J. Sci. Univ. Kelaniya 17 (2) (2024): 73 - 91 https://doi.org/10.4038/josuk.v17i2.8109

A model for predicting confirmed leptospirosis cases in Sri Lanka

W.E.M.D.T. Ekanayake[1] and L.S. Nawarathna[2]*

[1]Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka
[2]Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Peradeniya, Sri Lanka

*Corresponding author: lakshikas@sci.pdn.ac.lk

https://orcid.org/0000-0001-9665-5471

Received: 23.04.2024 Revised: 11.09.2024 Accepted: 04.11.2024 Online: 15.12.2024

Abstract Leptospirosis is a significant health issue in Sri Lanka, but current studies have only examined some districts using time series analysis. There is no comprehensive leptospirosis prediction model for the entire country. This study aims to develop a model to predict leptospirosis distribution across Sri Lanka. We collected reported leptospirosis cases in all districts from 2011 to 2022 from the Epidemiology Department. We tested the normality of the data using the Shapiro-Wilk test. Also, we evaluated various models, including linear regression, exponential regression, polynomial regression, Autoregressive Integrated Moving Average (ARIMA) time series, generalised additive models (parametric), Cox proportional hazards regression model (semi-parametric), and median-based linear model (non-parametric). The best-fitting model was selected based on the minimum Akaike information criterion (AIC) and Bayesian Information Criteria (BIC), and its accuracy was verified using the Root Mean Square Error (RMSE). We found that the ARIMA model was the best-fitting model for Galle, Mullaitivu, Batticaloa, and Puttalam. The exponential regression model was the best-fitting model for Kalmunai, while the polynomial regression model was the best-fitting model for Matale and Jaffna. The generalised additive model was the best-fitting model for Gampaha, Hambantota, Matara, Mannar, Vavuniya, Trincomalee, Anuradhapura, Badulla, Monaragala, and Ratnapura. In contrast, the median-based linear model was the best-fitting model for Colombo, Kalutara, Kandy, Nuwara Eliya, Kilinochchi, Ampara, Kurunegala, Polonnaruwa, and Kegalle. The fitted models were then used to forecast leptospirosis incidents in Sri Lanka from 2024 to 2026. For future research, we recommend Bayesian hierarchical models to identify high-risk clusters of leptospirosis in Sri Lanka.

Keywords: Disease forecasting, Epidemiology, Generalised additive model (GAM), Health prediction, Leptospirosis, Prediction models, Time series analysis

Introduction

Leptospirosis has been identified as the most widespread zoonosis and a major tropical disease affecting humans and creatures (Adler, 2015). It is caused by a spirochete (a type of spiral-shaped bacteria) of the genus Leptospira of the family Leptospiraceae (Erdinc et al., 2006). This is infected through occupational, recreation or domestic directly or via water or soil contaminated with the urine of carrier animals. There were annually 1.03 million cases and 58,900 deaths widespread due to leptospirosis. Adult males aged 20-49 were met with many cases and deaths (Costa et al., 2015).

Sri Lanka has one of the highest cases of human leptospirosis worldwide, related to monsoons and flooding. It is one of the central communicable diseases (Gnanapragasam, 2018). The highest cases of leptospirosis were reported from paddy workers than non-paddy workers, who got infected through cattle and household animals (Schønning et al., 2019). The reported cases are primarily male, and the exposure risk is for people engaged in farming or exposed to Paddy fields (Warnasekara & Agampodi, 2017). Paddy workers get infected through animal urine (Schønning et al., 2019; Robertson et al., 2012). A wider range of risk exposure includes recreational activities such as whitewater rafting in urban populations (Agampodi et al., 2014).

Annual deaths due to leptospirosis were approximately 730, and the kidney, liver and heart were the most common organs involved (Warnasekara et al., 2019). In Sri Lanka, there was a large outbreak of leptospirosis in 2008, with more

than 7000 reported cases (Robertson et al., 2012). These cases and deaths have not been distributed nationwide (Warnasekara & Agampodi, 2017). The most striking feature of this outbreak (2008, 2011 and 2013) was the higher prevalence of hypotension and bradycardia in patients (Ramsey et al., 2015).

According to a study done from 2007 to 2019, the wet zone had the highest number of leptospirosis incidents, and there was no difference between leptospirosis incidents in the dry zone and the highlands. Leptospirosis cases positively fluctuated with rainfall, relative humidity, and the number of rainy days/month, and they negatively fluctuated with temperature and solar radiation. Both meteorological parameters and leptospirosis cases are changing across different geo-locations, and the association between leptospirosis cases and the meteorological parameters could be different in the three regions mentioned above (Desvars et al., 2011; Sanjeewani, 2017). Leptospirosis outbreaks usually occur after heavy rainfall; such as floods in the endemic wet zones. Non-endemic dry zones of Sri Lanka in 2011 suggest a variety of biological, clinical, and molecular characteristics compared to previously reported leptospirosis outbreaks in the endemic areas after Molecular investigations with the quantification of post-flood leptospirosis (Agampodi et al., 2014). There is a correlation between the temporal pattern of reported leptospirosis cases and meteorological conditions (temperature, relative humidity, rainfall), which was investigated using time series analysis combined with spectral analysis and the least squares method (Sumi et al., 2017). ARIMA models can show the trend in leptospirosis cases to seasonal patterns and climate factors (Chadsuthi et al., 2012). The risk of leptospirosis disease in the districts is seasonal, with a small number of cases occurring from March to May and many cases occurring in October and December.

Many prediction models have been created to predict leptospirosis in Sri Lanka. The integer-valued generalised autoregressive conditional heteroscedasticity (INGARH) model, which included current and previous rainfall covariates and regression on previous cases of leptospirosis of a local and seasonal time scale, was selected as the best-performing model. This model identified that rainfall did not significantly correlate with leptospirosis incidents island-wide. The family of INGARH models developed forecasted leptospirosis incidents and provided early warning (Plouffe, 2016). Wavelet analyses were done for Leptospirosis cases in the Kandy district from 2004 to 2019. It was found that El Nino Southern Oscillation (ENSO), Indian Ocean Dipole (IOD), and ENSO Modoki indices correlated with leptospirosis Incidents (LI) of Kandy with 1.9 -11.5 months lags. Indices of ENSO showed two correlation patterns with Kandy leptospirosis incidents. Time-logged Detrended Cross-Correlation Analysis class (DCCA) results showed that all indices of the three teleconnections were significantly correlated with the Leptospirosis incidents of Kandy with 2-5 months lag periods (Ehelepola et al., 2021). Time series analysis and cross-correlation were done for leptospirosis incidents in Hambantota district, Sri Lanka, from 2008 to 2017, and it was found that through wavelet analysis, there was no significant correlation between rainfall and leptospirosis incidents in the correlation analysis. There was no clear correlation between both Nino indices and leptospirosis. Peaks of rainfall followed by peaks of leptospirosis incidents and Minimum, maximum, mean temperature, soil temperature, evaporation rate, and the duration of sunshine were followed by peaks in LI and peaks of rainfall followed by peaks of LI, all after lag periods (Ehelepola et al., 2021). A time series prediction model for Leptospirosis was developed in the dry zone of Sri Lanka using the January 2008 to December 2018 monthly leptospirosis incidents and monthly rainfall, rainy days, temperature, and relative humidity for the fitting model. The first 72 months (55%) were used to fit the model, and the model was validated using the subsequent 60 months (45%). Best fitted univariate model was ARIMA (1,0,0) (0,1,1)12 (AIC-1.08, MAPE-19.8) ARDL (1,3,2,1,0),(AIC-2.04,MAPE-30.4) was the best fitted ARDL model. In several patients, rainfall, rainy days, and temperature were positively associated, while relative humidity was negatively associated with temperature (Warnasekara et al., 2022). The disease is endemic in Sri Lanka, and Kalutara district is one of the most endemic areas because it has a large population and agricultural communities. ARIMA model was fitted by leptospirosis cases from January 2010 to December 2020 to predict monthly Leptospirosis cases in Kalutara district. ARIMA (3,1,1) was the best-fitted model to predict the future outbreak of leptospirosis (Thanusika et al.,

2022). More than 4000 cases per year were reported in Sri Lanka from 2010 to 2017. Colombo, Kalutara, Gampaha, Ratnapura, Galle, Matara, Kegalle, Matale and Kurunegala were named leptospirosis endemic districts in Sri Lanka since outbreaks were reported in all districts throughout these years. A time series model was created to forecast Leptospirosis cases in Colombo, Gampaha, Kalutara, Ratnapura and Kurunegala districts. The best-fitted model was identified with 92% accuracy, and based on this model, it was predicted that the leptospirosis cases in the above districts in 2018 will be 1200 (Gnanapragasam, 2018). A model was fitted for leptospirosis cases in the western province of Sri Lanka using time series analysis to study the behaviour of provinces-wise and district-wise distribution of Leptospirosis cases. Western province (28.41%) was the most affected part of the island by human leptospirosis, and Gampaha (10.78%), Kalutara (9.59%) and Colombo (8.04%) districts were ranked in the top 5 places. The accuracy of this fitted Seasonal Autoregressive Integrated Moving Average (SARIMA) (1,0,0) (0,1,1) is over 85%. According to the model, the forecasted number of leptospirosis cases in the western province was 1168 in 2017 (Gnanapragasam, 2017). However, there were 700 actual cases in the western province.

