Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Application of zero-truncated count data regression models to air-pollution disease

To cite this article: Z I Zulki Alwani *et al* 2021 *J. Phys.: Conf. Ser.* **1988** 012096

View the article online for updates and enhancements.

Application of zero-truncated count data regression models to air-pollution disease

Z I Zulki Alwani[1], **A I N Ibrahim**[1,*], **R M Yunus**[1] **and F Yusof** [2]

[1]Institut Sains Matematik, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
[2]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia

[*]**Email**: adrianaibrahim@um.edu.my

Abstract. Count data consist of non-negative integers that have many applications in various fields of studies. To handle count data, there are various statistical models that can be employed corresponding to the properties of the count data studied. Poisson regression model (PRM) is mostly used to model data with equidispersion, while negative binomial regression model (NBRM) is a model that is regularly employed to model over-dispersed count data. On the other hand, the usual count data regression models may not able to handle strictly positive counts. In this case, the appropriate model for the analysis of such data would be models truncated at zero. We are interested to study the relationship between pollution related disease with influential factors such as air pollution and climate variables in Johor Bahru, Malaysia, using these zero-truncated models, where the number of disease cases are strictly positive. In particular, the zero-truncated PRM and NBRM are used to determine the association between the number of dengue patients and their influential factors. From the study, zero-truncated NBRM is found to be the best model amongst the two models to model the relationship between the number dengue cases and air pollution and climate. Air pollution factors that significantly affect the number of cases for dengue are particulate matter (PM_{10}) and sulfur dioxide. Also, humidity and temperature are the climate factors that significantly affect the number of dengue cases.

1. Introduction

Discrete data or count data is one type of data that only contains non-negative integers. Count number of events is an example of count data. Several distributions are used to model count data. Poisson distribution is widely used for count data. For Poisson distribution, the characteristic is the mean and variance must be the same. However, real data are usually under-dispersed and over-dispersed, thus they do not satisfy the characteristic of Poisson distribution. To deal with over-dispersed data, negative binomial distribution is commonly used. The negative binomial is usually used by placing a Gamma prior distribution [1].

Besides that, most data is over-dispersed. Thus, Saffari (2011) stated that “Response variable in such cases is truncated for some outliers or large values” [2]. For the truncated dataset, analysis using basic count data regression such as Poisson regression may lead to incorrect estimation. For example, even though the data does not contain any zero counts, Poisson regression will try to predict zero counts. Thus, truncated regression is a suitable alternative approach to analyze truncated datasets such as zero-truncated Poisson regression model (PRM) and zero-truncated negative binomial regression model (NBRM). Zero-truncated PRM had been used to estimate the number of fishing trips a household takes in Alaska [3] and the recreational fishing trip demands in Brazilian Pantanal [4]. Besides that, the zero-truncated NBRM is used to estimate demand for economic value to participants in mountain biking in Moab, Utah [5] and recreation demand in Parks Canada between June and September 2004 [6].

Dengue becomes a global burden where it is estimated to have more than 50 million cases every year in more than 100 countries in Africa, America and Southeast Asia ([7], [8]). Dengue fever is caused by four antigenically different single-stranded positive RNA viruses belonging to the Flaviviridae family which are DEN-1, DEN-2, DEN-3 and DEN-4 ([9], [10],

[11]). Dengue occurred during the intermittent pandemic that affecting Asia and America in the 18th and 19th centuries [10]. Originally, the spread of dengue was slow. However, it is changing dramatically during and after World War II.

Dengue is a mosquito-borne viral disease that is transmitted between humans and mosquitoes that can be found in tropical and sub-tropical regions around the world. Dengue is been transmitted by *Aedes* mosquitoes species mainly by *Aedes aegypti* and *Aedes albopictus*. *Aedes aegypti* is the first vector for dengue originally from Africa. In the 15th to 19th century, it emerged from Africa due to the slave trade and spread to Asia in the 18th to 19th century due to commercial trade [12]. Then, *Aedes albopictus* also known as (Asian) forest mosquito is a mosquito native in the tropical and subtropical region in Southeast Area is the secondary vector for dengue. However, due to globalisation of trade such as transportation and travel in the past decades, *Aedes albopictus* had been spread to many countries.

