Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 3, 2021, page 203-213

Log-Linear Poisson Autoregressive Model in Modelling Confirmed Cases of Coronavirus Disease in Malaysia

[1]Yee Wei Jet [a], Norhaiza Ahmad [b]*
[a,b]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia.
*Corresponding author: norhaiza@utm.my

Abstract
The purpose of this study is to apply the log-linear Poisson autoregressive model in modelling the number of confirmed COVID-19 cases in Malaysia from September 2020 until March 2021. Two such models using different parameter mean are developed using the maximum likelihood method. Model evaluation and validation are done using the Pearson residuals. The adequate model with better performance is applied to predict the number of confirmed cases in April 2021. Satisfying results are obtained from the better model but the unstable number of confirmed cases and its extreme changes might affect the accuracy of the model.
Keywords: Poisson autoregressive model; COVID-19; overdispersion

1 Introduction

The outbreak of the coronavirus disease 2019 (COVID-19) started in December 2019 and quickly spreads across the globe. As the number of confirmed cases in Malaysia increased in March 2020, the government implemented the Movement Control Order (MCO) to restrain the spread of the disease. Nonetheless, the number of confirmed cases in Malaysia rises again in September 2020 as the relaxation of MCO.

According to the data provided by the Ministry of Health, the number of confirmed cases in different state varies. This variation causes difficulties in measuring and contrasting the difference in the spread of COVID-19 between states. Hence, modelling the number of confirmed cases in each state using a suitable model is crucial. Various models have been used to model the spread of COVID-19 in various countries, namely the time series model [1,2,3], compartmental model [4,5,6], and Poisson regression model [7]. Particularly, the application of the Poisson regression model can be restrictive due to its equidispersed assumption, where the mean is assumed to be equal to the variance. In practice, the variance of the observations often increases more rapidly than the mean and leads to overdispersion that invalidating the Poisson distribution. Hence, it seems to be unrealistic to follow the assumption, especially in epidemiological studies, where the variance changing is natural corresponding to many important processes [8].

Agosto and Giudici [9] propose to employ the log-linear Poisson autoregressive model in modelling the number of confirmed COVID-19 cases. Their results show that the model can

be applied to any country or region and in any period [9]. This is suitable for modelling the number of confirmed cases in each state of Malaysia. Studies also found that inferences on the future spread of COVID-19 can be drawn based on the values of the parameters in the model [9, 10]. Furthermore, the model is able to handle overdispersion in the data [11,12].

The objectives of this study are to examine the characteristic of the data of the confirmed COVID-19 cases in each state of Malaysia, to model the number of confirmed cases in each state of Malaysia using the log-linear Poisson autoregressive model and to evaluate the performance of the model.

2 Methodology

2.1 Log-Linear Poisson Autoregressive Model

Consider $\{Y_t\}$ as the time series of the number of confirmed COVID-19 cases, where $t \geq 0$ represents the time. The log-linear Poisson autoregressive model assumes the number of confirmed cases follows the Poisson distribution with mean λ_t. Besides, it accommodates both positive and negative correlation [12]. In this study, two such models are applied. For $t \geq 1$, the two log-linear Poisson autoregressive models are shown below:

$$Y_t \sim Poisson(\lambda_t), \qquad \ln \lambda_t = \omega + \alpha \ln(1 + Y_{t-1}) + \beta \ln \lambda_{t-1} \tag{1}$$

$$Y_t \sim Poisson(\lambda_t^*), \qquad \ln \lambda_t^* = \omega + \alpha \ln(1 + Y_{t-1}) + \beta \ln \mu_{t-1} \tag{2}$$

For the sake of clarity, model (1) and model (2) are referred as Model 1 and Model 2 respectively. Model 1 is the model applied in the study of Agosto and Giudici [9], whereas Model 2 is similar to Model 1 with a slight modification at the last term on the right-hand side. Instead of λ_{t-1} obtained from the model, Model 2 uses the actual mean number of confirmed cases at time $t - 1$, denoted as μ_{t-1}.

In both models, the value of Y_0, λ_0 and μ_0 are assumed to be fixed. Particularly, Y_0 takes the value of the first observed number of confirmed cases from the data. While the value of λ_0 in Model 1 is set to be 1. Note that different choices of initial λ_0 do not affect the results [12]. On the other hand, the value of μ_0 in Model 2 is the mean number of the confirmed cases at time $t = 0$. For some states in Malaysia, however, no confirmed cases are observed at the beginning of the period and result in $\mu_t = 0$ for some $t \geq 0$. In this case, $\mu_t = 1$ is used since the value of $\ln 0$ is undefined.

