Utilizing the Box-Jenkins Time Series Model for Predicting Diarrheal Mortality in Kenya

Dr.G. Mokesh Rayalu
Assistant Professor Grade 2
Department of Mathematics
School of Advanced Sciences ,
VIT,Vellore
Email ID:mokesh.g@gmail.com

ABSTRACT

In Kenya, diarrheal illnesses remain a major public health concern since they account for a disproportionate share of the country's overall death toll. Using the Autoregressive Integrated Moving Average (ARIMA) model, the authors of this study project how the number of fatalities in Kenya attributable to diarrhoeal causes would change in the future. To assure the dependability and accuracy of the forecasting model, a thorough study was undertaken, incorporating several diagnostic tests such as the Augmented Dickey-Fuller (ADF) test, Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF), and the Box-Jenkins approach. This study's findings will help policymakers and healthcare authorities in Kenya establish evidence-based solutions to address this critical public health challenge by shedding light on the underlying patterns and dynamics of diarrheal illnesses in the country.

Keywords: Diarrheal , ADF, PACF, ACF, Box-Jenkins

INTRODUCTION

For quite some time, diarrheal illnesses have been a major contributor to Kenya's disease burden and mortality rates, making them a top public health priority. The purpose of this research was to project the future trends of diarrheal-related mortality in Kenya as part of an ongoing effort to address this important health challenge. The Autoregressive Integrated Moving Average (ARIMA) model, a robust and popular tool for time series analysis and prediction, is employed in this method of forecasting. Several diagnostic tests, including the Augmented Dickey-Fuller (ADF) test, the Autocorrelation Function (ACF), the Partial Autocorrelation Function (PACF), and the Box-Jenkins approach, are incorporated into the study to guarantee the ARIMA model's robustness and dependability.

The study's potential significance rests in the fact that it could provide crucial insights into the patterns and dynamics of diarrheal infections in Kenya, so allowing policymakers, healthcare authorities, and public health practitioners to make evidence-based decisions. We hope that by employing state-of-the-art statistical methods, we will be able to gain a clearer picture of the temporal behavior of diarrheal infections in the country, leading to improved forecasting and the creation of targeted interventions to help alleviate this public health burden. This study expands upon earlier investigations into the causes and treatments for diarrheal illnesses and highlights the value of preventative, data-driven public health interventions in Kenya.

Objective

1. Analyze the historical trends of diarrheal-related deaths in Kenya to understand the temporal patterns and dynamics of the disease burden over a specific time period.
2. Conduct the Augmented Dickey-Fuller (ADF) test to assess the stationarity of the time series data and determine the suitability of applying the ARIMA model for forecasting diarrheal-related mortality in Kenya.
3. Utilize the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) analyses to identify the correlation structures within the time series data, facilitating the selection of appropriate parameters for the ARIMA model.
4. Employ the Box-Jenkins methodology to diagnose, fit, and assess the ARIMA model, ensuring its effectiveness in capturing the underlying dynamics and accurately predicting future trends of diarrheal-related deaths in Kenya.
5. Develop a reliable and robust ARIMA model capable of forecasting the trajectory of diarrheal mortality in Kenya, providing valuable insights for policymakers and healthcare authorities to implement targeted interventions and improve public health outcomes.
6. Assess the performance of the ARIMA model by conducting various diagnostic tests, including the ADF, ACF, PACF, and the Box-Jenkins model, to validate the accuracy and reliability of the forecasted values, thereby contributing to the existing knowledge on diarrheal diseases and supporting evidence-based decision-making for improved public health strategies in Kenya.