World Health Organization stated that “Dengue fever is a flu-like illness that can affect the infants, children and even the adults” [11]. Dengue can be classified into two types which are dengue (with or without symptoms) and severe dengue. Until now, there is no specific treatment or vaccine for dengue fever.

Various studies had been conducted to investigate the relationship between dengue cases with air pollution and climate factors. In tropical and subtropical areas, dengue cases all year-round and is exceptionally associated with temperature, rainy season, and vector fluctuation with seasonality [13]. Xu et al. (2017) studied the effect of climates which are temperature and rainfall on the dengue in Guangzhou, China using a structural equation model with generalised additive model [14]. The study showed that temperature has a positive effect on dengue incidence. Carneiro et al. (2017) studied the effect of the environment on dengue incidence and showed that humidity had a positive correlation while PM10 had a negative correlation on dengue cases [15]. Jamaludin et al. (2017) stated that humidity and temperature are significant factors to the number of dengue cases in Johor Bahru, Malaysia [16].

In our study, we are interested to study the relationship between dengue cases with influential factors such as air pollution and climate factors in Johor Bahru, Malaysia. Previously, Jamaludin et al. (2017) had used NBRM to model this relationship [16]. However, since the number of dengue disease cases is strictly positive, the zero-truncated models are good alternatives. In particular, we shall use the zero-truncated PRM and NBRM to determine the association between the number of dengue patients and their influential factors.

2. Material and method

2.1 Description of data

The data is taken from Johor Bahru, one of Malaysia's states located at the southernmost city of Peninsular Malaysia. Johor Bahru is one of the urban cities apart from Kuala Lumpur, Selangor and Pulau Pinang. As one of the main urban centers in Peninsular Malaysia, many places have become industrial areas in Johor Bahru such as Pasir Gudang and Tanjung Langsat.

In this study, we are interested to study the relationship between dengue cases with the level of environmental pollution and climate from 1 January 2012 to 31 December 2013. Data on the dengue disease were obtained from the Department of Information and Informatics, Johor Bahru District Health Office. The data of level of environment pollution obtained are Ground Level Ozone (GLO), Nitrogen Dioxide (NO2), Particulate Matter (PM10) and Sulfur Dioxide (SO2). Rainfall, temperature and humidity are the climate variables used in this study. Rainfall data were obtained from the Department of Irrigation and Drainage Malaysia, whereas temperature and humidity data were obtained from the Meteorological Department.

2.2 Methodology

This study is interested in assessing the relationship between dengue disease and the level of environmental pollution and climate using the weekly data. We assume that there are possibilities that the number of cases may not be influenced by the result of the pollution and climate variable in the same week but by the values in earlier weeks. Thus, the lag time is built for Lag 0 (same week), Lag 1 (first week) to Lag 20 (20th week), following Jamaludin et al. (2017) [16]. Pearson correlation [17] is used to determine the association between the number of dengue cases and the explanatory variables at various lags. Then, using the significant lags, zero-truncated PRM and zero-truncated NBRM are fitted to the data.

2.2.1 Zero-truncated PRM. The probability function of zero-truncated Poisson can be expressed as

$$f(y_i\colon \mu_i|y_i > 0) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!\,(1 - e^{-\mu_i})} \tag{1}$$

where μ_i is the mean for observation y_i. In this case, the zero-truncated PRM in equation (1) follows ordinary PRM where the canonical link is log link, such that $\log(\mu_i)$ defines the linear combination of the explanatory variables. (For further information, see [3], [18].)

