Public Health Analysis: An Assessment of Polynomial Distributed Lag in Mosquito-Born Disease Risk Modeling

Jessica Conrad*

Tulane University, New Orleans, LA 70118

April 24, 2018

Abstract

Using data for Ceará, Brazil, we construct a polynomial distributed lag model under different truncation lag criteria to predict reported dengue cases. Accurately predicting dengue cases provides the framework to develop forecasting models, which would provide public health professionals time to create targeted interventions for areas at high risk of dengue outbreaks. Others have shown that variables of interest such as temperature and vegetation can be used to predict dengue cases. These models did not detail how truncation lag criteria was chosen for their respective models when polynomial distributed lag was used. We explore current truncation lag selection methods used widely in the literature (simple, marginal, and minimized AIC) and determine which of these methods works best for our given dataset. While minimized AIC truncation lag selection produced the best fit to our data (Adjusted R^2=0.9996), this method used substantially more data to inform its prediction and resulted in a 21.46% increase in the Adjusted R^2 compared to the marginal truncation lag selection method (Adjusted R^2=0.7298). Finally, the following variables were found to be significant predictors of dengue in this region: mean normalized difference vegetation index, percent cloudy pixels, and Google Health Trends data.

Key Words: Polynomial distributed lag models, dengue, Brazil, mosquito born diseases, delayed effect, satellite indices

Running Title: Analysis of delayed effect of weather on dengue in Ceará, Brazil

*Corresponding author. E-mail address: jconrad4@tulane.edu

Contents

1 Background

Dengue cases are on the rise in Ceará, Brazil. Creating disease control programs to help prevent dengue are dependent on our understanding of the mechanisms involved with dengue transmission, and the region at risk.

1.1 Dengue in Ceará, Brazil

Mosquito born diseases are a major burden to public health, accounting for more than 17% of all infectious disease cases globally.[30] Approximately 40% of the world is at risk for dengue infection, with that burden even higher in endemic regions of Africa, Asia, and the Americas.[10, 14] Dengue virus in particular is a leading cause of death in tropical zones, such as Brazil.

Dengue virus originated in Africa or Southeast Asia and was geographically restricted until the mid-20th century. Cargo shipments during and after World War II are suspected to be the cause of the global spread of *Aedes* mosquitos around the world, as well as the diseases they carry.[6, 14] In 1967, *Aedes aegypti*, the main vector for dengue, was introduced to Brazil, and Brazil responded quickly by launching *Aedes* mosquito control programs. Despite these efforts, the mosquito spread rapidly across Brazil. By 1998, over half a million dengue cases were reported annually in Brazil and more than 1.5 million cases annually are reported today. [7, 13, 14]

The majority of all dengue cases in Brazil are reported in the Southeast and Northeast regions.[6] Ceará is a state in the Northeast region of Brazil. Dengue was introduced to Ceará in the mid 1980s[6, 13], with the first major outbreak reported in 1986. Since this time, there have been annual outbreaks of dengue in Ceará. In 1994, Ceará alone was responsible for 84% of all reported dengue cases, with the majority of these cases being reported in the city of Fortaleza.[7] Compared to dengue epidemics in the 1980s where fewer than a hundred cases a year were reported in Ceará, we are now seeing no less than a thousand cases or more each year. [13] Figure 1 shows the reported case data by state for the Northeast region of Brazil, in which Ceará is included, by epidemic week, from January 3, 2010 to the week of July 3, 2016. We can observe that Ceará often has one of the highest disease burdens of all states in this region in any given year, with an average reported case count of a over 1,000 per year.

Figure 2 shows the reported case data by mesoregion for Ceará for the same time span of epidemic weeks. We can observe from the raw data that the majority of the disease burden is in the mesoregion cluster Metropolitana de Fortaleza, or the City of Fortaleza for any given year, indicating a high disease burden in urban areas and heterogeneity of disease burden in this region.

As the number of dengue cases in Ceará and across Brazil continues to rise, it is important to create new innovative tools for informing disease control. We will discuss prediction methods which can inform disease control specialists on future cases of dengue.

