Open Science Journal of Statistics and Application
2018; 5(3): 47-57
http://www.openscienceonline.com/journal/osjsa
ISSN: 2381-4950 (Print); ISSN: 2381-4969 (Online)

Bivariate Negative Binomial Modelling of Epidemiological Data

Shenali Maryse Fernando, Marina Roshini Sooriyarachchi*

Department of Statistics, University of Colombo, Colombo, Sri Lanka

Email address

shenalifernando@yahoo.com (S. M. Fernando), roshinis@hotmail.com (M. R. Sooriyarachchi)

*Corresponding author

To cite this article

Shenali Maryse Fernando, Marina Roshini Sooriyarachchi. Bivariate Negative Binomial Modelling of Epidemiological Data. *Open Science Journal of Statistics and Application*. Vol. 5, No. 3, 2018, pp. 47-57.

Received: April 16, 2018; **Accepted**: May 15, 2018; **Published**: August 16, 2018

Abstract

Dengue fever and Leptospirosis (rat fever) are two of the most common zoonotic diseases in countries with tropical or subtropical climates. Both these diseases can develop into an epidemic situation. Many similar characteristics such as the variation of incidence with climatic variables and comparable clinical manifestation in the diseases can be seen in dengue and rat fever. The life threatening nature of the two diseases and the widespread nature of the diseases across Sri Lanka, have caused much concern amongst the society. This study was carried out with the objective of determining the bivariate distribution of the counts of dengue fever and rat fever, and identifying the determinants with regard to climatic factors. (Rainfall, humidity, temperature and their first two lag values). Generalized linear mixed models (GLMM) within the 'Glimmix' procedure on 'SAS' software was used to model the incidence of the two diseases. The study was based on data of the counts of the two diseases and the climatic variables obtained from three districts of the Western province of Sri Lanka, for the period year 2010- year 2016. This study showed that the bivariate modelling of the incidence of dengue fever and rat fever could be adequately done using a GLMM with a Negative Binomial distribution. A cluster effect was assumed within districts. Responses were also believed to be correlated over time. The correlation structure was accommodated using an autoregressive procedure of order one. Rainfall and its 2^{nd} lag and the 2^{nd} lag of humidity were associated with dengue fever, while the 2^{nd} lag of humidity were associated with rat fever, in the joint model. Further, both these diseases showed a changing pattern over time. The internal and external validation showed that the model predicts well.

Keywords

Bivariate Model, Generalized Linear Mixed Model, Negative Binomial Distribution, SAS, Correlation

1. Introduction

1.1. Relationship Between Dengue Fever and Leptospirosis and the Meteorological Variables

Both dengue fever and rat fever are dependent on climatic factors. These diseases have been found to be associated with variables such as rainfall, temperature and humidity. Global climatic change is predicted to accelerate over the next few decades. An increased frequency, intensity and duration of extreme climatic events is more likely, hence affecting the transmission of dengue fever and rat fever. Thereby, this becomes a global public health priority. Improved understanding of the relationship that these two diseases have with the climate is an important step towards finding ways to mitigate the impact of them on communities. [22, 24, 14].

The relationship that dengue fever and rat fever have with climatic variables, have been analysed in a univariate manner by many researchers in the past. [17, 12, 23, 25], have shown that rainfall, temperature and humidity and their lag terms play an important role in the incidence of both diseases. Favourable temperature for mosquito growth may accelerate the metabolic process in the *Aedes* mosquitoes. Low

humidity causes mosquitoes to feed more freequently to compensate for the dehydration. Rainfall influences the density of the mosquitoes by increasing their breeding places [2]. Thus, incidence of dengue fever is influenced by the climatic factors. [20] has indicated that, increase of the 2nd lag of rainfall caused the incidence of rat fever to rise. A study conducted in Reunion islands showed a significant positive correlation between temperature recorded 0, 1 and 2 months previously and the monthly number of rat fever patients [4]. A study conducted in Sri Lanka has shown the relationship that the incidence of rat fever has with rainfall, relative humidity and temperature. [3]

1.2. Medically Plausible Association Between Dengue Fever and Rat Fever

Dengue and Leptospirosis are two deceases with many common factors. Both diseases are endemic in countries with subtropical or tropical climates and have epidemic potential. Throughout the year cases of dengue and leptospirosis were seen. The incidence of both diseases peaks during the monsoon, leading to concurrent epidemics. The clinical manifestations of leptospirosis and dengue range from a mild self-limiting febrile illness to a severe and potentially fatal illness. When acute co-infection is present clinical diagnosis becomes quite challenging. Physicians who treat, face this challenge due to the vast overlapping spectrum of symptomatic manifestations of dengue and leptospirosis. [31]

Both diseases show similar clinical manifestations during the initial phase. It is expected that acute dual infections may occur due to simultaneous transmission during the rainy season. Such co-infections have been reported rarely as an uncommon occurrence. However, the results of the study done by [29] show that co-infections are not uncommon in Chandigarh and suburbs of India, but rather the diagnosis of leptospirosis is often overshadowed as outbreaks of dengue is common and for longer periods as compared to leptospirosis. During a confirmed Leptospirosis outbreak a reverse scenario may occur. Dual infection may possibly change the clinical spectrum to a more prominent one, presenting a diagnostic dilemma. [19, 29]

The above facts suggests medical plausibility of coinfection but no statistical studies have been done on this. Therefore, a bivariate model should be used in this study.

1.3. Objectives of the Study

This study was conducted to achieve two main objectives,

1. To determine the bivariate distribution of the counts of dengue fever and leptospirosis.
2. To identify the determinants of dengue fever and leptospirosis.

1.4. Data for the Study

The monthly counts of dengue fever and leptospirosis were obtained from the Epidemiology Unit, Medical Statistics Bureau, Colombo, 10. This data set consists of district-wise details (Colombo, Gampaha and Kalutara, all in the Western province) of dengue and leptospirosis patients reported from private and government hospitals during the period 2010 - 2016. The districts were selected based on the count of patients of the two diseases.