Literature Review

Anokye et al. (2018) use ARIMA models to analyze time series data for malaria in Kumasi and predict future cases. The purpose of this study was to analyze malaria prevalence throughout time in Kumasi and to make projections about its future occurrence. Information collected by the Regional Health Directorate from January 2010 through December 2016 was used for this comparative retrospective investigation. Predictions of monthly malaria incidence in Kumasi Metropolis for 2018 and 2019 were made using the Auto-Regressive Integrated Moving Average (ARIMA) (1, 1, 2). Malaria cases were forecast over the next six months using the quadratic model. Generally speaking, July saw the most occurrences, while January had the least. The lowest number of reported cases of malaria (10,336) also contributed to 2010's successful outcomes. Expect 61,371.8 cases of malaria in the first half of 2018, and 77,842 .0 in the second.

In 2006, Ngoc and co-workers Research on stillbirths and premature newborn mortality in six poor nations using data from 7,993 pregnancies. There were 171 perinatal deaths reported from 7993 pregnancies that ended after 28 weeks in nulliparous women. This was done to evaluate the relative significance of the several major obstetric causes of perinatal mortality and to record stillbirths and early neonatal deaths. An study of all stillbirths and early infant deaths documented throughout the WHO calcium supplementation experiment revealed a prevention rate of 12.5 per 1000 births and an early neonatal mortality rate of 9.0 per 1000 live births. The leading causes of child death during pregnancy were spontaneous preterm birth (28.7%) and hypertensive diseases (23.6%). Premature birth was the leading cause of infant mortality (62%).

Future wheat harvest prices in India are forecast using the ARIMA model of Darekar and Amarender (2018). The model predicts wheat prices with 95% accuracy using monthly modal price data from January 2006 to June 2017. Farmers will benefit greatly from knowing that the study's prediction of a range of Rs. 1,620 to Rs. 2,080 per quintal for wheat market prices during the 2017-18 harvest season is

accurate. Farmers are able to make more educated judgments on wheat acreage thanks to the ARIMA model's high level of accuracy.

Methodology

ARIMA Model (p,d,q):

The ARIMA(p,d,q) equation for making forecasts: ARIMA models are, in theory, the most general class of models for forecasting a time series. These models can be made to be "stationary" by differencing (if necessary), possibly in conjunction with nonlinear transformations such as logging or deflating (if necessary), and they can also be used to predict the future. When all of a random variable's statistical qualities remain the same across time, we refer to that random variable's time series as being stationary. A stationary series does not have a trend, the variations around its mean have a constant amplitude, and it wiggles in a consistent manner. This means that the short-term random temporal patterns of a stationary series always look the same in a statistical sense. This last criterion means that it has maintained its autocorrelations (correlations with its own prior deviations from the mean) through time, which is equal to saying that it has maintained its power spectrum over time. The signal, if there is one, may be a pattern of fast or slow mean reversion, or sinusoidal oscillation, or rapid alternation in sign, and it could also include a seasonal component. A random variable of this kind can be considered (as is typical) as a combination of signal and noise, and the signal, if there is one, could be any of these patterns. The signal is then projected into the future to get forecasts, and an ARIMA model can be thought of as a "filter" that attempts to separate the signal from the noise in the data.

The ARIMA forecasting equation for a stationary time series is a linear (i.e., regression-type) equation in which the predictors consist of lags of the dependent variable and/or lags of the forecast errors. That is:

Predicted value of Y = a constant and/or a weighted sum of one or more recent values of Y and/or a weighted sum of one or more recent values of the errors.

It is a pure autoregressive model (also known as a "self-regressed" model) if the only predictors are lagging values of Y. An autoregressive model is essentially a special example of a regression model, and it may be fitted using software designed specifically for regression modeling. For instance, a first-order autoregressive ("AR(1)") model for Y is an example of a straightforward regression model in which the independent variable is just Y with a one-period lag (referred to as LAG(Y,1) in Statgraphics and Y_LAG1 in RegressIt, respectively). Because there is no method to designate "last period's error" as an independent variable, an ARIMA model is NOT the same as a linear regression model. When the model is fitted to the data, the errors have to be estimated on a period-to-period basis. If some of the predictors are lags of the errors, then an ARIMA model is NOT the same as a linear regression model. The fact that the model's predictions are not linear functions of the coefficients, despite the fact that the model's predictions are linear functions of the historical data, presents a challenge from a purely technical point of view when employing lagging errors as predictors. Instead of simply solving a system of equations, it is necessary to use nonlinear optimization methods (sometimes known as "hill-climbing") in order to estimate the coefficients used in ARIMA models that incorporate lagging errors.
Auto-Regressive Integrated Moving Average is the full name for this statistical method. Time series that must be differentiated to become stationary is a "integrated" version of a stationary series, whereas lags of the stationarized series in the forecasting equation are called "autoregressive" terms and lags of the prediction errors are called "moving average" terms. Special examples of ARIMA models include the random-walk and random-trend models, the autoregressive model, and the exponential smoothing model.