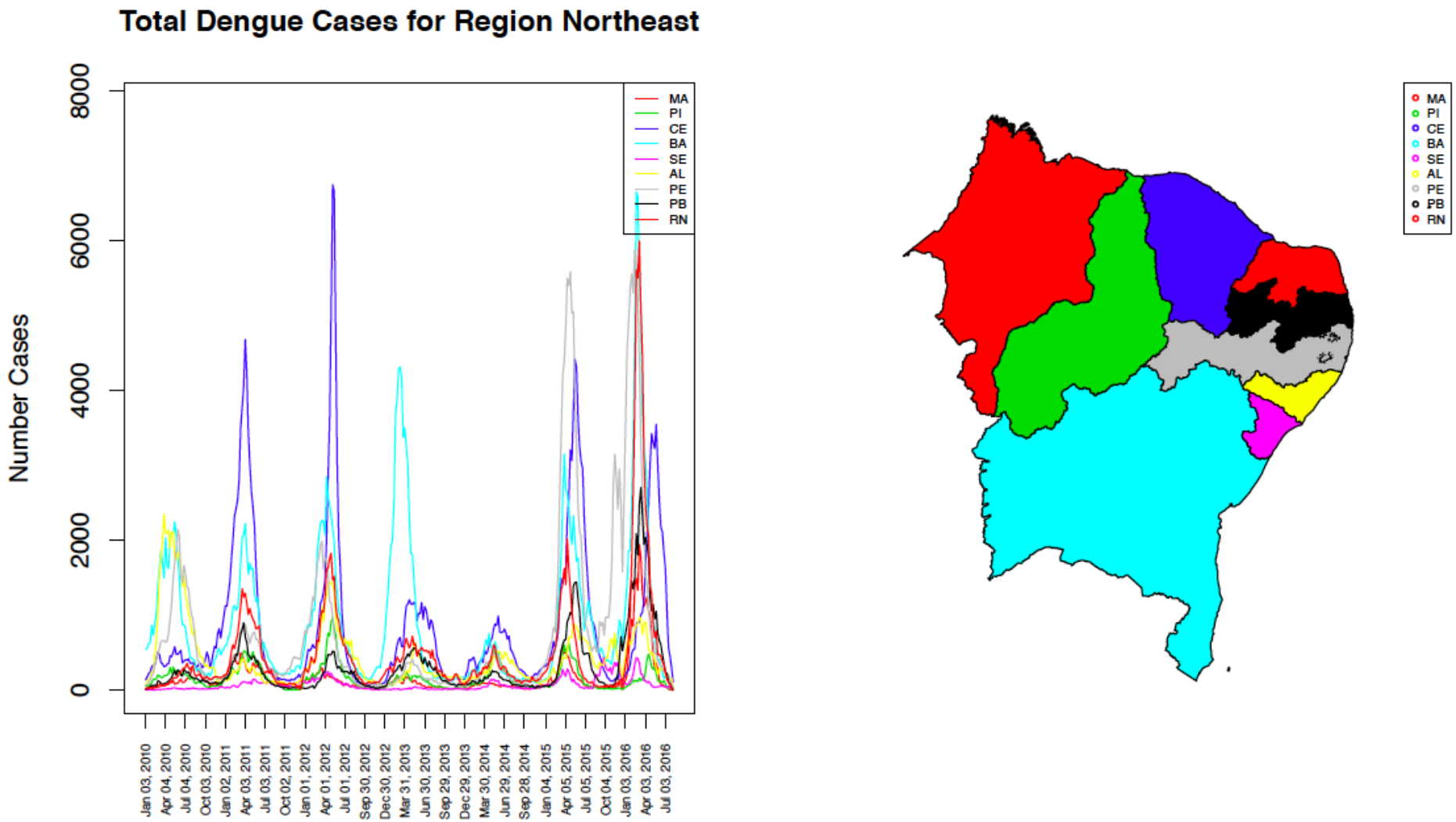

Figure 1: **Left:** The raw weekly case data for Ceará, Brazil, (in royal blue) compared to other states in the Northeast region of Brazil for January 2010 through July 2016. **Right:** Map of the Northeast region of Brazil, color code for each state matching the raw case data shown at left. The raw reported case data was provided by the Ministério da Saúde, also known as the Ministry of Health of Brazil.