Climatic data were obtained from the Meteorology Department, Colombo 7. The data set consists of three climatic variables, total rainfall for the month (mm), mean relative humidity for the month (%) and mean temperature for the month (°C) of the three selected districts for the period 2010-2016. First two lag values of each of the climatic variables were also considered for the study.

Annual district wise population data for the period 2010 - 2016 were obtained from the Registrar General's Department, Sri Jayawardenapura Kotte. The final data set consists of two response variables, nine climatic variables and two random effects and an offset variable. The detailed description of the data is given in table 1.

***Table** 错误!文档中没有指定样式的文字。. Description of data.*

Variable	Notation
District	District
Year	Year
Month	Month
Count of Dengue	CountD
Count of Rat fever	Countrf
Rainfall (mm)	Rain
Log (rainfall) (mm)	Lrain
Rainfall lag1 (mm)	Rainl1
Rainfall lag2 (mm)	Rainl2
Log (Rainfall) lag1 (mm)	Lrainl1
Log (Rainfall) lag2 (mm)	Lrainl2
Humidity (%)	Humid
Humidity lag1 (%)	Humidl1
Humidity lag2 (%)	Humidl2
Temperature (°C)	Temp
Temperature lag1 (°C)	Templ1
Temperature lag2 (°C)	Templ2
expected number of disease cases	Exp_{ijk1} (dengue) Exp_{ijk2} (rat fever)

2. Methodology

Initially the variables for the study were selected based on the study objectives, considering the existing literature and expert advice. [12],[23]. Based on the findings from [25] and other literature, the lag effects of the climatic variables were taken into account. Thus the impact of the climatic variables in the current month and their first two lag values were used in the analysis.

Having done a descriptive analysis, a modification of the Zhang and Boos test was used to assess the association between the categorical explanatory variables and the response variables in a univariate manner.

An advanced analysis was done to model dengue and leptospirosis counts in a univariate and bivariate manner. The data were considered to be clustered within groups (districts). Joint modeling was used via 'Generalized Linear Mixed Models'. The estimates were obtained using the PROC GLIMMIX procedure in SAS software. An autoregressive procedure was used to adjust for the fact that the responses were correlated over time within districts.

2.1. Generalized Linear Mixed Models (GLMM)

The general linear mixed model is of the form

$$y = X\beta + Z\gamma + \varepsilon \quad (1)$$

Where y is a (n x 1) column vector, the outcome variable; X is a (n x p) design matrix of the p predictor variables; β is a (p x 1) column vector of the fixed effects regression coefficients. Here Z is the (n x q) design matrix of the q random effects; γ is a (q x 1) vector of r random effects and ε is a (n x 1) column vector of errors (the part of y that is not explained by the model).

In classical statistics $\gamma \sim N(0, G)$ is assumed. Also y is considered to have a normal distribution, where G is the variance-covariance matrix of the random effects.

Extending this model to responses from any distribution of the exponential family, Generalized Linear Mixed models are obtained. These models are of the form

$$E[Y|\gamma] = g^{-1}(X\beta + Z\gamma) \quad (2)$$

Where g (.) is a differentiable monotonic link function and $g^{-1}(.)$ is its inverse.

The GLMM contains a linear mixed model inside the inverse link function. This model component is referred to as the linear predictor,

$$\eta = X\beta + Z\gamma \quad (3)$$

The most common residual covariance structure is

$$R = I\Sigma_{\varepsilon}^{2} \quad (4)$$

Where I is the identity matrix and Σ_{ε}^{2} is the residual variance. This structure assumes a homogenous residual variance for all (conditional) observations and that they are (conditionally) independent. Other structures can be assumed such as compound symmetry or autoregressive. The G terminology is common in SAS, and also leads to referring to G-side structures for the variance covariance matrix of the random effects and R-side structures for the residual variance covariance matrix. [10]

2.2. Fitting a Negative Binomial Model for Clustered, time Correlated Data

Count data in epidemiological studies assumes only non-negative integer values. (i.e. 0, 1, 2…). If a Normal model is fitted, there is a possibility that it will produce negative predicted counts. Hence, count data is usually modeled using the Poisson distribution, which is characterized by having equal mean and variance of the response variable (Y_i). Thereby, $E(Y_i) = Var(Y_i) = \mu_i$. However, in instances where over-dispersion is present, i.e. when $E(Y_i) < Var(Y_i)$, the Poisson model will no longer be appropriate. In such situations, the Negative Binomial model can be used. The Negative Binomial model is denoted by $Y_i \sim NB(\mu_i, \mu_i + \alpha \mu_i^2)$, where $E(Y_i) = \mu_i$, $V(Y_i) = \mu_i + \alpha \mu_i^2$ and α controls for the over-dispersion. [29]

Initially, modelling was done using a Poisson model, in this study. However, as the data showed signs of over-dispersion, it was decided to use a Negative Binomial model. [25, 29]

2.3. Offset Variable

Each geographic unit considered (District) that is prone to have dengue and rat fever will have a different population size. Therefore an additional parameter known as the 'offset' will be used to incorporate the rate of the disease reported rather than the count. [13], [26].

The offset is calculated as follows [13]

$$Offset = log\ (base\ e)\ (expected\ number\ of\ disease\ cases) \quad (5)$$

This quantity depends on the population size of each district for a given year.

The expected number if everyone has the same exposure (i=month, j=year, k=district indexed cell)

$$Rate = Total\ count/Total\ exposed \quad (6)$$

$$(Exp)_{ijk} = (Exposed)_{ijk} * Rate \quad (7)$$

Due to the unavailability of monthly population sizes of each district, and using the prior knowledge that the population sizes for a given district does not change significantly over a period of one year, the same "offset" was used for all the months of the year for a given district.

2.4. The Negative Binomial Regression Model (NBP)

Several parameterizations are available for bivariate negative binomial regression. Two well-known models are NB-1 and NB-2. [15, 31, 18, 8]. Recently, the functional form of NB regression has been extended and introduced as NB-P regression, where both NB-1 and NB-2 regressions are special cases of NB-P when P=1 and P=2 respectively. [32, 11]. The advantage of using NB-P is that it parametrically nests both NB-1 and NB-2, and therefore, allowing statistical tests of the two functional forms against a more general alternative.