Most of the past studies used time series analysis for some selected districts in Sri Lanka. Since there is an inadequate prediction model for Leptospirosis in Sri Lanka, we plan to study using regression modelling and time series analysis. The primary motivation of this study is to identify the best-fitted model for Leptospirosis distribution in Sri Lanka by analysing parametric modelling, semi-parametric modelling, and non-parametric modelling by using Linear Regression, Polynomial Regression, Exponential Regression, Cox proportional hazards Regression, ARIMA Time series, Generalized Additive Model and Median Based Linear Model. The study also achieves the following objectives: plot spatial graphs of Leptospirosis distribution in Sri Lanka and propose a predictive model for Leptospirosis distribution in Sri Lanka based on the identified best-fitted model. These models will be used to predict the future Leptospirosis distribution. The best-fitted model for Leptospirosis distribution in Sri Lanka was selected using the lowest AIC and BIC values. The model was validated using Theil's coefficient and the Ljung-Box–Box Test. The best model will be identified using the Root Mean Square Error (RMSE).

Materials and Methods

This study used leptospirosis cases reported in Sri Lanka from 2011 to 2022, collected from the Epidemiology Unit. Spatial location and population in each district were collected from the National Spatial Data Infrastructure Department in Sri Lanka. The Sri Lankan Geodetic system reference coordinate system was used to locate the confirmed leptospirosis cases from 2011 to 2022, using the postal code to locate them geographically within a map of Sri Lanka.

Initially, a preliminary analysis was carried out to understand the data trends. Leptospirosis distribution by district was analysed using spatial graphs for the past 12 year period from 2011 to 2022. This estimate was stratified by district and mapped to determine the observed spatial distribution of the disease in the country. The distribution of interannual leptospirosis incidence was determined and plotted. Distribution of the data was observed by box plots, also known as box and whisker plots (Merriam-Webster) (Aggarwal, 2017).

The Shapiro–Wilk normality test is applied to check whether the model is parametric or non-parametric. The best model for predicting confirmed leptospirosis cases in Sri Lanka is developed using Linear Regression(LR), Polynomial Regression(PR), Exponential Regression(ER), Cox proportional hazards Regression(SPR), ARIMA Time series(TS), Generalized Additive Model(GAM) and Median based Linear Model. The best model is identified using the minimum value of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Vrieze, 2012).

Linear Regression Model

There are parameters β_0, β_1, and σ^2, in the simple linear regression model such that for any fixed value of the independent variable *x*, the dependent variable is a random variable related to *x* through the model equation can be given as:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad ; \quad i = 1,2,3,\dots,n \qquad (1)$$

The quantity ε in the model equation is the "error", a random variable assumed to be symmetrically distributed with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$, x is called the design matrix β is the vector of parameters, ε is the error vector and Y is the response vector (Su et al., 2012).

Linear Regression Model

There are parameters β_0, β_1, and σ^2, in the simple linear regression model such that for any fixed value of the independent variable x, the dependent variable is a random variable related to x through the model equation can be given as:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad ; \quad i = 1,2,3,\dots,n \quad (1)$$

The quantity ε in the model equation is the "error", a random variable assumed to be symmetrically distributed with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$, x is called the design matrix β is the vector of parameters, ε is the error vector and Y is the response vector (Su et al., 2012).

Polynomial Regression Model

A polynomial function is a function that can be expressed as a polynomial. A polynomial is generally represented as $P(x)$. The highest power of the variable of $P(x)$ is known as its degree. The degree of a polynomial function is significant as it tells us about the behaviour of the function $P(x)$ when x becomes very large. The domain of a polynomial function is entire real numbers (R) (Heiberger et al., 2009).

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 \quad (2)$$

For $x \gg 0$ or $x \ll 0$, $a_n \neq 0$

Exponential Regression Model

The natural exponential function $\exp:\mathbb{R} \to \mathbb{R}$ can be characterised in various ways. It is commonly defined by the following power series (Gao et al., 2023).

$$\exp(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots \quad (3)$$

Using the power series, the constant e can be defined as

$$e = \exp(1) = \sum_{k=0}^{\infty} \frac{1}{k!} \quad (4)$$

The term-by-term differentiation of this power series reveals that

$$\frac{d}{dx}\exp(x) = \exp x \quad (5)$$

for all real x, leading to another common characterisation of $\exp x$ as the unique solution of the differential equation

$$\dot{y}(x) = y(x) \quad (6)$$

has satisfied the initial condition

$$y(0) = 1 \quad (7)$$

ARIMA Model

Identifiable patterns and random noise (error) are two basic parts of Time series. There are different models in time series, namely, autoregressive (AR), moving average (MA), autoregressive moving average (ARMA), and autoregressive integrated moving average (ARIMA). The most commonly used model is the Box-Jenkins ARIMA model, which is successfully applied in economic time series prediction (Helail, 2015). Seasonal time series data have seasonal ARIMA models, also known as SARIMA.

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right)(1 - B)^d Y_t = \delta + (1 - \theta_1 B - \cdots - \theta_q B^q) a_t \quad (8)$$

where $\left(1 - \phi_1 B - \cdots - \phi_p B^p\right)$ is the AR operator of order p, $\left(1 - \theta_1 B - \cdots - \theta_q B^q\right)$ is the MA operator of order q, δ is the constant term, and a_t is the shock element at time t.

Forecasting of the ARIMA model

Time series prediction uses the previous values and error terms to forecast the value. ARMA (p, q) model for h-period on the forecast as follows (Adhikari & Agrawal, 2013; Helail, 2015),

$$\hat{y}_{t+h} = \hat{\sigma} + \hat{\emptyset}_1 y_{t+h-1} + \cdots + \hat{\emptyset}_p y_{t+h-p} + e_{t+h} - \hat{\theta}_1 e_{t+h-1} - \ldots - \hat{\theta}_q e_{t+h-q} \quad (9)$$

Generalised Additive Model

A generalised additive model (GAM) is a model in which the linear response variable depends linearly

on unknown smooth functions of some predictor variables, and interest focuses on inference about these smooth functions. The model relates a univariate response variable, *Y*, to some predictor variables x_i. An exponential family of distributions is specified for *Y*, such as Normal, Binomial or Poisson distributions along with a link function $g(E(Y))$ as the identity or log functions, relating the expected value of *Y* to the predictor variables via a structure such as (Hastie, 2017):

$$g(E(Y)) = \beta_0 + f_1(x)_1 + f_2(x)_2 + \cdots + f_m(x_m) \quad (10)$$

β_0 is the intercept.
$f_1(x_1)$, $f_2(x_2)$,…,$f_p(x_p)$ are the smooth functions of the predictor variables x_1, x_2.. x_p

The Cox proportional hazards regression model

The Cox proportional hazards regression model can be written as follows:

$$h(t) = h_0(t)\exp(b_1x_1 + b_2x_2 \ldots + b_px_p) \quad (11)$$

where *h(t)* is the expected hazard at time *t*, and $h_0(t)$ is the baseline hazard that represents the hazard when all of the predictors (or independent variables) x_1, x_2, x_p are equal to zero (Härdle et al., 2004).

Median Based Linear Model

Non-parametric models assume that the data distribution cannot be defined in terms of such a finite set of parameters. However, they can often be defined by assuming an infinite dimension θ. Usually, θ as a function. The amount of information captured about a *D* can grow as data grows (Orbanz & The, 2010).

