Development and Validation of Statistical and Deterministic Models Used to Predict Dengue Fever in Mexico

A thesis presented by

Aditi Hota

to the Applied Mathematics Department

in partial fulfillment of the honors requirements

for the degree of

Bachelor of Arts

Harvard College

Cambridge, Massachusetts

Adviser: Mauricio Santillana

April 1 2014

Abstract

Many epidemiological approaches have been proposed to forecast the incidence of infectious diseases such as influenza, malaria, or dengue fever. However, little has been done in the literature to thoroughly understand the accuracy of the forecasts produced by these approaches. Thus, in general the extent to which one model may be more effective than another at forecasting specific diseases is not clear. Additionally, further investigation is required to understand, whether gathering local information such as climate data or connectivity and human mobility across geographic regions will significantly improve predictions across models. In this thesis, two modeling approaches were developed, implemented, and validated to predict dengue fever incidence in Mexico. Namely, a set of statistical models and a dynamic deterministic model were designed and implemented to identify their abilities to forecast dengue. A comparative analysis was performed amongst the constructed statistical and deterministic models to understand their forecasting power. While findings may be expected to change across geographic locations at finer spatial scales, the results of this national analysis indicate that the autoregressive and deterministic approaches are capable of predicting the dengue incidence very well, and producing comparable forecasts for a one month time window. While extensions of the autoregressive models that included seasonality further improved model forecasts, it was found that adding climate covariates, such as temperature and precipitation, as predictors in our autoregressive models did not show consistent improvements. This suggests that further work needs to be pursued in order to understand the potential implications of climate changes in dengue fever incidence.

Acknowledgements

Looking back on the development of my thesis project truly allows me to appreciate the analytical and critical thinking skills that I have gained from the applied mathematics field over the past four years. This project would not have been possible without the guidance, expertise, and motivation from my mentor, Dr. Mauricio Santillana. He helped me develop this project alongside members from the HealthMap organization at Children's Hospital Boston, Dr. John Brownstein and Dr. Michael Johansson. I am grateful for all the ideas that we shared during this research process.

Additionally, I would like to thank the members of the Applied Mathematics department. The knowledge and scientific inquiry skills that I have learned from their classes during my undergraduate career has culminated in this project. Finally, thank you to all my friends and family for their support and encouragement. This project is dedicated to my undergraduate career and all of the people who have played a role in its growth.

Contents

Chapter 1

Introduction

1.1 Motivation

John Snow, the physician credited with the removal of a water pump on Broad Street in London, is widely recognized in the field of public health. During September of 1854, in the vicinity of Broad Street, 127 people had died from cholera. Within the next 10 days, the number of deaths had reached to an astounding 500. Snow took the initiative to gather information about the use of particular water pumps in the neighborhood and connected the deaths of individuals to the infamous pump on Broad Street; thus establishing a causal relationship between polluted water and cholera transmission. [30]

Using data analysis as a tool to track diseases has been around for over 100 years. The significance behind Snow's story relates to using data, the number of cholera incidents within the vicinity of the Soho district in London, to target and eliminate the spread of an infectious disease. Since 1854, we have come a long way with the development of sophisticated epidemiological models only to be met with the challenge of further mutations and development of the diseases we are trying to monitor.

In this research project, we aim to use quantitative analysis to understand the patterns of dengue fever in Mexico. Dengue, an illness transmitted by mosquitoes, is endemic to many South Asian and Latin American countries and affects millions of people every year. Using effective modeling approaches to monitor and forecast dengue trends would allow us to measure epidemics before they occur – giving public health officials time to prepare with the necessary supplies and task force. Multiple modeling techniques, such as statistical and deterministic approaches, are capable of capturing dengue trends efficiently. Certain models incorporate climate variables and the interaction of mosquito vectors to enhance their predictions. However, to what extent is one model more effective than the other? To address this question, it is necessary to construct, implement, and perform a comparative analysis of the predictive models.

In literature, there exists a plethora of dengue models applied to many geographic regions. Comparing all these models is a large task and outside the scope of this thesis; therefore, we wish to explore our question by specifically focusing on the country of Mexico. The goal of this project is to use Mexico as a case study to construct and evaluate the predictive abilities of two complementary epidemiological models for dengue fever; namely the deterministic and statistical approaches.

1.2 Biological Information about Dengue

Dengue fever is a common vector borne illness that afflicts approximately 50-100 million people per year and its incidence has grown by 30-fold over the past 50 years – making it one of the most important viral disease spread by mosquitoes. [31] Symptoms of this mosquito-borne infection may range from a severe flu-like illness to hemorrhagic fever, which could potentially lead to death. [31] The *Aedes aegypti* and *Aedes albopictus* species of the mosquito are known to be the principal vectors

responsible for spreading the dengue virus. [4] Dengue is found in 4 different strains (DENV 1, DENV 2, DENV 3, DENV 4) which are spread to humans through the mosquito vector. Additionally, an infected human has the potential to spread the disease to other locations if an uninfected mosquito bites the individual and becomes an infected vector. This infected vector can continue the cycle by biting other uninfected humans. Once a human has been infected by one specific strain, he or she is immune to future infections from this particular strain. However, repeated infections from different dengue virus strains can put individuals at risk of developing a severe case of dengue hemorrhagic fever (DHF). [4]

Dengue fever and DHF are commonly present in areas such as Latin America and Thailand, where the suitable environmental conditions allow for the growth and reproduction of the mosquito vectors. This includes many areas with high levels of precipitation, humidity, and temperature. [22, 23, 2] These climate factors influence the conditions that help to foster breeding grounds for vectors capable of spreading the dengue virus. [11] In addition to the climate effects on dengue incidences, geographic locations play a role in the number of dengue strains present, which can determine the severity of dengue in a particular location. For example, since the 1950s there have been dengue strains specific to areas such as the Caribbean Islands, Mexico, and South Asia. However, over time, these dengue strains are also expanding their reach by invading other geographical regions where they did not originally exist. [17]

Without a vaccine capable of preventing dengue, the only tactic used to prevent infections is to decrease the risk of mosquito bites. [4] It is essential to monitor the prevalence of dengue over time, especially in countries with a history of dengue epidemics since we lack an option to curtail future infections. Much research is being conducted to develop predictive models to monitor the spread of dengue fever in particular geographical areas. Multiple approaches can be taken to develop these

models; however, it is also important to evaluate and question whether the results are accurate and provide us with useful information.

1.3 Dataset Information

Obtaining an extensive set of reliable dengue cases is important for building and validating models. Therefore, to address the main research question of a comparative analysis of dengue prediction models, we used the best available set of dengue cases. Mexico's Ministry of Health provided monthly counts of dengue cases for various states from January of 1985 to December of 2011. Using this dataset, we assessed the quality of two different forecasting strategies: autoregressive deterministic modeling. The autoregressive models were additionally extended to include covariate data such as: temperature, precipitation, and relative humidity. Climate data from January of 1985 to December of 2011 were obtained from the National Oceanic and Atmospheric Administration (NOAA). Monthly temperature averages were measured in Celsius and obtained over a gridded area of Mexico. Precipitation, measured as average millimeters per day, and relative humidity, measured as a percentage, were obtained in the same manner as temperature. Finally, statistical models were built using R statistical software while deterministic models were built using MATLAB.

1.4 Summary of Model Analysis

In Chapter 2, we construct 32 different autoregressive models and their covariate extensions to study whether there is a significant improvement in their predictive ability. Generally, complex autoregressive models, which factor in seasonality, showed the greatest improvement in model prediction. Then to draw comparisons to the deterministic models, we developed a system of differential equations based on our own set of assumptions, found in Chapter 3. The result of the deterministic model is com-

parable to the simpler autoregressive models, indicating that further modifications of the deterministic model could lead to improved predictions. The autoregressive or statistical approach requires a few years worth of data to construct and train a model, where as the deterministic approach only requires the initial values to generate dynamic predictions for the immediate future. We will touch more upon the evaluation of the two methods in Chapter 4 of this work.

Chapter 2

Autoregressive Models

The current chapter presents the construction of 32 different simple and complex autoregressive models in addition to their covariate extensions. Section 2.1 provides background information about the applications of autoregressive models to predict dengue cases in different regions of the world. In Sections 2.2 - 2.4, we continue by describing components of an autoregressive model and the steps taken to construct them. Finally, in 2.5, we present the results of the regular autoregressive models and their covariate extensions.