The Univariate Negative Binomial-P regression model (NB-P) is given by

$$pr(y_{i,}) = \left(\frac{\Gamma(y_i + \alpha^{-1}\mu_i^{2-p})}{y_i!\,\Gamma(\alpha^{-1}\mu_i^{2-p})}\right)\left(\frac{\alpha^{-1}\mu_i^{2-p}}{\alpha^{-1}\mu_i^{2-p}+\mu_i}\right)^{\alpha^{-1}\mu_i^{2-p}}\left(\frac{\mu_i}{\alpha^{-1}\mu_i^{2-p}+\mu_i}\right)^{y_i} \quad (8)$$

Where $v_i^{-1} = \alpha$ is the dispersion parameter. The mean and the variance of NBR are $E(y_i) = \mu_i$ and $var(y_i) = v_i = (\mu_i + \alpha \mu_i^2)$.

Bivariate Negative Binomial-P regression model (BNBR-P) can be derived from the product of two NB-P marginals and a multiplicative factor parameter. (P is the functional parameter). The probability mass function of BNBR-P is,

$$\mathrm{pr}(y_{i1},y_{i2}) = \left[\prod_{t=1}^{2}\left(\frac{\Gamma(y_{it}+\alpha_t{}^{-1}\mu_{it}{}^{2-p})}{y_{it}!\,\Gamma(\alpha_t{}^{-1}\mu_{it}{}^{2-p})}\right)\left(\frac{\alpha_t{}^{-1}\mu_{it}{}^{2-p}}{\alpha_t{}^{-1}\mu_{it}{}^{2-p}+\mu_{it}}\right)^{\alpha_t{}^{-1}\mu_{it}{}^{2-p}}\left(\frac{\mu_{it}}{\alpha_t{}^{-1}\mu_{it}{}^{2-p}+\mu_{it}}\right)^{y_{it}}\right].[1+\Phi(e^{-y_{i1}}-c_{i1})(e^{-y_{i2}}-c_{i2})] \quad (9)$$

Where α_t, t = 1, 2, are the dispersion parameters. P = 1 or 2, is the functional parameter, Φ is the multiplicative factor (correlation) parameter, and

$$c_{it} = E\,(e^{-yit}) = \left(\frac{\alpha_t{}^{-1}\mu_{it}{}^{2-p_t}}{\alpha_t{}^{-1}\mu_{it}{}^{2-p_t}+e^{-1}+1}\right)^{\alpha_t{}^{-1}\mu_{it}{}^{2-p_t}}, \; t = 1, 2$$

Note that the most appropriate link function for the negative binomial distribution is the log link. [6, 24]

3. Results from the Analysis

3.1. The Best Univariate Model for Dengue Fever

In the modelling procedure the log of rainfall was used instead of rainfall for convergence reasons.

The parameter estimates, standard errors of the estimates, degrees of freedom, t-value and the associated p-value of the final model are given in table 2.

Table 2. *Parameter estimates of the best model for dengue fever.*

Effect	year	Estimate	Standard error	Degrees of freedom	t value	Pr > |t|
Intercept		-8.6349	1.838	2	-4.7	0.0424
lrainl2		0.1655	0.02215	455	7.47	<.0001
Temp		0.01316	0.04708	455	0.28	0.7800
Lrain		-0.04432	0.0222	455	-2	0.0465
lrainl1		0.07819	0.0187	455	4.18	<.0001
templ1		0.2021	0.04718	455	4.28	<.0001
Humid		-0.03825	0.01049	455	-3.65	0.0003
Year	2010	0.2119	0.06871	455	3.08	0.0022
Year	2011	0.05857	0.07026	455	0.83	0.4049
Year	2012	0.1847	0.07116	455	2.6	0.0098
Year	2013	0.1249	0.07059	455	1.77	0.0775
Year	2014	0.1883	0.06939	455	2.71	0.0069
Year	2015	0.08997	0.0698	455	1.29	0.1981
Year	2016	0	.	.	.	.
humidl2		0.02325	0.009883	455	2.35	0.0191

Note:

i. Year 2016 was considered to be the base level

ii. Estimation technique used was residual pseudo-likelihood (RSPL) [27]

Interpretation of the Parameter Estimates of the Best Univariate Model for Dengue Fever

The fitted model is given in equation (10).

$$Log(\mu_{ijk}) = -8.6349 + 0.1655\,(lrainl2)_{ijk} + 0.01316\,(temp)_{ijk} - 0.04432\,(lrain)_{ijk} + 0.07819(lrainl1)_{ijk} + 0.2021(templ1)_{ijk} - 0.03825\,(humid)_{ijk} + 0.2119\,(Year_{2010}) + 0.05857(Year_{2011}) + 0.1847(Year_{2012}) + 0.1249\,(Year_{2013}) + 0.1883(Year_{2014}) + 0.08997\,(Year_{2015}) + 0.02325(humidl2)_{ijk} \quad (10)$$

μ_{ijk}= Expected number of dengue patients in the i^{th} month, j^{th} year and k^{th} district. The parameter estimates give the contribution of the explanatory variables to the log of the expected number of dengue fever cases recorded in each month, year and district. [13]. The parameter estimates which are significant at the 5% level of significance are interpreted below. Only the variable 'year' was considered as a discrete factor from the fixed effects. The variable year consists of 7 levels with the base level being 2016.