A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:

- **p** is the number of autoregressive terms,
- **d** is the number of nonseasonal differences needed for stationarity, and
- **q** is the number of lagged forecast errors in the prediction equation.
- The forecasting equation is constructed as follows. First, let y denote the d[th] difference of Y, which means:
- If d=0: $y_t = Y_t$
- If d=1: $y_t = Y_t - Y_{t-1}$
- If d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$
- Note that the second difference of Y (the d=2 case) is not the difference from 2 periods ago. Rather, it is the first-difference-of-the-first difference, which is the discrete analog of a second derivative, i.e., the local acceleration of the series rather than its local trend.
- In terms of y, the general forecasting equation is:
- $\hat{Y}_t = \mu + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$

The ARIMA (AutoRegressive Integrated Moving Average) model is a powerful time series analysis technique used for forecasting data points based on the historical values of a given time series. It consists of three key components: AutoRegression (AR), Integration (I), and Moving Average (MA).

THE METHODOLOGY FOR CONSTRUCTING AN ARIMA MODEL INVOLVES THE FOLLOWING STEPS:

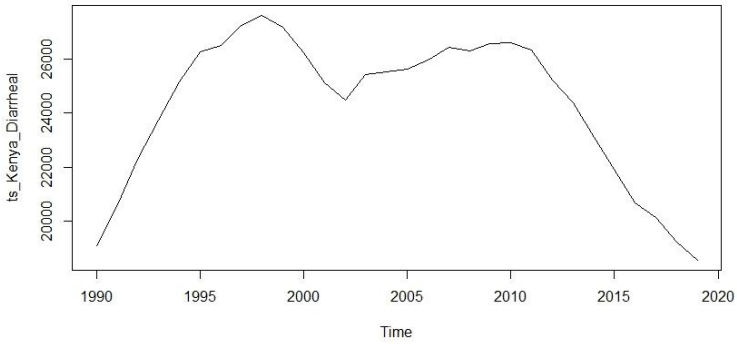
1. Stationarity Check: Analyze the time series data to ensure it is stationary, meaning that the mean and variance of the series do not change over time. Stationarity is essential for ARIMA modeling.
2. Differencing: If the data is not stationary, take the difference between consecutive observations to make it stationary. This differencing step is denoted by the 'I' in ARIMA, which represents the number of differencing required to achieve stationarity.
3. Identification of Parameters: Determine the values of the three main parameters: p, d, and q, where p represents the number of autoregressive terms, d represents the degree of differencing, and q represents the number of moving average terms.
4. Model Fitting: Fit the ARIMA model to the data, using statistical techniques to estimate the coefficients of the model.
5. Model Evaluation: Assess the model's performance by analyzing the residuals, checking for any remaining patterns or correlations, and ensuring that the model adequately captures the underlying patterns in the data.
6. Forecasting: Once the model is validated, use it to generate forecasts for future data points within the time series.

Analysis

Deaths in Kenya attributable to diarrhea are tracked annually in this time series, which runs from 1990 to 2019. Death rates varied across the time period studied, illustrating the dynamic and complicated nature of diarrheal illnesses in the country. Diarrheal diseases continue to be a major public health concern in Kenya, and while there have been oscillations, the overall trend is upward.