2.1 Overview of Autoregressive Models

Autoregressive models are applicable to phenomena where future predictions are influenced by lagged terms of a particular variable of interest. Under the assumption that present time points are dependent on past time lags, autoregressive models have been used to study a wide range of topics from seismo-volcanic activity to meteorological patterns. [14, 16] Specifically, in the field of public health, incidence of dengue fever has been modeled using autoregression and its extensions. [19, 9, 27, 28, 21] Extensions of these models incorporate external meteorological covariates to improve the accuracy of dengue case predictions. [9]

This project uses the main assumption of autoregressive analysis to construct regular autoregressive models (AR), seasonal autoregressive models (SARIMA), and autoregressive models with climate covariates by utilizing reported monthly dengue cases in Mexico from January 1985 to December 2011. Overall, the SARIMA model has been found to generate the most accurate predictions in various countries since the additional seasonal component is able to capture the cyclic pattern of dengue cases across many years. [19, 27, 28] Although studies are able to construct the best fitting models for each country, there is a large discrepancy amongst these optimal models. There is no universal autoregressive model that can be applied to many countries or even smaller geographic regions within a country. For example, in both Thailand and Indonesia monthly dengue cases were used to develop SARIMA models for particular regions of the country; however, the best-fitting models changed along with the specific geographic regions within each country. [27, 28]

The latter models indicate that autoregression is sensitive to geographic locations, since we focus our efforts on predicting dengue cases at the national level, we do not explore the potential geographic heterogeneities in models that may arise at finer spatial scales. Therefore, we aim to compare the predictive abilities of different AR and SARIMA models and their extensions applied to national dengue cases in Mexico.

2.2 Components of Autoregressive Models

2.2.1 Lagged Terms

Autoregressive models assume that correlations exist between the value of a variable at a particular time point and lagged terms of the same variable. In the same manner used to construct linear regression models, the response variable is regressed against lagged terms of the variable itself. While constructing these models, it is important to methodologically establish whether or not there exists a relationship

between the explanatory and response variables by screening for the correlation of the variables against its lagged terms. In section 2.2.2, we illustrate how the autocorrelation function can be used to determine the number of lagged terms, denoted by the letter $\boldsymbol{p}$. The notation AR($\boldsymbol{p}$) denotes a p^{th} order autoregressive model consisting of p lagged terms.

$$y_t = \sum_{i=1}^{n} \beta_i y_{t-i} + \epsilon_t \tag{2.1}$$

In the equation above, y_t represents the variable at time t as a representation of the combination of i time lags given by $\beta_i y_{t-i}$ and an error term, ϵ_t. Data is used to solve the system of equations to determine each β_i coefficient using some version of the least squares approach. For example, if we are constructing an AR(2) model, the system of linear equations in matrix notation would be as follows:

$$\begin{bmatrix} y_{t-1} & y_{t-2} \\ y_{(t+1)-1} & y_{(t+1)-2} \\ \cdots & \cdots \\ y_{(t+n)-1} & y_{(t+n)-2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} y_t \\ y_{t+1} \\ \cdots \\ y_{n+1} \end{bmatrix}$$

The equation for this model would be:

$$y_t = \sum_{i=1}^{2} \beta_i y_{t-i} + \epsilon_t \tag{2.2}$$

Many statistical approaches such as maximum likelihood, method of least squares, and the Whittle estimation are used for estimating the parameters of time series models; we utilized the conditional-sum-of-squares (CSS) approach to determine the values for the coefficients of our autoregressive model terms. [25]

2.2.2 Moving Average

Autoregressive models can further incorporate a moving average term to describe lagged random errors as part of a time series model. Similar to an AR model, it is a linear regression model of the variable of interest against lagged white noise error terms. The number of lagged moving average terms is denoted by the letter $\boldsymbol{q}$. The notation MA($\boldsymbol{q}$) denotes q^{th} order moving average model consisting of q lagged error terms.

$$y_t = \sum_{i=1}^{n} \alpha_i \psi_{t-i} + \mu \tag{2.3}$$

In the equation above, y_t represents the variable at time t as a representation of the combination of i time lags of its random error given by $\alpha_i \psi_{t-i}$ and the expected mean, μ. Data is used to solve the system of equations to determine each α_i coefficient using some version of a least squares approach. For example, if we are constructing an MA(2) model, we would have the following equation:

$$y_t = \sum_{i=1}^{2} \alpha_i \psi_{t-i} + \mu \tag{2.4}$$

2.2.3 Differencing Term

Combining autoregression and moving average terms, we can build an autoregressive moving average (ARMA) model. This model is able to describe a stationary time series where the data is centered around a constant mean and variance. [7] However, with the case of modeling dengue fever, there are many fluctuations in the data over time – corresponding to epidemic and endemic periods. Dengue trends have seasonal effects since mosquitoes effectively reproduce under suitable conditions leading to an increase in their ability to spread the disease. [4] To account for the non-stationary nature of dengue cases, we use the autoregressive integrate moving average model

(ARIMA) by incorporating a differencing term which removes the trend in the time series and provides a stationary mean and variance. In the case of 1st order differencing, we have: $y'_t = y_t - y_{t-1}$. If we use D to represent the differencing factor, as a backshift operator then we have:

$$y'_t = y_t - y_{t-1} = y_t - Dy_t = (1 - D)y_t \tag{2.5}$$

This differencing term then allows us to define a particular variable based on a specified time lag. Utilizing the autoregressive, moving average, and differencing terms to build a model that captures the time series of lagged terms of a particular variable, we can develop the following ARIMA($\boldsymbol{p}$,$\boldsymbol{d}$,$\boldsymbol{q}$) model with autoregressive lags of p, difference terms d, and moving average lags of q. [3]

$$(1 - \phi_1 D - \phi_2 D^2 - ... - \phi_p D^p)(1 - D)^p y_t = c + (1 - \psi_1 D - \phi_2 D^2 - ... - \psi_q D^q)\epsilon_t \tag{2.6}$$

2.3 Extensions

2.3.1 Seasonality

Utilizing a seasonal component may help to capture the trends in our data effectively because the dengue season in Mexico starts around August of one calendar year and extends into January of the next calendar year. In constructing a SARIMA model, the seasonal autoregressive and moving average terms are multiples of the lagged terms according to the seasonality factor selected. In the case with monthly dengue counts, a seasonality factor of 12 for an AR(2) model would include time lags from x_{t-12} and x_{t-24}. A SARIMA model incorporates the original non-seasonal components defining the model, in addition to the new seasonal terms based on the periodicity selected.

2.3.2 Covariates

In many endemic areas, a constant seasonal dengue trend is disrupted by a larger unexpected epidemic during a particular year. SARIMA models depend only on dengue cases and may not provide the best estimate for years with abnormally high number of cases. For certain datasets, SARIMA models have been further improved by incorporating time lags that reflect the climate-disease transmission system. Although there has been a noticeable improvement in the predictive abilities, the results were not statistically significant than the initial SARIMA models without covariates. [18, 10, 19]

Covariates can also be included in the autoregressive models to enhance their predictive abilities. In general cases, other time series or seasonal trends such as climate or atmospheric data can be appended as a linear regression variable to construct predictions based on autoregression of the variable of interest and additional predictor variables. Specifically with diseases such as dengue or influenza, climate variables play a role because they influence the rate at which infections can spread. [29] One prime example of dengue modeling that incorporates seasonality and covariates is proposed by Gharbi et al. as a three month forecast that included lagged covariates of humidity and temperature to improve the accuracy for a surveillance system in Guadeloupe. [9]

2.4 Model Construction

2.4.1 Dividing the Dataset

To construct our models, we separated our original dataset of dengue cases from 1985 - 2011 into two separate groups: a training data set, consisting of dengue cases from January 1985 to December 1999 and a testing data set, consisting of dengue

cases from January 2000 to December 2011. All of the autoregressive models constructed used the same initial training period data set. This means that for any model constructed, the regression included dengue cases from 1985 to 1999 in addition to dengue cases up to the prior month of prediction. For example, if predicting for March 2003, the regression model built used data from January 1985 to February 2003. Thus, predictions were generated using dynamic models that incorporated the most current information to forecast into the future. Therefore, as the autoregressive model incorporates each subsequent value of dengue cases, it constructs a different autoregressive model to account for the new data point, but also maintains the information from the training period used to construct the model. Relative to a static model, using dynamic models will provide the most accurate models because they incorporate the most recent and accurate case counts released by the Mexican Ministry of Health. If static models were used, predictions well into the future, for example six months out, would rely on predictions made for the prior five months leading to an aggregation of errors for forecasts far into the future. For the remainder of the autoregressive models discussion, we use the notation AR($\boldsymbol{p}$,$\boldsymbol{d}$)S($\boldsymbol{p}$,$\boldsymbol{d}$) to refer to AR models with p lagged terms and d differencing terms and SARIMA models with p lagged seasonal terms and d differencing seasonal terms.

2.4.2 Autocorrelation Function

As stated earlier, our primary assumption for developing AR models is that we expect future counts of dengue cases to be related to observed counts from previous time lags. To identify the particular time lags, we apply the autocorrelation function(ACF) to the data. The ACF visually represents the coefficients of correlation between the time series variable and its lags. Default correlations out to 25 time steps were determined by using the acf function in R. Even though the ACF plot provides useful information about selecting significant time lags, they do not control for the

propagated correlations that accumulate in higher order lags. The partial autocorrelation function (PACF) corrects for this to generate partial correlations between the variables and its lags. At particular lags, it calculates the difference between the actual correlation at that lag and the expected correlation due to the propagation of correlations at that lag. [26]

Figure 2.1, below, shows the ACF of the national dengue cases from January 1985 to December 1999 in Mexico. The single peaks are the correlation coefficients and the blue dashed lines indicate the threshold for significance. Correlation values below this line are not considered to be statistically significant. The dampened sinusoidal pattern of the correlation values suggests including a seasonality component to effectively capture the trends in the data.