2.2.2 Zero-truncated NBRM. The probability function of zero-truncated negative binomial can be expressed as

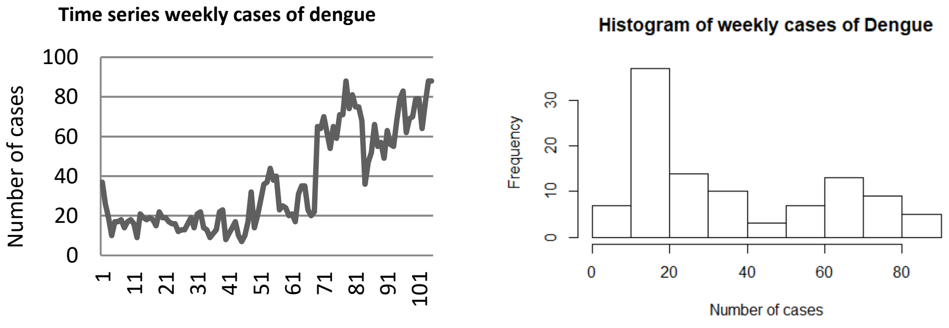
$$f(y_i\colon \alpha, \mu_i|y_i > 0) = \frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha)\,\Gamma(\,y_i + 1)} \left(\frac{\alpha}{\mu_i +\ \alpha}\right)^{\alpha} \left(1 - \frac{\alpha}{\mu_i +\ \alpha}\right)^{y_i} \tag{2}$$

with mean, μ_i, and shape parameter, α, and $\Gamma(.)$ is the gamma function. The zero-truncated NBRM in equation (2) also has the log link as the canonical link, similar to the zero-truncated PRM in equation (1). (For further information, see [3], [18].)

3. Results and discussion

3.1 Descriptive data

In this study, weekly number of dengue cases from 2012 to 2013 is used. The number of weeks included in this study is 105 weeks starting from January 2012 to December 2013. Figure 1(a) shows the time series plot of the total number of dengue cases from 2012 to 2013. In general, the number of cases for dengue recorded shows an increasing trend from week 69 (April 2013) and week 86 (August 2013). The maximum number of dengue cases is 88 cases recorded in December 2013 (week 104 and week 105) and the minimum reading recorded is 7 in November 2012 (week 45) followed by 8 cases in September 2012 (week 40). Figure 1(a) also shows an increasing pattern for the total number of dengue cases in 2012 and the total number of dengue cases in 2013. 933 dengue cases were reported in 2012 while in 2013, the total dengue cases are 2902. This means that dengue cases in 2013 are three times larger compared to dengue cases in 2012. In total, the dengue cases in Johor Bahru from 2012 to 2013 are 3835 cases.



(a) (b)

Figure 1. Graph of weekly dengue cases from January 2012 to December 2013: (a) Time series of weekly dengue cases; (b) Histogram of weekly dengue cases.

3.2 Descriptive statistic for dengue cases

Table 1 shows the descriptive statistics for dengue cases in Johor Bahru from 2012 to 2013. The mean of dengue cases is 37, median 23, standard deviation 24.7072, skewness 0.6555, kurtosis 1.9251 and variance 610.4441. From the result of descriptive statistic in table 1, it is clearly shown that dengue cases are over-dispersed since the variance is much larger than the mean. Then, the skewness shows that dengue is skewed to the right and it is shown in figure 1 (b).

Table 1. Descriptive statistics of dengue cases.

Min	Max	Mean	Median	Standard Deviation	Skewness	Kurtosis	Variance
7	88	37	23	24.7072	0.6555	1.9251	610.444

3.3 Effect of air pollution and climate on number of disease

Weekly data of air pollution and climate variables from 2012 to 2013 are used to find the effect of these variables on the number of weekly dengue cases. Air pollution used in this study are Ground Level Ozone (GLO), Nitrogen Dioxide (NO2), Particulate Matter (PM10) and Sulfur Dioxide (SO2). The climates used are humidity, temperature and rainfall.

Particular matter (PM) is particulate that suspended in the air comes with various sizes such as 10 micrometer diameter called PM10 and 2.5 micrometer diameter called PM2.5. PM comes from natural activities such as volcano, dust storm or sea spray and from human activities such as road dust, burning fossil fuels in vehicles and industrial industry. In figure 2(a), PM10 has consistent pattern with mean 42.7020. However, the index of PM10 dramatically increases in week 76 to 77 and decreases in week 77 to 78. From the figure, PM10 might not be significant to dengue cases.

GLO can occur in any place and any time but the concentration peak in the afternoon when the sunlight is most intense. GLO may affect human health negatively as when inhaled, within the tract, it react chemically with many biological molecule. GLO in figure 2(b) show decreasing pattern in the early and end of the year. However, GLO does not show any significant effect on the dengue cases.