The best model for predicting confirmed leptospirosis cases in Sri Lanka was developed using Linear Regression, Polynomial Regression, Exponential Regression, Cox proportional hazards Regression, ARIMA Time series, Generalized Additive Model and Median based Linear Model. This was performed using the following steps:

1. The Shapiro–Wilk normality test is applied to check whether the model is parametric or non-parametric (Hanusz et al., 2016).
2. Test of assumption was done by Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling tests (Berk & Freedman, 2003)
3. The best-fitted model for Leptospirosis distribution in Sri Lanka for each district was identified using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Vrieze, 2012)
4. Model validation was done by using Theil's Coefficient and Ljung–Box Test (Bliemel, 1973; Hassani & Yeganegi, 2020)
5. Finally, the cross-validation was done by using Root Mean Squared Error (RMSE) (Chai & Draxler, 2014)

In summary, Polynomial Regression, Exponential Regression, ARIMA, Generalized Additive Model, and median-based linear models were used to identify predicting models of Leptospirosis cases for all districts in Sri Lanka.

Results and Discussion

Figure 1 shows the total number of incidents of leptospirosis in Sri Lanka from 2011 to 2022. There is a high leptospirosis distribution in Ratnapura, Kalutara and Galle. Figure 2 shows the leptospirosis distribution in each year from 2011 to 2022. Data frequency is mentioned in Table 1.

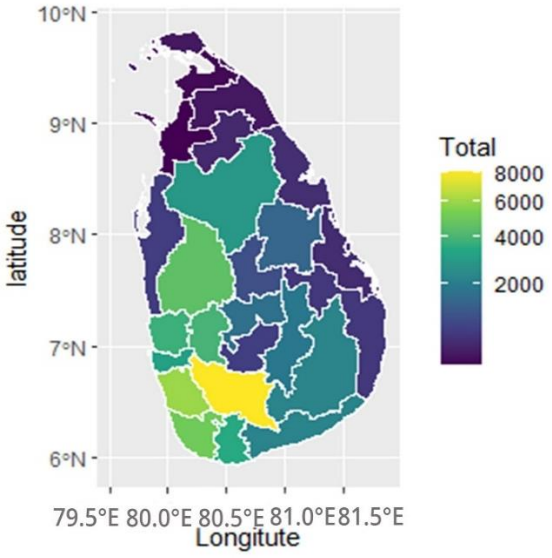


Figure 1: The distribution of leptospirosis cases in Sri Lanka from 2011-2022.

Table 1: Data frequency of leptospirosis distribution in Sri Lanka from 2011-2022

District	Year											
	2011	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
Colombo	507	220	222	225	316	292	194	241	306	425	236	293
Gampaha	557	329	489	463	450	338	118	233	160	314	337	421
Kalutara	400	307	456	375	445	438	388	712	677	1113	729	566
Kandy	192	85	99	85	136	118	64	124	117	336	298	217
Matale	173	52	78	47	65	92	38	119	63	135	95	141
Nuwara Eliya	55	43	34	37	54	70	55	54	69	141	73	97
Galle	252	148	278	246	292	377	546	436	597	1214	788	636
Hambantota	517	101	181	96	174	106	70	93	287	288	290	338
Matara	392	213	177	128	269	209	288	292	555	624	359	368
Jaffna	3	3	10	17	21	24	38	22	44	38	33	28
Kilinochchi	2	4	10	1	7	17	6	13	23	26	60	15
Mannar	16	30	15	4	9	11	3	1	2	12	32	45
Vavuniya	54	19	51	10	21	19	33	52	63	53	25	35
Mullaitivu	10	3	38	13	22	29	29	15	30	33	35	44
Batticaloa	29	12	44	17	34	56	35	63	55	40	57	78
Ampara	72	29	44	24	24	26	28	71	68	116	75	138
Trincomalee	104	43	63	20	19	41	43	64	27	32	5	42
Kurunegala	1576	160	402	161	372	172	121	379	347	334	573	363
Puttalam	125	42	49	65	48	53	29	59	64	74	34	83
Anuradhapura	263	102	345	177	410	277	131	251	247	365	272	288
Polonnaruwa	88	75	187	85	149	91	86	192	135	175	156	184
Badulla	84	38	63	56	92	135	157	189	266	413	342	276
Monaragala	194	79	212	107	196	175	207	422	189	1	453	362
Ratnapura	645	315	435	463	424	638	622	777	1245	1591	919	1156
Kegalle	371	196	315	310	349	188	243	368	348	662	648	715
Kalmunai	7	10	11	3	13	22	10	15	37	25	22	34

Descriptive Analysis

Leptospirosis distribution by districts was analysed using all the available data in the form of line graphs and thematic maps (2011-2022). This estimate was stratified by district and mapped to determine the observed spatial distribution of the disease in the country. The distribution of interannual leptospirosis incidence was determined and plotted.

The normality test was applied for all districts in Sri Lanka, as mentioned in Table 2. Colombo, Kalutara, Kandy, Nuwara Eliya, Kilinochchi, Ampara, Kurunegala, Polonnaruwa and Kegalle were not normally distributed, and these districts can be categorised as non-parametric modelling. Gampaha, Matale, Galle, Hambantota, Matara, Jaffna, Mannar, Vavuniya, Mullaitivu, Batticaloa, Trincomalee, Puttalam, Anuradhapura, Badulla, Monaragala, Ratnapura and Kalmunai can be categorised as parametric modelling (Hanusz et al., 2016).

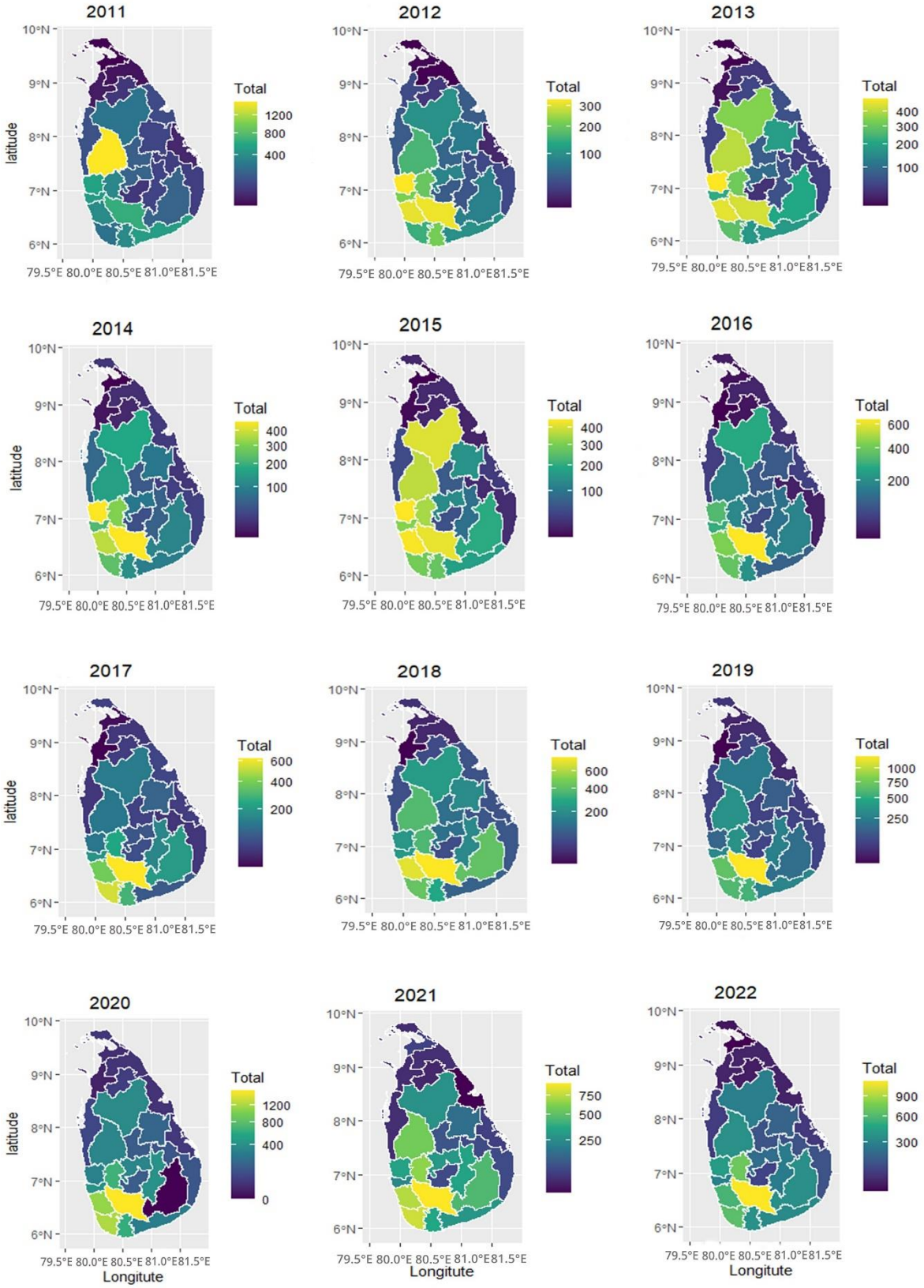


Figure 2: Annual leptospirosis distribution in Sri Lanka from 2011-2022.