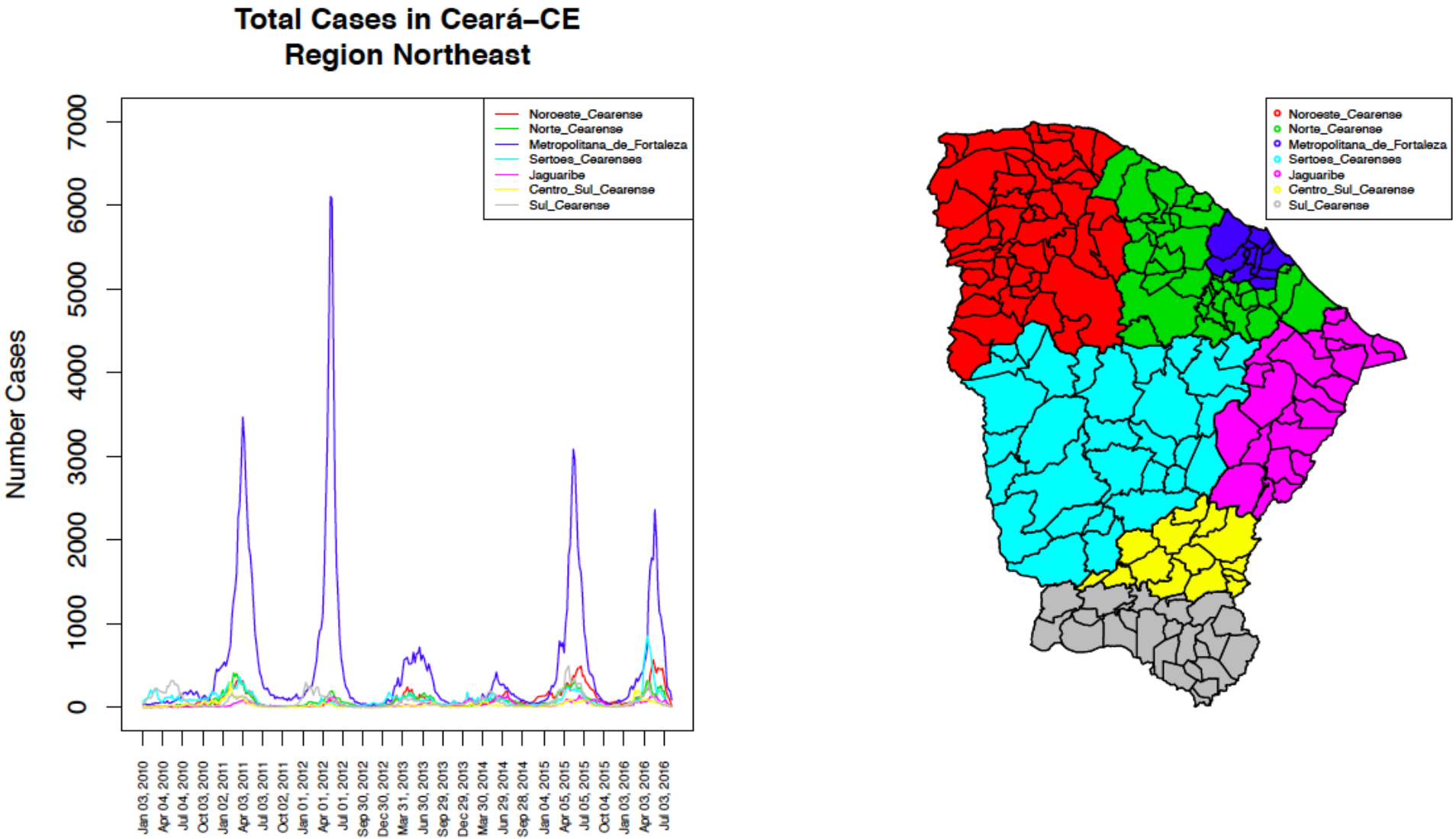

Figure 2: **Left:** The raw weekly case data for Ceará, Brazil, by mesoregion clusters of municipalities for January 2010 through July 2016 (7 geopolitical clusters of the 44 municipalities). Mesoregions are geopolitical clusters of municipalities, grouped by proximity and common characteristics by the Instituto Brasileiro de Geografia e Estatística in 1999 for statistical analysis purposes.[28] The raw reported case data was provided by the Ministry of Health of Brazil. **Right:** Map of Ceará, Brazil, color coded for each mesoregion matching the raw case data shown at left.

1.2 Transmission Dynamics of Dengue

In order to create targeted, successful intervention campaigns for dengue in Brazil, it is important to understand the underlying dynamics of the disease. In this section, we will review what dengue is, the transmission pathway of dengue, the life cycle of *A. aegypti*, and available intervention measures.

Dengue virus is transmitted via the bite of a mosquito to a human. Mosquitos of the *Aedes* family are the main vector for dengue, with *A. aegypti* responsible for the majority of all dengue transmission.[6, 10, 14] Typically, symptoms of dengue fever develop 5 days after infection [10, 14] and can last up to 10 days. During this time, there is a 5-day period where the viral load is high enough for humans to transmit dengue virus to a mosquito via a bite. Most cases are asymptomatic [14, 31], but symptoms include undifferentiated febrile illness (viral syndrome), dengue fever (DF), dengue hemorrhagic fever (DHF), or dengue shock syndrome (DSS). It is important to note that dengue infections are most likely higher than reported, with an estimated 15% reporting rate in 1999,[22] which may be due in part to the large number of asymptomatic cases of dengue driving a silent epidemic.[14, 31] The global burden of dengue is far worse than is currently assumed.

Vaccines are in development for dengue, but as of yet do not have widespread use. This is due in large part to issues arising from interactions between dengue serotypes. There are four major serotypes of denuge (DENV-1, DENV-2, DENV-3, and DENV-4), all of which are found in Brazil.[32] Certain combinations of these serotypes can increase or decrease risk of dengue hemorrhagic fever,[35] complicating treatment and development of vaccines, which must take this into consideration.

Transmission of dengue by mosquitos generally follow this pathway [10, 14]:

- An infected mosquito bites an uninfected (susceptible) human
- The virus multiplies in the blood over the course of 4 days
- A susceptible mosquito bites the now infected human (5-12 days since first infection of human)
- The pathogen develops in the gut of the mosquito over 8-12 days (extrinsic incubation period)
- The mosquito is now infected and the cycle repeats

Infected mosquitos can continue to transmit dengue virus for the duration of their lifetime, usually 3-4 weeks. This cycle assumes that there are adult mosquitos to participate in the transmission pathway. Therefore, the disease burden on humans is correlated to the abundance of adult mosquitos.

Abundance of *A. aegypti* depends on their breeding habits: mosquitos lay their eggs into containers with water, where they can remain for months, and eggs hatch after a rain or flooding. Larvae develop into pupa within a week, and into adult mosquitos within another two days. This process takes 8-10 days at room temperature.[14] From Figures 1 and 2, we can see there is a peak

in dengue cases by April in Brazil of each year, though the magnitude of the peak varies year to year. This is a general seasonal trend across Brazil, which corresponds to the rainy season and warmer spring temperatures.[6, 14, 24]

Upon understanding the life cycle of the mosquito, we can identify multiple influencing factors that effect the development of the mosquito, therefore abundance of dengue cases. Influencing factors include the following:

- **Urban Environment:** For the eggs to be laid, there need to be suitable containers for breeding. The vector itself prefers urban environments for breeding, which also enhances the spread of the virus.[6, 10, 14] Urban environments offer close contact with a blood source (humans) and easy breeding grounds. This tendency lends to the geographic heterogeneity that can be observed in the reported dengue cases. As seen in Figure 2, the mesoregion containing Fortaleza, the major city of Ceará, reported the most dengue cases in any given year.

- **Water Sources/Rainfall:** Eggs hatch after there has been rain or flooding, and therefore are sensitive to events involving water. The amount of rainfall is a known indicator of dengue prevalence.[14, 29]

- **Temperature:** The developmental and reproductive cycles of *A. aegypti* are sensitive to temperature variation, which is what results to the overall seasonal variance of dengue that is observed on a macro level. [6, 14] Higher temperatures reduce the time it takes for dengue virus to develop in the gut of the mosquito, increasing the infectious period of the mosquito.[14]

- **Vegetation:** Male and female mosquitos feed on plant nectars, fruit juices, and other plants sugars as their main energy source, making vegetation another factor important to their survival. [10] Abundance of vegetation depends on factors such as rainfall, temperature, and humidity.

In the absence of a reliable vaccine, public health professionals have focused on vector control measures to reduce dengue incidence. Vector control measures take advantage of the mosquito life cycle. For example, larval fish target the larval stage of mosquito development, while insecticide sprays target adult mosquitos. Education on mosquito breeding grounds have also been used to try to reduce mosquito abundance in urban areas. [14] But such measures require time and expensive resources to implement. Figures 1 and 2 show that the disease burden of dengue in Brazil is highly heterogeneous geographically and changes year to year. To reduce costs and increase effectiveness, control programs should be more targeted to specific areas of high risk.