Effect of humidity on the response

The coefficient of humidity was negative, indicating that the increase in the humidity will lead to a decrease in the count of dengue fever patients. Suppose the humidity of a particular district, year and month increases by 1 unit while all other effects remain constant and the expected number of dengue cases before and after this increment are μ_{ijk1} and μ_{ijk2} respectively.

$$Log(\mu_{ijk1}) = \beta_0 + \beta_1(lrainl2)_{ijk} + \beta_2(temp)_{ijk} + \beta_3(lrain)_{ijk} + \beta_4(lrainl1)_{ijk} + \beta_5(templ1)_{ijk} + \beta_6(humid)_{ijk} + \beta_7(Year_{2010}) + \beta_8(Year_{2011}) + \beta_9(Year_{2012}) + \beta_{10}(Year_{2013}) + \beta_{11}(Year_{2014}) + \beta_{12}(Year_{2015}) + \beta_{13}(humidl2)_{ijk} \quad (11)$$

$$Log(\mu_{ijk2}) = \beta_0 + \beta_1(lrainl2)_{ijk} + \beta_2(temp)_{ijk} + \beta_3(lrain)_{ijk} + \beta_4(lrainl1)_{ijk} + \beta_5(templ1)_{ijk} + \beta_6(humid+1)_{ijk} + \beta_7(Year_{2010}) + \beta_8(Year_{2011}) + \beta_9(Year_{2012}) + \beta_{10}(Year_{2013}) + \beta_{11}(Year_{2014}) + \beta_{12}(Year_{2015}) + \beta_{13}(humidl2)_{ijk} \quad (12)$$

$\log\left(\frac{\mu_{ijk2}}{\mu_{ijk1}}\right) = \beta_6 = -0.03825$ by the calculation (12) – (11)

$$\left(\frac{\mu_{ijk2}}{\mu_{ijk1}}\right) = \exp(\beta_6) = \exp(-0.03825) = 0.9625$$

$$\mu_{ijk2} = 0.9625(\mu_{ijk1} \quad (13)$$

This result implies that the expected number of dengue fever patients of a particular district, month and year decrease by a ratio of approximately 0.96, as a result of 1 unit increment in the humidity of that district, month and year.

Using similar calculations, it was found that,

The expected number of dengue patients of a particular district, month and year,

1. increase by a ratio of 1.18, as a result of 1 unit increment of log(2nd lag of rainfall) [increment of 2.718 units of 2nd lag of rainfall] of that district, month and year
2. decrease by a ratio of 0.96, as a result of 1 unit increment of log(rainfall) of that district, month and year
3. increase by a ratio of 1.08, as a result of 1 unit increment in the log(1st lag of rainfall) of that district, month and year
4. increase by a ratio of 1.22, as a result of 1 unit increment in the 1st lag of temperature of that district, month and year.
5. increase by a ratio of 1.02, as a result of 1 unit increment of the 2nd lag of humidity of that district, month and year
6. expected number of dengue patients of a particular district and month from the year 2010, 2012 and 2014 are 1.24, 1.20 and 1.21 times higher than that in the year 2016 respectively.

3.2. Best Univariate Model for Rat Fever

The parameter estimates, standard errors of the estimates, degrees of freedom, t-value and the associated p-value of the final model are given in table 3.

Table 3. *Parameter estimates of the best model for rat fever.*

Effect	year	Estimate	Standard error	Degrees of freedom	t value	Pr > |t|
Intercept		-2.9956	0.2854	2	-10.5	0.0090
lrainl2		0.1167	0.01933	461	6.04	<.0001
Year	2010	-0.05865	0.06494	461	-0.9	0.367
Year	2011	-0.08041	0.06578	461	-1.22	0.2222
Year	2012	-0.2514	0.06785	461	-3.71	0.0002
Year	2013	-0.1447	0.06717	461	-2.15	0.0317
Year	2014	-0.1678	0.06677	461	-2.51	0.0123
Year	2015	-0.1177	0.0668	461	-1.76	0.0786
Year	2016	0	.	.	.	.

Note:

i. Year 2016 was considered to be the base level

ii. Estimation technique used was residual pseudo-likelihood (RSPL) [28]

Interpretation of the Parameter Estimates of the Best Univariate Model for Rat Fever,

The fitted model is given in equation (14).

$$Log(\mu_{ijk}) = -2.9956 + 0.1167(\text{lrainl2})_{ijk} - 0.05865\ (\text{year\_2010}) - 0.08041(\text{year}_{2011}) - 0.2514(\text{year}_{2012}) - 0.1447(\text{year}_{2013}) - 0.1678(\text{year}_{2014}) - 0.1177(\text{year}_{2015}) \quad (14)$$

μ_{ijk}= Expected number of dengue patients in the i^{th} month, j^{th} year and k^{th} district.

The parameter estimates gives the contribution of the explanatory variables to the log of the expected number of rat fever cases recorded in each month, year and district. [13]

Using a similar method to what was used in section 3.1 the parameter estimates that are significant at 5% level of significance were interpreted as follows,

The expected number of rat fever patients of a particular district, month and year increase by a ratio of 1.12, as a result of 1 unit increment of the log (2^{nd} lag of rainfall) of that district, month and year

The expected number of rat fever patients of a particular district and month from the year 2012, 2013 and 2014 are 0.78, 0.87 and 0.85 times lower than that in the year 2016

3.3. Best Bivariate Model for Dengue Fever and Rat Fever

The parameter estimates, standard errors of the estimates, degrees of freedom, t-value and the associated p-value of the final model are given in table 4.

Table 4. *Parameter estimates of the best bivariate model for dengue fever and rat fever.*

Effect	Distribution	Estimate	Standard error	Degrees of freedom	t value	Pr > |t|
Dist	NEGBIN1	-6.0638	1.1313	461	-5.36	<.0001
Dist	NEGBIN2	1.1578	1.2224	461	0.95	0.3441
lrainl2*dist	NEGBIN1	0.1866	0.03947	461	4.73	<.0001

Effect	Distribution	Estimate	Standard error	Degrees of freedom	t value	Pr > |t|
lrainl2*dist	NEGBIN2	0.07768	0.0423	461	1.84	0.0669
humidl2*dist	NEGBIN1	0.03619	0.01495	461	2.42	0.0159
humidl2*dist	NEGBIN2	-0.05223	0.01616	461	-3.23	0.0013
lrain*dist	NEGBIN1	-0.0795	0.03303	461	-2.41	0.0165
lrain*dist	NEGBIN2	0.05455	0.03351	461	1.63	0.1042

Note

i. dist = distribution

ii. Estimation technique used was Laplace maximum-likelihood [28]