The PACF plot, Figure 2.2, eliminates the mutually correlated terms and indicates statistically significant peaks at lags 1,2,10,11,12, and 13. The extension of the first two lags into lags 10-13 indicates that future dengue cases have a relationship with case counts approximately a year earlier. To account for this, the best approach to modeling our data is to start off with 1-2 autoregressive terms and extend it to a seasonal model.

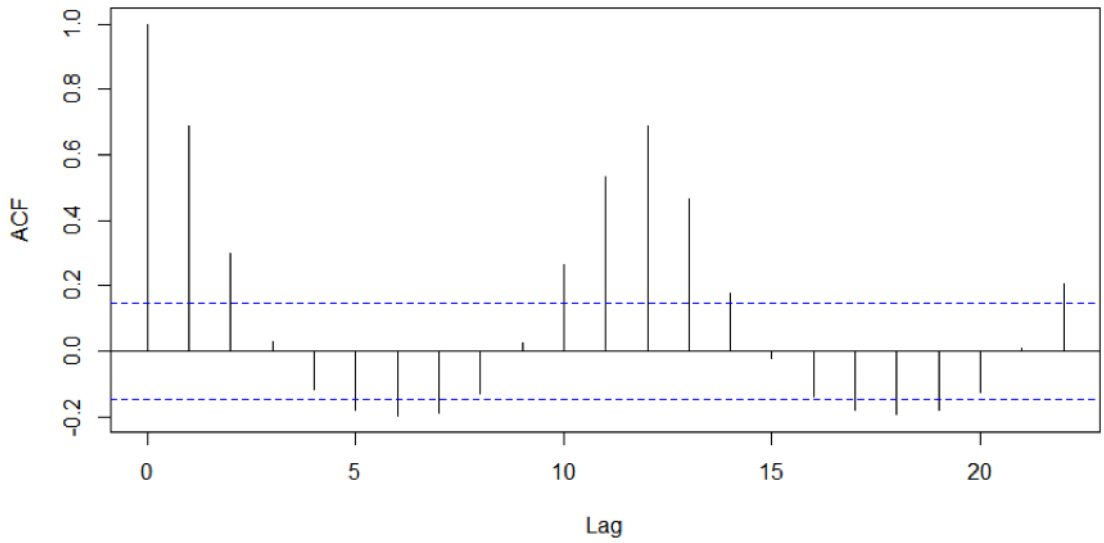

Figure 2.1: Autocorrelation Function Output for Dengue Cases 1985-1999

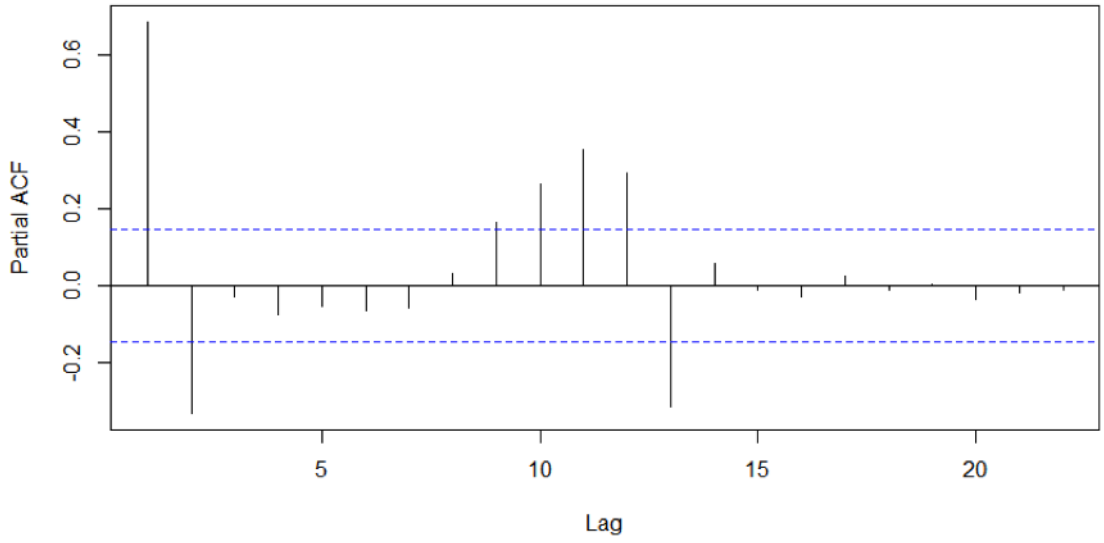

Figure 2.2: Partial Autocorrelation Function Output for Dengue Cases 1985-1999

The results of the ACF and PACF plots indicate that autoregressive terms of 1 or 2 lags will be suitable for analysis of the data as there is a strong correlation between

time steps of 1 and 2 lags. Although our initial analysis suggests that 1-2 lags will be the most suitable for the data set, we explored the effects of these lags out to four for the non-seasonal terms and out to 6 lags for the seasonal terms. Additionally, we decided to explore the effects of adding a 1 order of seasonal differencing assuming that this may remove any trends in the time series. In all, we considered 32 models ranging from simple autoregressive models with 1 lagged term to seasonal autoregressive models with 6 lagged terms.

2.4.3 Model Evaluation

To determine the best AR and SARIMA models, root mean square error (RMSE) and correlations (COR) of the predictions from 2001 to 2011 were recorded for each model. All predictions were compared to the recorded data obtained from the Mexican Ministry of Health. For assessment, the best models will have a low RMSE value and high COR values indicating that prediction errors are minimized and follow the general trend of observed dengue cases over time, respectively.

2.5 Results for AR and SARIMA Models

2.5.1 Shorter Lagged Terms

We first fitted an autoregressive model with 1 lag term to determine the prediction of dengue cases from January 2000 to December 2011. Figure 2.3 below displays the results of the predictions 1 month, 3 months, and 6 months into the future. It is noticeable that the more short-term predictions produce accurate results as indicated by higher correlation values of 0.85, for the 1-month predictions, versus 0.04 for the 6-month predictions and lower root mean square values of 2843 for the 1 month prediction, versus 5185 for the 6 month prediction. This indicates that the autoregressive models produce better immediate predictions for the succeeding month

rather than predictions of six months into the future. Figure 2.3 indicates the decline in accuracy with the blue line representing the 1 month predictions, yellow line as the 3 month predictions, and the red line as the 6 month predictions. All of these results can be compared to the observed values noted by the black line.

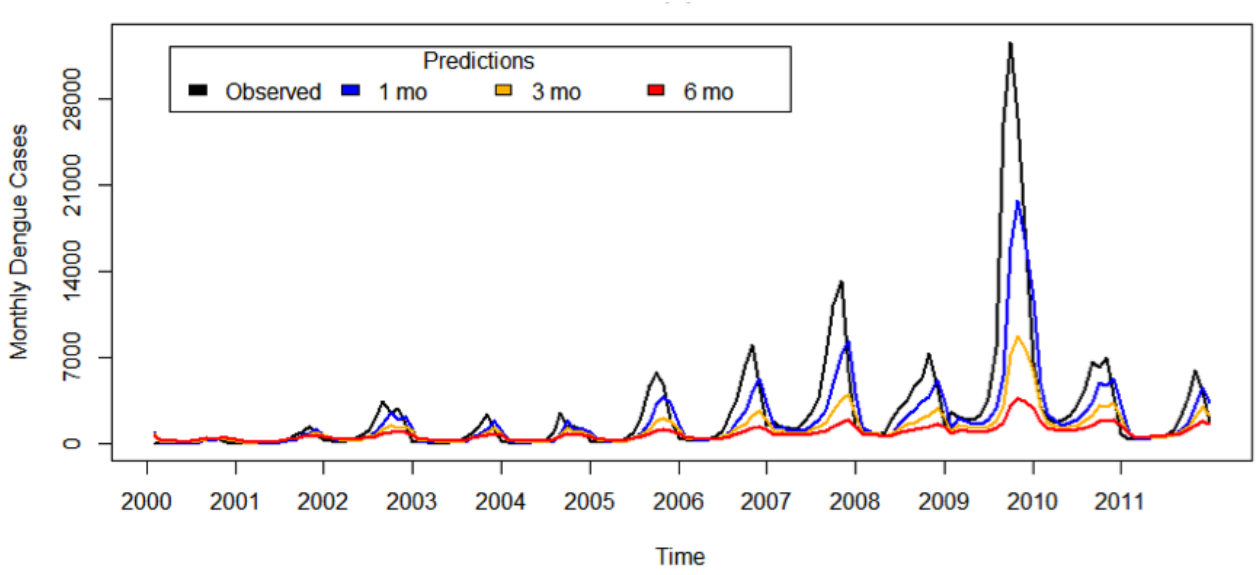

Figure 2.3: Monthly Predictions for AR(1) Model from Jan 2000 to Dec 2011

To improve the initial AR(1) and AR(2) models, we implemented seasonal autoregressive and seasonal integrated autoregressive models. This extension improved the initial models as the RMSE values decreased relative to the AR(1) and AR(2) models and there was an increase in the correlation values. Extending the regular autoregressive models to include a seasonal component helps to improve our initial AR(1) model, which is represented in the decrease of the RMSE values and an increase of the COR values of the seasonal models. Specifically, the AR(1)S(1,1) and AR(1)S(2,1) produced results with the lowest RMSE values along with AR(1)S(2) and AR(1)S(2,1) producing the highest correlation values. Table 2.1 below presents the results of the RMSE and COR values for the initial set of AR and SARIMA mod-

els. Each row indicates the number of AR, SAR, and differencing terms and whether all coefficients in the model were statistically significant at the 0.05 level.