2 Introduction

Upon understanding some of the factors of the life cycle of *A. aegypti* outlined in Section 1.2, we can see that mosquito abundance is dependent on multiple factors: access to standing water, temperature, urbanization, vegetation, and rainfall. All of these variables effect the development of the mosquito from egg to larvae to adult. There is a time delay from weather factors to increased mosquito abundance, and therefore from weather factors to increased risk of infection from mosquitos.

We will use polynomial distributed lag models to forecast risk of dengue cases in Ceará, Brazil, using the following predictor variables described in Section 2.1.

2.1 Variables with Delayed Effect on Mosquitos

Since we are focusing on dengue transmission in Brazil, we will consider the life cycle of *A. aegypti* specifically. This mosquito thrives in tropical and subtropical regions where the winter is no colder than 10°C. This is due to the fact that their reproductive cycle is highly dependent on temperature, as previously mentioned in Section 1.2. [6, 10, 14] This and other factors were outlined in detail in Section 1.2. We want to look at data that describes these effects as predictors for dengue caseloads.

Satellite indices data were provided by the Descartes Labs, and a description of this data in Table 1. For NDVI, Green NDWI, SWIR NDWI, and NBR, the maximum, mean, and minimum value collected each epidemic week was reported. Percent cloudy pixels were collected to account for weeks when the indices could not be collected due to cloud cover. This may be an indicator of rainy seasons or fog. It is unclear from the data. When the satellite data is sampled at a higher frequency than once per week, the average value for each epidemic week is reported. Data was recorded from January 3, 2010 to the week of July 3, 2016 to match reported case data, provided by the Ministry of Health of Brazil. Reported case data does not include serotype information for a municipality or state level.

Index Name	**Abr.**	**Formula**	**Description**
Normalized Difference Vegetation	NDVI	$\frac{NIR-Red}{NIR+Red}$	Indicator of healthy, green vegetation
Normalized Difference Water (Green)	Green NDWI	$\frac{Green-NIR}{Green+NIR}$	Indicator of water content in leaves
Normalized Difference Water (SWIR)	SWIR NDWI	$\frac{NIR-SWIR_1}{NIR+SWIR_1}$	Indicator of water content in water bodies
Normalized Burn Ratio	NBR	$\frac{SWIR_1-SWIR_2}{SWIR_1-SWIR_2}$	Indicator of burned areas and fire severity
Percent Cloudy Pixels	–	–	Percent of total pixels covered by clouds

Table 1: Table with description of satellite indices provided by Descartes Lab database. Green, red, near infrared (NIR), and short wave infrared (SWIR) refer to spectral reflectance measurements of wavelengths used in the calculations.

Additionally, temperature and relative humidity data was collected from National Oceanic and

Atmospheric Administration (NOAA). Data was averaged across space and time such that maximum, mean, and minimum temperature and relative humidity, reported weekly at a municipality level for Ceará. Google Health Trends data for the state of Ceará was also included as a possible predictor. Google Health Trends data used in this analysis was for the relative number of searches by state for the term "dengue."

The following variables of interest were considered as possible predictors of future dengue cases: Maximum, mean, and minimum NDVI; Maximum, mean, and minimum Green NDWI; Maximum, and minimum SWIR NDWI; Maximum, mean, and minimum NBR; Maximum, mean, and minimum temperatures in Celsius; Relative humidity; Percent cloudy pixels; Google Health Trends data. Mean SWIR NDWI is missing from the data sets that were provided to the author. Summary statistics for the data used in this analysis is shown in Table 2.

The data is provided at a municipality level scale. There were 5564 municipalities in Brazil at the start of this data set in 2010, with these organized into 136 mesoregions within 27 states. The state of Ceará has 7 mesoregions which contain 184 municipalities.

Variable	**N**	**Mean**	**Std Dev**	**Min.**	**Max.**
Reported Dengue Cases	344	886.0581395	1084.53	61	6754
Max NDVI	344	0.0047152	0.00036208	0.0032693	0.0054419
Mean NDVI	344	0.0013696	0.000141389	0.000839939	0.0019117
Min NDVI	344	-0.0047575	0.0014459	-0.0067328	-0.0015583
Max Green NDWI	344	0.0048181	0.0014178	0.0016656	0.0067616
Mean Green NDWI	344	-0.0011883	0.000143239	-0.0016603	-0.000678846
Min Green NDWI	344	-0.0042328	0.000294099	-0.0048271	-0.002906
Max SWIR NDWI	344	0.0058034	0.000416952	0.0044158	0.0070395
Min SWIR NDWI	344	-0.0030609	0.0012145	-0.0054567	-0.000539747
Max NBR	344	0.005248	0.000955511	0.0033392	0.0068565
Mean NBR	344	0.0011671	0.0002766	-0.000305496	0.0015328
Min NBR	344	-0.0035686	0.0013242	-0.005861	-0.000699305
Percent cloudy pixels	344	0.3279851	0.1213258	0.0715466	0.6041456
Relative humidity	344	0.5329018	0.0429924	0.4382377	0.6041358
Max temperature	344	0.2269754	0.0082925	0.2080857	0.2454742
Mean temperature	344	0.1941354	0.0062933	0.1784672	0.2102962
Min Temperature	344	0.1568894	0.0181173	0.0994845	0.1729983
Google Health Trends	344	2540.45	2148.93	314.6121608	12358.35

Table 2: Table with summary statistics for all variables.