The meaning behind each parameter ω, α and β have been explained by Agosto and Giudici [9]. For instance, ω denotes the intercept term, while α expresses the short-term dependence on yesterday's confirmed cases. Note that $\ln(1 + Y_{t-1})$ is used instead of $\ln(Y_{t-1})$ for handling zero values. Whereas β expresses the long-term dependence on the historical confirmed cases of the previous days. Studies show the values of α and β can reveal the underlying trend of the spread of COVID-19 [9,10]. If $\beta > \alpha$, then there exists an increasing trend. Conversely, a decreasing trend is expected if $\alpha > \beta$. In addition, the parameters α and β subject to the conditions where $|\alpha + \beta| < 1$ if both α and β have the same sign, or $\alpha^2 + \beta^2 < 1$ if α and β have different sign [12].

Fokianos and Tjøstheim [12] have demonstrated that the parameters ω, α and β in Model 1 and Model 2 can be estimated using the method of maximum likelihood. Take Model 1 as an example, suppose $\boldsymbol{\theta} = (\omega, \alpha, \beta)$ represents the three-dimensional vector of the parameters and the starting value of λ_0 is given. Then the likelihood function for $\boldsymbol{\theta}$ in terms of the observations $Y_1, Y_2, \ldots, Y_n$ is as followed:

$$L(\boldsymbol{\theta}) = \prod_{t=1}^{n} \frac{\exp\left(-\lambda_t(\boldsymbol{\theta})\right) \lambda_t^{Y_t}(\boldsymbol{\theta})}{Y_t} \tag{3}$$

Taking logarithm at both sides of (3):

$$\ln L(\boldsymbol{\theta}) = \ln\left[\prod_{t=1}^{n} \frac{\exp\left(-\lambda_t(\boldsymbol{\theta})\right) \lambda_t^{Y_t}(\boldsymbol{\theta})}{Y_t}\right]$$

$$\ln L(\boldsymbol{\theta}) = \sum_{t=1}^{n} \ln \exp\left(-\lambda_t(\boldsymbol{\theta})\right) + \ln \lambda_t^{Y_t}(\boldsymbol{\theta}) - \ln Y_t$$

$$\ln L(\boldsymbol{\theta}) = \sum_{t=1}^{n} Y_t \ln\left(\lambda_t(\boldsymbol{\theta})\right) - \lambda_t(\boldsymbol{\theta}) - \sum_{t=1}^{n} \ln Y_t \tag{4}$$

Note that $\sum_{t=1}^{n} \ln Y_t$ in (4) is a constant. Hence, (4) is maximized by maximizing:

$$l(\boldsymbol{\theta}) = \sum_{t=1}^{n} Y_t \ln\left(\lambda_t(\boldsymbol{\theta})\right) - \lambda_t(\boldsymbol{\theta})$$

The estimation of parameters in Model 2 is done in the same manner.

2.2 Model Evaluation and Validation

The difference between each actual and estimated number of confirmed cases from the model is measured using Pearson residual. The Pearson residual of Model 1 is given by $\mathrm{e_t} = \frac{Y_t - \lambda_t}{\sqrt{\lambda_t}}$, where Y_t is the actual number of confirmed cases and λ_t is the estimated number of confirmed cases from the model. For Model 2, λ_t^* is used instead of λ_t.

The performance of each model in fitting the number of confirmed cases in each state of Malaysia is then compared using the mean squared Pearson residual. The mean squared Pearson residual is given by $\frac{\sum_{t=1}^{N} {e_t}^2}{N-p}$, where e_t is the Pearson residual, N is the number of observances and p is the number of parameters in the model, including the intercept. A lower mean squared Pearson residual indicates the model performs better than the other. The comparison of the model's performance, for each state, is made to determine the best model that can represent the number of confirmed cases in the particular state.

Besides, the Pearson residuals are also used to check the model's adequacy. The model is adequate if the sequence of its Pearson residuals is a sequence of white noise with constant variance [11,12]. The cumulative periodogram is used to examine whether the Pearson residuals consist of white noise. The Pearson residuals is a white noise sequence if the cumulative periodogram lies within the 95% confidence interval.

3 Results and Discussion

3.1 Exploratory Data Analysis

The data of the number of confirmed COVID-19 cases in each state and each federal territory of Malaysia is retrieved from the official website of the Ministry of Health Malaysia. For the sake of brevity, the federal territories (Kuala Lumpur, Putrajaya and Labuan) are also referred as the states of Malaysia throughout the rest of this study. The data ranged from 1 September 2020 until 31 March 2021, a total of 212 daily observations for each state. Table 1 shows some descriptive statistics of the number of confirmed cases in each state. The table is sorted based on the maximum cases in descending order.

According to Table 1, the maximum number of confirmed cases for each state ranges from 29 to 3285, indicating different severity in the states. Observe that the variance is greater than the mean in all states. To have a better insight, the index of dispersion, D for each state is included in the last column of the table, where $D = \frac{\sigma^2}{\mu}$. If $D = 1$, then the data is equidispersed; if $D > 1$, then the data is overdispersed; if $0 \leq D < 1$, then the data is underdispersed. Clearly, the data for all states are overdispersed.

