.53	-0.56	-0.03	3.4451**	44.5040**
Kapit	35.51	0.54	-0.42	-3.8292**	111.0600***
Samarahan	33.30	0.32	-0.75	-3.6815**	64.6250***

NOTE: ** *p*<0.01, *<0.001**

The performance evaluation results of the model based on Air Pollution Index value only. To evaluate the performance of the models, the data were tested by using ADF. This is to ensure the data is stationary before running the GARCH family model. Hence, the performance for INGARCH (1,1), NBINGARCH (1,1) and INARCH (1,0) is being compared by using AIC and BIC. The lowest AIC indicates the best model. The best model for ILP Miri and Bintulu is INARCH (1,0) model. Furthermore, INGARCH (1,1) model prove that it shows the best model for Kapit only. Besides, majority location for Kimanis, Limbang, Kapit and Samarahan. indicates that NBINGARCH (1,1) is the best model. By comparing the results shown in Table 2, we can see that INGARCH and NBINGARCH (1,1) model is in good agreement with the actual values (p-value > 0.05).

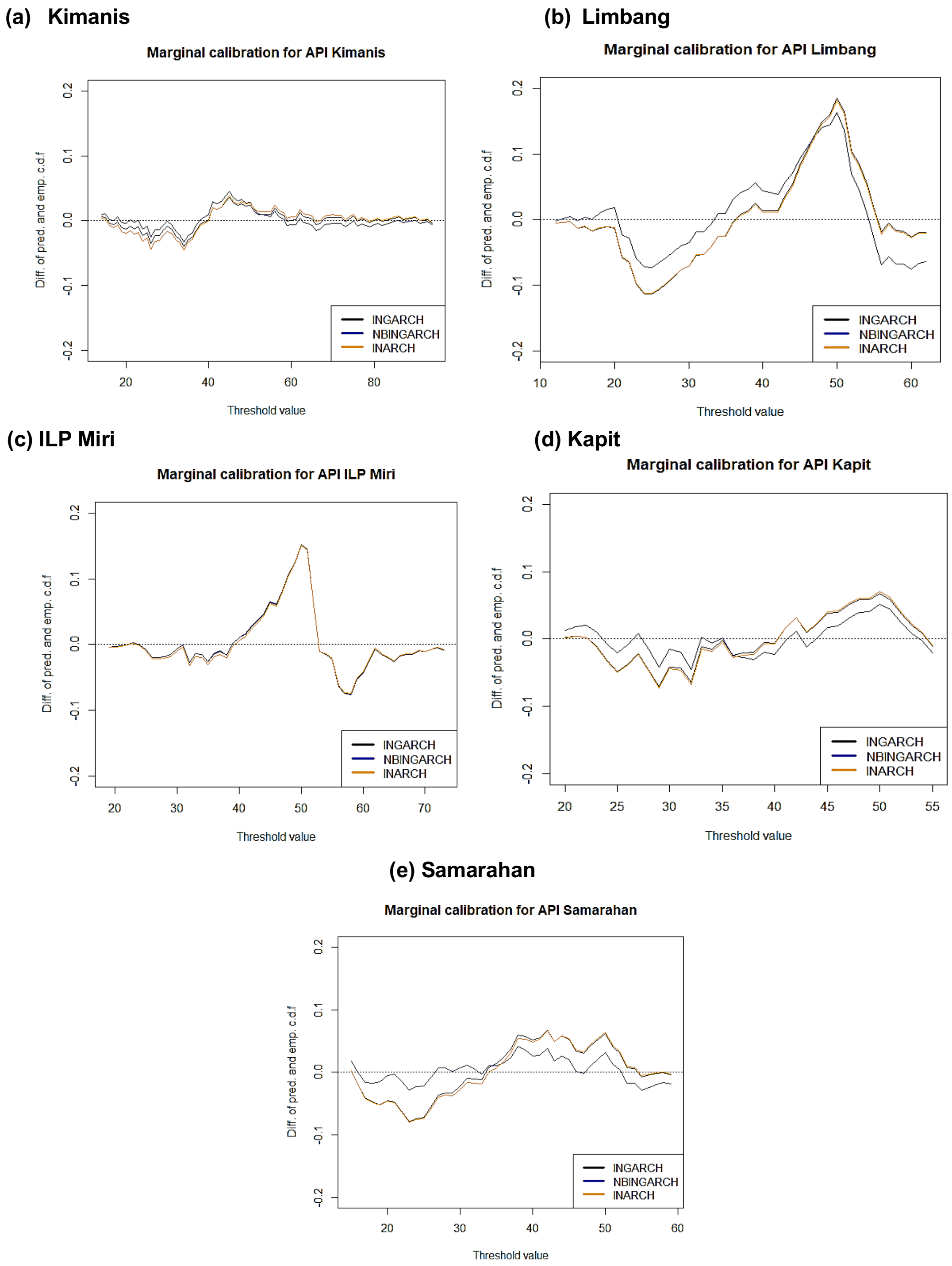
The autocorrelation function of respond residuals are identical for the INGARCH (1,1) NBINGARCH (1,1) and INARCH (1,0) model. Their empirical autocorrelation functions, does not exhibit any serial and correlation or seasonality which has not been taken into account by the models. Marginal Calibration can be assessed by taking the difference between the average predictive cumulative distribution function (c.d.f.) and the empirical c.d.f. of the observations. Minor fluctuations about zero are expected if the marginal calibration hypothesis is true. Therefore, to assess marginal calibration and sharpness of the prediction, the marginal calibration plot is constructed for these data. Figure 4 depicts the marginal calibration plots for the air pollution index data which indicate that the NBINGARCH (1,1) model approach to zero for all location.

