Departmemt of Mathematics and Statistics

CAPSTONE PROJECT LITERATURE REVIEW

Sustainable and Intelligent Time-Series Models for Epidemic Disease Forecasting and Analysis

Michael Carnival

Supervisor: **Dr. Pu Shusen**

September 1, 2024

Contents

CHAPTER 1

Summary and Analysis

In this paper, the researcher attempted to tackle forecasting and predicting the trends of several epidemic diseases like Monkeypox, HIV, COVID-19, and influenza using ARMIA, Polynomial Regression, SARIMA, Holt's, FB-prophet time-series models. Forecasting the trends of diseases is important in equipping the healthcare system or governments worldwide to act appropriately to prevent and control the spread of these outbreaks. It improves warning signals to the public early on. Up to July 2022, the COVID-19 pandemic has infected more than 500 million people, of which 6 million have died in over 100 countries. Apart from COVID-19, Monkeypox and HIV epidemic disease are still of concern to the WHO. The Monkeypox virus was also in the recent headline, and the WHO declared it a public health emergency with 20,000 cases in total as of July 2022 (Gupta et al., 2023a). HIV has and continues to wreak havoc, claiming 40.1 million lives as of 2021. The models mentioned above will provide a robust framework for forecasting the onset and spread of several epidemics. The performance metrics Root Mean Squared Error (RMSE), Mean Squared Error (MSE), mean absolute percentage error (MAPE), and R2 (R2 score) were used to evaluate the best models for use in predicting and forecasting these epidemics.

The objective of this study is to develop a sustainable framework to assist governments and healthcare industries in effectively controlling the spread of epidemics through forecasting. This framework is particularly useful for regions where weak institutions and political instability flourish, which can exacerbate the impact of epidemics on

lower-middle-income countries. In addition, the increase in urbanization and overuse of the natural environment all contribute to an increased risk for disease outbreaks (Di Giulio and Eckburg, 2004). A more sustainable framework is imperative to ensuring public health safety for low-middle-income countries.

Having a sustainable intelligent framework for forecasting disease outbreaks is crucial for lower-middle income countries by strengthening the early warning message to communities, timely interventions and resource allocation to critical infrastructure like hospitals and testing facilities, and reducing infections, deaths, and the burden on the healthcare system.

CHAPTER 2

Methodology

The methods proposed in this research include data collection and data preprocessing, data analysis and visualization, partition of data, model selection, model fitting and hypertuning, model evaluation, comparison of models, and future prediction using the best models. The data collection was based on authenticity, minimum void/null values, maximum period, and consistency. The data preprocessing followed the steps of converting the date column to datetime64 python object; the selected attributes were total cases, deaths, and daily new cases of various epidemic diseases around the world, created a new data frame containing the attributes of epidemic disease and dates corresponding to it. The partitioning of the data followed a 98 %train and 2% test sequentially since the dataset is sufficiently large enough to yield statistically meaningful results. The data visualization provides interpretability of various epidemic diseases and gives insights into the trajectory of where current diseases are heading in the future. The data analysis will support the underlying meaning of these visualizations by providing data relationships and data-driven insights, making the data more understandable. The model selection, fitting, and hypertuning will include the auto-regressive integrated moving average (ARIMA), Fb-Prophet, Holt's linear, Holt's winter, and polynomial regression models in finding the best model for forecasting the disease mentioned. The evaluation metrics used are MSE, RMSE, MAPE, and R2 scores to assess and to compare various models to find the optimal algorithm for each disease.