Table 2: Shapiro Wilk normality test statistics for Sri Lankan Districts.

District	Test Statistics (w)	p value-Shapiro-Wilk normality test
Gampaha	0.962	0.812
Matale	0.942	0.519
Galle	0.871	0.068
Hambantota	0.874	0.074
Matara	0.925	0.332
Jaffna	0.952	0.664
Mannar	0.873	0.072
Vavuniya	0.914	0.242
Mullaitivu	0.958	0.760
Batticaloa	0.977	0.971
Trincomalee	0.865	0.056
Puttalam	0.889	0.113
Anuradhapura	0.959	0.777
Badulla	0.912	0.229
Monaragala	0.919	0.281
Ratnapura	0.908	0.204
Kalmunai	0.929	0.372
Colombo	0.837	0.025
Kalutara	0.849	0.036
Kandy	0.854	0.041
Nuwara Eliya	0.836	0.025
Kilinochchi	0.779	0.006
Ampara	0.855	0.043
Kurunegala	0.634	0.000
Polonnaruwa	0.859	0.048
Kegalle	0.844	0.031

Figure 3 shows the boxplots of leptospirosis cases in Sri Lankan districts. There are outliers in Galle, Trincomalee and Puttalam and Outlier handling was performed for these three districts. The data points lesser than the 10^{th} percentile were replaced with the 10^{th} percentile value and those greater than the 90^{th} percentile were replaced with the 90^{th} percentile value in Galle, Trincomalee & Puttalam districts.

Model Identification

For parametric modelling, linear, nonlinear, semi-parametric and time series were fitted for all districts. Linear Regression as the linear model, Exponential Regression, Polynomial Regression as the nonlinear, Cox proportional hazards regression model as the semi-parametric model and ARIMA as the time series model were fitted. For non-parametric modelling, a median-based linear model was fitted for leptospirosis cases from 2011 – 2022 for each district.

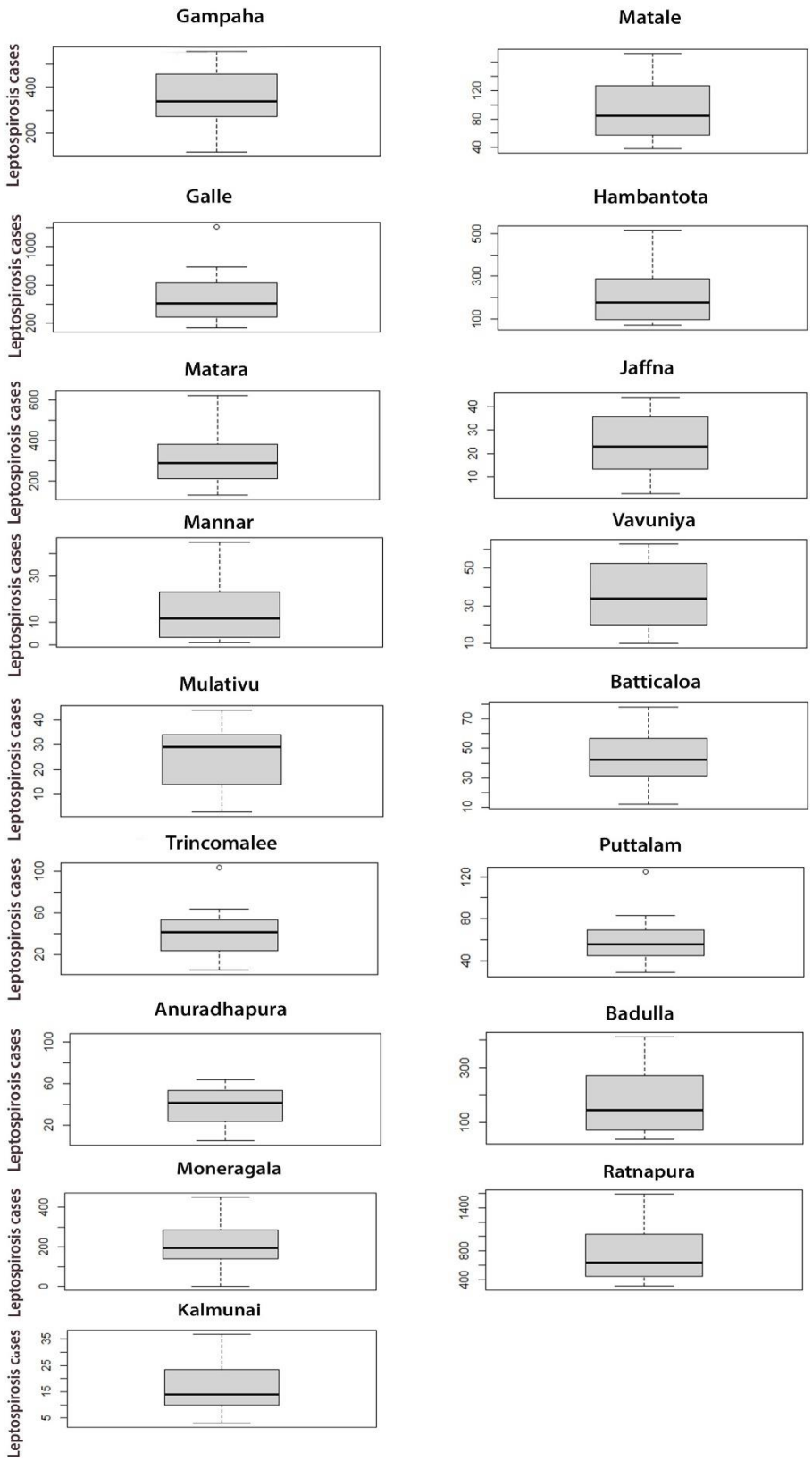
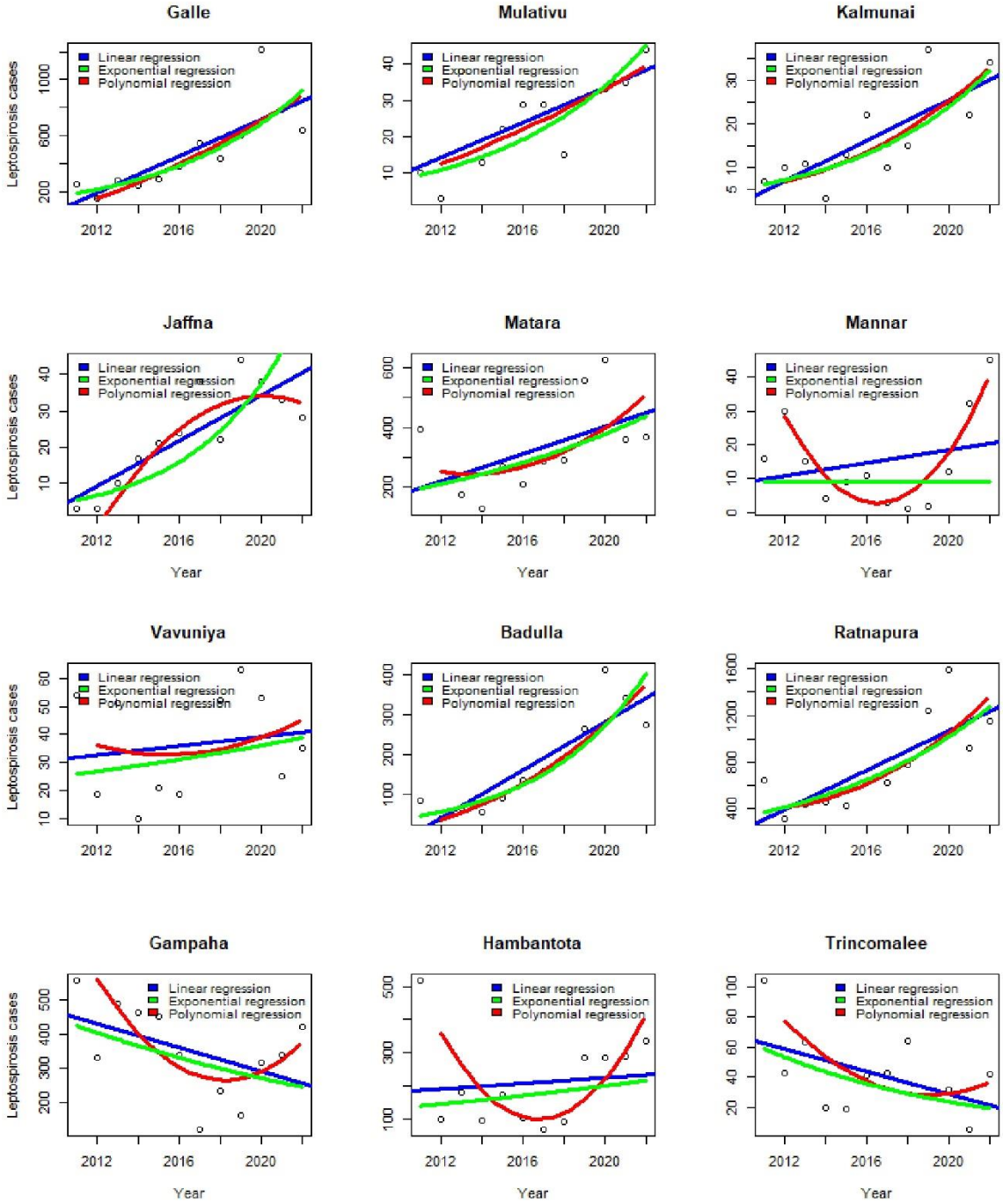