3.3.1. Interpretation of the Parameter Estimates of the Bivariate Model for Dengue Fever and Rat Fever

The model selected can be represented as,

$$log\left(\frac{\mu_{ijk1}}{\mu_{ijk2}}\right) = -6.0638\,(dist1) + 1.1578(dist2) + 0.1866(lrainl2 * dist1)_{ijk} + 0.07768\,(lrainl2 * dist2)_{ijk} + 0.03619\,(humidl2 * dist1)_{ijk} - 0.05223\,(humidl2 * dist2)_{ijk} - 0.0795\,(lrain * dist1)_{ijk} + 0.05455(lrain * dist2)_{ijk} \quad (15)$$

where, dist1= Negative Binomial 1 and dist2=Negative Binomial 2

$\left(\frac{\mu_{ijk1}}{\mu_{ijk2}}\right)$ = Expected number of patients of the two diseases in the i^{th} month, j^{th} year and k^{th} district. (1-dengue fever, 2-rat fever)

The parameter estimates give the contribution of the explanatory variables to the joint distribution of the number of dengue and rat fever patients recorded in each month, year and district. [13]

Interpretation of the parameter estimates that are significant at the 5% level of significance of the joint model should be carried out for the two marginal models separately.

Using a similar method to what was used in section 3.1, the interpretation of the two marginal models separately are as follows,

According to the results from the Negative Binomial 1 model, the expected number of dengue fever patients of a particular district, month and year,

1. increase by a ratio of 1.21, as a result of 1 unit increment of the log (2^{nd} lag of rainfall) of that district, month and year
2. increase by a ratio of 1.04, as a result of 1 unit increment of the 2^{nd} lag of humidity of that district, month and year
3. decrease by a ratio of 0.92, as a result of 1 unit increment of the log(rainfall) for that district, month and year.

According to the results from the Negative Binomial 2 model, the expected number of rat fever patients of a particular district, month and year,

1. decrease by a ratio of 0.95, as a result of 1 unit increment of the 2^{nd} lag of humidity of that district, month and year.

3.3.2. Residual Analysis of the Bivariate Model

It is important to do a residual analysis to check whether the model developed is suitable. Therefore, the residual analysis was performed by plotting the residuals against the exponent of the predicted value.

The graph in figure 1 indicates that majority of the residuals are between -1 and 2. Only 6 residuals are above the value 2. The majority of the points are evenly distributed vertically around 0 and there is no clear pattern in them. Therefore, it can be said that the residual analysis suggests that the bivariate model developed is adequate.

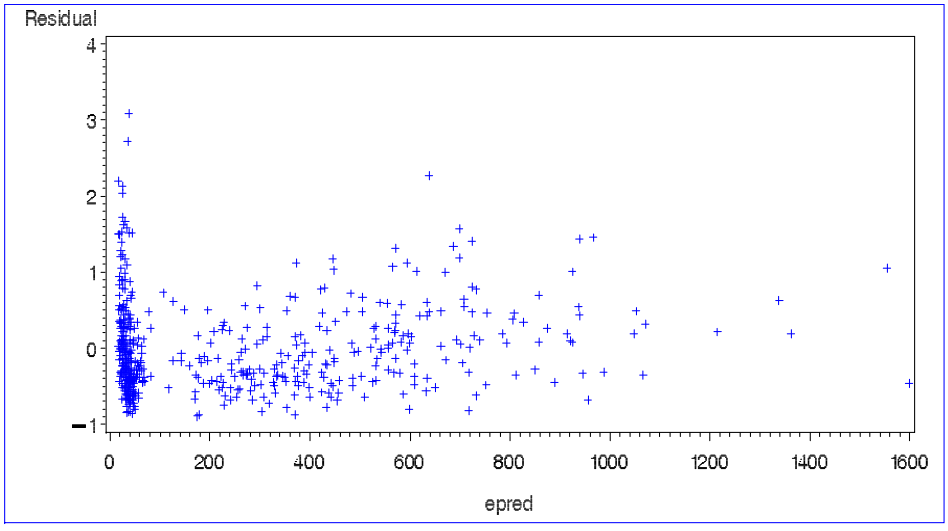

Figure 1. *Graph of the residuals vs exp (predicted values).*

4. Validation

4.1. Constructing the Receiver Operator Characteristic Curve

The rate of occurrence of dengue fever is much higher than the rate of occurrence of rat fever. Therefore dengue fever was used to identify the optimal cut-off point for classification. The variable 'count of dengue fever' was converted into a binary variable. This was done by classifying the count of dengue fever patients into two groups using the mean (532). (If the count of dengue<532, then it was coded as 1 and if the count of dengue >532, then it was coded as 2). Then a Logistic model was used to model the relationship between the count of dengue fever and the other variables that were significant in the bivariate model constructed in section 3.3 and the count of rat fever (explanatory variables including the count of rat fever), with the objective of developing a classification rule. Then using the logistic model developed, predictor values $\hat{Y}$ for the existing data set was obtained. If $\hat{Y}$>mean (532), then the unit is said to have a high count of dengue, if not the unit is said to have a low count of dengue. Using threshold values ranging from 0 to 1, the sensitivity and specificity were calculated for each of these cases. The analysis was done using SAS 9.2.

$$Sensitivity = \frac{Total\ predicted\ no.of\ units\ with\ 'high\ level'\ counts\ of\ dengue\ fever}{Total\ number\ of\ units\ with\ 'high\ level'\ counts\ of\ dengue\ fever} \tag{16}$$

$$Specificity = \frac{Total\ predicted\ no\ of\ units\ with\ 'low\ level'\ counts\ of\ dengue\ fever}{Total\ number\ of\ units\ with\ 'low\ level'\ counts\ of\ dengue\ fever} \tag{17}$$

When sensitivity increases specificity decreases, and vice versa. In an ideal classification test both sensitivity and specificity should be high [9] However, in practice the ideal situation rarely occurs, especially in medical related studies [27]. Therefore, based on the objective of the classification test a cut-off point should be selected.