Table 1: Summary of data statistics by state

State	Maximum cases	Mean	Variance	Index of dispersion
Selangor	3285	537.297	295190.760	549.399
Negeri Sembilan	1392	76.571	14208.493	185.560
Perak	1215	59.939	10961.243	182.874
Sabah	1199	256.033	39795.075	155.429
Johor	1103	191.057	61767.902	323.296
Kuala Lumpur	783	166.165	34811.693	209.501
Sarawak	426	74.014	9973.502	134.751
Kedah	397	38.769	2620.065	67.582
Melaka	344	29.623	3362.653	113.516
Penang	337	75.156	5311.848	70.678
Pahang	288	18.377	1000.663	54.451
Kelantan	257	28.726	1458.948	50.788
Terengganu	179	16.410	911.456	55.541
Labuan	105	11.208	350.146	31.242
Putrajaya	39	4.929	58.019	11.770
Perlis	29	1.406	15.873	11.292

3.2 Model Evaluation

Model 1 and Model 2 are fitted to the data of the number of confirmed cases from 1 September 2020 until 31 March 2021, for every state in Malaysia. The models are estimated using the maximum likelihood method and the results are available by the authors. Generally, both models fit the number of confirmed cases well, except for some extreme changes in the number of confirmed cases in some states. Overall, most of the estimations of both models are close to each other. Hence, the model with better performance is determined by comparing the mean squared Pearson residual of the two models. The respective mean squared Pearson residual is shown in Table 2; the lower mean squared Pearson residual in each state is shaded in green.

Table 2: Mean squared Pearson residual of each model for the respective state

	Perlis	Kedah	Penang	Perak	Selangor	Kuala Lumpur	Putrajaya	Negeri Sembilan
Model 1	5.003	26.021	37.149	63.148	104.334	66.126	2.134	112.148

Model 2	5.945	24.597	46.117	79.112	109.441	94.331	2.843	116.816

	Melaka	Johor	Kelantan	Terengganu	Pahang	Sarawak	Labuan	Sabah
Model 1	39.926	39.042	12.205	8.298	33.525	14.532	20.468	31.292
Model 2	41.919	50.655	13.866	10.001	33.432	16.504	24.264	34.700

Based on Table 2, Model 1 achieves a lower mean squared Pearson residual in most states. While Model 2 has a lower mean squared Pearson residual in Kedah and Pahang only. However, the respective mean squared Pearson residual of both models in Kedah and Pahang does not differ much. These suggest Model 1 has better performance than Model 2.

3.3 Model Validation

Figure 1 shows the cumulative periodogram of Model 1 in Perlis and the cumulative periodograms of Model 1 in the rest of the states are shown in Figure 2-Figure 5. Based on Figure 1, part of the cumulative periodograms of Model 1 in Perlis lie outside the 95% confidence interval, indicating Model 1 is not adequate in Perlis. While the cumulative periodograms of Model 1 in the rest of the states lie within the 95% confidence interval, indicating its adequacy in the states. While Model 2 is only adequate for 6 out of 16 states, the results are available by the authors.

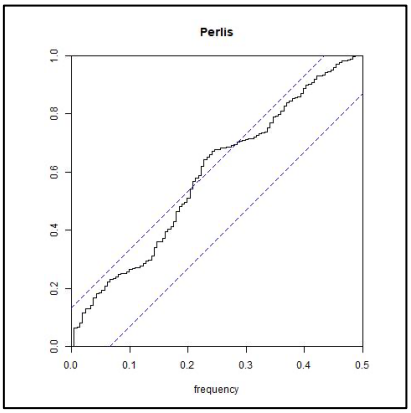


Figure 1: Cumulative periodogram of Model 1 in Perlis

3.4 Model Prediction

Model 1 is then used to predict the number of confirmed cases in April 2021. The predictions of Model 1 (red line) for the states except Perlis are illustrated in Figure 2-Figure 5. The actual and fitted number of confirmed cases from January until March 2021 are included to observe the trend. The parameters of Model 1 and the trend prediction for the respective state are shown in Table 3.

Generally, the predictions from Model 1 are close to the actual number of confirmed cases. For instance, Model 1 is able to predict the number of confirmed cases in Kedah, Selangor, Kuala Lumpur, Melaka, Johor and Sabah. For Perak and Negeri Sembilan, Model 1 also gives a satisfying result of prediction despite fails to fit the sudden spikes during the stage of model estimation. Furthermore, Model 1 has less accurate predictions at some spikes of the number of confirmed cases in Penang, Putrajaya, Terengganu, Pahang and Sarawak. While the predictions on Kelantan are generally lower than its actual number of confirmed cases.