Terms					RMSE			COR		
	AR	SAR	Diff.	Sig.	1 Mo.	3 Mo.	6 Mo.	1 Mo.	3 Mo.	6 Mo.
AR(1)	1	0	0	Yes	2843	4797	5185	0.85	0.31	0.04
AR(2)	2	0	0	No	2609	4971	5280	0.92	0.49	0.18
AR(1)S(1)	1	1	0	Yes	2357	4112	4467	0.90	0.60	0.50
AR(2)S(1)	2	1	0	No	2349	4134	4496	0.90	0.60	0.50
AR(1)S(1,1)	1	1	1	No	1852	3338	3668	0.92	0.72	0.65
AR(2)S(1,1)	2	1	1	No	1965	3277	3567	0.91	0.73	0.67
AR(1)S(2)	1	2	0	Yes	1976	3666	4055	0.93	0.73	0.65
AR(2)S(2)	2	2	0	Yes	2007	3533	3882	0.92	0.74	0.67
AR(1)S(2,1)	1	2	1	No	1828	3215	3494	0.93	0.75	0.69
AR(2)S(2,1)	2	2	1	No	1938	3148	3353	0.91	0.76	0.72

Table 2.1: Values of RMSE and COR for 1 month, 3 month, and 6 month predictions for shorter lagged autoregressive models

Figures 2.4 and 2.5 compare the RMSE and COR values of the top 3 models along with the basic AR(1) model. There is a clear improvement in the RMSE values for the 1 month predictions as it decreases from an error of 2843 for the AR(1) model to an error of 1852 for the AR(1)S(1,1) model. Likewise, the correlation values increase from 0.85 to 0.92 for the latter models. The black line represents the AR(1) model, blue line is the AR(1)S(1,1) model, yellow line is the AR(1)S(2) model, and the red line is the AR(1)S(2,1). The results also indicate that there is a significant improvement from the regular AR models to the SARIMA models with a decrease of approximately 1000 dengue cases for the RMSE values and an increase of approximately 0.07 for the COR values. However, amongst the SARIMA models, there is a small discrepancy in

the RMSE values within the range of 50-100 cases and the range of 0.01 for the COR values.

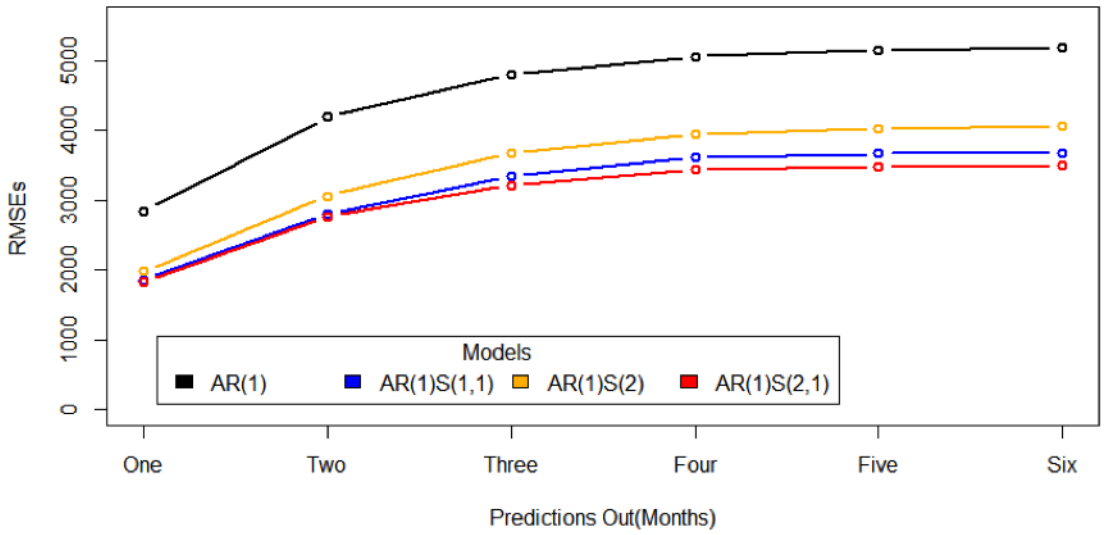

Figure 2.4: Comparison of RMSE values from Jan 2000 to Dec 2011 for Autoregressive Models

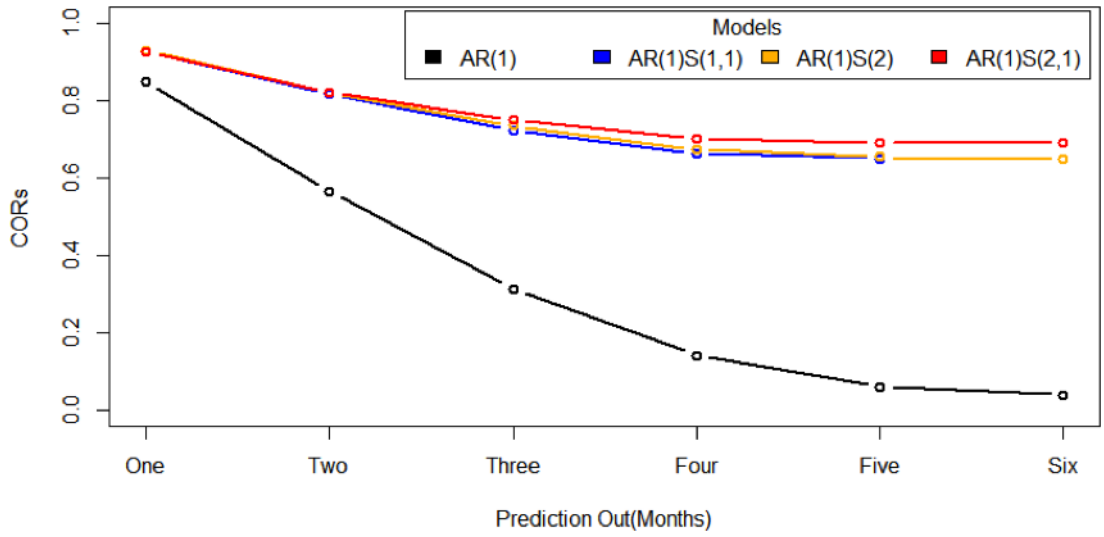

Figure 2.5: Comparison of Correlation from Jan 2000 to Dec 2011 for Autoregressive Models

In Figure 2.6 below, we are visually able to represent the comparisons of the AR and SARIMA models to indicate the increase in the accuracy of the predictive ability of the SARIMA models. The red line, representing the AR(1)S(2,1) model, captures the trend of the observed dengue values (in black) rather closely compared to the other AR(1)S(1,1) model represented by the yellow line and the AR(1) model represented by the blue line.