The main purpose of developing a classification rule in this study is to identify the areas that have a high rate of occurrence of the disease so that necessary measures could be taken by the relevant authorities. This is evaluated using the sensitivity associated with the classification rule. Hence, the sensitivity should be high. On the other hand specificity should also be of a value that is not too low, since incorrectly classifying a unit as one with a high level of occurrence of the disease will result in a waste of time and resources in conducting further investigations. [27]. Therefore in order to satisfy both criteria, cut-off probability of 0.55097 was chosen as the best cut-off point or the best trade off. It can be seen that the selected cut off gives sensitivity (True positive rate - TPR) of approximately 80% and specificity (True Negative Rate - TNR) of 52%. That is probability of correctly identifying units with high level counts of dengue fever is 0.80 while correctly identifying units with low counts of dengue fever is 0.52. The area under the curve (AUC) is 0.711 and it can be interpreted as 'acceptable discrimination' according to the rule of thumb by Hosmer & Lemeshow, 2000.

The logistic model used to identify the probability of the point that gives the best trade off includes many climatic variables that are of continuous type, hence the model becomes complex. Computing the value that should be used to classify the count of rat fever as high or low becomes tedious with the presence of other climatic variables. Therefore, the probability value that was identified using the more accurate model is applied to a simpler logistic model. The simpler logistic model includes only the count of rat fever as an explanatory variable and the binary variable created for the count of dengue fever was used as the response variable. Table 5 gives the parameter estimates of the fitted logistic model.

Table 5. *Parameter estimates of the simpler logistic model.*

Parameter	Estimate
Intercept	-0.0962
Count of rat fever	0.0191

Using these parameter estimates an approximation to the value that could be used to classify the count of rat fever as high or low was computed as follows,

$$log\left(\frac{0.55097}{1-0.55097}\right) = 0.20459$$

$$\frac{0.20459+0.0962}{0.0191} = 15.75 \tag{18}$$

Hence, it was decided to use 15.75 to classify the count of rat fever as high or low. That is: if the count of rat fever for a given unit is < 15.75, it is classified as 1, if the count of rat fever for a given unit is > 15.75, it is classified as 2.

4.2. Internal Validation

Observed and the predicted counts from the fitted bivariate negative binomial model were computed in order to validate the fitted model using the existing data set. (2010 – 2016). Internal predictions for the existing data set were calculated using the software SAS 9.2. The classification of the observed and the predicted values were done in the following manner,

1. If the count of dengue <532 and count of rat fever <15.75 then classified as 1
2. if the count of dengue <532 and count of rat fever >=15.75 then classified as 2
3. If the count of dengue >=532 and count of rat fever <15.75 then classified as 3
4. If the count of dengue >=532 and count of rat fever >=15.75 then classified as 4

Internal Predictive Accuracy of the Final Bivariate Model is as follows,

Table 6 gives the prediction results for the internal validation.

$$Internal\ predictive\ accuracy\ of\ the\ model = \frac{(127+56)}{252} \text{X}\ 100\ = 72.6\% \sim 73\% \quad (19)$$

Table 6. *Internal predictive accuracy of the final bivariate model.*

		Actual results		Total
		2	4	
Predicted results	1	9	4	13
	2	127	18	145
	3	5	20	25
	4	13	56	69
Total		154	98	252

4.3. External Validation

In order to test the predictive accuracy of the developed model, it is important to test the model on a new set of data (external data). Therefore, a new set of data was obtained from the epidemiology unit and the meteorology department of Sri Lanka for the first 8 months of 2017.

External predictions of the developed model was computed using the software SAS 9.2. The classification of the observed and the predicted counts were done in the same way as before (section 6.4)

External Predictive Accuracy of the Final Bivariate Model is as follows,

Table 7 gives the prediction results for the external validation.

$$External\ predictive\ accuracy\ of\ the\ model = \frac{17}{24} \text{X}\ 100 = 70.83\% \sim 71\% \quad (20)$$

Table 7. *External predictive accuracy of the final bivariate model.*

		Actual results	Total
		3	
Predicted results	2	1	1
	3	17	17
	4	6	6
Total		24	24

Since both internal and external validation methods indicate that the prediction accuracy of the model is at a reasonably high level, it could be said that the final bivariate model for dengue fever and rat fever is performing well.

5. Discussion

5.1. Comparison of Univariate and Bivariate Models Based on Akaike's Information Criterion (AIC) and Bayes Information Criterion (BIC)

Laplace maximum-likelihood estimation method was used to calculate the AIC's and BIC's of all three models. A model with a lower AIC and BIC is considered to be a better model. [1]. Table 8 and 9 gives these results respectively.

Table 8. *AIC of the models.*

Model	AIC	Total AIC
Univariate dengue fever	6547.66	10555.43
Univariate rat fever	4007.77	
Bivariate dengue and rat fever	5623.65	5623.65

Table 9. *BIC of the models.*

Model	BIC	Total BIC
Univariate dengue fever	6530.35	10527.3
Univariate rat fever	3996.95	
Bivariate dengue and rat fever	5612.83	5612.83

The bivariate model shows a significant reduction in AIC and BIC than the addition of the two AIC's and BIC's respectively of the univariate models. The reduction in AIC is 4931.78 (10555.43 - 5623.65), while the reduction in BIC is 4914.47 (10527.3 - 5612.83). Therefore, the bivariate model is more suitable than having two univariate models.

5.2. Comparison of Univariate and Bivariate Models Based on the Stand Errors of the Variance Parameter

Table 10 gives the results.

Table 10. *Comparison of models using the standard errors of the covariance parameters.*

Covariance parameter	S. E in univariate dengue model	S. E in univariate rat fever model	S. E. in bivariate model
Variance	0.01178	0.05074	0.008057

The standard error of the variance is lower in the joint model than in the two univariate models. Therefore the bivariate model is more suitable than having two univariate models. [7]

AIC, BIC and the standard error of the variance of the distribution suggests that the bivariate model is better than having two univariate models.

Univariate modelling of dengue fever

There is an increment in the incidence of dengue fever when the 1st and 2nd lag of log (rainfall) increases. The results of [25], also suggests that the incidence of dengue escalated with the increase of rainfall of the previous two months. The present study shows a decrease in the incidence, when log (rainfall) of the current month increases. [2] also shows similar results, where dengue fever has a negative correlation with rainfall. Heavy rainfall may not provide favourable conditions to mosquitoes as it washes away the mosquito eggs and larvae, thus reducing the density.