Table 2. Performance GARCH model comparison

Location	Parameter	INGARCH (1,1)	NBINGARCH (1,1)	INARCH (1,0)
Kimanis	μ	4.73***	4.73***	5.29***
	α	1.77^{-5}**	1.77^{-5}**	-
	β	0.89**	0.89**	0.88**
	AIC	1252.21	1243.73	1250.88
	BIC	1261.77	1256.48	1257.26
Limbang	μ	11.4***	11.4***	11.14***
	α	2.76^{-5}**	2.76^{-5}**	-
	β	0.714**	0.71**	0.72**
	AIC	1399.071	1324.02	1397.10
	BIC	1408.63	1336.77	1430.475
ILP Miri	μ	10.62***	10.60***	11.62***
	α	0.01**	0.01**	-
	β	0.77**	0.77**	0.76
	AIC	1207.14	1205.01	1203.23
	BIC	1214.57	1219.89	1209.61
Kapit	μ	9.73***	9.73***	15.80***
	α	0.25**	0.25**	-
	β	0.47**	0.47**	0.56**
	AIC	1220.52	1243.47	1224.38
	BIC	1230.08	1253.03	1230.76
Samarahan	μ	6.59***	6.59***	11.92***
	α	0.30**	0.30**	-
	β	0.50**	0.50**	0.64**
	AIC	1286.38	1251.58	1292.46
	BIC	1295.94	1264.33	1298.83

NOTE: ** *p*<0.01, *<0.001**

According to the Pearson Standard Error (ε_t) results shown in Table 3, NBINGARCH(1,1) model outperformed the INARCH(1,0) and INGARCH (1,1) model. This is because NBINGARCH(1,1) model for Kimanis, Limbang, ILP Miri, Kapit and Samarahan has the lowest value of εt. However, INARCH (1,0) and INGARCH(1,1) could only perform better in ILP Miri and Kapit alternatively. This two methods showed inconsistent results in the best forecast for API in different types of background area. The graph for NBINGARCH (1,1) forecasting API data for stations Kimanis, Limbang, ILP Miri, Kapit, and Samarahan are shown in Figure 4 . Meanwhile, Figure 5 show graphically comparison of model fit for API data between model INARCH(1,0), NBINGARCH(1,1) and INGARCH(1,1). This prove that the best forecasting model is NBINGARCH(1,1) because the NBINGARCH line fitted to actual line.