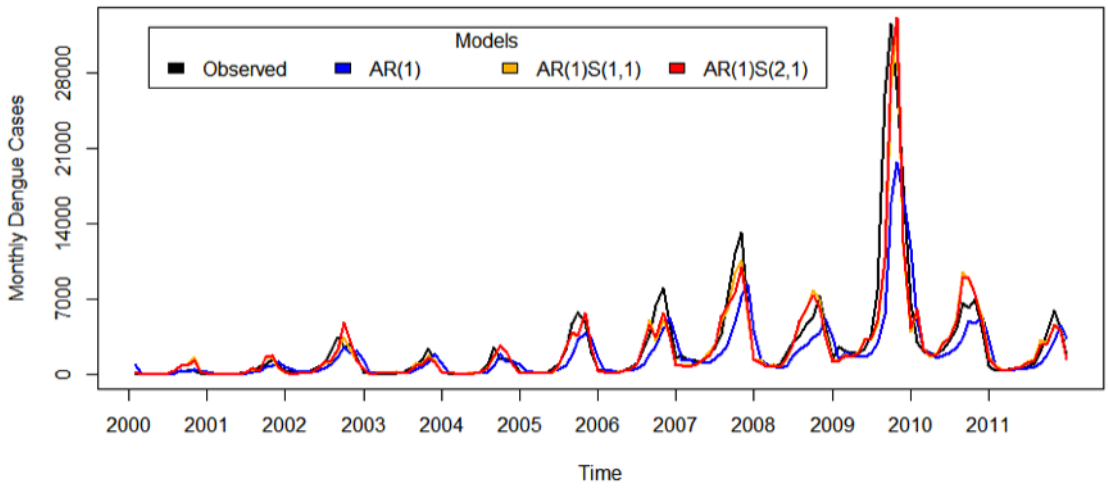

Figure 2.6: Comparison of Top Autoregressive Models to AR(1)

2.5.2 Covariate Extension

We further explored whether the addition of climate explanatory variables such as temperature, precipitation, and relative humidity could improve the original autoregressive models – under the expectation that climate variables influence the mosquito population, which in turn can influence the number of dengue cases. For constructing these models, the optimal lags of the covariates were determined by selecting the covariate lags that corresponded to the highest correlation against the dengue cases from 1985 to 1999. The optimal lags are as follows: temperature of lag 4, precipitation of lag 2, and humidity of lag 6. Although these optimal lags have the highest correlations, we also tested for models with 1 and 2 monthly lags under the assumption that more recent climate data could have more influence on the future predictions of dengue. Overall, the inclusion of climate variables does not significantly improve the original AR and SARIMA models. In Tables 2.2-2.4, we are specifically comparing the original AR(1) and the top two models - AR(1)S(1,1) and AR(1)S(2,1)

to their covariate counterparts Climate AR(1), Climate AR(1)S(1,1), and Climate AR(1)S(2,1).

The RMSE values of the climate AR(1) models have decreased approximately 200-400 cases across all the 3 covariates. This suggests an improvement in the predictive abilities of the climate AR(1) model. Likewise, the correlation values for the lagged climate variables have improved for all models except for the relative humidity at lag 6. Table 2.2 below indicates that the temperature and precipitation models contained all significant terms, as they are highlighted in bold.

Covariate AR(1)								
	Temp (4)							
		RMSE	2138	3734	4635	5110		
		COR	0.93	0.69	0.38	0.08		
	Precip (2)							
		RMSE	2657	4000				
		COR	0.84	0.58				
	Rel Hum (6)							
		RMSE	2460	3799	4731	5149	5306	5379
		COR	0.86	0.64	0.32	0.08	-0.06	-0.15

Table 2.2: Comparison of RMSE and COR Values for AR(1) Covariate Models with Temperature, Precipitation, and Relative Humidity

When comparing the RMSE and COR values from the regular AR(1)S(1,1) model to its covariate model, there is no improvement in the predictive ability. Comparing Table 2.1 to Table 2.2 shows that RMSE values have increased and correlation is consistent across the models. This indicates that for complex SARIMA models, adding the climate covariates does not improve the predictive ability of the models. Although comparing the RMSEs of the climate AR(1)S(1,1) to the regular AR(1)

model, there is a decrease of approximately 900 cases. This may suggest improved predictability; however, it is not clear if it is in relation to the addition of seasonal lags or climate variables. These same results also extend to the comparison between the regular AR(1)S(2,1) model and its climate model, please see Tables 2.3 and 2.4.

Covariate AR(1)S(1,1)								
	Temp (4)							
		RMSE	1860	2820	3368	3651		
		COR	0.92	0.81	0.72	0.66		
	Precip (2)							
		RMSE	1954	2849				
		COR	0.91	0.81				
	Rel Hum (6)							
		RMSE	1946	2858	3409	3673	3751	3801
		COR	0.91	0.81	0.72	0.65	0.63	0.62

Table 2.3: Comparison of RMSE and COR Values for AR(1)S(1,1) Covariate Models with Temperature, Precipitation, and Relative Humidity

Covariate AR(1) S(2,1)								
	Temp (4)							
		RMSE	1859	2876	3379	3714		
		COR	0.92	0.80	0.74	0.68		
	Precip (2)							
		RMSE	1916	2787				
		COR	0.92	0.82				
	Rel Hum (6)							
		RMSE	1895	2796	3277	3520	3589	3646
		COR	0.92	0.82	0.742	0.70	0.68	0.67

Table 2.4: Comparison of RMSE and COR Values for AR(1)S(2,1) Covariate Models with Temperature, Precipitation, and Relative Humidity

Although climate variables have been found to improve certain autoregressive models in literature, our results indicate that the improvement was only found in the simplest AR(1) model. [9] Therefore, adding climate variables to the best models AR(1)S(1,1) and AR(1)S(2,1) from the set of initial autoregressive models does not indicate any improvement suggesting that utilizing only the seasonality factor may be sufficient to capture a significant portion of the dengue trends.

2.5.3 Longer Seasonal Lagged Terms

Building autoregressive models involves more exploration than utilizing only the number of lagged terms determined from the ACF and PACF plots. In this project, we decided to extend the seasonal lags out to 6 terms. Doing so produced results with significant model improvement where RMSE values decreased by approximately 100 cases while the correlations or predictions further out (3 months and 6 months) improved by approximately 0.1. Table 2.5 represents the results of the longer sea-

sonal lagged models. Although the extended lags look promising, the only concern is evaluating the simplicity of the model balanced with its improvement in predictive abilities. Since this project has the availability dengue data from 1985, it may be worth considering models with longer seasonal lags because it may be better able to use the plethora of data to construct a better fitting model. In Table 2.5, Each row indicates the number of AR,SAR, and differencing terms and whether all coefficients in the model were statistically significant at the 0.05 level.

Terms					RMSE			COR		
	AR	SAR	Diff.	Sig.	1 Mo.	3 Mo.	6 Mo.	1 Mo.	3 Mo.	6 Mo.
AR(1)S(4)	1	4	0	No	1742.30	3307.77	3525.25	0.94	0.76	0.72
AR(1)S(5)	1	5	0	No	1739.00	2735.52	3284.44	0.94	0.76	0.71
AR(1)S(6)	1	6	0	No	1722.89	3238.05	3456.61	0.94	0.76	0.71
AR(1)S(3,1)	1	3	1	No	1711.35	3190.20	3307.04	0.94	0.75	0.73
AR(1)S(4,1)	1	4	1	No	1709.64	3195.13	3358.21	0.94	0.75	0.72

Table 2.5: Values of RMSE and COR for 1 month, 3 month, and 6 month predictions for longer lagged autoregressive models

2.5.4 Model Evaluations

The results from this analysis indicate that the seasonal autoregressive models are able to capture the trends of the dengue cases reasonably well. The graphical representation of the predictions in Figure 2.6 indicate how closely the AR(1)S(2,1) model follows the trend of the observed dengue cases from 2000 to 2011. When peak dengue cases are on the scale of 7000 to 32000 an error of 1700-1800 is reasonable to guide public health officials with an idea of the expected number of dengue cases during a particular season. Likewise, all one month prediction correlations are above 0.9 with the exception of the AR(1) model. This indicates that there is a strong

relationship between the comparisons of predictions to their corresponding observed values. Additionally, the autoregressive models applied to Mexican dengue data help to qualify the assumption that dengue cases in the future are dependent of lagged cases in the past. As we saw that modification of this assumption to include seasonal terms helps to improve its predictive ability and adding climate covariates do not seem to improve the predictive ability in any substantial way.

Furthermore, we applied these models to selected states in Mexico and discovered that there is an inconsistency in the best models across the examined states. It is not surprising that adding climate variables to our national autoregressive model did not improve the models because it is difficult to accurately apply the effects of climate to a country as a whole when there is a wide range of climates specific to different geographical regions across the country. We believed that applying climate variables to specific states may show improvement in predictive abilities since we have localized the climate factors and dengue case counts. Surprisingly, the results indicated that there was no significant improvement in the model indicating that the dengue trends are sufficiently captured with the autoregression and seasonal factors.

Chapter 3

Deterministic Modeling

In the previous chapter, we presented the applications of autoregressive models to forecast dengue cases in Mexico. In the following sections we discuss the applications an alternative approach – namely the deterministic model to forecast dengue. First, we explain the elements of the SIR model, a common deterministic model applied to infectious diseases. Then, we modify elements of this general model to build assumptions applicable to the dengue transmission cycle. Finally, we present results indicating that the deterministic model is comparable to simpler autoregressive models. This suggests that deterministic models have the same predictive power as autoregressive models, without the need of a training period.