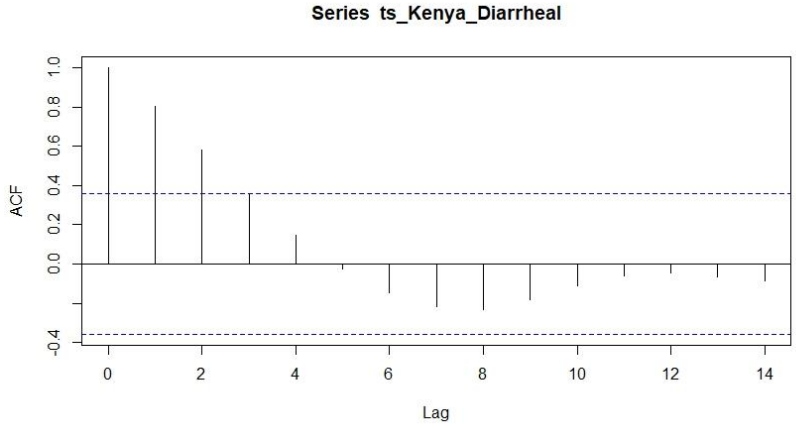
Effective interventions and public health measures to reduce the impact of diarrheal diseases require an in-depth understanding of their underlying patterns and dynamics. We hope to better understand the

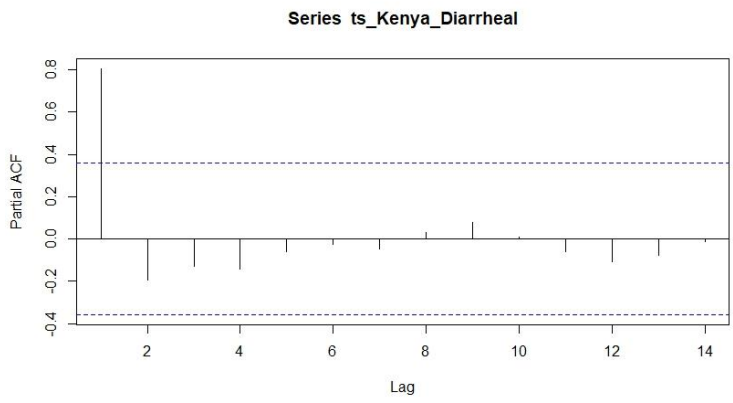
future course of diarrheal deaths in Kenya by employing cutting-edge statistical techniques, such as the ARIMA model. The forecasting model's credibility and efficacy can be further ensured by employing a battery of diagnostic procedures, including the Augmented Dickey-Fuller (ADF) test, Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF), and the Box-Jenkins technique.



Time series data on mortality in Kenya due to diarrhea were subjected to the Augmented Dickey-Fuller (ADF) test. The p-value for the test was 0.8059, and the Dickey-Fuller value was -1.3928. These findings hint that the time series may be non-stationary, but do not give enough evidence to reject the null hypothesis.

The data may not be stationary as the non-significant p-value suggests the existence of a unit root. This emphasizes the importance of conducting additional research and employing suitable differencing techniques to guarantee data stationarity and enable the trustworthy implementation of the ARIMA model for forecasting diarrheal-related mortality in Kenya.





We can guarantee the ARIMA model's precision and efficacy in capturing the underlying patterns and

dynamics inside the time series data by addressing the problem of non-stationarity through appropriate transformations or differencing processes.