2.2 Distributed Lag Models

All our data is presented in a time series format. In statistics and economics, distributed lag models are used for analysis of time series data. Distributed lag models assume that the effect of a predictor x on an output y occurs over an interval of time, $t \in [0, -L]$, rather than all at once. In other words, the output y depends on the effect of x at time $t, t-1, \ldots, t-l, \ldots, t-L$, where L is the truncation lag time. We can write this model as the following:

$$Y(t) = c_0 + \beta_0 x(t) + \beta_1 x(t-1) + \cdots + \beta_l x(t-l) + \ldots \beta_L x(t-L) \tag{1}$$

$$= c_0 + \sum_{l=0}^{L} \beta_l x(t-l) \tag{2}$$

Traditionally, this could be considering the delayed effect of new income taxes on the income of suppliers or of a new policy on the economy.[19] In terms of health, we can consider more generally the delayed effect of an exposure on a given response. For example, the effect of particulate matter air pollution to lung health [34] for environmental health, the effect of Gross National Product (GNP) to health [16] for health policy, and the effect of climate on mosquitos [8, 21] for vector born diseases.

We will focus on polynomial distributed lag models, which were first proposed by Shirley Almon in 1965.[3] Consider for example $x(t)$ and $x(t-1)$ of Equation 1. Since these variables are correlated, their coefficients β_0 and β_1 are also correlated, leading to issues with multicollinearity. This can lead to two issues: (a) if the model is identifiable, the covariance matrix of the estimated coefficients β_l may be nearly singular leading to numerical problems and (b) since there are $L+2$ coefficients to estimate, the degrees of freedom will quickly be exhausted. So, we instead assume that all the β_l follow some continuous function such that $\beta_l = f(l)$. For polynomial distributed lag, we construct $f(l)$ as some polynomial function of at most degree D in lag length l. Using Equation 2, we can now write[17]:

$$\beta_l = f(l) = \alpha_0 + \alpha_1 l + \cdots + \alpha_D l^D \qquad = \sum_{d=0}^{D} \alpha_d l^d \qquad \text{for all } l. \tag{3}$$

$$Y(t) = c_0 + \sum_{l=0}^{L} f(l) x(t-l) \qquad = c_0 + \sum_{l=0}^{L} \sum_{d=0}^{D} \alpha_d l^d x(t-l)\ . \tag{4}$$

Since β_l is now determined by a polynomial function, we have resolved our issues: (a) The covariance matrix of the β_l is no longer nearly singular and (b) Now there are only $D+2$ coefficients to estimate instead of $L+2$. We can limit the size of D such that we do not exhaust our degrees of freedom.

We can further simplify the problem by setting $s_d = \sum_{l=0}^{L} l^d x(t-l)$, and rewrite $Y(t)$ to find:

$$Y(t) = c_0 + \alpha_0 s_0 + \alpha_1 s_1 + \cdots + \alpha_D s_D \qquad = c_0 + \sum_{d=0}^{D} \alpha_d s_d\ . \tag{5}$$

From here, we can solve the linear system for α_d to reconstruct β_l. Since there may still be issues with correlation among the new coefficients α_d, we choose our polynomial basis functions $f(l)$ such that they are orthogonal. Rather, $o_d(l)$ is some orthogonal polynomial of degree d in lag length l, such that $s_d = \sum_{l=0}^{L} o_d(l)x(t-l)$. [20] Orthogonal polynomials can be constructed using the methodology laid out by Emerson et al. as follows: [12, 20]

$$\sum_{i=1}^{n} w_i o_j(i) o_k(i) = \begin{cases} 1 \text{ if } j = k \\ 0 \text{ if } j \neq k \end{cases}, \tag{6}$$

where w_i is a weighting factor, $n = d + 1$, and $o_j(i)$ is the j^{th} degree of the orthogonal polynomial in lag length i.

2.3 Objectives

We will apply a polynomial distributed lag model to the data provided for Ceará, Brazil, under various truncation lag selection criteria. In particular, we will analyze the prediction capabilities of each of the variables of interest described in Section 2.1, identifying which contribute most to the risk of dengue.

3 Methodology

As described in Section 2.1, we were provided with the data sets for 17 different predictor variables of interest at a municipality level for Ceará and the reported number of dengue cases. Since we want to do analysis for the state-level of Ceará, we aggregated the data to the state level after weighting for geographic scale. The data used in this analysis contains the average value for each predictor for each epidemic week from January 3, 2010 to July 3, 2016 (344 epidemic weeks) only for the state-level of Ceará. No mesoregion level or municipality level information was considered in this analysis. Summary statistics for this data can be found in Table 2.

Previous distributed lag models for mosquito borne diseases do not provide an explanation on how the truncation lag time L is selected, or how the degree of the polynomial D is selected. We will analyze 3 different methods for selecting the truncation lag time, L, and set the degree of the polynomial such that $D = 8$ since this has been used before in the literature. [8]

We will consider a few methods for selecting the truncation lag time L of the polynomial distributed lag model, summarized here:

- **Simple lag selection:** All variables have a truncation lag time of 9 weeks, or about 2 months. This is the minimum number of weeks that must be provided to create an 8^{th} degree polynomial.

- **Marginal lag selection:** All variables have a custom truncation lag time depending on when the marginal lag coefficient becomes statistically insignificant for $\alpha = 0.05$ T test. [1, 3, 5, 25]
- **Minimized AIC lag selection:** All variables have a custom truncation lag time depending on when the Akaike's Information Criteria (AIC) score is minimized for our given dataset. [1, 3, 5]

For the marginal lag selection, we select progressively higher truncation lag times starting with $L = 9$ as our minimum acceptable truncation lag time until the marginal β_L coefficient is statistically insignificant based on a T-test with $\alpha = 0.05$. To ensure that at least one β_l coefficient is statistically significant, we first allow truncation times to increase until at least one β_l coefficient is statistically significant then continue to increase the truncation lag time until the marginal β_L coefficient is statistically insignificant. We repeat this analysis for all variables to find the best truncation lag time. The SAS code for minimizing the AIC score can be found in Section 8.3.