2.1 ARIMA method

While maintaining the data processing steps, the various models ARIMA, Fb-Prophet, Holt's linear, Holt's winter, and polynomial regression address the problem of prediction and forecasting the spread of diseases in different ways. The ARIMA model can be broken down into AR, I, and MA components. The AR(p) component determines the lagged series, where p determines the number of lagged series. Each subsequent series is dependent on the past

$$y_t = \alpha + \omega_t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_q y_{t-q} \tag{2.1}$$

$y_{t-1}, y_{t-2}, y_{t-p}$ are the lags (past values). $\phi_1, \phi_2, ..., \phi_p$ are lag coefficients calculated from the model; ω_t is the white noise, $alpha = 1 - \sum_{i=1}^{N} \phi_i$ μ and α is . μ is the mean of all the values. The MA(q) component is the moving average model, where q is the number of lagged forecasting error terms in the prediction. It accounts for the relationship between the observation and residual error in the moving average model applied to the lag observation (Wang et al., 2018). Its equation is defined below (Pathoee et al., 2022).

$$y_t = \alpha + \omega_t + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + ... + \theta_q \omega_{t-q} \tag{2.2}$$

The $\omega_t, \omega_{t-1}, ..., \omega_{t-q}$ are the model's error terms with respect to lags $Z_t, Z_{t-1,}, Z_{t-2}, ..., Z_{t-q}$. The integrated (I) component indicates any differencing in the series such that any series not stationary are brought to stationary. In time series datasets, having stationary property is critical in order to make best prediction on future values (Gupta et al., 2022a). This integrated component is especially important when common statistical properties (mean, sd, median, var) in forecasting epidemics disease constantly vary. The equation for integrated component is shown below.

$$y_t = (Z_t - Z_{t-1}) - (Z_{t-1} - Z_{t-2}) = Z_t - 2Z_{t-1} + Z_{t-2} \tag{2.3}$$

The full ARIMA equation is

$$y_t = \alpha + \sum_{i=i}^{p} \phi_i y_{t-i} + \omega_t + \sum_{j=i}^{p} \phi_j \omega_{t-j} \tag{2.4}$$

While the ARIMA model is suitable for time-series datasets, it is noted that the data satisfies stationary as opposed to non-stationary data. Non-stationary data leads to

poor forecasting accuracy and unreliable results (Wei et al., 2023). In order to check for stationarity, the common method used is the Augmented-Dickey Fuller (ADF) test. The output of the ADF will result in a p-value of which it has to be less than 0.05. If the p-value is greater than 0.05, the integrated (I) component gets implemented to make the data stationary.

2.2 Fb method

The open-source Fb-prophet model developed by Facebook focuses on factors where seasonality and holiday impacts fit with nonlinear patterns on a daily, monthly, and yearly basis. These factors provide a powerful tool for time series forecasting in analyzing whether these diseases are affected during holidays or particular seasons. The advantage of the prophet model is its ability to manage missing data, changes within trends, the effects of outliers, and the ability to add exogenous variables. A decomposable time-series model used by Prophet is shown.

$$y(t) = A(t) + B(t) + C(t) + E(t) \tag{2.5}$$

where,

$$A(t) = \text{trend}(\text{logistic}/\text{linear}) \tag{2.6}$$

$$B(t) = \text{Periodic change}/\text{seasonality} = \sum_{i=1}^{N} \left(a_n \cos\left(\frac{2\pi nt}{p}\right) + b_n \sin\left(\frac{2\pi nt}{p}\right)\right) \tag{2.7}$$

The p is the expected constant period of the time series where p= 365 is the annual data or p =7 for weekly data. The T and N must be calculated for every value of t in the historical data and projected data. C(t) is the effect of holiday_i for each holiday, D_i is the set of past and future dates for holiday shown

$$Z(t) = [1(t \in D_1), ..., 1(t \in D_l)] \tag{2.8}$$

The E(t) is excluded in the prophet model. A Hamiltonian Monte Carlo algorithm efficiently samples from the posterior distribution of the model parameters and aids in calculating parameter uncertainty (Kirchgassner et al., 2012).

2.3 Polynomial Regression Method

The polynomial regression of degree between (2 to 10) models is another tool that can be of use for forecasting epidemics since the nature of the data for epidemics is nonlinear. Polynomial regression is a type of multiple regression, where one independent variable x and the dependent variable y are modeled to nth degree polynomial.

$$y = f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_d x^d + \epsilon \tag{2.9}$$

In addition to the preceding equation, the polynomial equation may express as

$$y = f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_1 + ... + \beta_d x_d + \epsilon \tag{2.10}$$

Where $x_1 = x, x_2 = x^2, ..., x_n = x^n$, and the Least Square applied will give an estimate to the response variable (Sulasikin et al., 2020).