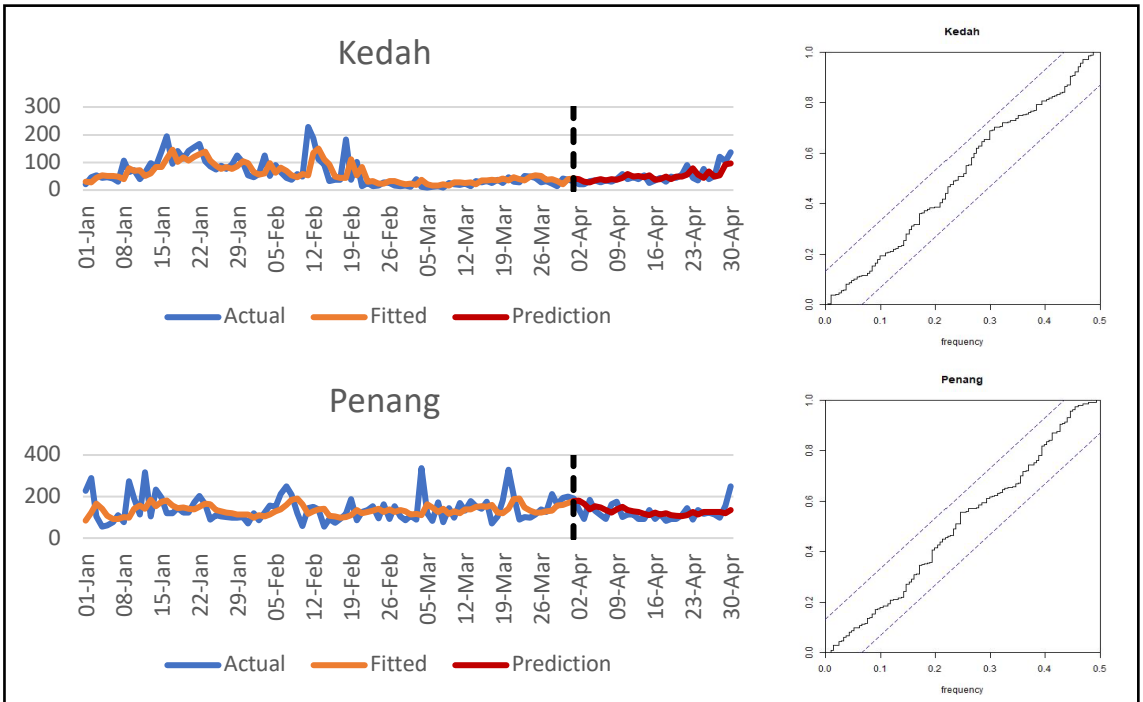
Moreover, it seems that Model 1 has poor predictions on the number of confirmed cases in Labuan. The unstable number of confirmed cases in Labuan and its extreme changes during a short period might affect the accuracy of the model. Nevertheless, the predictions of the states are able to follow the actual data.

On the other hand, the trend prediction made by comparing the values of α and β is not accurate for some states. In fact, some states do not show an obvious increasing or decreasing trend. Hence, comparing the values of the parameters α and β might not be appropriate to predict the spreading trend of COVID-19 in the states of Malaysia.

Table 3: Parameters of Model 1 and trend prediction for the respective state

	Kedah	Penang	Perak	Selangor	Kuala Lumpur	Putrajaya	Negeri Sembilan	Melaka
ω	0.6563	0.4268	0.2050	0.4514	0.3345	0.0000	0.2797	0.4670
α	0.6056	0.3198	0.3038	0.4539	0.2961	0.3805	0.1436	0.4815
β	0.2406	0.5961	0.6622	0.4821	0.6487	0.6114	0.8045	0.4227
Trend	Decrease	Increase	Increase	Increase	Increase	Increase	Increase	Decrease

	Johor	Kelantan	Terengganu	Pahang	Sarawak	Labuan	Sabah
ω	0.2018	0.1276	0.0528	0.4230	0.2324	0.1443	0.3305
α	0.3467	0.3264	0.4838	0.5274	0.4373	0.1562	0.4769
β	0.6227	0.6484	0.5070	0.3655	0.5224	0.8138	0.4680
Trend	Increase	Increase	Increase	Decrease	Increase	Increase	Decrease