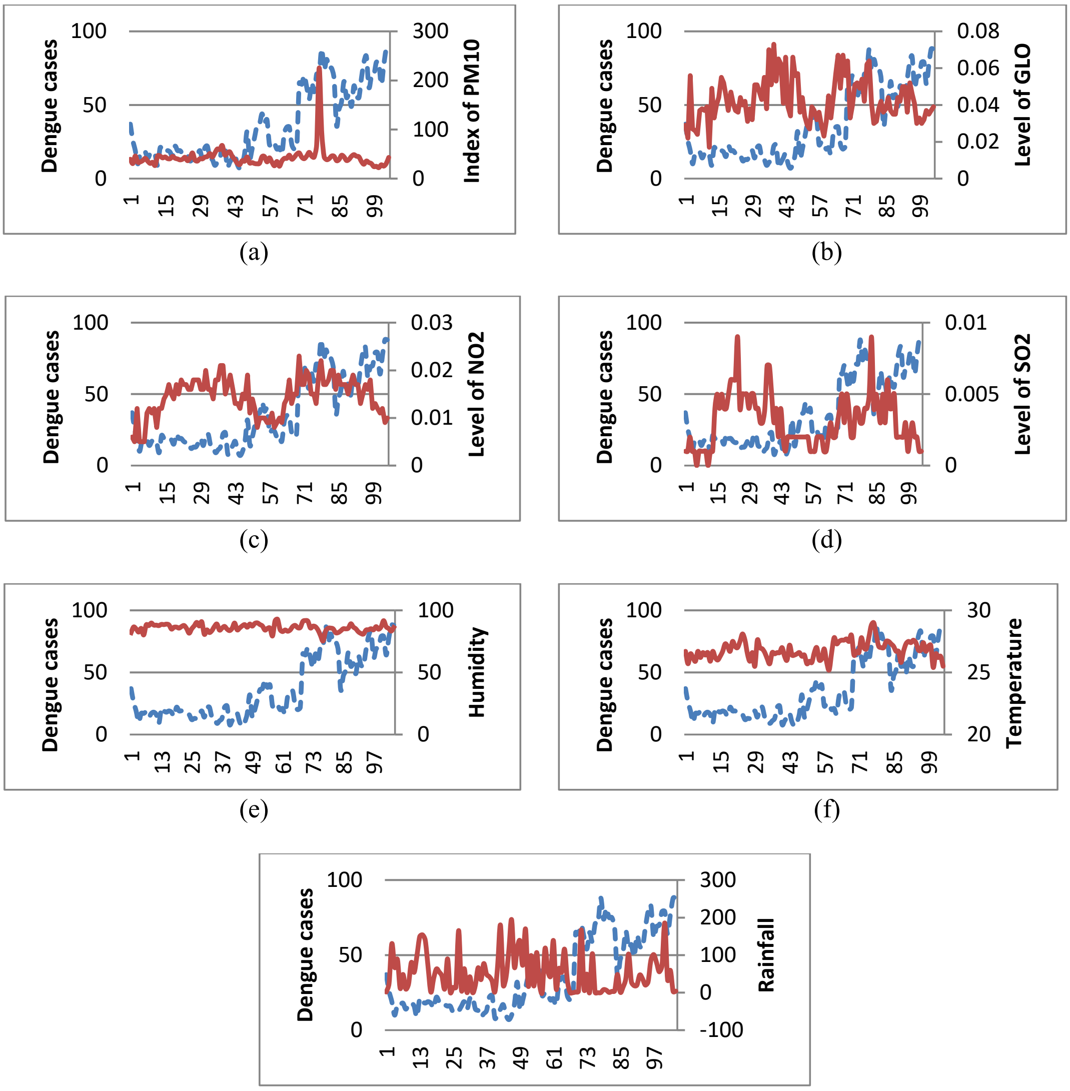
NO2 will irritate the human airways of the respiratory system as when breathing with high concentration NO2 and exposure over short period of NO2 can develop respiratory related disease. It believes that NO2 come to the air from burning of fuel and vehicle exhaust. From figure 2(c), dengue cases increase when level of NO2 decreases.

The primary source of SO2 in the air is burning of foil and industrial activities. Difficulties in breathing and harmful for human respiratory are the effect of short exposure to SO2. Based on figure 2(d), the level of SO2 show the decreasing pattern in the early and end of the year. When level of SO2 decreases, the number of dengue cases increases.