2.4 Holt's Linear Method

Holt's linear model employs the exponential smoothing for times series (Al-Rashedi and Al-Hagery, 2023). This feature of the Holt's model is popular among trend-driven data forecasting (Cong et al., 2019), which has three independent equations. The first equation,

$$a_t = \alpha y_t (1 - \alpha)(a_{t-1} + b_{t-1}) \tag{2.11}$$

modifies the most recent smoothed value for the trend of the previous period. The second equation

$$b_t = \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1}, \tag{2.12}$$

expresses the trend as a variation between the most recent two smoothed values and is updated throughout time (Yang et al., 2021). The final forecast is produced using the third equation

$$y_{t+k} = a_t + b_t k, \tag{2.13}$$

where two parameters are adjusted for trend smoothing α and the other for overall smoothing β. The α is in between $0 \leq \alpha \leq 1$, β is in between $0 \leq \beta \leq 1$, and k is the number of time periods for which the forecast is created (Hyun et al., 2021).

2.5 Holt's Winter Method

The Holt's winter model is beneficial when the data contain seasonality and tread. This model handles it by computing the central value and then adding or multiplying it with the slope and seasonality (Priyadarshini et al., 2023). There are two variations of this procedure, and the components in each are different. The first variation is when the seasonal fluctuations are constant throughout the series, and the use of the additive method comes into play. When the seasonal fluctuations change proportionally to the level of the series, the multiplicative method is utilized for the trend (Hasri et al., 2021). In the additive method, the observed is scaled to express the seasonal component in absolute terms, the level equation adjusts the observed series for the season by deducting the seasonal component, and the sum of seasonal elements within each year will result in around 0. the component form is shown below.

$$\hat{X}_{t+h|t} = d_t + hb_t + C_{t+h-m(k+1)} \tag{2.14}$$

$$d_t = \alpha(x_t - C_{t-m}) + (1-\alpha)(d_t - 1 + b_t - 1) \tag{2.15}$$

$$b_t = \beta^*(bd_t - d_{t-1} + (1-\beta^*)b_{t-1} \tag{2.16}$$

k is the integral part of $(h-1)/m.(x_t - C_{t-m})$ are the seasonally adjusted values and $(d_t - 1 + b_t - 1)$ the non-seasonal forecast for time t. For the multiplicative method, the seasonal component in percentages and seasonally adjusting the series by dividing by its component (Gaur, 2020). The seasonal component will sum up to m, where m denotes the frequency of the seasonality. The figure 2.1 below demonstrates the methodology employed for tackling the research question.

2.6 Noteworthy Methods

The models in the above section are all interesting, especially Holt's winter method. This method is the first time I have heard used for time-series forecasting. I have heard of the common ARIMA method, SARIMA, and Fb-prophet in other research papers, but not Holt's linear or Holt's winter method. These two methods consistently perform well

on the various epidemic disease data fed into the model to make predictions. The tables shown below 2.2 for COVID-19, Moneypox, HIV, and Influenza give all the details of each model performance