For the minimized AIC lag selection, we select progressively higher truncation lag times starting with $L = 9$ as our minimum acceptable truncation lag time until the AIC score is minimized for our time set, with a maximum possible truncation lag of $L = 343$ since we have 344 weeks in our dataset. The SAS code for minimizing the AIC score can be found in Section 8.4.

All polynomial distributed lag analysis was performed in SAS to provide robust analysis, AIC statistics, t-tests for the β coefficients, and other outputs. All models were run with $D = 8$ degrees of freedom for the polynomial defining the coefficient of each variable. It is also important to note that SAS assumes a weighting factor of $w_i = 1$ when constructing orthogonal polynomials in Proc PDLREG.[20] The code related to this process can be found in Section 8.2.

After selecting our truncation lag times, we construct univariate polynomial distributed lag models for each of the variables of interest. For each of these models we calculate Total R^2, a measure of the amount of variance in output variable y explained by the variable of interest, x. Based on the Total R^2, we will select the best 8 predictors for each selection criteria and build a multivariable model such that

$$Y(t) = c_0 + \sum_{i}^{8} \sum_{l=0}^{L} \sum_{d=0}^{D} \alpha_d l^d x_i(t-l) \ .$$

To compare these models, we again look at the Total R^2. The Total R^2 is not a good measure to compare however because the value of Total R^2 always increases as more parameters are included in a model. Instead we should use the Adjusted R^2 value to compare between groups.

$$R^2_{adj} = 1 - \frac{(1-R^2)(N-1)}{N-p-1} \ ,$$

where N is the total number of observations and p is the number of parameters estimated by the model. The Mean Square Error (MSE) for each model is also recorded, as this is sometimes used as a selection criteria in place of the AIC score.

4 Results

In this section, we present the results, using the methodology described in Section 3. Please note that the t-tests for the β coefficients used for the marginal lag selection procedure are not shown in this paper.

Summarized in Table 3, Table 4, and Table 5 are the summary results for univariate analysis. We can use the univariate analysis to select the variables of interest that appear to contribute the most to predicting dengue cases in Ceará.

For the simple lag selection criteria (Table 3), the following 8 variables account for most of the variation in dengue cases: mean NDVI, mean Green NDWI, percent cloudy pixels, relative humidity, maximum temperature, mean temperature, minimum temperature, and Google Health Trends. For the marginal lag selection criteria (Table 4), the following 8 variables account for most of the variation in dengue cases: maximum NDVI, mean NDVI, mean Green NDWI, percent cloudy pixels, relative humidity, maximum temperature, mean temperature, and Google Health Trends. For the minimized AIC lag selection criteria (Table 5), the following 8 variables account for most of the variation in dengue cases: mean NDVI, minimum NDVI, minimum Green NDWI, minimum SWIR NDWI, maximum NBR, mean NBR, minimum NBR, percent cloudy pixels, and Google Health Trends. These results are summarized in Table 6.

Using the variables that contribute the most information, we build multivariable models for predicting dengue cases for each of the three truncation lag criteria. Each multivariable model contains the top 8 predictor variables for each of the three truncation lag criteria. These models have Total R^2 values of 0.6329, 0.8441, and 0.9996 for the simple, marginal, and minimized AIC selection criteria respectively. To calculate the Adjusted R^2, we use $N = 344$ and $p = 1 + 5m = 41$, given $m = 8$ variables of interest in the model. Therefore, the Adjusted R^2 values for each of the models are 0.5831, 0.8229, and 0.9995 for the simple, marginal, and minimized AIC selection criteria respectively. We find that the minimized AIC truncation lag selection yields the best prediction model.

Variable	**Trunc. Lag**	**AIC**	**MSE**	**Total R^2**	**Top 8 Predictors**
Max NDVI	9	56486	1196693	0.0280	
Mean NDVI	9	5602.72509	1042653	0.1531	*
Min NDVI	9	5656.01352	1222428	0.0071	
Max Green NDWI	9	5656.0773	1222660	0.0069	
Mean Green NDWI	9	5609.27173	1063229	0.1364	*
Min Green NDWI	9	5652.58196	1209970	0.0172	
Max SWIR NDWI	9	5632.64173	1140050	0.0740	
Min SWIR NDWI	9	5652.87109	1211014	0.0164	
Max NBR	9	5654.76983	1217898	0.0108	
Mean NBR	9	5650.76026	1203408	0.0225	
Min NBR	9	5651.71766	1206888	0.0197	
Percent cloudy pixels	9	5500.96624	769519	0.3750	*
Relative humidity	9	5508.04866	785961	0.3616	*
Max temperature	9	5518.25164	810268	0.3419	*
Mean temperature	9	5549.6948	890003	0.2771	*
Min Temperature	9	5623.74816	1110182	0.0983	*
Google Health Trends	9	5524.56899	825692	0.3293	*

Table 3: Table with output from SAS Proc PDLREG using simple lag selection for all variables of interest.

Variable	**Trunc. Lag**	**AIC**	**MSE**	**Total R^2**	**Top 8 Predictors**
Max NDVI	34	5189.72066	1057514	0.1935	*
Mean NDVI	29	5176.52079	777852	0.3996	*
Min NDVI	32	5267.45823	1218920	0.0660	
Max Green NDWI	32	5268.45477	1222819	0.0631	
Mean Green NDWI	28	5186.58308	762400	0.4100	*
Min Green NDWI	34	5203.63104	1106047	0.1565	
Max SWIR NDWI	26	5342.89357	1124370	0.1254	
Min SWIR NDWI	22	5414.02303	1138584	0.1050	
Max NBR	41	5094.05975	1134559	0.1497	
Mean NBR	26	5376.22542	1248621	0.0287	
Min NBR	28	5358.56841	1166632	0.0971	
Percent cloudy pixels	23	5253.26275	727064	0.4300	*
Relative humidity	21	5284.23871	723261	0.4299	*
Max temperature	25	5185.44743	651217	0.4922	*
Mean temperature	17	5408.74143	866756	0.3093	*
Min Temperature	35	5197.03646	1143019	0.1303	
Google Health Trends	13	5448.90084	801559	0.3551	*

Table 4: Table with output from SAS Proc PDLREG using marginal lag selection for all variables of interest. T-test for marginal β_L coefficient was done for $\alpha = 0.05$.