Figure 3: The boxplots of each district for leptospirosis distribution

Figure 4 shows a linear regression fitted model (Su et al., 2012), An Exponential Regression fitted model (Gao et al., 2023), and a polynomial regression fitted model (Heiberger et al., 2009) for leptospirosis cases in all districts, which were categorised as parametric modelling. Figure 5 shows the fitted and median-based linear models for Colombo, Kalutara, Kandy, and Nuwara Eliya, categorised as non-parametric modelling (Orbanz & The, 2010).



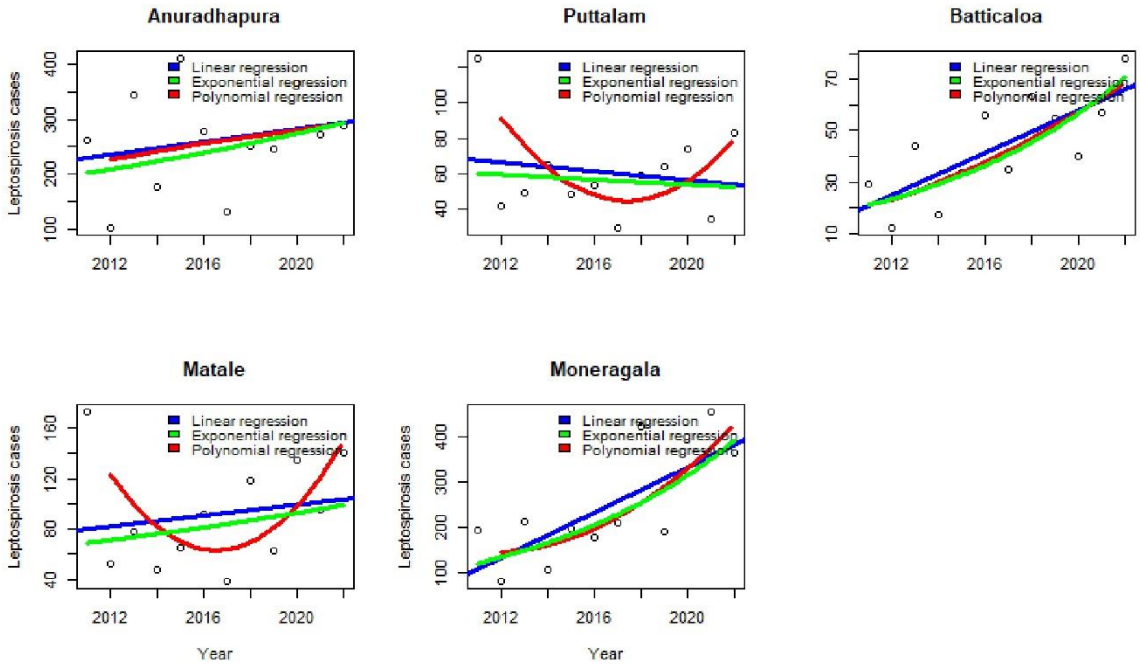


Figure 4: Fitted Linear regression, Exponential regression and Polynomial regression models of each district in Sri Lanka.

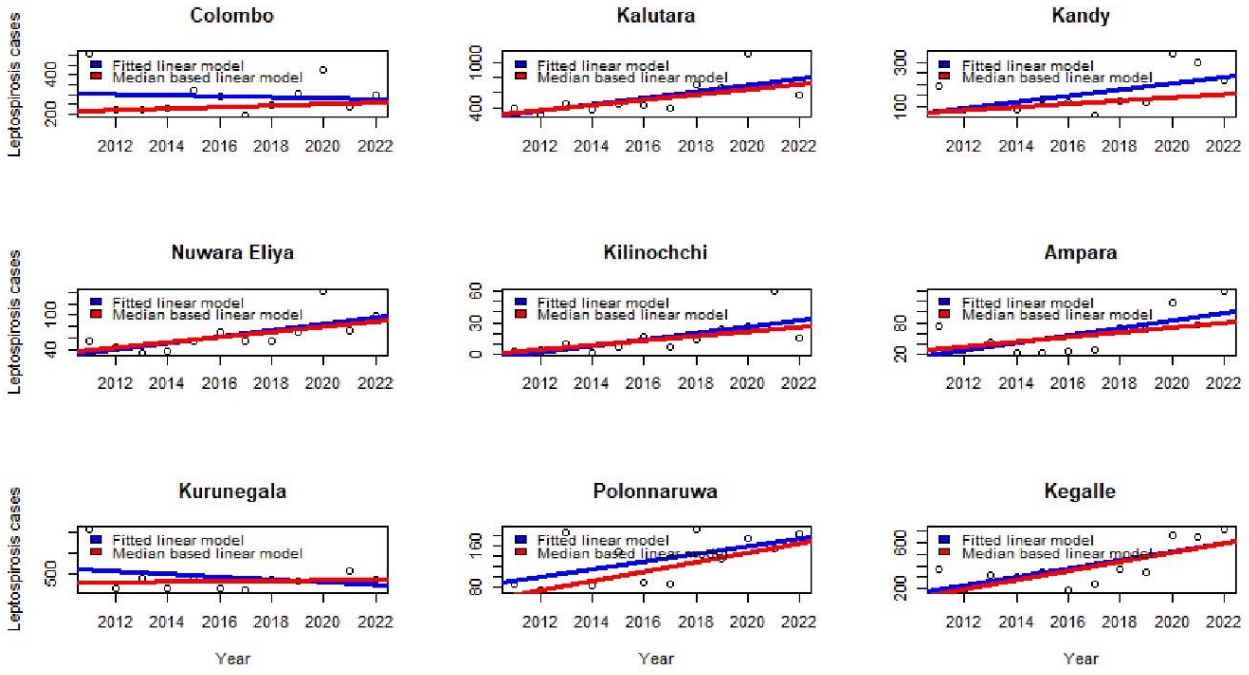


Figure 5: Fitted linear and median-based linear models of each district in Sri Lanka.

Model Selection

Table 3 exhibits the AIC and BIC values of each district for the Linear regression model, the Polynomial regression model, the Exponential regression model, the semi-parametric model, the Time series model and the Generalised additive model. The best model has the minimum value of AIC and BIC. Therefore, the ARIMA model was the best model for Galle, Mullaitivu, Batticaloa, and Puttalam. A Polynomial regression model was selected as the best model for Matale and Jaffna. The Exponential regression model was fitted as the best model for Kalmunai. The generalised Additive model was the best model for Gampaha, Hambantota, Matara, Mannar, Vavuniya, Trincomalee, Anuradhapura, Badulla, Monaragala and Ratnapura. Table 4 shows the AIC and BIC values for all ARIMA models in Galle, Mullaitivu, Batticaloa and Puttalam districts. The best ARIMA model has a minimum value of AIC and BIC. Therefore, for Galle, the ARIMA (1,1,0) model; for Mullaitivu, the ARIMA (2,1,0) model; for Batticaloa, the ARIMA (2,1,0) model; and Puttalam, the ARIMA (0,0,0) model were selected. Table 5 shows that the p-value of the Augmented Dickey-fuller Test is more significant than 0.05 for all ARIMA models. Hence, it can be concluded that the model is significant for the above four districts (Vrieze, 2012).