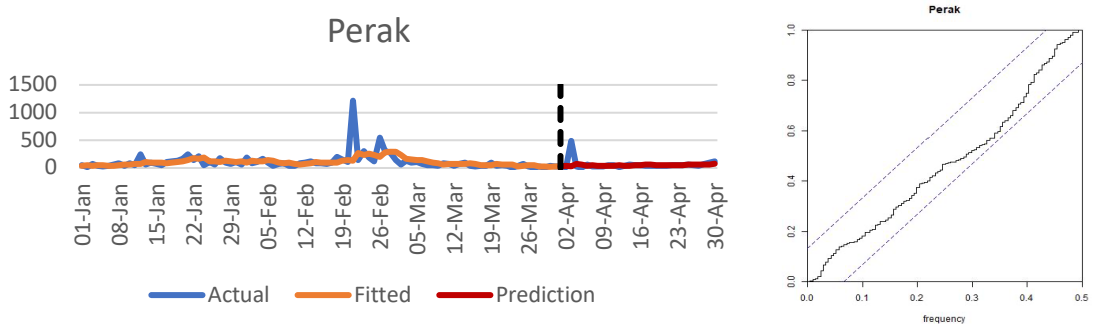
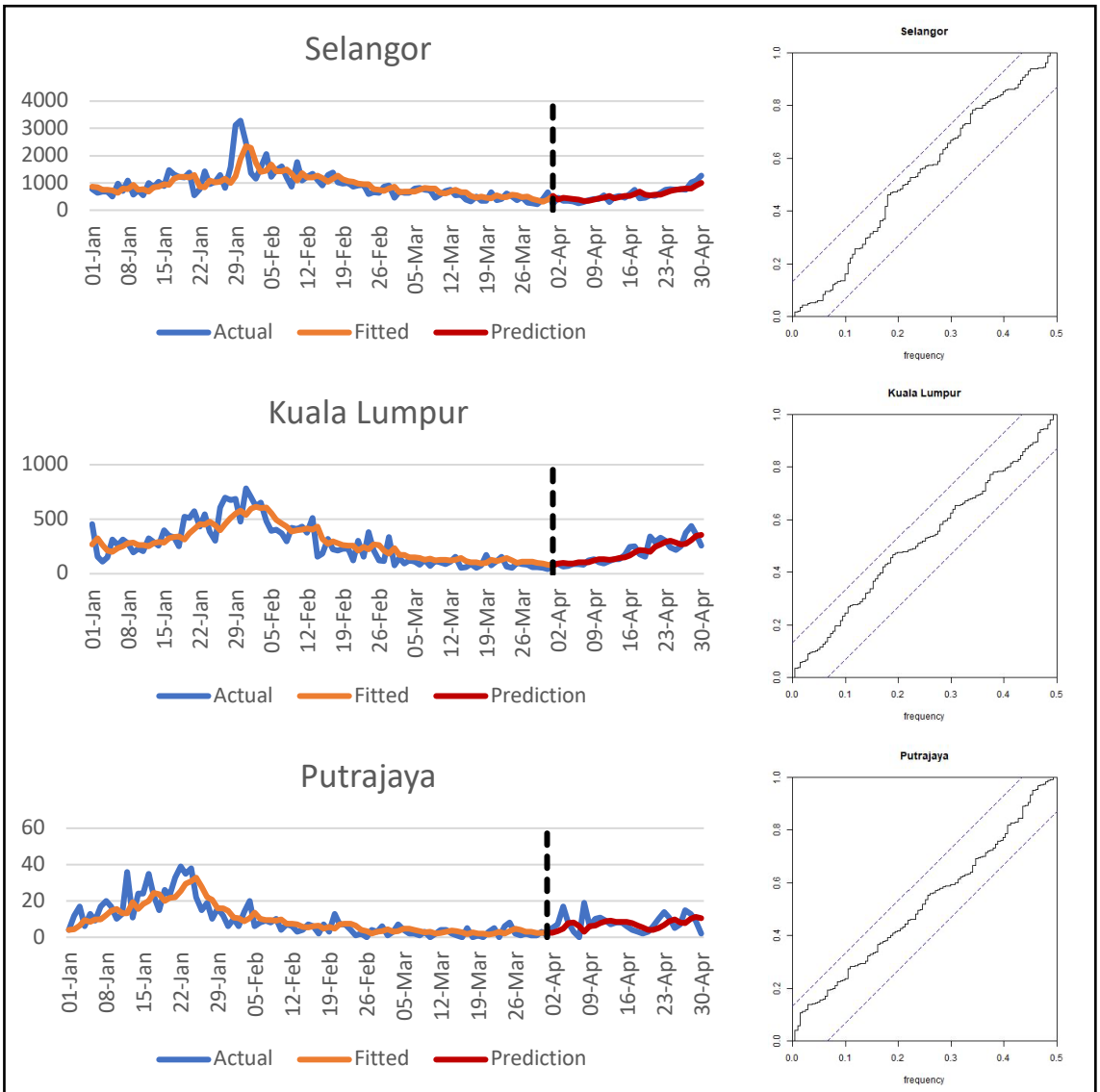


Figure 2: Prediction and cumulative periodogram of Model 1 in Kedah, Penang and Perak