3.1 Background to Deterministic Models

The Kermack-McKendrick Model, also known as the SIR model, was originally used as a means to monitor and determine the magnitude of the incidences of infectious diseases. Since its origin in 1927, it has been applied to model various infectious diseases including dengue, influenza, malaria, and varicella. [13, 20, 5] The original SIR model is based on a fixed population group divided into three smaller subgroups: susceptible, infected, and recovered. To determine the number of individuals in these

three subgroups at a particular time point, it is necessary to solve a system of differential equations that models the rates of change for the number of individuals in each group. Figure 3.1 below is a visual representation of the transmission cycle used by the Kermack-McKendrick model. Individuals are able to move forward from the susceptible to the infected stage once they have become infected with the disease. Likewise, infected individuals can only move forward to the recovered stage once they are cured from the disease. A recovered individual is assumed to be fully immune to the disease and permanently remains in this group.

Figure 3.1: Representation of the Stages of a Simple SIR Model

Under these assumptions, the rates of change for individuals within each subgroup at a particular time point can be represented with the following set of equations:

$$\frac{dS}{dT} = -\beta S(t)I(t) \tag{3.1}$$

$$\frac{dI}{dT} = \beta S(t)I(t) - \gamma I(t) \tag{3.2}$$

$$\frac{dR}{dT} = \gamma I(t) \tag{3.3}$$

In these sets of equations, we have a fixed population N which at every time point is equivalent to $S(t) + I(t) + R(t)$. $S(t)$ represents the number susceptible at time t, $I(t)$ represents the number infected at time t, and $R(t)$ represents the number recovered at time t. β is the rate of transmission or infection of the disease. It can be regarded as the average number of transmissions from an infected person during

a particular time period. Therefore, its units would be in terms of per people-time. γ represents the rate of recovery for an infected individual with units of per time. [15] Since there is no simple analytic solution to the system of Equations 3.1-3.2, numerical methods need to be implemented to obtain an approximate solution. For this specific project, we use the ode45 solver in MATLAB to determine our dengue case estimates.

This basic model does not take into account realistic events such as dynamic changes in the population in the form of births and deaths or the chance that death may be an additional stage resulting from infection. Under alternative assumptions, modifications to this model have extended to include vector-to-host transmission cycles, which require another set of differential equations describing the susceptible, infected, and recovered populations for the individuals in the vector population. [13, 1]

3.1.1 Applications to Dengue

As mentioned in the Introduction, dengue fever consists of 4 different strains. Infection from one dengue strain, for instance DENV-1, does not result in immunity to the other 3 dengue strains. Therefore individuals who have not been exposed to the DENV-2, DENV-3, or DENV-4 strains will be classified as susceptible. Additionally, repeated infections place humans at a higher risk of contracting the dengue hemorrhagic fever and resulting in death. [4] Complexities such as the ones mentioned above can be amended to the simple SIR model by incorporating additional stages of susceptibility and infection. Figure 3.2 is an example of a multistage SIR models that can be applied to the dengue cycle. A review by Johansson et al. provides an exhaustive set of approaches taken to model the dengue transmission cycle. The advanced models are able to incorporate the human-vector transmission cycle by modeling the additional mosquito vector population. [13]

Figure 3.2: Representation of the Stages of an Advanced SIR Model

One of the major challenges with developing deterministic models is calculating accurate parameter estimations. With the case of dengue, we need to define the values for β, the rate of infection, and for γ, the rate of recovery. Without appropriate tools to measure these quantities, many models use parameterization to obtain a range of reasonable values for these terms. [1] For the purposes of this thesis, we used the ranges of parameters from Andraud to build our predictive models. Values of β ranged from 0.9 to 22.5 $\frac{\text{people}}{\text{time}}$ and values of γ ranged from 2.14 to 10 $\frac{1}{\text{month}}$.

3.2 Deterministic Model Construction

By imposing additional assumptions to the original Kermack-McKendrick model, we were able to construct the following system of equations to represent the change over time for the three subgroups of susceptible, infected, and recovered populations. As with the original model, $S(t)$ represents the number susceptible at time t, $I(t)$ are the number infected at time t, and $R(t)$ are the number recovered at time t. The parameters β and γ represent the rate of infection and the rate of recovery, respectively. Section 3.2.1 provides a list of assumptions for this particular model.

$$\frac{dS}{dT} = \frac{-\beta S(t) I(t)}{N(t)} \tag{3.4}$$

$$\frac{dI}{dT} = \frac{\beta S(t)I(t)}{N(t)} - \gamma I(t) \tag{3.5}$$

$$\frac{dR}{dT} = \gamma I(t) \tag{3.6}$$

3.2.1 Assumptions

1. Change in population dynamics can be implemented directly into the SIR model by adding an additional parameter to account for the fluctuations. In equation 3.7, μ is added to the original susceptible rate to account for the birth in the population. Likewise to handle the death in a population a constant population death rate term, α, can be added to all three groups in the population since death can occur during any stage of the the dengue transmission cycle. [12]

$$\frac{dS}{dT} = \frac{-\beta S(t)I(t)}{N(t)} + \mu - \alpha \tag{3.7}$$

$$\frac{dI}{dT} = \frac{\beta S(t)I(t)}{N(t)} - \gamma I(t) - \alpha \tag{3.8}$$

$$\frac{dR}{dT} = \gamma I(t) - \alpha \tag{3.9}$$

Instead of adding an additional parameter, which would require further estimation, we were able to account for the change in population by mimicking the population growth in Mexico. Population trends in Mexico from 1985 to 2011, provided by the World Bank, indicated that Mexico's population growth has an upward quasi-linear trend. [33] Using this as the framework, we constructed a linear model with an intercept equal to the 5 times the maximum number of individuals infected in 1986 and a slope of $\frac{2042}{12}$. The population growth rate was determined by constructing a linear regression model with population data from 1986 to 1999, as shown in equation 3.10. The idea behind utilizing the linear

growth model to calculate the total population at a particular time removed the obstacle of parameterizing the unknown μ and α rates.

$$N(t) = 31000 + \frac{2042}{12} \times \text{Months Elapsed from January 1986} \tag{3.10}$$

2. To develop and compare the optimal estimate of the β and γ parameters, we needed to make the units of the quantities comparable to the ones in published literature. [1] This can be modified by dividing the original $-\beta S(t)I(t)$ by the total population $N(t)$. Therefore, the units of β are per time and there is no need to adjust for the units of γ.

3. The model does not account for the 4 strains of dengue because we did not have access to incidence data to validate our approach. Instead, we aggregated infections from all dengue strains as one instance of a dengue infection. [8] Since dengue serotypes range from one state to the next in Mexico, and individuals rarely develop immunity to all 4 strains, we chose to let all recovered individuals become susceptible at any given point in time. [8] Computationally, every month we determined the initial number susceptible $S(t)$ by subtracting the number of infected from the total population: $S(t) = N(t) - I(t)$. In other words, we solved for the number of recovered individuals, $R(t)$, within the month, but we always reset it to 0 at the beginning of a new prediction period. All calculations, were performed using the built-in Matlab command: ode45.

4. Finally, we use a host-to-host transmission cycle to best predict dengue on a national scale. For example, if an individual from one region of a country was to travel to another region and become bitten by an uninfected mosquito, then this vector will become infected and is able to spread the virus to other uninfected humans. In a larger population it is difficult to control for the movement of infected and uninfected individuals; therefore to simplify this phenomena, we

assume that there is host-to-host transmission. If we were to model a local region within Mexico, it is reasonable to integrate a vector-to-host transmission cycle because it may be easier to control for the movement of the infected individuals by adding an additional removal rate for these individuals to our deterministic equations. However, when incorporating any vector interaction, we need to implement the subgroups of the vector population, giving rise to the obstacle of estimating more parameter values.

3.2.2 Model Forecasting

Similar to autoregressive models, the deterministic approach also uses a dynamic model by incorporating the most recent information from the latest counts of dengue cases. The key parallel between the autoregressive and deterministic models is that the number of dengue cases at a particular time point for the autoregressive models is equivalent to the number of infected individuals at a particular time point in the deterministic model. One major difference of the deterministic model is that it does not require a training period for model construction. Therefore, in order to compare the two approaches, we chose to analyze their performance only for the years 2000 to 2011.

As stated before, one of the major obstacles for constructing a deterministic model, using differential equations, is to accurately determine parameters that best describe the rate of infection, β and the rate of recovery, γ. The best way to approach this problem was to first construct a model where the units of β and γ are comparable to the ones in literature and to optimize with respect to the parameter values for each predictive time step. Optimization can be done in multiple fashions; however, we decided that minimizing the RMSE of the predicted and the observed cases per time step with respect to the β and γ parameters would be the best route. In short, we chose to solve this problem as a non-linear least squares problem.

To better understand how the predictions from this model are constructed, the algorithm below walks through the steps to determine the predicted number of dengue cases for March 1986.