Table 3: Model selection values (AIC and BIC) for parametric models

District	Model	AIC	BIC
Gampaha	LR	153.361	154.816
	PR	150.362	152.302
	ER	175.474	176.928
	SPR	153.361	154.816
	TS	153.860	155.310
	GAM	**145.357**	**149.721**
Matale	LR	128.573	130.028
	PR	**124.764**	**126.703**
	ER	161.427	162.882
	SPR	128.573	130.028
	TS	126.990	127.960
	GAM	129.141	134.475
Galle	LR	137.786	139.241
	PR	139.772	141.712
	ER	162.567	163.887
	SPR	137.786	139.241
	TS	**131.730**	**132.930**
	GAM	139.764	143.159
Hambantota	LR	156.575	158.030
	PR	148.400	150.339
	ER	180.829	182.283
	SPR	156.575	158.030
	TS	154.720	155.690
	GAM	**143.363**	**147.727**
Matara	LR	154.691	156.145
	PR	155.864	157.804
	ER	174.544	175.998
	SPR	154.691	156.145
	TS	141.630	142.030
	GAM	**130.580**	**135.914**
Jaffna	LR	87.599	89.054
	PR	**82.894**	**84.833**
	ER	92.927	94.381
	SPR	87.599	89.054
	TS	83.900	85.100
	GAM	84.111	87.020
Mannar	LR	101.294	102.749
	PR	89.182	91.122
	ER	108.068	109.523
	SPR	101.294	102.749
	TS	94.190	96.610
	GAM	**71.138**	**75.502**
Vavuniya	LR	107.597	109.051
	PR	109.340	111.280
	ER	109.652	111.107
	SPR	107.597	109.051
	TS	105.930	106.900
	GAM	**81.004**	**86.823**
Mullaitivu	LR	91.570	93.024
	PR	93.542	95.481

	ER	130.84	132.295
	SPR	91.570	93.024
	TS	**89.625**	**91.217**
	GAM	97.280	100.674
Batticaloa	LR	99.632	101.086
	PR	101.502	103.441
	ER	100.698	102.153
	SPR	99.632	101.086
	TS	**95.108**	**96.700**
	GAM	104.255	107.649
Trincomalee	LR	100.822	102.276
	PR	102.663	104.602
	ER	105.344	106.799
	SPR	100.822	102.276
	TS	100.440	101.410
	GAM	**79.914**	**85.733**
Puttalam	LR	101.426	102.881
	PR	102.620	104.560
	ER	99.720	101.175
	SPR	101.426	102.881
	TS	**99.463**	**100.433**
	GAM	99.649	105.468
Anuradhapura	LR	146.557	148.012
	PR	148.556	150.495
	ER	170.874	172.329
	SPR	146.557	148.012
	TS	145.221	146.191
	GAM	**97.791**	**103.610**
Badulla	LR	135.665	137.119
	PR	136.603	138.543
	ER	159.717	161.171
	SPR	135.665	137.119
	TS	124.545	124.943
	GAM	**116.480**	**121.814**
Monaragala	LR	153.993	155.448
	PR	155.609	157.548
	ER	109.133	110.587
	SPR	153.993	155.448
	TS	154.746	155.715
	GAM	**45.262**	**51.081**
Ratnapura	LR	170.732	172.187
	PR	171.805	173.745
	ER	178.442	179.897
	SPR	170.732	172.187
	TS	159.265	159.663
	GAM	**133.056**	**138.875**
Kalmunai	LR	84.148	85.603
	PR	85.654	87.593
	ER	**21.108**	**22.562**
	SPR	84.148	85.603
	TS	84.876	85.274
	GAM	23.505	29.324

Table 4: Model selection values (AIC & BIC) for each district for Time series models

District	**Time series Model**	**AIC**	**BIC**
Galle	ARIMA(0,1,0)	134.721	135.517
	ARIMA(1,1,0)	**131.733**	**132.927**
	ARIMA(0,1,0)	133.993	134.391
	ARIMA(2,1,0)	133.162	134.754
	ARIMA(1,1,0)	134.616	135.412
Mullaitivu	ARIMA(0,1,0)	94.632	95.428
	ARIMA(1,1,0)	91.198	92.392
	ARIMA(0,1,0)	93.096	93.494
	ARIMA(2,1,0)	**89.625**	**91.217**
	ARIMA(3,1,0)	90.415	92.405
	ARIMA(2,1,0)	91.303	92.497
Batticaloa	ARIMA(0,1,0)	102.173	102.969
	ARIMA(1,1,0)	99.354	100.548
	ARIMA(0,1,0)	100.658	101.056
	ARIMA(2,1,0)	**95.108**	**96.700**
	ARIMA(3,1,0)	97.005	98.994
	ARIMA(2,1,0)	98.091	99.285
Puttalam	ARIMA(0,0,0)	**99.463**	**100.433**
	ARIMA(1,0,0)	99.800	101.255
	ARIMA(0,0,0)	132.344	132.829

Table 5: Significance of the best Time series models

District	Time series Model	Augmented Dickey-Fuller Test	Lag Order	p-value
Galle	ARIMA(1,1,0)	-1.556	2	0.742
Mullaitivu	ARIMA(2,1,0)	-0.996	2	0.921
Batticaloa	ARIMA(2,1,0)	-1.629	2	0.714
Puttalam	ARIMA(0,0,0)	-2.576	2	0.353

P value of the box test, Ljung box values > 0.05 of time series plots for Galle, Mullaitivu, Batticaloa and Puttalam. Therefore, it can be concluded that the autocorrelation of residuals differs from zero. Hence, the selected Time series models are significant (Hassani & Yeganegi, 2020). To perform the cross-validation, randomly split the data into a training and test data set. The training data set was fitted by the models, and then values were predicted with the validation set. The RMSE value was calculated for all districts using the training and test data sets. Both training and Test data sets for each district are approximately similar, so this can be concluded as the models are validated (Chai & Draxler, 2014).

Table 6 shows the P values of the Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling tests. The p values indicated that the residuals are normally distributed (Hanusz et al., 2016).

Table 6: Test statistics for residuals in each district

Districts	p values			
	Shapiro-Wilk	**Kolmogorov-Smirnov**	**Cramer-von Mises**	**Anderson-Darling**
Gampaha	0.4591	0.5840	0.0011	0.2903
Matale	0.4156	0.9078	0.0011	0.4471
Galle	0.0084	0.2910	0.0011	0.0116
Hambantota	0.0672	0.8148	0.0011	0.1189
Matara	0.0116	0.2662	0.0000	0.0044
Jaffna	0.5090	0.9211	0.0018	0.4240
Mannar	0.3248	0.7300	0.0011	0.4209
Vavuniya	0.1217	0.6260	0.0011	0.0941
Mullaitivu	0.2288	0.7638	0.0025	0.1780
Batticaloa	0.1406	0.8366	0.0018	0.2281
Trincomalee	0.8913	0.9949	0.0018	0.9087
Puttalam	0.3066	0.9288	0.0062	0.3789
Anuradhapura	0.8386	0.9810	0.0011	0.8191
Badulla	0.0942	0.7788	2.00E-04	0.0957
Monaragala	0.3966	0.6655	0.0018	0.3840
Ratnapura	0.0116	0.1492	0.0000	0.0054
Kalmunai	0.4768	0.9150	0.0018	0.5023

The final model

According to the parameter estimation results in Table 7, the final ARIMA models can be expressed as the following equations.