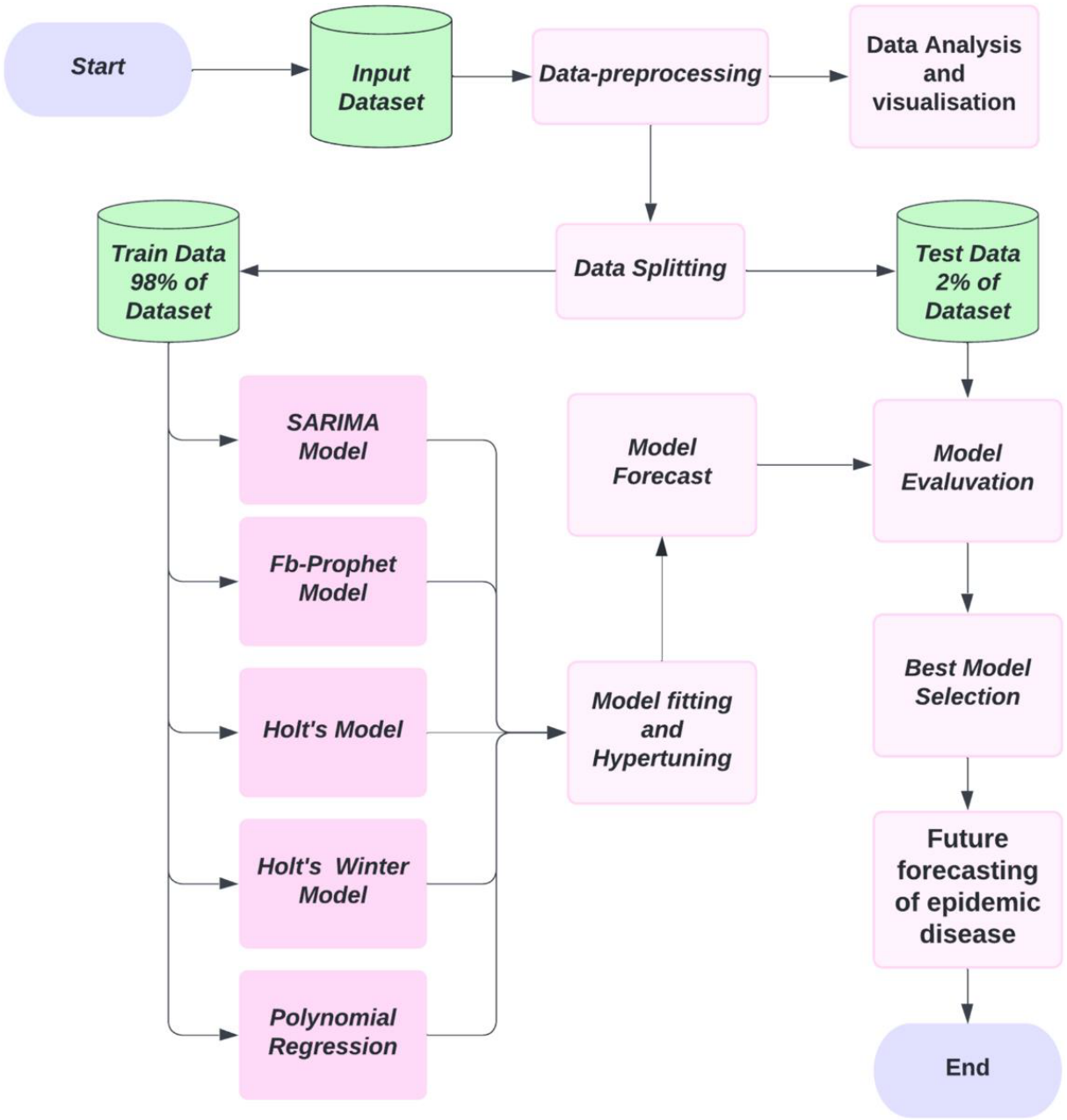

Figure 2.1: Proposed Framework for Epidemic diseases analysis and prediction.

Table 2. Error values generated for COVID-19 Dataset.

Model	RMSE	MSE	MAPE	R2 Score
FB-Prophet	1.652168e+06	2.729660e+12	0.002686	0.890732
ARIMA Model	2.195603e+06	4820674e+12	0.003334	0.807029
Holts Model	7.267927e+05	5.282276e+11	0.001146	0.978855
Holts Winter Model	7.164886e+05	5.133559e+11	0.001127	0.979450
Polynomial Regression	2.052482e+07	4.212683e+14	0.036630	−15.863304

Table 4. Error Values generated for Monkeypox.