1. Set the initial conditions of I_0 equal to the number of dengue cases from February 1986 and N_0 equal to 5 times the maximum number of infected cases in 1986 plus $\frac{2042}{12}$ times the number of months elapsed from January of 1986, in this case: 2. With these two initial conditions, we can set $R_0 = 0$ and $S_0 = N_0 - I_0$

2. Provide the initial parameter values for β and γ. When assigning the initial parameters for March of 1986, we use the optimal values of β and γ that were obtained from minimizing the RMSE between the predicted and observed value from February 1986. Note: When starting the model initially, we do not have any information to set the parameter values for the January 1986 prediction so we set the values to be 10 for both β and γ.

3. Using the ode45 solver in MATLAB, we determine the predicted value (number of infected individuals) for March 1986 by entering the initial conditions from step 1 and parameter values from step 2 to run the deterministic model.

4. The fmincon function in MATLAB was used to minimize the RMSE between the predicted value of dengue cases in March 1986 and the recorded number of dengue cases from March 1986. With optimization, it is necessary to provide an upper and lower bound on the β and γ parameters to get reasonable values. These bounds were determined from the range of parameter values by Andraud 2012.

5. We expect that the optimal value from the RMSE should be as close to 0 as possible since we are optimizing at every single time point. Therefore, we relaxed the upper bound on the γ parameter if the optimal RMSE value was

above 2. During this process, we recorded the optimal values of both β and γ and noticed that in many instances, the γ values tend to reach its upper bound. Relaxing the upper limit on gamma by 10 units helped to account for this problem.

6. The optimal β and γ values determined from the fmincon search for the March 1986 prediction will be stored as the initial parameter values for the generating the prediction for April 1986. This process continues for each subsequent monthly prediction.

3.3 Model Evaluations

The deterministic model is a well-defined approach taken to model dengue cases in Mexico. This preliminary model with crude assumptions is comparable to the simple autoregressive models. Although parameterization proves to be a challenge, the benefits of the deterministic model is that we do not require a large data set to train it; rather, we are able to construct our model from the first set of initial values and predict for every subsequent time step. Further enhancement of the deterministic model to include the host-to-vector interaction or multiple stages of susceptibility may enhance its predictive ability.

Figure 3.3 below presents a comparison of the predicted values, in red, from the deterministic model to the observed dengue cases, in black, from 2000 to 2011. Overall, the dynamic deterministic model is able to capture the trend of the dengue cases over time. The RMSE for the evaluation period is 2750 cases and the COR is 0.84. This is comparable to the simple AR(1) model constructed earlier with a RMSE value of 2843 and a COR value of 0.85 along with the AR(2) model with an RMSE value of 2609 and a COR value 0.92.

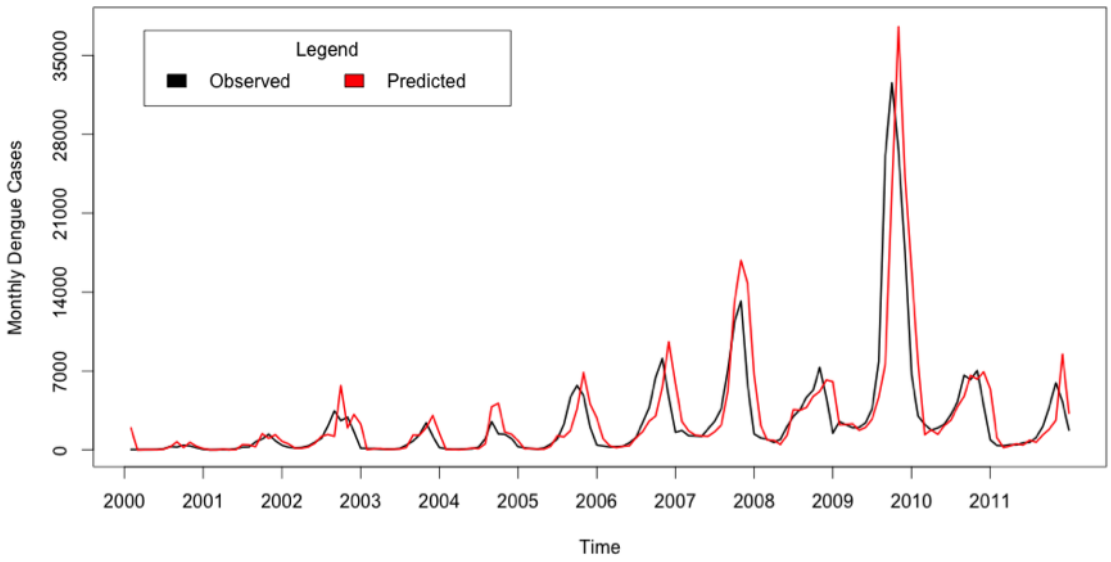

Figure 3.3: Prediction of Dengue Cases from Jan 2000 to Dec 2011 for Deterministic Model

Without the use of any seasonality factor or climate covariates in our deterministic approach, we were able to construct models that are comparable to the simple autoregressive models. However, extending the results from the deterministic model for comparison to the climate and seasonal models may not be strictly fair. Future work could involve the expansion of the simple deterministic model to incorporate seasonality, climate covariates, or interactions with the vector population to further reduce the RMSE.

Chapter 4

Conclusion

In this thesis, we showed that the statistical and deterministic approaches can be used to predict monthly dengue cases in Mexico from January 2000 to December 2011. For the autoregressive modeling, a training data set from 1985 to 1999 was used to develop predictive models for the years 2000 to 2011. However, for the deterministic model, we only required information from the previous time step, December 1999, to determine our predictions from 2000 to 2011. With these two approaches, we have shown that their RMSE values are within the same order of magnitude and comparable to one another. In this section, we will briefly summarize the significance of this project and provide a thorough comparative analysis of the two modeling approaches.

4.1 Significance

Dengue is a widespread virus that puts approximately more than one-third of the world's population at risk for infection. [4] Although there have been many campaigns to curtail the spread of dengue in regions across Asia, Brazil, and Mexico, it is difficult to determine the years when an epidemic will occur. [6, 24, 32] However, if we were able to forecast years when high volumes of dengue cases were expected, it would alleviate

much of the concern for public health officials. Prior knowledge would allow them to allocate their resources efficiently and plan interventions accordingly.

Effective modeling approaches are helpful for monitoring and forecasting dengue trends. Currently, there are numerous models from the statistical and deterministic approaches used to predict future dengue cases; however, there is not a clear answer to the relative effectiveness of the models to one another. A thorough understanding of the effectiveness of predictive models is more involved since it would require analyzing the multiple models across many geographic locations where dengue is prevalent. This project aims to explore the main question by focusing on the country of Mexico and comparing the applications of the autoregressive to deterministic based models there.

4.2 Comparative Analysis

Chapter 2 indicates that AR and SARIMA models are plausible methods that can be used to forecast dengue fever in Mexico. After using a training period data set from 1985-1999 we were able to construct a dynamic model to make predictions from 2000 - 2011. We understand that immediate predictions, such as 1 month predictions, are more accurate than 6 month predictions due to the aggregation of errors as we predict further into the future. Utilizing seasonal components in our models reduced the error of dengue cases by approximately 1000 cases compared to the simple AR(1) and AR(2) models. When we added climate covariates, we expected to further improve our estimates; however, we were surprised that there was no significant improvement. This may be due to a significant portion of the dengue trends being captured by the seasonality of the models. When we further pushed our exploration to use longer lags – out to 6 terms – the models only improved with a decrease of approximately 100 cases for the RMSE, relative to the shorter lag SARIMA models. This may indicate that for the purposes of obtaining forecasts using only 1-2 lag terms may be sufficient.

For our deterministic model, we only required the previous time point from December 1999 to start making our predictions from January 2000 onwards. One advantage to the deterministic model is that we did not require an extensive data set to train the model; however, we are faced with the challenge of estimating parameters. Figure 4.1 below presents a visual comparison of the predictions from the deterministic approach, in blue, as well as the top two SARIMA models - AR(1)S(1,1), in yellow, and AR(1)S(2,1), in red.

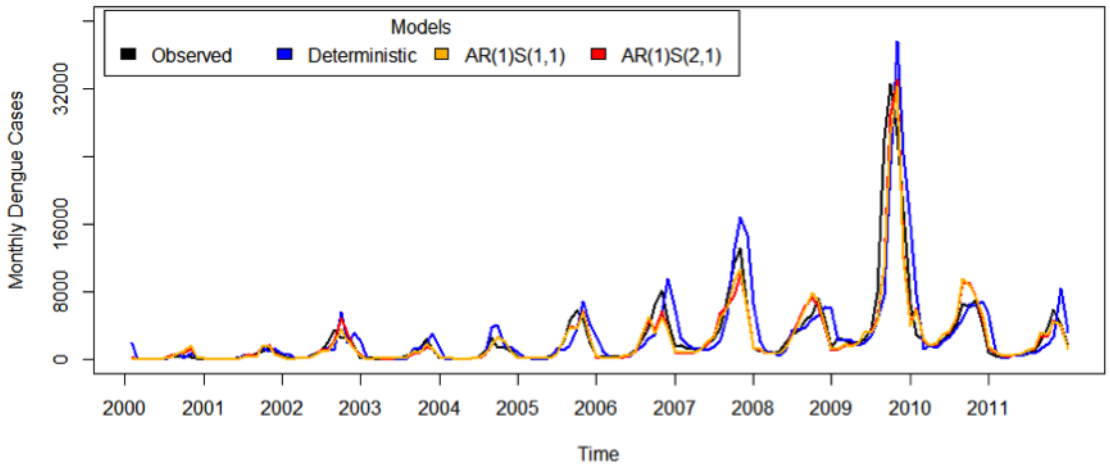

Figure 4.1: Comparison of Prediction for Dengue Cases from Jan 2000 - Dec 2011 for Deterministic, AR(1)S(1,1) and AR(1)S(2,1) Models

All three models indicate that they are able to capture the peaks of the dengue epidemics pretty well; however, the deterministic approach tends to overestimate these peaks more, relative to the SARIMA models. In Section 4.3 below, we discuss the further improvements that can be extended from this project.