The final model was ARIMA (1,1,0) for Galle. This model can be expressed in the following form:

$$\widehat{Y_t} - Y_{t-1} = \mu + \emptyset_1(Y_{t-1} - Y_{t-2}) \quad (12)$$

$$\widehat{Y_t} = \mu + Y_{t-1} + \emptyset_1(Y_{t-1} - Y_{t-2}) \quad (13)$$

According to the parameter estimation results which is in Table 7, the final polynomial models can be expressed as the following equations:

$$P(x) = a_2x^2 + a_1x + a_0 \quad (14)$$

The final exponential model can be expressed as the following equations according to the parameters estimation results in Table 7.

$$y = ab^x \quad (15)$$

Table 7: Parameter estimation of the best Time series models, Polynomial Regression models and Exponential regression models

District	Model	Parameters	S.E.	p-value
Galle	ARIMA(1,1,0)	$\emptyset_1$ -0.65	0.24	<0.05
		μ 38.72	13.71	<0.05
Mullaitivu	ARIMA(2,1,0)	$\emptyset_1$ -0.98	0.24	<0.05
		$\emptyset_2$ -0.65	0.26	<0.05
		μ 2.55	1.17	<0.05
Batticaloa	ARIMA(2,1,0)	$\emptyset_1$ -1.08	0.21	<0.05
		$\emptyset_2$ -0.73	0.20	<0.05
		μ 3.66	1.38	<0.05
Puttalam	ARIMA(0,0,0)	$\emptyset_1$ 53.73	3.73	<0.05
Matale	Polynomial Regression	a_0 9.54e+06	4.03e+06	0.0422
		a_1 -9.47e+03	4.00e+03	0.0421
		a_2 2.35e+00	9.92e-01	0.0421
Jaffna	Polynomial Regression	a_0 -1.84e+06	7.05e+05	0.0285
		a_1 1.82e+03	6.99e+02	0.0287
		a_2 -4.50e-01	1.73e-01	0.0289
Kalmunai	Exponential Regression	a 296.5084	83.8816	0.0054
		b 0.1484	0.0416	0.0051

According to the parameter estimation results, which is in Table 8, the final generalised additive models can be expressed as the following equations:

$$g\big(E(Y)\big) = \beta_0 + f_1(x)_1 + f_2(x)_2 + f_3(x)_3 + f_4(x)_4 + f_5(x)_5 + f_6(x)_6 + f_7(x_7) \quad (16)$$

Table 8: Parameter estimation of the generalised additive models

	District				
Coefficient	**Gampaha**	**Hambantota**	**Matara**	**Mannar**	**Vavuniya**
Intercept	532.283	486.194	385.938	17.401	53.886
ns(year,df=n)1	65.706	-293.384	-305.83	-27.641	30.205
ns(year,df=n)2	-183.076	-319.552	-108.592	4.341	-84.794
ns(year,df=n)3	-429.389	-535.653	-201.977	-19.62	-2.215
ns(year,df=n)4	-325.974	-162.079	-56.139	-19.768	-55.88
ns(year,df=n)5	-152.072	-76.028	-180.138	-2.447	6.847
ns(year,df=n)6	-353.438	-605.436	507.52	34.748	0.427
ns(year,df=n)7	20.888	136.641	-50.506	18.519	18.551
ns(year,df=n)8			-208.793		-25.719
ns(year,df=n)9			89.382		-75.461
ns(year,df=n)10					13.851

	District				
Coefficient	**Trincomalee**	**Anuradhapura**	**Badulla**	**Monaragala**	**Ratnapura**
Intercept	48.109	262.771	82.61	194.026	645.994
ns(year,df=n)1	15.015	303.736	-7.656	150.828	-59.711
ns(year,df=n)2	-58.26	-355.018	-25.696	-247.349	-288.578
ns(year,df=n)3	-4.002	526.638	53.917	192.451	-141.375
ns(year,df=n)4	-18.047	-418.233	96.704	-216.337	43.392
ns(year,df=n)5	19.824	132.439	57.631	254.272	-16.332
ns(year,df=n)6	-35.612	-158.991	338.416	252.662	243.284
ns(year,df=n)7	4.899	202.169	313.976	-528.58	1438.656
ns(year,df=n)8	-60.363	50.173	174.54	496.078	121.382
ns(year,df=n)9	-34.954	-248.221	224.406	86.232	-88.076
ns(year,df=n)10	0.196	194.607		282.605	789.74

Table 9: Parameter estimation of the Non-parametric models

District		**Model Coefficient**	**2.50%**	**97.50%**
Colombo	Intercept	-7124.70	-16164.93	8809.56
	Slope	3.65	-4.24	8.14
Kalutara	Intercept	-68169.31	-120772.29	-49024.41
	Slope	34.06	24.55	60.18
Kandy	Intercept	-14753.50	-36910.25	-2057.30
	Slope	7.38	1.07	18.38
Nuwara Eliya	Intercept	-9011.80	-13052.00	-7001.00
	Slope	4.50	3.92	10.33
Kilinochchi	Intercept	-4187.25	-6037.50	-3687.83
	Slope	2.08	1.83	3.00
Ampara	Intercept	-9005.47	-20274.05	-4006.00
	Slope	4.49	2.00	10.08
Kurunegala	Intercept	-10421.00	-43601.00	17905.23
	Slope	5.33	-8.76	21.75
Pollonaruwa	Intercept	-18033.00	-20055.00	-7138.70
	Slope	9.00	3.60	10.00
Kegalle	Intercept	-79166.22	-104091.56	-36193.07
	Slope	39.44	18.11	51.78

Model accuracy using Box test, Ljung Box Values & RMSE

Table 10 shows the P value of the box test, Ljung box values of time series plots for each district. If the P value of Ljung box Test is greater than 0.05, the model is significant. If it is below 0.05, the model is not significant. So, the Time series model is significant.

To perform cross validation, randomly split the data into a training data set and a test data set from year 2017 to 2023. The training data set was fitted by the models and then predicted values with the validation set. The RMSE value was calculated for a few districts for the training and the test data set.

Table 11 shows the RMSE values of the Training and the Test data set. Both values for each district are approximately similar so this concludes that the models are accurate. The RMSE value was calculated for a few districts for the training and the test data set.

Table 10: box test, Ljung box values of time series plots for each district

	Box test, Ljung Box		
District	**X-squared**	**df**	**p-value**
Galle	0.773	2	0.68
Mulativu	0.511	2	0.774
Batticaloa	0.783	2	0.676
Puttalam	1.953	2	0.377

Table 11: RMSE values of Train data set and Test Data set

	RMSE value	
District	**Train Data**	**Test Data**
Gampaha	32.98	34.36
Mannar	0.2	0.46
Vavuniya	11.61	11.72
Trincomalee	9.48	9.88
Matale	20.56	20.63
Nuwara Eliya	16.68	16.92
Kilinochchi	7.25	7.87
Colombo	39.97	40.68
Kalutara	0.48	0.49
Kandy	5.41	5.44

Table 12 shows the predicted Leptospirosis incidents of Galle, Mullaitivu, Batticaloa, and Puttalam from 2024 to 2026. Cases will increase in Galle, Mullaitivu, and Batticaloa.

Table 12: Next three years Leptospirosis Predictions of Galle, Mullaitivu, Batticaloa and Puttalam using ARIMA models

District	**Year**	**Point Forecast**	**Lo 95**	**Hi 95**
Galle	2024	693.053	528.359	857.747
	2025	742.199	538.587	945.810
	2026	774.191	556.174	992.208
Mullaitivu	2024	44.820	23.603	66.036
	2025	49.581	27.211	71.950
	2026	48.891	22.509	75.273
Batticaloa	2024	74.801	48.112	101.489
	2025	79.986	51.711	108.260
	2026	69.006	35.472	102.539
Puttalam	2024	53.729	27.282	80.176
	2025	53.729	27.282	80.176
	2026	53.729	27.282	80.176

Conclusions

The Shapiro-Wilk normality tests separated all districts into parametric and non-parametric models. Linear regression, exponential regression, polynomial regression, ARIMA time series, generalised additive models as the parametric model, Cox proportional hazards regression model as the semi-parametric model, and a median-based linear model as the non-parametric model were fitted. The ARIMA model was the best-fitted model for Galle, Mullaitivu, Batticaloa and Puttalam. The exponential regression model was the best-fitted model for Kalmunai. The polynomial regression model was the best-fitted model for Matale and Jaffna. The generalised additive model was the best-fitted model for Gampaha, Hambantota, Matara, Mannar, Vavuniya, Trincomalee, Anuradhapura, Badulla, Monaragala and Ratnapura. The median-based linear model was the best-fitted model for Colombo, Kalutara, Kandy, Nuwara Eliya, Kilinochchi, Ampara, Kurunegala, Polonnaruwa and Kegalle. Leptospirosis forecasts for Sri Lanka can be calculated using the proposed models.