(a) Kimanis

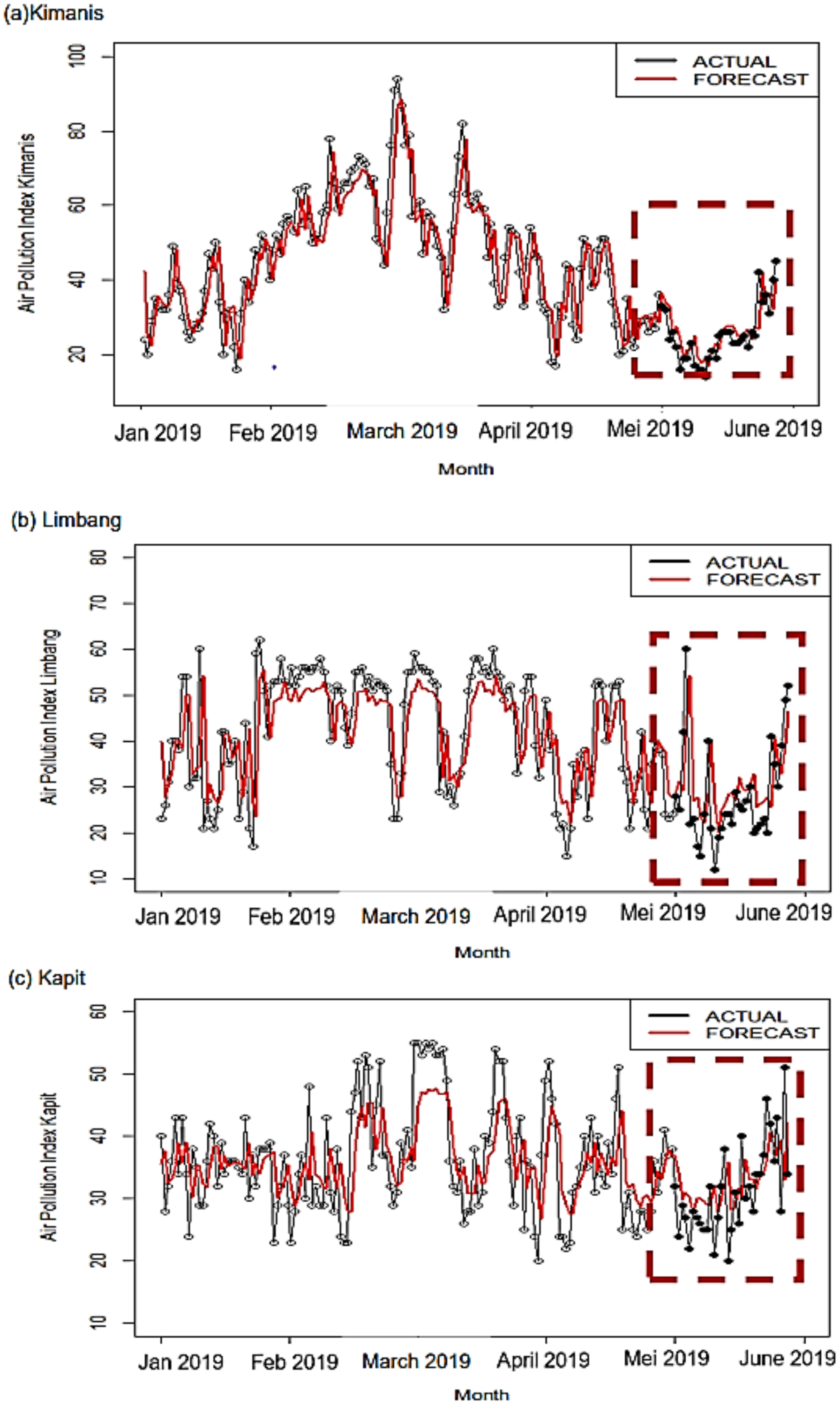
(b) Limbang

(c) ILP Miri

(d) Kapit

(e) Samarahan

Figure 4. Marginal calibration plot after model fitting to API data from January 2019 until June 2019 in (a) Kimanis, (b) Limbang, (c) ILP Miri, (d) Kapit and (e) Samarahan.