Increase in the humidity of the current month lead to a decrease in the incidence of dengue. [2] also shows that there

is a negative correlation between dengue fever and humidity. Low humidity causes mosquitoes to feed more frequently to compensate for dehydration. Thereby, a reduction in the incidence of dengue occurs when humidity increases. The current study shows that the increment in the 2nd lag of humidity leads to an increase in the incidence of the disease. A positive correlation between the 2nd lag of humidity and the incidence of dengue fever was seen in a study conducted in China. [16]. Thus, the results of the current study are in line with the results presented in the literature.

The incidence of dengue fever increases when the 1st lag of temperature increases. [25], also shows that the incidence of dengue fever increases when the temperature of the previous month increases. Temperature is known to affect dengue incidence by exerting a sizeable influence on the population dynamics of the dengue mosquitoes. Temperature can impact the conditions for egg laying, stimulation of egg hatching, and the abundance of *Aedes* larvae and pupae. [21]

Univariate modelling of rat fever

The results of the current study reveal that the incidence of rat fever increases when the log (2nd lag of rainfall) increase. A study conducted in Reunion Islands found a significant positive correlation between Leptospirosis cases and monthly cumulated rainfall and the highest correlation was found with average monthly rainfall recorded two months previously. [24], [5]. The study conducted by [24] has not found any statistically significant correlation between climatic factors and the incidence of dengue. However, a significant correlation has been found between rainfall and Leptospirosis in majority of the districts with high incidence rates. The districts considered in the current study are ones with high incidence rates of Leptospirosis. Thus, indicating that the results of the two studies tally. This fact suggests that, rainfall plays a more important role in Leptospirosis epidemics than the endemic transmission in Sri Lanka. [20], has shown that increase in the 2nd lag of rainfall causes the incidence of the disease to rise. "The lag period of 1–2 months between heavy rainfall and cases is consistent with the probable effect of flooded land and water-soaked soils on leptospiral organism survival (1 to 2 months) and an average incubation period for leptospirosis of 1 to 3 weeks. In many parts of the world heavy rainfall and flooding can lead to outbreaks of leptospirosis, especially in tropical countries since transmission is often indirect in these areas". [20]

Bivariate modelling of dengue fever and rat fever

According to the results from the Negative Binomial 1 model, the expected number of dengue fever cases of a particular district, month and year increases when the log (2nd lag of rainfall) and 2nd lag of humidity increase. The incidence of dengue fever decreases when the log (rainfall) increases. [25], has also shown that the incidence of dengue fever increases when the 2nd lag of rainfall increases. A positive correlation between the 2nd lag of humidity and the incidence of dengue fever was seen in a study conducted in China. [16]

According to the results from the Negative Binomial 2 model, the expected number of rat fever cases of a particular district, month and year decreases when the 2nd lag of humidity increase. Although previous studies that modelled the incidence of rat fever (univariate) showed considerable correlations between the number of Leptospirosis cases and rainfall, relative humidity and temperature [3], when the bivariate model is considered only the 2nd lag of humidity becomes significant with respect to the incidence of rat fever. The above study [3] pertains to only the Gampaha district and has considered only one year's data, whereas the current study has considered all three districts of the western province for 7 years. Thus, this factor would have contributed towards the differences in results. [24], also presents that temperature does not have a significant impact on the incidence of rat fever, which tallies with current study's result.

Previous studies suggest to use an autoregressive structure to account for the correlation over time of observations of a given district [17], [25]. However, results of the current study showed that the G matrix of the time aspect gives very small correlation between months, indicating that it was not necessary to use an AR (1) adjustment for this problem [26]. This would have made the analysis simpler and avoided the huge 12x12 matrix.

The internal and external validations for the bivariate model developed indicated that the model predicts well.

6. Conclusion

Incidence of dengue fever and rat fever can be jointly modelled using a negative binomial distribution.

Joint modelling yields better results than modelling the two diseases in a univariate manner.

Internal and external predictive accuracy of the joint model is high.

Cluster effect should be considered as the incidence rate changes according to locality.

Rainfall and its 2nd lag and the 2nd lag of humidity is significantly associated with the incidence of dengue fever according to the joint model.

2nd lag of humidity is significantly associated with the incidence of rat fever according to the joint model.

Climatic conditions that lead to high levels of incidence of the two diseases can be identified using the model developed.

Resources could be allocated in a more effective way to control the incidence of the two diseases by using the model developed.

References

[1] Akaike, H. (1974). A New Look at the Statistical Model Identification. IEEE Transaction on Automatic Control, AC–19, 716–723.

[2] Alsheri, M. S. (2013). Dengue fever Outburst and its Relationship with ClimaticFactors. World Applied Scienced Journal, 22 (4), 506-515. doi: 10.5829/idosi.wasj.2013.22.04.443

[3] Denipitiya, D. T., Chandrasekharam, N. V., Abeyewickreme, W., Viswakula, S., & Hapugoda, M. D. (2016). Spatial and seasonal analysis of human leptospirosis in the District of Gampaha, Sri Lanka. Sri Lankan Journal of Infectious Diseases, 6 (2), 83-93.

[4] Descloux, E., Mangeas, M., Menkes, C., Lengaigne, M.,..., Leroy, A., & De, L. X. (2012). Climate-based models for understanding and forecasting dengue epidemics. PLoS Negl Trop Dis. doi: 10.1371/journal.pntd.0001470

[5] Desvars, A., Jego, S., Chiroleu, F., Bourhy, P., Cardinale, E., & Michault, A. (2011). Seasonality of Human Leptospirosis in Reunion Island (Indian Ocean) and its Association with Meteorological Data. PLoS ONE, 6 (5). Retrieved from https://doi.org/10.1371/journal.pone.0020377

[6] Faroughi, P., Karimi, M. S., Ismail, N., & Karimi, A. (Summer, 2017). Estimation of Count Data using Bivariate Negative Binomial Regression Models. Quarterly Journal of Quantitative Economics, 14 (2): 143-166.