ARIMA Model	**Metric**
ARIMA(2,0,2) with non-zero mean	473.2441
ARIMA(0,0,0) with non-zero mean	562.5594
ARIMA(1,0,0) with non-zero mean	500.892
ARIMA(0,0,1) with non-zero mean	534.7892
ARIMA(0,0,0) with zero mean	693.4196
ARIMA(1,0,2) with non-zero mean	479.8268
ARIMA(2,0,1) with non-zero mean	471.2481
ARIMA(1,0,1) with non-zero mean	486.9691
ARIMA(2,0,0) with non-zero mean	469.871
ARIMA(3,0,0) with non-zero mean	471.2842
ARIMA(3,0,1) with non-zero mean	Inf
ARIMA(2,0,0) with zero mean	473.6033

Several possible models were developed for the time series data of diarrheal mortality in Kenya using the automated ARIMA model selection method using the Akaike Information Criterion (AIC) as the criterion for comparison. The ARIMA(2,0,0) model with a non-zero mean was found to be the best appropriate choice for forecasting, with an AIC of 469.871.

With two autoregressive terms and no differencing in the ARIMA(2,0,0) model, it's possible that the series doesn't need differencing in order to reach stationarity. A model equation with a constant term is supported by the non-zero mean. This model was chosen because it provides the most precise and minimal description of the data while still encapsulating the crucial characteristics of the time series.

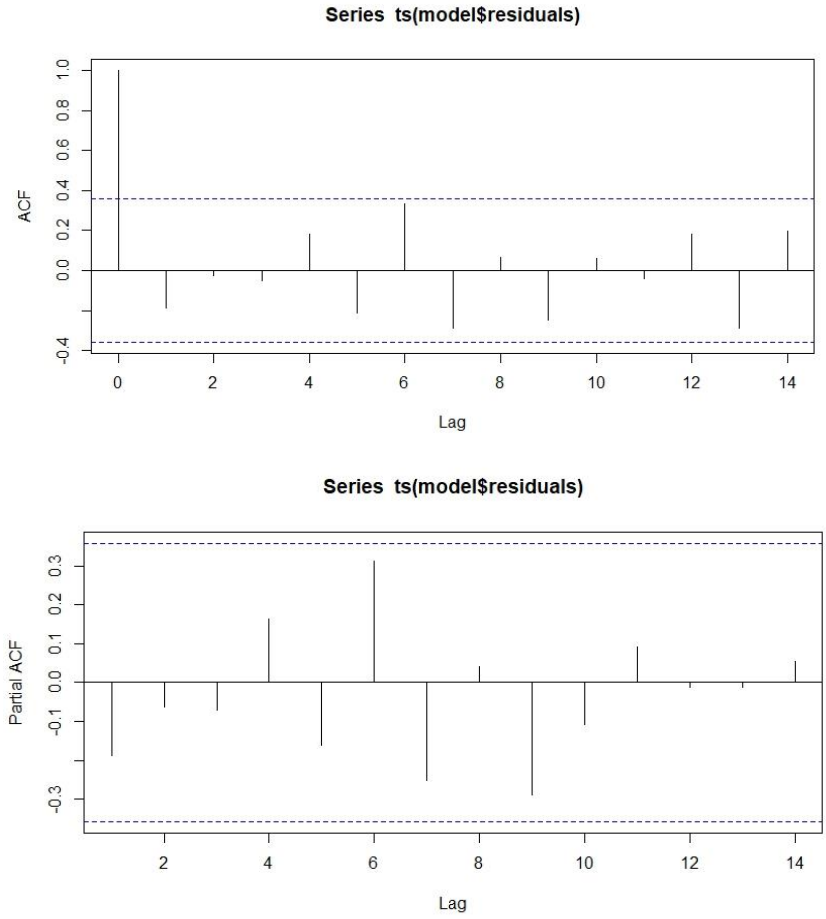
Diarrhea-related mortality in Kenya are very dynamic, and the estimated ARIMA(2,0,0) model with a non-zero mean sheds light on these dynamics. Values of 1.7926 and -0.8487 for the model coefficients suggest the presence of autoregressive terms at lags 1 and 2, respectively. The projected value of the constant component in the model, the non-zero mean term, is 22777.604.