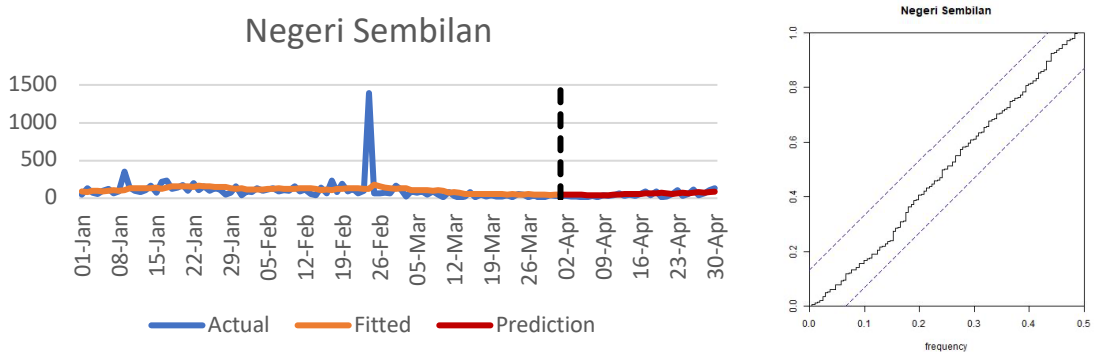
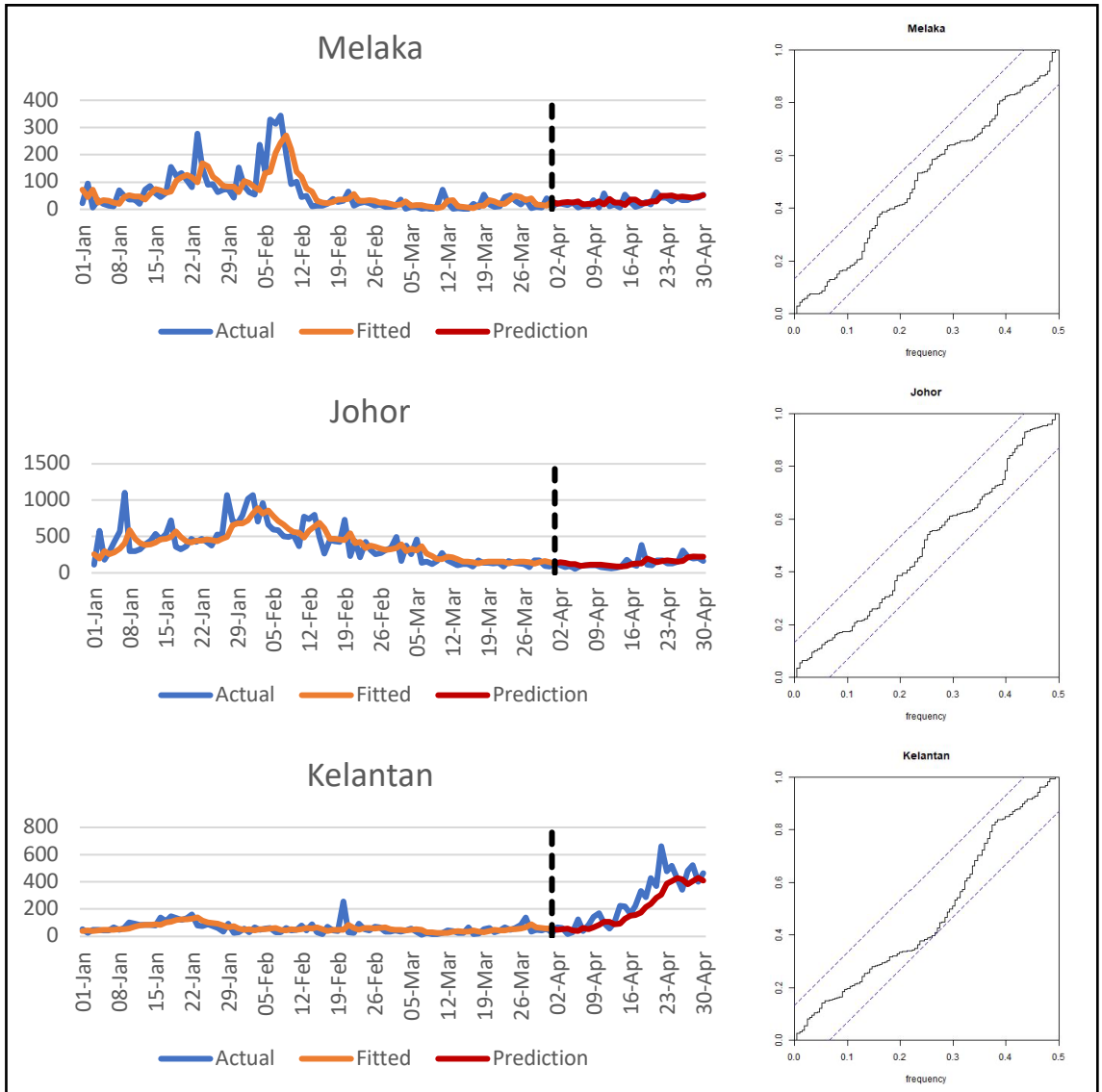


Figure 3: Prediction and cumulative periodogram of Model 1 in Selangor, Kuala Lumpur, Putrajaya and Negeri Sembilan