[7] Hapugoda, J. C., Sooriyarachchi, M. R., Kalupahana, R. S., & Satharasinghe, D. A. (2017). Joint Modeling of Mixed Responses - An application to Poultry Industry. International Conference on Computational Mathematics, Computational Geometry & Statistics (CMCGS), (pp. 182-185). Singapore. Retrieved 2017.

[8] Hilbe, J. (2007). Modeling count data. (M. Lovric, Ed.) New York: Sptinger. Retrieved from https://www.encyclopediaofmath.org/images/2/2a/Modeling_count_data.pdf

[9] Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. United States of America: John Wiley & sons, Inc. doi: 10.1002/0471722146

[10] Institute for Digital Research and Education. (2018, January 10). Introduction to Generalized Linear Mixed Models. Retrieved from https://stats.idre.ucla.edu/other/mult-pkg/introduction-to-generalized-linear-mixed-models/

[11] Ismail, N., & Zamani, H. (2013). Estimation of Claim Count Data Using Negative Binomial, Generalized Poisson, Zero-Inflated Negative Binomial and Zero-Inflated Generalized Poisson Regression Models. Casualty Actuarial Society E-Forum.

[12] Jaroensutasinee, K., Jaroensutasinee, M., & Promprou, S. (2005). Climatic Factors Affecting Dengue Haemorrhagic Fever Incidence in Southern Thailand. Dengue Bulletin, 29, 41-48.

[13] Jayanetti, W., & Sooriyarachchi, M. R. (2013). A Multilevel Study of Dengue Epidemiology in Sri Lanka Modelling Survival and Incidence.

[14] Karim, M. N., Munshi, S. U., Anwar, N., & Alam, M. (2012). Climatic factors influencing dengue cases in Dhaka city: A model for dengue prediction. The Indian Journal of Medical Research, 136 (1), 32-39.

[15] Lawless, J. F. (1987). Negative Binomial and Mixed Poisson Regression. Can J Stat, 15: 209-225. doi: 10.2307/3314912

[16] Lu, L., Lin, H., Tian, L., Yang, W., Sun, J., & Liu, Q. (2009). Time series analysis of dengue fever and weather in Guangzhou, China. BMC Public Health, 9, 395. Retrieved from https://doi.org/10.1186/1471-2458-9-395

[17] Martínez-Bello, D. A., López-Quílez, A., & Torres-Prieto, A. (2017, June 3). Bayesian dynamic modeling of time series of dengue disease case counts. PLOS Neglected Tropical Diseases, 11 (7), e0005696. doi: 10.1371/journal.pntd.0005696

[18] McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (2 ed.). London: Chapman & Hall.

[19] Mishra, B., Sighal, L., Sethi, S., & Ratho, R. K. (2013). Leptospirosis Coexistent with Dengue Fever: A Diaagnostic Dilemma. Journal of Globall Infectious Diseases, 5 (3), 121-122. doi: 10.4103/0974-777X.116878

[20] Mohan, A. R., Cumberbatch, A., Adesiyun, A. A., & Chadee, D. D. (2009). Epidemiology of human leptospirosis in Trinidad and Tobago, 1996–2007: A retrospective study. Acta Trop, 112 (3), 260-265. doi: 10.1016/j.actatropica.2009.08.007

[21] Moore, C. G., Cline, B. L., Ruiz-Tiben, E., Lee, D., Romney-Joseph, H., & al., e. (1978). Aedes aegypti in Puerto Rico: environmental determinants of larval abundance and relation to dengue virus transmission. Am J Trop Med Hyg, 1225-1231.

[22] Naish, S., Dale, P., Mackenzie, J. S., McBride, J., Mengerson, K., & Tong, S. (2014). Climate change and dengue: a critical and systematic review of quantitative modelling approaches. BMC Infectious Diseases, 14: 167.

[23] Nakhapakorn, K., & Tripathi, N. K. (2005). An information value based analysis of physiscal and climatic factors affecting dengue fever and dengue hemorrhagic fever incidence. International Journal of Health Geographics, 4, 13-35.

[24] Palihawadana, P., Amarasekare, J., S, G., Gamage, D., Jayasekara, S. A., & Dayananda, M. D. (2014). The climatic factors associated with incidence of Leptospirosis in Sri Lanka. JOURNAL OF THE COLLEGE OF COMMUNITY PHYSICIANS OF SRI LANKA, 19, 29-33.

[25] Perera, H. L., & Sooriyarachchi, M. R. (2008). Fitting Generalized Linear Models in the Presence of Correlated Data: An Application to an Epidemiological Study of Dengue Fever. Sri Lankan Journal of Applied statistics, 9 (special issue), 159-175.

[26] Rashbash, J., Steele, F., Browne, W., & Goldstein, H. (2004). A user's guide to MLwiN, version 2.10.

[27] Rathnayake, G. I., & Sooriyarachchi, M. R. (2014). Automated Statistical Information System (ASIS) for Diagnosis and Prognosis of Life-threatening Viral Diseases. Sri Lankan Journal of Applied Statistics, 15-3.

[28] SAS Inc. Institute. (2009). SAS/STAT 9. 2 User's Guide, Second Edition. Cary: SAS Pub.

[29] Sharma, K. K., Latha, P. M., & Kalawat, U. (2012). Coinfection of leptospirosis and dengue fever at a tertiary care centre in South India. Scho Res, 6.

[30] Trinidade, D. D., Ospina, R., & Amorim, L. D. (2015). Choosing the right strategy to model longitudinal count data in Epidemiology: An application with CD4 cell counts. Epidemiology Biostatistics and Public Health, 12 (4).

[31] Wijesinghe, A., Gnanapragash, N., Ranasinghe, G., & Ragunathan, M. K. (2015). Fatal co-infection with leptosirosis and dengue in a Sri Lankan male. BMC Research Notes, 8: 348.

[32] Winkelmann, R. (2008). Econometric analysis of count data. Verlag, Heidelberg: Springer.

[33] Zamani, H., & Ismail, N. (2012). Functional form for the Generalized Poisson Regression Model. Commun Stat Theor M, 41: 3666-3675. doi: 10.1080/03610926.2011.564742.