4.3 Future Work

To fully understand and evaluate predictive dengue models for comparison, it is necessary to do a thorough analysis by building and comparing the predictive abilities of various models across multiple geographic locations. In general literature, autoregressive and deterministic modeling are two common approaches taken to model infectious disease, and this work was able to construct preliminary models to apply to dengue incidence in Mexico. This research project was only able to cover a portion of the analysis required to answer the question of the effectiveness of models. It may be worthwhile exploring the climate autoregressive models by modeling specific geographic regions rather than states within Mexico. States are a convenient way to select particular locations within Mexico; however there still exists a geographical variability within each state since it is just an arbitrary political boundary. Since dengue may not be found in remote areas of Mexico where the population is scarce, it could be insightful to focus these geographic specific models on urban areas where dengue has a higher prevalence.

The deterministic approach has the potential to be further improved by considering stricter assumptions with the dengue transmission cycle. As mentioned previously, forming stricter assumptions requires further parameter estimations for the model. Although this may be resolved by experimentally estimating the parameters, it may be an extensive process to determine these values. One improvement may be to include multiple stages of infected and susceptible populations or add the interaction of the mosquito vector population. Additionally, accounting for connectivity and mobility within the larger population could enhance the accuracy of these models.

This thesis project has been able to provide a portion of the analysis for comparing the predictive abilities of models used to forecast dengue fever in Mexico. The results suggest that deterministic and statistical models are capable of predicting dengue fever which can be helpful to public health officials. Understanding the

plethora of models constructed and comparing their effectiveness is essential to us as mathematicians to consider whether the models we are constructing are improving our prior knowledge or comparable to the knowledge that we already have. By considering these types of analyses can we seek to further improve the predictive abilities of epidemiological models.

Bibliography

[1] Mathieu Andraud, Niel Hens, Christiaan Marais, and Philippe Beutels. Dynamic epidemiological models for dengue transmission: a systematic review of structural approaches. *PloS one*, 7(11):e49085, January 2012.

[2] E Barclay. Is climate change affecting dengue in the Americas? *Lancet Review*, 371(9617):973–974, 2007.

[3] P. Brockwell and R. Davis. ARIMA Models for Nonstationary Time Series. In *Introduction to Time Series and Forecasting, Second Edition*, pages 180–193. Springer, 2002.

[4] CDC. Dengue Entomology & Ecology, 2010.

[5] Brian J Coburn, Bradley G Wagner, and Sally Blower. Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1). *BMC medicine*, 7:30, January 2009.

[6] Clàudio Csillag. Brazil Launches Dengue-Eradication Campaign. *Lancet*, 349(9051):1997, 1997.

[7] R Dahyot. Time Series and Applied Forecasting.

[8] Francisco J Díaz, William C Black, José a Farfán-Ale, María a Loroño Pino, Kenneth E Olson, and Barry J Beaty. Dengue virus circulation and evolution in Mexico: a phylogenetic perspective. *Archives of medical research*, 37(6):760–73, August 2006.

[9] Myriam Gharbi, Philippe Quenel, Joël Gustave, Sylvie Cassadou, Guy La Ruche, Laurent Girdary, and Laurence Marrama. Time series analysis of dengue incidence in Guadeloupe, French West Indies: forecasting models using climate variables as predictors. *BMC infectious diseases*, 11(1):166, January 2011.

[10] Kensuke Goto, Balachandran Kumarendran, Sachith Mettananda, Deepa Gunasekara, Yoshito Fujii, and Satoshi Kaneko. Analysis of effects of meteorological factors on dengue incidence in Sri Lanka using time series data. *PloS one*, 8(5):e63717, January 2013.

[11] Simon Hales, Neil de Wet, John Maindonald, and Alistair Woodward. Potential effect of population and climate changes on global distribution of dengue fever: an empirical model. *Lancet*, 360(9336):830–4, September 2002.

[12] H. Hethcote, L. Gross, T.G. Hallam, and S.A. Levin. Three Basic Epidemiological Models. In *Applied Mathematical Ecology*, pages 119–144. Springer-Verlag, 1989.

[13] Teri Johnson and Barry McQuarrie. Mathematical Modeling of Diseases: Susceptible-Infected-Recovered (SIR) Model. Technical report, University of Minnesota, Morris, Math 4901 Senior Seminar, 2009.

[14] R. Katz and R. Skaggs. On the Use of Autoregressive-Moving Average Processes to Model Meteorological Time Series. *American Meteorological Society*, 109(3):479–484, 1981.

[15] W. O. Kermack and a. G. McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721, August 1927.

[16] Philippe Lesage, François Glangeaud, and Jérôme Mars. Applications of autoregressive models and time–frequency analysis to the study of volcanic tremor and long-period events. *Journal of Volcanology and Geothermal Research*, 114(3-4):391–417, May 2002.

[17] M a Loroño Pino, C B Cropp, J a Farfán, a V Vorndam, E M Rodríguez-Angulo, E P Rosado-Paredes, L F Flores-Flores, B J Beaty, and D J Gubler. Common occurrence of concurrent infections by multiple dengue virus serotypes. *The American journal of tropical medicine and hygiene*, 61(5):725–30, November 1999.

[18] Rachel Lowe, Trevor C Bailey, David B Stephenson, Tim E Jupp, Richard J Graham, Christovam Barcellos, and Marilia Sá Carvalho. The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in Southeast Brazil. *Statistics in medicine*, 32(5):864–83, February 2013.

[19] Paula M Luz, Beatriz V M Mendes, Claudia T Codeço, Claudio J Struchiner, and Alison P Galvani. Time series analysis of dengue incidence in Rio de Janeiro, Brazil. *The American journal of tropical medicine and hygiene*, 79(6):933–9, December 2008.

[20] Sandip Mandal, Ram Rup Sarkar, and Somdatta Sinha. Mathematical models of malaria–a review. *Malaria journal*, 10(1):202, January 2011.

[21] E. Z Martinez, E. A. S. da Silva, and A. L. D. Fabbro. A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil. *Revista Da Sociedade Brasileira de Medicina Tropical*, 44(4):436–440, 2011.

[22] Kanchana Nakhapakorn and Nitin Kumar Tripathi. An information value based analysis of physical and climatic factors affecting dengue fever and dengue haemorrhagic fever incidence. *International Journal of Health Geographics*, 4(1):14, 2005.

[23] S Promprou, M Jaroensutasinee, and K Jaroensutasinee. Climatic Factors Affecting Dengue Haemorrhagic Fever Incidence in Southern Thailand. *Dengue Bulletin*, 29:41–48, 2005.

[24] RedCross. Mexico : Dengue Outbreak, 2009.

[25] P M Robinson. Conditional-sum-of-squares estimation of models for stationary time series with long memory, 2006.

[26] E Root. Lecture 3: Time Series Analysis: Stochastic Processes [Powerpoint slides], 2010.

[27] T Silawan, P Singhasivanon, J Kaewkungwal, S Nimmanitya, and W. Suwonkerd. Temporal Patterns and Forecast of Dengue Infection in Northeastern Thailand. *Southeast Asian Journal of Tropical Medicine and Public Health*, 39(1):90–98, 2008.

[28] MS Sitepu, J Kaewkungwal, N Luplerdlop, N Soonthornworasiri, T Silawan, and S Poungsombat. Temporal Patterns and a Disease Forecasting Model of Dengue Hemorrhagic Fever in Jakarta Based on 10 Years of Surveillance Data. *Southeast Asian Journal of Tropical Medicine and Public Health*, 44(2):206–217, 2013.

[29] RP Soebiyanto and RK Kiang. Modeling Influenza Transmission Using Environmental Parameters. *Southeast Asian Journal of Tropical Medicine and Public Health*, 38(8):330–334, 2010.

[30] Judith Summers. Broad Street Pump Outbreak, 1989.

[31] WHO. Dengue Control.

[32] WHO. Action Against Dengue, 2011.

[33] World Bank. Population (Total), 2013.