Furthermore, future researchers should conduct Bayesian hierarchical models to identify high-risk clusters of Leptospirosis incidents in Sri Lanka.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

Adhikari, R. and Agrawal, R.K. (2013). An Introductory Study on Time Series Modeling and Forecasting. Lap Lambert Academic Publishing. https://doi.org/10.48550/arXiv.1302.6613

Adler B. (2015). History of Leptospirosis and Leptospira. *Current topics in microbiology and immunology*, *387*, 1–9.

Agampodi, S. B., Dahanayaka, N. J., Bandaranayaka, A. K., Perera, M., Priyankara, S., Weerawansa, P., ... & Vinetz, J. M. (2014). Regional differences of Leptospirosis in Sri Lanka: observations from a flood-associated outbreak in 2011. PLoS Neglected Tropical

Diseases, 8(1), e2626. https://doi.org/10.1371/journal.pntd.0002626

Agampodi, S. B., Karunarathna, D., Jayathilala, N., Rathnayaka, H., Agampodi, T. C., & Karunanayaka, L. (2014). Outbreak of Leptospirosis after white water rafting: sign of a shift from rural to recreational Leptospirosis in Sri Lanka? *Epidemiology & Infection*, *142*(4), 843-846. https://doi.org/10.1017/S0950268813001465

Aggarwal, C.C. (2017). An Introduction to Outlier Analysis. In: Outlier Analysis. Springer, Cham. https://doi.org/10.1007/978-3-319-47578-3_1

Altun, E. (2021). The log-weighted exponential regression model: alternative to the beta regression model. Communications in Statistics-Theory and Methods, 50(10), 2306–2321.

Berk, R. A., & Freedman, D. A. (2003). Statistical Assumptions. Punishment and social control, 235.

Bliemel, F. (1973). Theil's forecast accuracy coefficient: A clarification. https://doi.org/10.1177/002224377301000413

Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. Wiley Interdisciplinary Reviews: Computational Statistics, 11(3), e1460. https://doi.org/10.1002/wics.1460

Chadsuthi, S., Modchang, C., Lenbury, Y., Iamsirithaworn, S., & Triampo, W. (2012). Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time–-series and ARIMAX analyses. *Asian Pacific journal of tropical medicine*, *5*(7), 539–546. https://doi.org/10.1017/S095026881600203X

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific model development, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Costa, F., Hagan, J. E., Calcagno, J., Kane, M., Torgerson, P., Martinez-Silveira, M. S., ... & Ko, A. I. (2015). Global morbidity and mortality of leptospirosis: a systematic review. *PLoS Neglected Tropical Diseases*, *9*(9), e0003898. https://doi.org/10.1371/journal.pntd.0003898

Desvars, A., Jégo, 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), e20377.

Ehelepola, N. D. B., Ariyaratne, K., Aththanayake, A. M. S. M. C. M., Samarakoon, K., & Thilakarathna, H. M. (2021). The correlation between three teleconnections and leptospirosis incidence in the Kandy District, Sri Lanka, 2004–2019. Tropical medicine and health, 49(1), 1-14.

Ehelepola, N. D. B., Ariyaratne, K., & Dissanayake, D. S. (2021). The interrelationship between meteorological parameters and leptospirosis incidence in Hambantota district, Sri Lanka 2008–2017 and practical implications. PLoS One, 16(1), e0245366. https://doi.org/10.1371/journal.pone.0245366

Erdinc, F. S., Koruk, S. T., Hatipoglu, C. A., Kinikli, S., & Demiroz, A. P. (2006). Three cases of anicteric Leptospirosis from Turkey: Mild to severe complications. *Journal of Infection*, *52*(2), e45-e48. https://doi.org/10.1016/j.jinf.2005.05.015

Gao, Y., Mahadevan, S., & Song, Z. (2023). An over-parameterised exponential regression. arXiv preprint arXiv:2303.16504. https://doi.org/10.48550/arXiv.2303.16504

Gnanapragasam, S. R. (2017). An Empirical Study on Human Leptospirosis Cases in the Western Province of Sri Lanka. http://doi.org/10.4038/ouslj.v12i1.7354

Gnanapragasam, S. R. (2018). Outbreaks of Leptospirosis in selected districts of Sri Lanka. *Int J Acad Res Dev*, *3*, 27-32.

Hanusz, Z., Tarasinska, J., & Zielinski, W. (2016). Shapiro–Wilk test with known mean. REVSTAT-Statistical Journal, 14(1), 89-100.

Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). Non-parametric and semiparametric models (Vol. 1). Berlin: Springer.

Hassani, H., & Yeganegi, M. R. (2020). Selecting optimal lag order in Ljung–Box test. Physica A: Statistical Mechanics and its

Applications, 541, 123700. https://doi.org/10.1 016/j.physa.2019.123700

Hastie, T. J. (2017). Generalised additive models. In Statistical Models in S (pp. 249-307). Routledge

Heiberger, R. M., Neuwirth, E., Heiberger, R. M., & Neuwirth, E. (2009). Polynomial regression. R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics, 269-284.

Helail, M.S.(2015). A Comparative study on ARIMA and GARCH models in forecasting time series data applied study on currency exchange data(ILS/JOR). Master's thesis, Al-Azhar University-Gaza.Kaur H & Ahuja S. (2019). ARIMA Modeling for Forecasting the Electricity Consumption of a Health Care Building. International Journal of Innovative Technology and Exploring Engineering (IJITEE). 8, 2795-2799.

Janssen, P., & Stouffs, R. (2015, May). Types of parametric modelling. In Proceedings of the 20th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) (pp. 157-166).

Merriam-Webster. (n.d.). Outlier. In Merriam-Webster.com dictionary. Merriam. https://www.merriam-webster.com/dictionary/outlier

Orbanz, P., & Teh, Y. W. (2010). Bayesian Non-parametric Models. Encyclopedia of machine learning, 1, 81–89.

Plouffe, C. C. (2016). Space-time modelling of emerging infectious diseases: Assessing leptospirosis risk in Sri Lanka.

Ramsey, A., Rubin-Smith, J., Norwich, A., Katumuluwa, S., Hettiarachchi, A., Wimalage, S., ... & Agampodi, S. (2015). Outbreak investigation of Leptospirosis in Padaviya, Sri Lanka. *Anuradhapura Medical Journal*, *9*(2).

Robertson, C., Nelson, T. A., & Stephen, C. (2012). Spatial epidemiology of suspected clinical Leptospirosis in Sri Lanka. *Epidemiology & Infection*, *140*(4), 731–743. https://doi.org/10.1017/S0950268811001014

Sanjeewani, R. M. S. S. (2017). Temporal and Spatial Trends of Leptospirosis Cases in Sri Lanka.

Schønning, M. H., Phelps, M. D., Warnasekara, J., Agampodi, S. B., & Furu, P. (2019). A case-control study of environmental and occupational risks of Leptospirosis in Sri Lanka. *EcoHealth*, *16*(3), 534-543.

Su, X., Yan, X., & Tsai, C. L. (2012). Linear Regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3), 275-294. https://doi.org/10.1002/wics.1198

Sumi, A., Telan, E. F. O., Chagan-Yasutan, H., Piolo, M. B., Hattori, T., & Kobayashi, N. (2017). Effect of temperature, relative humidity and rainfall on dengue fever and leptospirosis infections in Manila, the Philippines. *Epidemiology & Infection*, *145*(1), 78-86. https://doi.org/10.1017/S09 5026881600203X

Thanusika, T., Selvabavitha, V., & Anuthrika, T. (2022). Time series model to predict the Leptospirosis outbreak in Kalutara district, Sri Lanka.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychological methods, 17(2), 228. https://doi.org/10.1037/a0027127

Warnasekara, J. N., & Agampodi, S. B. (2017). Leptospirosis in Sri Lanka. http://dx.doi.org/10.4038/sljid.v7i2.8155

Warnasekara, J., Agampodi, S., & Abeynayake R, R. (2021). Time series models for prediction of leptospirosis in different climate zones in Sri Lanka. PLoS One, 16(5), e0248032. https://doi.org/10.1371/journal.pone.0248032

Warnasekara, J., Agampodi, S., & NR, A. (2022). SARIMA and ARDL models for predicting Leptospirosis in Anuradhapura district Sri Lanka. *PloS one*, *17*(10), e0275447. https://doi.org/10.1371/journal.pone.0275447

Warnasekara, J., Koralegedara, I., & Agampodi, S. (2019). Estimating the burden of Leptospirosis in Sri Lanka; a systematic review. *BMC Infectious Diseases*, *19*(1), 1-12.