Model	RMSE	MSE	MAPE	R2 Score
FB-Prophet	1794.839315	3.221448e+06	0.069247	−0.027517
ARIMA Model	708.323696	5.017225e+05	0.025078	0.839970
Holts Model	531.171254	2.821429e+05	0.017937	0.910007
Holts Winter Model	379.843170	1.442808e+05	0.012907	0.953980
Polynomial Regression	507.078716	2.571288e+05	0.016952	0.917986

Table 6. Error Values generated for HIV Dataset.

Model	RMSE	MSE	MAPE	R2 Score
FB-Prophet	14139.507407	1.999257e+08	0.148528	−0.256543
ARIMA Model	6791.600427	4.612584e+07	0.067477	0.710097
Holts Model	8245.154159	6.798257e+07	0.083077	0.572726
Holts Winter Model	2312.498370	5.347649e+06	0.024852	0.966390
Polynomial Regression	4475.562581	2.003066e+07	0.041066	0.874106

Figure 2.2: Forecasting Epidemic Diseases Results.

CHAPTER 3

Significance of the Work

After comparing the various algorithms for each disease. Holt's Winter model performed the best in all diseases. It is worth noting that polynomial regression did perform exceptionally well on Monkeypox, HIV, and Influenza diseases, where the degrees were equal to 3 and 7 concerning monkeypox and influenza. Therefore, Holt's winter and Polynomial regression models are well suited for forecasting diseases mentioned above with the least errors.

The results from the models are significant in the epidemiology context by providing a powerful prediction algorithm to forecast potential future diseases cases. From my recent dive into research on the topic of time-series analysis for COVID-19, models such as ARIMA and SARIMA are prevalent algorithms for making these forecasts models. We now know that Holt's winter model is a powerful tool for forecasting epidemics and should be exploited more often because having a better forecasting model will result in better preventative spread, early warning signals, help healthcare prepare, and allow countries to act readily and quickly to dampen the effects of the epidemic for lower-middle-income countries. Not only can these models be utilized in managing the spread of outbreaks, these models can also translate into other fields containing time-series elements. Econometrics is a great candidate for using some of the models mentioned to predict high-stakes problems for social, financial, environmental, and commercial needs.

CHAPTER 4

Connection to Other Work

When comparing the result section of this paper with other paper (A. Sulasikin et al. 2020) using similar methods, there are some inconsistencies to forecasting epidemic for covid-19. In this paper, Holt's linear and ARIMA model underperformed in all of the metrics compared to the Holt's winter, and polynomial regression. For the other paper, ARIMA model was the best model. The differences in the performance of ARIMA might be the type of time-frame, the train-test splits, and hypertunning. Nevertheless, both papers do support the use of these time-series model for future epidemic forecasting. Influential papers cited by the authors include the time series forecasting of COVID-19 (Chimmula et al. 2020), estimation of HIV incidence in the United States (Hall, H. et al. 2008), Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks (Kane, M. et al. 2014), introduction to modern time series analysis (KirchgÃ¤ssner, G et al. 2012).

CHAPTER 5

Relevance to Capstone

The content of this study is relevant to our capstone project because we are trying to understand and apply forecasting models to epidemic diseases. Learning various time-series models will be beneficial in helping us decide which models work well in different types of diseases setting. Knowing the best model for a particular disease forecasting is crucial when deciding where healthcare resources should focus and motivating the government to create policies to dampen the effects of these diseases. The methods of consideration for our capstone project are Holt's linear, Holt's winter method, ARIMA, and polynomial regression. These models seemed robust at predicting values. Plus, it would be worthwhile to see if these models claim what they do in forecasting. If these result metrics on various diseases were true, it would be substantial to have these used in our capstone project in terms of additional models. The potential areas where the study does not come to fruition are Holt's linear and Holt's winter methods, which correspond to the other papers' different results in comparison to this paper.