(a)Kimanis

(b) Limbang

(c) Kapit

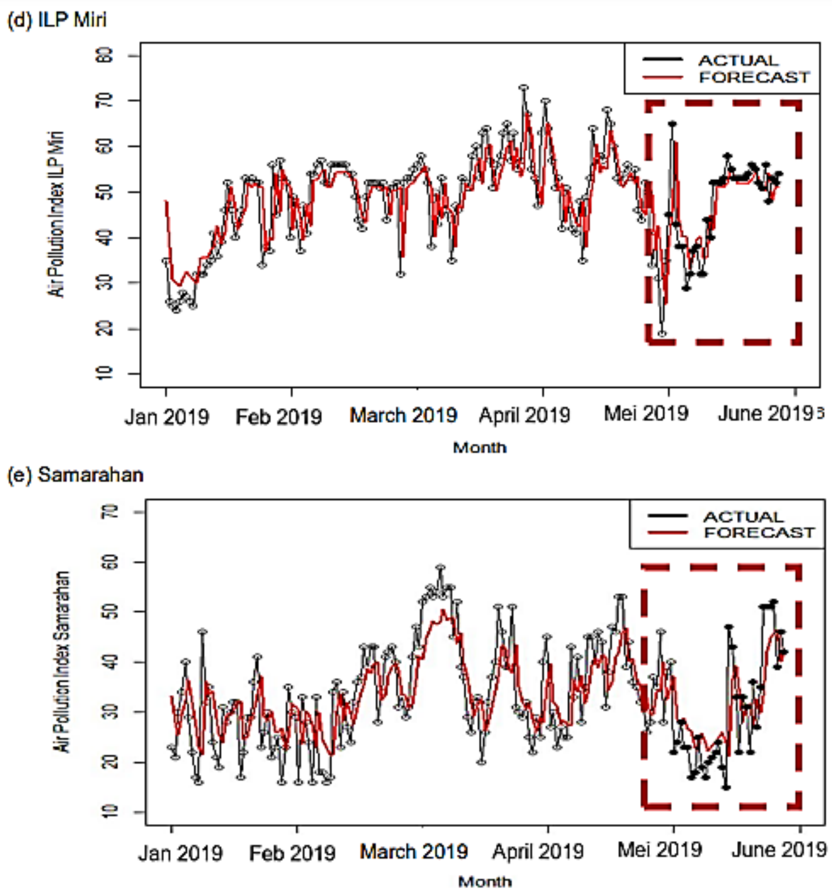


Figure 5. The graph show the actual value (black line) and the forecast value (red line) according to NBINGARCH(1,1) model (best forecasting model). The prediction interval (red square) are designed to ensure a global coverage rate of 90% (Forecasting accuracy) in (a) Kimanis, (b) Limbang, (c) ILP Miri, (d) Kapit and (e) Samarahan.

Table 3. MSE and ε_t Performance for comparison with different forecasting methods

Model	MSE	ε_t
Kimanis		
INGARCH (1,1)	55.64	1.32
NBINGARCH (1,1)	55.64	**0.98**
INARCH (1,0)	55.31	1.29
Limbang		
INGARCH (1,1)	85.61	2.20
NBINGARCH (1,1)	85.61	**0.98**
INARCH (1,0)	85.58	2.21
ILP Miri		
INGARCH (1,1)	42.21	0.89
NBINGARCH (1,1)	42.21	**0.89**
INARCH (1,0)	42.08	0.92
Kapit		
INGARCH (1,1)	41.54	1.14
NBINGARCH (1,1)	41.54	**0.98**
INARCH (1,0)	41.86	1.14
Samarahan		
INGARCH (1,1)	47.49	1.47
NBINGARCH (1,1)	47.49	**0.98**
INARCH(1,0)	47.89	1.48

Conclusions

This paper discussed the model comparison, namely INGARCH (1,1), NBINGARCH (1,1) and INARCH (1,0). The primary goal of this study was to identify the most efficient time series methods in air quality forecasting model using GARCH family model concerning the daily API in five locations in Sarawak, Malaysia. The results from this study shows that NBINGARCH (1,1) is the best model for Kimanis, Limbang and Samarahan which is due to it low value of AIC and BIC compared to INGARCH(1,1) and INARCH (1,0) . Besides, INARCH (1,0) and INGARCH (1,1) are the best model for ILP Miri and Kapit respectively. Hence, this prove that the NBINGARCH (1,1) model was capable of treating modelling and forecasting index values of API. NBINGARCH (1,1) proved to be a flexible and intelligent forecasting method that is a useful and effective tool for modelling the complex and poorly understood processes. Although with univariate model, whereby the input was from the best both INARCH (1,0) and INGARCH (1,1) lags, the NBINGARCH (1,1) will be able to give more accurate predictions for the observed API at all five sites compared to the conventional ARCH family INARCH (1,0) and INGARCH (1,1). Therefore, we can suggest that the simplest NBINGARCH (1,1) can be used for future forecasting of air pollutants for univariate integer value since it is good to predict fluctuating series, which contain trend and seasonality, such as air quality data.

Data availability

The data is retrieved from the open-source data of the Air Pollution Index from the Department of Statistic Malaysia website. https://www.data.gov.my/

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

The authors are indebted to the Department of Environment Malaysia for providing the air pollution data that made this paper possible.