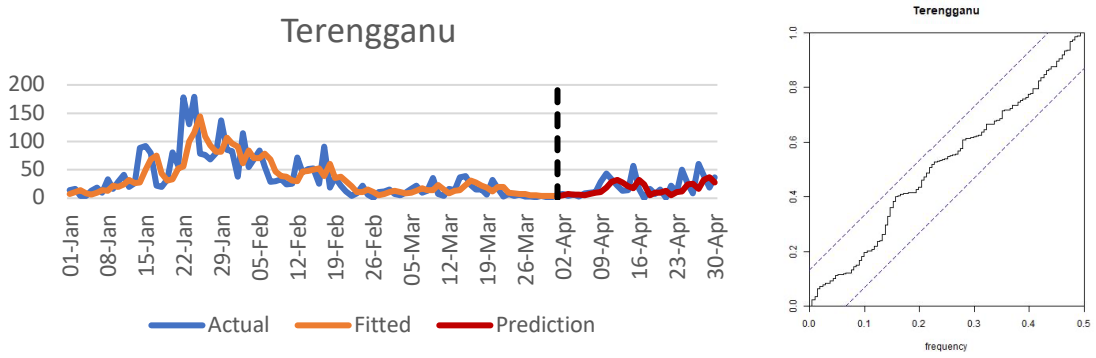
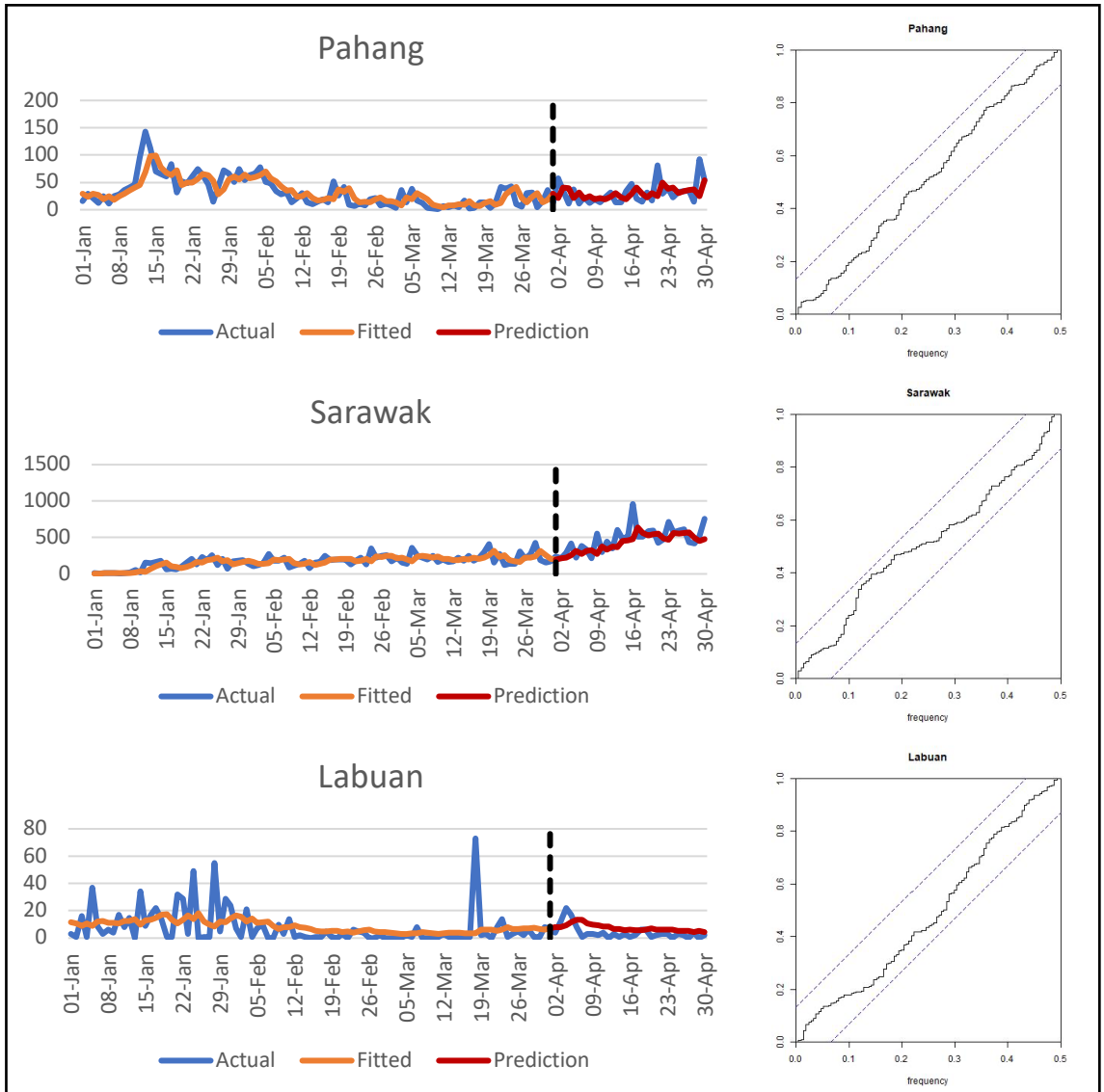


Figure 4: Prediction and cumulative periodogram of Model 1 in Melaka, Johor, Kelantan and Terengganu