All year round, Johor Bahru has consistent rainfall, humidity and temperature. However, in Malaysia every year there are two monsoon seasons: 1) North-East Monsoon take place between December to February with heavy rain and wind, and 2) Southwest Monsoon take place between June to August with drought and windblown. Other than that, the climate of Johor Bahru is moderate rain.

In figure 2(e), humidity shows consistent pattern and might affect the number of dengue cases. Then, temperature and rainfall also show fluctuating pattern through the year based on figures 2(f) and 2(g). At the end of the year, when the temperature decreases, the number of dengue cases increases. However, rainfall does not show any particular effect on the dengue cases.

Note that the relationship between disease and climate variables and pollutants cannot be clearly identified only through a plot of time series. Therefore, further analysis using regression model shall be done.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

Figure 2. Time series of dengue and air pollution and climate (**dashed blue** = dengue, **solid red** = air pollution or climate variable).

3.4 Relationship between the dengue disease and pollution and climate

Pearson correlation was used to find the correlation between dengue disease and pollution and climate variables. The duration used is from the same week (Lag 0) to week 20 (Lag 20). The time lag with the highest value of Pearson correlation between each explanatory variable and the response is chosen to be the significant lag. The results of Pearson correlation between dengue disease and influential factors are shown in table 2.

From table 2, GLO, NO2, PM10, SO2 and temperature have a positive association with the dengue cases while rainfall and humidity have a negative association with the cases, at different lags. This means GLO affects the number of dengue cases after 13 weeks (lag 13), NO2 after 8 weeks (lag 8) and so on. We can also see that dengue cases shows moderate association with temperature and the lowest association with SO2. The significant lags for each variable will be used in the modeling approach that will be explained in the next section.

Table 2. Result of Pearson correlation of the dengue disease with its influential factors at the significant lags.

Influential factor		Significant lag	Pearson correlation
Air pollution	GLO	13	0.0641
	NO2	8	0.1620
	PM10	18	0.2105
	SO2	11	- 0.0609
Climate	Rainfall	13	- 0.2638
	Humidity	15	- 0.2477
	Temperature	5	0.3788

3.5 The selection of model

Previously, Jamaludin et al. (2017) had used the NBRM to model the relationship between the dengue cases and the pollution and climate factors. For our dengue data, an over-dispersion test has been done. From this test, the *p*-value is less than 0.05, thus the data set is over-dispersed. However, since the weekly number of dengue cases shown in figure 1(a) does not contain any zero, zero-truncated PRM and zero-truncated NBRM might be appropriate alternative models for this data.

First, all the variables with the corresponding significant lags are fitted in the selected model. Then, only the significant variables are included in the final model after using stepwise selection based on the Akaike information criterion (AIC) [17]. Finally, adequacy checking for the assumption of normality for the final model is done based on Q-Q plot, Shapiro-Wilk (SW) test, Jarque-Berra (JB) test and D' Agostino (DA) test. All the analyses are done using Excel and R. For zero-truncated PRM and zero-truncated NBRM, the package VGAM in R is used where iterative method is applied to obtain the maximum likelihood estimates [19].

The results of the final two models after stepwise selection are displayed in tables 3 and 4. From table 3, for zero-truncated PRM, the significant variables are NO2, PM10, SO2, rainfall, humidity and temperature. Zero-truncated NBRM has significant variables which are PM10, SO2, humidity and temperature. Based on significant variables for each model, PM10, SO2, humidity and temperature significantly affect the number of dengue cases.

Based on the result of normality checking and the Q-Q plot in table 4, the residuals for zero-truncated PRM is normal. For the zero-truncated NBRM, the normality assumption of residual is approximately satisfied since the null hypothesis that residuals are normally distributed is rejected for the SW test (*p*-value is less than 0.05; however, still larger than 0.01) but not rejected for the DA and JB tests.

For further selection of the best model amongst the two models, the model deviance is used [17]. Looking at the values of the deviance presented in table 3, zero-truncated NBRM has much lower deviance compared to zero-truncated PRM. Thus, this indicates that zero-truncated NBRM might be the better model to model the effect of the air pollution and climate on the number of dengue cases in Johor Bahru, Malaysia.