Variable	Trunc. Lag	AIC	MSE	Total R^2	Top 8 Predictors
Max NDVI	127	3292.33084	204076	0.7554	
Mean NDVI	190	2309.76637	179490	0.8330	*
Min NDVI	196	2214.2396	172478	0.8448	*
Max Green NDWI	197	2219.35809	197638	0.8228	
Mean Green NDWI	190	2316.32149	187295	0.8266	
Min Green NDWI	127	3269.23168	195383	0.7659	
Max SWIR NDWI	150	2911.39525	183349	0.7934	
Min SWIR NDWI	276	990.430203	108252	0.9424	*
Max NBR	275	993.369358	91628	0.951	*
Mean NBR	276	988.821444	105721	0.9437	*
Min NBR	194	2223.91416	150825	0.8632	*
Percent cloudy pixels	275	997.108919	96731	0.9483	*
Relative humidity	275	990.268283	87601	0.9531	
Max temperature	198	2239.04167	250694	0.776	
Mean temperature	196	2254.01681	225661	0.797	
Min temperature	159	2791.91279	198705	0.7801	
Google Health Trends	272	997.774975	53737	0.97	*

Table 5: Table with output from SAS Proc PDLREG using minimized AIC lag selection for all variables of interest.

Variable	Simple	Marginal	Minimized AIC
Max NDVI		*	
Mean NDVI	*	*	*
Min NDVI			*
Max Green NDWI			
Mean Green NDWI	*	*	
Min Green NDWI			*
Max SWIR NDWI			
Min SWIR NDWI			*
Max NBR			*
Mean NBR			*
Min NBR			
Percent cloudy pixels	*	*	*
Relative humidity	*	*	
Max Temp.	*	*	
Mean Temp.	*	*	
Min Temp.	*		
Google Health Trends	*	*	*

Table 6: Summary of top 8 variables found to be significant for each of the truncation lag selection cases. The * indicates the variable was found to be significant for the given truncation lag selection case.

5 Discussion

First, we analyze which selection type for the truncation lag time L created the best multi-variable model of the given data set for Ceará, Brazil. Since, as stated in Section 4, Total R^2 always increases with the introduction of more parameters, we choose to compare the models produced by the three selection methods using the Adjusted R^2. Based on the Adjusted R^2, minimized AIC lag selection produced the best model of the data provided. It is important to note however, that minimized AIC lag selection also consistently chose a higher truncation lag time than the two other selection criteria considered. This means that more data was used to inform the prediction for the minimized AIC truncation lag model. This type of truncation lag criteria may be inefficient for larger datasets as it implies a higher data use.

The marginal truncation lag model compared to the minimized AIC truncation lag model used on average 229.1429 fewer weeks of information and had a 17.67% smaller Adjusted R^2 value. This type of truncation lag criteria is more efficient than the minimized AIC truncation lag model, using

much less data to inform a relatively strong prediction of the data.

Next, we analyze which variables of interest contribute the most to the risk of dengue. From in Table 6, we can see that for all truncation lag selection cases, the following variables were found to be significant: mean NDVI, percent cloudy pixels, and Google Health Trends. Since Mean NDVI is an indicator of vegetation, this indicates that there is a strong relationship between available vegetation and adult mosquito abundance. As discussed in Section 1.2, adult mosquitos feed on plant sugars as their main energy source, therefore this finding is consistent with the literature. This result is consistent with other papers that have quantified a relationship between vegetation and mosquito-born disease prevalence.[9, 11, 15, 26] Percent cloudy pixels, which was also found to be significant across all models, was included in our analysis under the assumption that it may be an indicator of rainy seasons or fog. Google Health Trends data, which reflects the relative volume of internet searches for "dengue," was significant across all selection criteria and has been shown in the literature to be a possible predictor of dengue cases. [18, 23, 27]

Simple and marginal truncation lag selection both found the following predictors to be the most significant: mean NDVI, mean Green NDWI, percent cloudy pixels, relative humidity, maximum temperature, mean temperature, and Google Health Trends. As expected, maximum and mean temperature were a significant indicator of dengue prevalence, which is consistent with the literature. Two of the five variables found to be significant across all models are indicators of water content in the environment (Green NDWI and relative humidity). This is consistent with mosquito development being highly dependent on water sources as described in Section 1.2.

The minimized AIC lag selection also exhibited another interesting trend: the majority of all significant variables of interest generated an ideal truncation lag time of approximately 275 weeks, or about 5 years. It has been shown in the literature that dengue occurs in 3 to 5 year period cycles.[4, 24] The minimized AIC lag selection criteria appears to identify this cyclic trend in the data. The variables that exhibited this trend were minimum SWIR NDWI, maximum NBR, mean NBR, percent cloudy pixels, and Google Health Trends. Minimum SWIR NDWI measures water content in water bodies, a variable that changes slowly over time. Maximum and mean NBR may be reflecting land development and not just fire damage due to how they are calculated. NBR specifically reflects a reduction in vegetation.[2] These variables may therefore be influencers of the more long term cyclic trends of dengue infection.