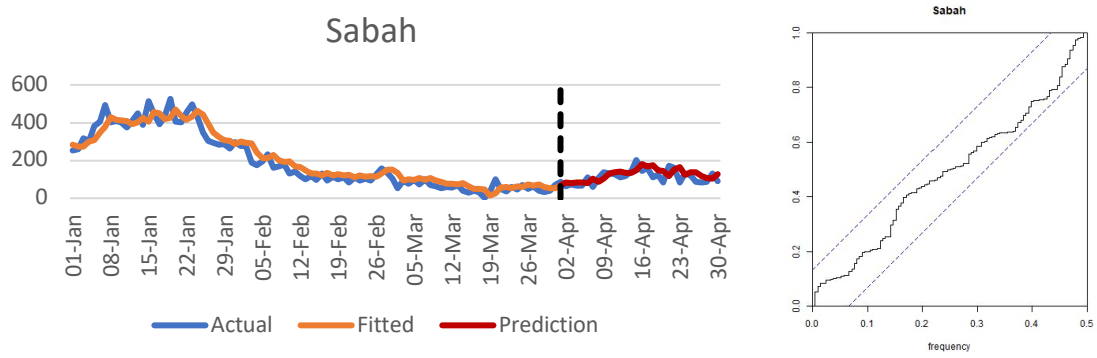


Figure 5: Prediction and cumulative periodogram of Model 1 in Pahang, Sarawak, Labuan and Sabah

4 Conclusion

To conclude, the data of the confirmed COVID-19 cases in each state of Malaysia is overdispersed and an ordinary Poisson model is not suitable for modelling the data. Thus, the log-linear Poisson autoregressive model is used to model the data.

Results find that Model 1 is adequate for modelling the number of confirmed COVID-19 cases in all states of Malaysia, except Perlis. Besides, Model 1 has generally a better performance than Model 2 in modelling the number of confirmed cases in the states of Malaysia during the period of 1 September 2020 until 31 March 2021. Therefore, Model 1 is applied to predict the number of confirmed cases in all states of Malaysia during April 2021, excluding Perlis as the model is not adequate for the state. Although the number of confirmed cases exhibits different behaviours between the training period (1 September 2020 until 31 March 2021) and the testing period (1 April until 30 April 2021), the model is able to provide reasonable predictions on the number of confirmed cases.

Nevertheless, the unstable number of confirmed cases and its extreme changes might affect the accuracy of the model. For instance, Model 1 gives less accurate results in fitting and predicting the number of confirmed cases in Labuan, which exhibited a great range of fluctuation in the number of confirmed cases. Results also show that comparing the values of the parameters α and β might not be appropriate to predict the spreading trend of COVID-19 in the states of Malaysia.

5 References

[1] Katris, C. A time series-based statistical approach outbreak spread forecasting: Application of COVID-19 in Greece. 2020. *Expert Systems With Applications*. 2020. Vol. 166.

[2] Takele, R. Stochastic modelling for predicting COVID-19 prevalence in East Africa Countries. *Infectious Disease Modelling*. 2020. 5: 598-607.

[3] Khan, F. M. & Gupta, R. ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. *Journal of Safety and Resilience*. 2020. 1: 12-18.

[4] Ariffin, M. R. K., Gopal, K., et al. Coronavirus disease 2019 (COVID-19) infectious trend simulation in Malaysia: A mathematical epidemiologic modelling study. 2020.

[5] Mbuvha, R. & Marwala, T. Bayesian inference of COVID-19 spreading rates in South Africa. *PLoS ONE*. 2020. 15(8).

[6] Binti Hamzah, F. A., Lau, C. H., Nazri, H., Ligot, D. V., Lee, G, Tan, C. L., et al. CoronaTracker: World-wide COVID-19 Outbreak data analysis and prediction. *Bulletin of the World Health Organization*. 2020.

[7] Chintalapudi, N., Battineni, G., Sagaro, G. G. & Amenta, F. COVID-19 outbreak reproduction number estimations and forecasting in Marche, Italy. *International Journal of Infectious Diseases*. 2020. 96: 327-333.

[8] Ferland, R., Latour, A. & Oraichi, D. (2006). Integer-valued GARCH process. Journal of Time Series Analysis, 27(6), 923-942.

[9] Agosto, A. & Giudici, P. A Poisson autoregressive model to understand COVID-19 contagion dynamics. *Risks*. 2020. 8(77).

[10] Kharroubi, S. A. Modeling the spread of COVID-19 in Lebanon: A Bayesian perspective. *Frontier in Applied Mathematics and Statistics*. 2020. 6(40).

[11] Fokianos, K., Rahbek, A. & Tjøstheim, D. Poisson autoregression. *Journal of the American Statistical Association*. 2009. 104(488): 1430-1439.

[12] Fokianos, K. & Tjøstheim, D. Log-linear Poisson autoregression. *Journal of Multivariate Analysis*. 2011. 102: 563-578.