4. Conclusion

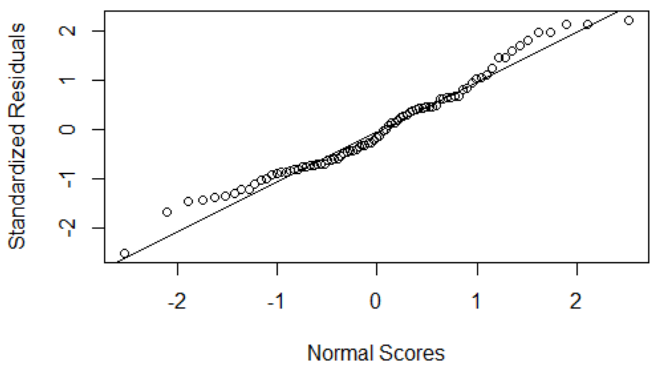
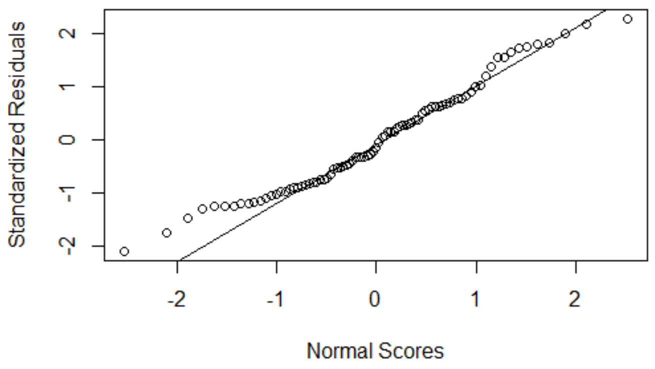
In this study, the number of air-pollution disease cases which is dengue fever in Johor Bahru from 2012 to 2013 has been studied. The data of the predictor variables which are the level of environmental pollution and climate variables are used. Pearson correlation is used to find the significant lags (weeks) between the number dengue cases and its explanatory variables.

Two regression models for count data have been used in this study which are zero-truncated PRM and zero-truncated NBRM. These two models were used to find the relationship between the number of dengue cases and its influential factors. We have found that zero-truncated NBRM is the best model amongst the two models to model the relationship between the number dengue cases and air pollution and climate. Air pollution variables that significantly affect the number of cases for dengue are PM10 and SO2. Also, humidity and temperature are the climate variables that significantly affect the number of dengue cases.

Table 3. Summary of dengue cases modelling using zero-truncated PRM and zero-truncated NBRM (final model).

Model		**Coeff.**	**Std. error**	**p-value**	**AIC**	**Deviance**
Zero-Truncated PRM	Intercept	-0.9703	0.778	0.212	1485.792	-1471.79
	NO2 (Lag 8)	16.39	5.837	0.005		
	PM10 (Lag 18)	3.621×10^{-3}	5.561×10^{-4}	< 0.001		
	SO2 (Lag 11)	-53.11	10.79	< 0.001		
	Rainfall (Lag 13)	-0.354×10^{-3}	3.953×10^{-4}	0.001		
	Humidity (Lag 15)	-4.797×10^{-2}	5.612×10^{-3}	< 0.001		
	Temperature (Lag 5)	0.320	2.530×10^{-2}	< 0.001		
Zero-Truncated NBRM	Intercept	-0.848	3.012	0.780	765.106	-753.11
	PM10 (Lag 18)	5.128×10^{-3}	2.848×10^{-3}	0.072		
	SO2 (Lag 11)	-54.419	36.668	0.138		
	Humidity (Lag 15)	-5.368×10^{-2}	2.129×10^{-2}	0.012		
	Temperature (Lag 5)	0.338	9.200×10^{-2}	< 0.001		

Table 4. Q-Q plot and normality test results for final model.

Model	**Q-Q plot**	**Normality test (*p*-value)**
Zero-Truncated PRM		SW : 0.088 DA: 0.371 JB: 0.368
Zero-Truncated NBRM		SW : 0.039 DA: 0.129 JB: 0.187

Note that this study has some limitation that is, this study only examines the effects of air pollution and climate on dengue cases based on the recorded data for a particular period of time. Therefore, the effects of these factors on patients beyond the study period still remain unknown.