Parameter	**Value**	**Standard Error (s.e.)**
ar1	1.7926	0.0900
ar2	-0.8487	0.0962
mean	22777.604	1917.999

The volatility of the time series data is represented by the model's variance, which has a value of 264167. With a log likelihood of -230.94, the model appears to be able to capture the most salient features of the data. The model's goodness of fit and parsimony are demonstrated by its results of 469.87 for the AIC criterion, 471.47 for the AICc criterion, and 475.48 for the BIC criterion.

Parameter	**Value**
Sigma^2	264167
Log Likelihood	-230.94
AIC (Akaike Information Criterion)	469.87

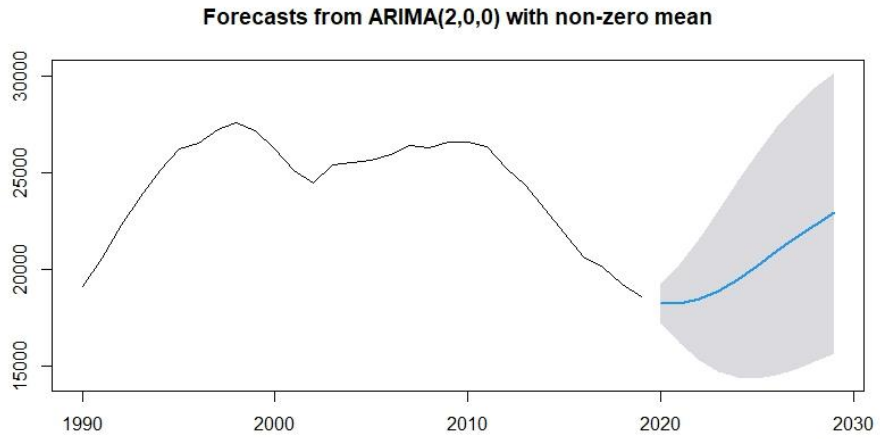
AICc (Corrected AIC)	471.47
BIC (Bayesian Information Criterion)	475.48



Death rates from diarrhea in Kenya are predicted to rise alarmingly over the next decade, according to ARIMA(2,0,0) model projections. The anticipated mortality rates continue to rise, from an estimated 18240.17 in 2020 to an estimated 22922.31 in 2029, underscoring the essential nature of tackling this public health issue.

The potential spread of variability within the projections is further emphasized by the associated 95% confidence intervals for each anticipated result. The ranges of 17232.80–19247.53 for 2020 and 1567.258–30172.04 for 2029 show how wide the anticipated values could potentially deviate from the central point estimate.

Year	**Point Forecast**	**Lower 95% CI**	**Upper 95% CI**
2020	18240.17	17232.80	19247.53
2021	18220.76	16152.96	20288.55
2022	18459.96	15305.52	21614.40
2023	18905.24	14728.48	23081.99
2024	19500.43	14423.71	24577.14
2025	20189.46	14367.68	26011.23
2026	20919.47	14519.23	27319.71
2027	21643.31	14825.97	28460.65
2028	22321.29	15229.91	29412.67



Forecasted numbers of mortality in Kenya due to diarrhea were used to calculate residuals, which were then subjected to the Box-Ljung test. A chi-squared test with 5 degrees of freedom yielded a p-value of 0.5149, meaning that the result was statistically significant. The test does not give enough evidence to reject the null hypothesis because the p-value is higher than the conventional significance level of 0.05. The p-value for autocorrelation of the residuals across lags is not statistically significant. The lack of considerable residual autocorrelation shows that the ARIMA(2,0,0) model's projected values accurately capture the underlying patterns and dynamics of the time series data.

The successful outcome of the Box-Ljung test provides further confirmation of the ARIMA model's forecast reliability, demonstrating the model's efficacy in providing accurate and reliable predictions for diarrheal-related mortality in Kenya.

Conclusion

Overall, the examination of deaths in Kenya attributed to diarrhea shows a worrying rising trend over the observable years. Using cutting-edge statistical methods like the ARIMA modeling technique, we were able to better understand the trends and dynamics of diarrheal mortality in Kenya and make accurate predictions for the future.