References

[1] J. K. Sethi and M. Mittal, "Analysis of air quality using univariate and multivariate time series models," *Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering*, pp. 823–827, 2020.

[2] M. Kampa and E. Castanas, "Human health effects of air pollution.," *Environmental Pollution*, vol. 151, no. 2, pp. 362–357, 2008.

[3] A. B. Kurt, A., Oktay, "Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7986–7992, 2010.

[4] World Resources Institute, "World Resources Institute: (2002) Rising Energy Use: Health Effects of Air Pollution.," *World Resources Institute. http://www.airimpacts.org (2002).*, vol. Accessed J, 2002.

[5] X. Y. Ni, H. Huang, and D. W, P, "Relevance analysis and short-term prediction of PM 2.5 concentration in Beijing based on multisource data," *Atmospheric Environment*, vol. 150, pp. 146–161, 2017.

[6] Chelani, B. Asha, and S. Devotta, "Air quality forecasting using a hybrid autoregressive and nonlinear model," *Atmospheric Environment 40*, vol. 10, pp. 1774–1780, 2006.

[7] A. Kumar and P. Goyal, "Forecasting of daily air quality index in Delhi," *Science of the Total Environment*, vol. 24, no. 409, pp. 5517–5523, 2011.

[8] Tsakiri, K. G, and Z. Igor, G, "Prediction of ozone concentrations using atmospheric variables," *Air Quality, Atmosphere & Health*, vol. 2, no. 4, pp. 111–120, 2011.

[9] E. M. Y. Wu, S. L. Kuo, and W. C. Liu, "Generalized autoregressive conditional heteroskedastic model for water quality analyses and time series investigation in reservoir watersheds," *Environmental Engineering Science*, vol. 29, pp. 227–237, 2012.

[10] Kadiyala, Akhil, and A. Kumar, "Multivariate time series models for prediction of air quality inside a public transportation bus using available software," *Environmental Progress & Sustainable Energy*, vol. 2, no. 33, pp. 337–341, 2014.

[11] Passamani, Giuliana, and M. Paola, "Local atmospheric pollution evolution through time series analysis," *Journal of Mathematics and Statistical Science 2*, vol. 12, pp. 781–788, 2016.

[12] K. I. Hoi, K. V. Yuen, and K. M. Mok, “Prediction of daily averaged PM10 concentrations by statistical time-varying model,” *Atmospheric Environment 43*, vol. 16, pp. 2579–2581, 2009.

[13] N. N. Mohamad, I. Mohamed, and N. K. Haur, “Moment properties and quadratic estimating functions for integer-valued time series models,” *Pakistan Journal of Statistics and Operation Research*, vol. 14, no. 1, pp. 157–175, 2018.

[14] R. B. Silva and W. Barreto-Souza, “Flexible and Robust Mixed Poisson INGARCH Models,” *Journal of Time Series Analysis*, vol. 40, no. 5, pp. 788–814, 2019.

[15] Q. Li, H. Chen, and F. Zhu, “Robust Estimation for Poisson Integer-Valued GARCH Models Using a New Hybrid Loss,” *Journal of Systems Science and Complexity*, vol. 34, no. 4, pp. 1578–1596, Aug. 2021.

[16] B. Kim, S. Lee, and D. Kim, “Robust estimation for bivariate poisson ingarch models,” *Entropy*, vol. 23, no. 3, Mar. 2021.

[17] World Air Quality Index, “World’s Air Pollution: Real-time Air Quality Index,” *World Air Quality Index*, 2020.

[18] R. Ferland, A. Latour, and D. Oraichi, “Integer-valued GARCH process,” *Journal of Time Series Analysis*, vol. 27, no. 6, pp. 923–942, 2006.

[19] C. Weiß, “Modelling time series of counts with overdispersion,” *Statistical Methods and Application*, vol. 18, pp. 507–519, 2009.

[20] T. Liboschik, K. Fokianos, and R. Fried, “Tscount: An R package for analysis of count time series following generalized linear models,” *Journal of Statistical Software*, vol. 82, no. November, 2017.

[21] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*, vol. 68, no. 22. 2016. doi: 10.1063/1.115817.

[22] T. Stadnitski, “Time series analyses with psychometric data,” *PLoS ONE*, vol. 15, no. 4, pp. 1–12, 2020.

[23] E. H. Glenn, “Getting Started in R,” pp. 9–14, 2016.