In future studies, we might consider other methods for choosing the time lag of air pollution and meteorological variables that affect the number of disease cases. We might also consider other count data models and/or methods such as generalised additive model (GAM) that can include the effect of time lag directly in the model.

Acknowledgements

This work was financially supported by the Ministry of Higher Education, Malaysia under the Fundamental Research Grant Scheme No: FP110-2019A.

Reference

[1] Avci E 2018 Using count regression models to determine the factors which effects the hospitalization number of people with schizophrenia *Journal of Data Science* **16(3)** 511-28

[2] Saffari S E, Adnan R, and Greene W 2011 Handling of over-dispersion of count data via truncation using Poisson regression model *Journal of Computer Science and Computational Mathematics* **1(1)** 1-4

[3] Grogger J T and Carson R T 1991 Models for truncated counts *Journal of applied econometrics* **6(3)** 225-38

[4] Shrestha R K, Seidl A F and Moraes A S 2002 Value of recreational fishing in the Brazilian Pantanal: a travel cost analysis using count data models *Ecological Economics* **42(1-2)** 289-99

[5] Chakraborty K and Keith J E 2000 Estimating the recreation demand and economic value of mountain biking in Moab, Utah: an application of count data models *Journal of Environmental Planning and Management* **43(4)** 461-9

[6] Martinez-Espineira R and Amoako-Tuffour J 2008 Recreation demand analysis under truncation, overdispersion, and endogenous stratification: an application to Gros Morne National Park. *Journal of Environmental Management* **88(4)** 1320-32

[7] World Health Organization, Special Programme for Research, Training in Tropical Diseases, World Health Organization. Department of Control of Neglected Tropical Diseases, World Health Organization. Epidemic, & Pandemic Alert 2009 *Dengue: guidelines for diagnosis, treatment, prevention and control*. World Health Organization

[8] Whitehorn J and Farrar J 2010 Dengue *British Medical Bulletin* **95(1)** 161-73

[9] Chambers T J, Hahn C S, Galler R and Rice C M 1990 Flavivirus genome organization, expression, and replication *Annu. Rev. Microbiol*. **44** 649-88

[10] Monath T P 1994 Dengue: the risk to developed and developing countries *Proceedings of the National Academy of Sciences* **91(7)** 2395-400

[11] World Health Organization 2014 *Dengue and severe dengue* (No. WHO-EM/MAC/032/E). World Health Organization. Regional Office for the Eastern Mediterranean

[12] Simmons C P, Farrar J J, van Vinh Chau N and Wills B 2012 Dengue *New England Journal of Medicine* **366(15)** 1423-32

[13] Jing Q and Wang M 2019 Dengue epidemiology *Global Health Journal* **3(2)** 37-45

[14] Xu L, Stige L C, Chan K S, Zhou J, Yang J, Sang S, ... and Stenseth N C 2017 Climate variation drives dengue dynamics *Proceedings of the National Academy of Sciences* **114(1)** 113-8

[15] Carneiro M A F, Alves B D C, Gehrke F D S, Domingues J N, Sá N, Paixão S, ... and Fonseca F 2017 Environmental factors can influence dengue reported cases *Revista da Associação Médica Brasileira* **63(11)** 957-61

[16] Jamaludin A R B, Yusof F, Lokoman R M, Noor Z Z, Alias N and Aris N M 2017 Correlational study of air pollution-related diseases (asthma, conjunctivitis, URTI and

dengue) in Johor Bahru, Malaysia *Malays. J. Fundam. Appl. Sci* **13** 354-61

[17] Montgomery D C Peck E A and Vining G G 2012 *Introduction to linear regression analysis* (New Jersey: John Wiley & Sons)

[18] Jing Zuur A F, Ieno E N, Walker N J, Saveliev A A and Smith G M 2009 GLM and GAM for count data *Mixed effects models and extensions in ecology with R* (New York: Springer) pp. 209-243

[19] Yee T W 2021 *VGAM: Vector Generalized Linear and Additive Models* R package version 1.1-5 https://CRAN.R-project.org/package=VGAM