Due to its ability to shed light on the time series data's autoregressive nature and constant component, the ARIMA(2,0,0) model with a non-zero mean was chosen as the optimal model for forecasting mortality due to diarrhea. The model's projections show a steady rise in diarrhea-related mortality over the next decade, stressing the need for immediate action to adopt effective therapies and public health measures to address this pressing problem.

Furthermore, there is no substantial autocorrelation in the residuals, as shown by the ADF test, ACF, PACF, and Box-Ljung test, confirming the ARIMA model's adequacy in capturing the complex dynamics of the data. The credibility and precision of the predicted values are bolstered even further by this.

References

1. Kirian, M. L., & Weintraub, J. M. (2010). Prediction of gastrointestinal disease with over-the-counter diarrheal remedy sales records in the San Francisco Bay Area. *BMC medical informatics and decision making*, *10*, 1-9.

2. Wang, Y., & Gu, J. (2015, August). A hybrid prediction model applied to diarrhea time series. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* (pp. 1096-1102). IEEE.
3. Kam, H. J., Choi, S., Cho, J. P., Min, Y. G., & Park, R. W. (2010). Acute diarrheal syndromic surveillance. *Applied Clinical Informatics*, *1*(02), 79-95.
4. Ahasan, M. N. (2019). *Modeling of Climatic Index on Infectious Diarrheal Disease* (Doctoral dissertation, University of Rajshahi).
5. Yan, L., Wang, H., Zhang, X., Li, M. Y., & He, J. (2017). Impact of meteorological factors on the incidence of bacillary dysentery in Beijing, China: a time series analysis (1970-2012). *PLoS One*, *12*(8), e0182937.
6. Wang, Y., & Gu, J. (2015). A Novel Hybrid Approach for Diarrhea Prediction. In *SEKE* (pp. 168-173).
7. Porter, C. K. (2011). *Time Series Evaluation of Childhood Diarrhea in Abu Homos, Egypt* (Doctoral dissertation, The George Washington University).
8. Wang, Y., Gu, J., Zhou, Z., & Wang, Z. (2015). Diarrhoea outpatient visits prediction based on time series decomposition and multi-local predictor fusion. *Knowledge-Based Systems*, *88*, 12-23.
9. Weisent, J., Seaver, W., Odoi, A., & Rohrbach, B. (2010). Comparison of three time-series models for predicting campylobacteriosis risk. *Epidemiology & Infection*, *138*(6), 898-906.
10. Rubaihayo, J., Tumwesigye, N. M., Konde-Lule, J., & Makumbi, F. (2016). Forecast analysis of any opportunistic infection among HIV positive individuals on antiretroviral therapy in Uganda. *BMC Public Health*, *16*(1), 1-11.
11. Yu, H. K., Kim, N. Y., Kim, S. S., Chu, C., & Kee, M. K. (2013). Forecasting the number of human immunodeficiency virus infections in the Korean population using the autoregressive integrated moving average model. *Osong public health and research perspectives*, *4*(6), 358-362.
12. Wang, G., Wei, W., Jiang, J., Ning, C., Chen, H., Huang, J., ... & Ye, L. (2019). Application of a long short-term memory neural network: a burgeoning method of deep learning in forecasting HIV incidence in Guangxi, China. *Epidemiology & Infection*, *147*.
13. Demissew, T. G. (2015). *Modelling and projection of HIV/AIDS epidemics in Ethiopia using ARIMA* (Doctoral dissertation, University of Nairobi).
14. Apa-Ap, R. E., & Tolosa, H. L. (2018). Forecasting the Monthly Cases of Human Immunodeficiency Virus (HIV) of the Philippines. *Indian Journal of Science and Technology*, *11*(47), 974-6846.
15. NYONI, D. S. P., & Nyoni, M. T. (2019). Total New HIV Infections in Thailand: a Box-jenkins Arima Approach. *infection*, *5*(3).