Bibliography

[1] A. Sulasikin, Y. Nugraha, J. Kanggrawan, and A. L. Suherman, Forecasting for a data-driven policy using time series methods in handling COVID-19 pandemic in Jakarta, *IEEE International Smart Cities Conference (ISC2)*, 2020.

[2] A. Al-Rashedi and M. A. Al-Hagery, Deep learning algorithms for forecasting COVID-19 cases in Saudi Arabia, *Applied Sciences*, vol. 13, no. 3, p. 1816, 2023.

[3] V. K. R. Chimmula and L. Zhang, Time series forecasting of COVID-19 transmission in Canada using LSTM networks, *Chaos, Solitons & Fractals*, vol. 135, p. 109864, 2020.

[4] D. B. Di Giulio and P. B. Eckburg, Human monkeypox: an emerging zoonosis, *The Lancet Infectious Diseases*, vol. 4, no. 1, pp. 15â25, 2004.

[5] A. Gupta, S. K. Singh, B. B. Gupta, et al., Evaluating the sustainable COVID-19 vaccination framework of India using recurrent neural networks, *Wireless Pers Commun*, 2023. DOI: 10.1007/s11277-023-10751-3.

[6] S. Gaur, Global forecasting of COVID-19 using ARIMA based FB-Prophet, *International Journal of Engineering Applied Sciences and Technology*, vol. 5, no. 2, pp. 463-467, 2020.

[7] H. I. Hall, R. Song, P. Rhodes, J. Prejean, Q. An, L. M. Lee, et al., Estimation of HIV incidence in the United States, *JAMA*, vol. 300, no. 5, pp. 520-529, 2008.

[8] H. Hasri, S. A. M. Aris, and R. Ahmad, Linear regression and Holtâs Winter algorithm in forecasting daily coronavirus disease 2019 cases in Malaysia: preliminary study, *2021 IEEE National Biomedical Engineering Conference (NBEC)*, pp. 157-160, 2021.

[9] S. M. Hyun, T. H. Hwang, and K. Lee, The prediction model for classification of COVID-19 infected patients using vital sign, *2021 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 678-681, 2021.

[10] G. KirchgÃ¤ssner, J. Wolters, and U. Hassler, *Introduction to Modern Time Series Analysis*, Springer Science & Business Media, 2012.

[11] K. Pathoee, D. Rawat, A. Mishra, V. Arya, M. K. Rafsanjani, and A. K. Gupta, A cloud-based predictive model for the detection of breast cancer, *International Journal of Cloud Applications and Computing (IJCAC)*, vol. 12, no. 1, pp. 1â12, 2022.

[12] I. Priyadarshini, P. Mohanty, R. Kumar, and D. Taniar, Monkeypox outbreak analysis: an extensive study using machine learning models and time series analysis, *Computers*, vol. 12, no. 2, p. 36, 2023.

[13] A. Sulasikin, Y. Nugraha, J. Kanggrawan, and A. L. Suherman, Forecasting for a data-driven policy using time series methods in handling COVID-19 pandemic in Jakarta, *2020 IEEE International Smart Cities Conference (ISC2)*, pp. 1-6, 2020.

[14] Y. Wang, M. Hu, Q. Li, X. P. Zhang, G. Zhai, and N. Yao, Abnormal respiratory patterns classifier may contribute to large-scale screening of people infected with COVID-19 in an accurate and unobtrusive manner, *ArXiv preprint arXiv:2002.05534*, 2020.

[15] W. Wei, G. Wang, X. Tao, Q. Luo, L. Chen, X. Bao, et al., Time series prediction for the epidemic trends of monkeypox using the ARIMA, exponential smoothing, GM (1, 1) and LSTM deep learning methods, *Journal of General Virology*, vol. 104, no. 4, p. 001839, 2023.

[16] Y. Yang, Y. Zhu, S. P. Tseng, L. Tang, Y. Chen, and X. Guo, Prediction and analysis of HIV/AIDS incidence based on ARIMA model in China, *2021 9th International Conference on Orange Technology (ICOT)*, pp. 1-4, 2021.