Similarly, for the marginal lag selection exhibited a trend where the majority of all significant variables of interest generated an ideal truncation lag time of approximately 25 weeks, or about half a year. The variables that exhibited this trend were mean NDVI, mean Green NDWI, percent cloudy pixels, relative humidity, and maximum temperature. This may indicate that the short term effects of these variables are significant for predicting reported dengue cases. We must note however that percent cloudy pixels was identified as a predictor of both short and long term effects.

This model framework is highly dependent on the quality of our outcome data, reported dengue cases. First, as mentioned in the Section 1, the reported case data does not include serotype

information. Instead, we are predicting all dengue cases, regardless of serotype. Secondly, many dengue cases are asymptomatic and therefore go unreported. Predictions and forecasts from a models such as these would primarily be of benefit only to healthcare providers to treat and track symptomatic, reported cases. These predictions may not fully reflect the disease burden of dengue in a given area.

6 Summary and Conclusions

We applied a polynomial distributed lag model to the data provided for Ceará, Brazil, using simple, marginal, and minimized AIC truncation lag selection criteria. It was found that the minimized AIC truncation lag model provided the best fit to the data, but was inefficient at using the data compared to the marginal lag model. For a larger dataset including all of the states of Brazil or a subset of municipalities for example, the marginal truncation lag selection criteria could be considered sufficient. These models were also developed on a state-level scale, though data was provided at a municipality-level scale. To more accurately express the spatial heterogeneity in disease burden, a higher resolution scale should be used now that it has been shown that the model is valid at a low resolution scale.

The strong Adjusted R^2 value for both the marginal lag and minimized AIC truncation lag models show that polynomial distributed lag models could be used to successfully predict reported dengue cases in Ceará, Brazil. Future work would be to use these models to forecast reported dengue cases, therefore giving public health professionals forewarning on the severity of upcoming epidemics to create targeted, efficient interventions.

Furthermore, the ideal truncation lag results for minimized AIC criteria appeared to identify the period of the cycle of dengue in Ceará. This periodicity in dengue epidemics has been shown to be dependent on seasonal variation in vector demography and the number of dengue serotypes present in the population. Patterns with 5 year cycles further indicate asymmetry in serotype virulence and temporary cross-immunity.[33] Therefore, while all four serotypes of dengue are present in Brazil,[32] this finding indicates that there may be asymmetry in their relative prevalence and virulence in Ceará. As such, more serotype-specific data is essential to gather in order to create targeted programs in Ceará.

The variables associated with this trend were indicators for large water sources, reduced vegetation, and increased internet searches for "dengue." Monitoring these variables and understanding their relationship with reported dengue cases would help improve interpretations of these models.

The ideal truncation lag results for marginal lag criteria identified predictors with short term influence on the reported number of dengue cases. The variables associated with this trend were indicators for healthy vegetation, water content in leaves, humidity, and temperature. These variables have a more immediate effect in the expected disease burden of dengue for a given year.

Of all the variables of interest provided, mean NDVI, percent cloudy pixels, and Google Health Trends were found to be the significant predictors of reported dengue cases for Ceará, Brazil, across all selection criteria. These findings were consistent with risk indicators found in the literature. It is unclear what the implications are of percent cloudy pixels, which included in our analysis under the assumption that it may be an indicator of rainy seasons or fog, and further research is needed to assess the implications of this indicator.

The models developed here do provide evidence that dengue, a mosquito-borne disease, can be predicted using satellite indicies and temperature data. Current epidemics with Zika and Chikungunia in South America could be predicted and forecasted in a similar manner with this same model and dataset, provided that reported case counts for at least 1 year are known. Since Zika is also spread by *A. aegypti*, coefficient estimates for dengue could be used to make rough predictions of the disease burden of Zika.

7 Acknowledgments

This work was supported by NSF SEES grant CHE-1314029, Descartes Labs, Los Alamos National Laboratory, and the New Mexico Consortium. Thank you to Christian Geneus for spending many hours sitting with me while I wrote this paper. Special thanks to Adrian Devitt-Lee for coding assistance and moral support.

References

[1] *Time Series Analysis*. Princeton University Press, Princeton, N.J., 1994.

[2] Normalized burn ratio. *http://gsp.humboldt.edu/OLM/Courses/GSP*$_2$16$_O$*nline/lesson*5 − 1/*NBR.html*, 2014.

[3] S. Almon. The distributed lag between capital appropriations and expenditures. *Econometrica*, 33(1), 1965.

[4] S. N. Bennet, A. J. Drummond, D. D. Kapan, M. A. Suchard, J. L. Muñoz Jordan, O. G. Pybus, E. C. Holmes, and D. J. Gubler. Epidemic dynamics revealed in dengue evolution. *Molecular Biology and Evolution*, 27(4):811–818, April 2010.

[5] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley Sons Inc., 5 edition, 1976.

[6] F. P. Câmara, R. L. G. Theophilo, G. T. dos Santos, S. R. Pereira, D. C. Câmara, and R. R. de Matos. Estudo retrospectivo (histórico) da dengue no brasil: características regionais e dinâmicas. *Revista da Sociedade Brasileira de Medicina Tropical*, 40(2), March 2007.

[7] R. V. da Cunha, M. P. Miagostovich, Z. Petrola, E. S. M. de Araújo, D. Cortez, V. Pombo, R. V. de Souza, R. M. R. Nogueira, and H. G. Schatzmayr. Retrospective study on dengue in fortaleza, state of ceará, brazil. 93(2), March 1998.

[8] J. K. Davis, G. Vincent, M. B. Hildreth, L. Kightlinger, C. Carlson, and M. C. Wimberly. Integrating environmental monitoring and mosquito surveillance to predict vector-borne disease: Prospective forecasts of a west nile virus outbreak. *PLoS Currents*, 9, 2017.

[9] M. Duik-Wasser, H. E. Brown, T. G. Andreadis, and D. Fish. Modeling the spatial distribution of mosquito vectors for west nile virus in connecticut, usa. *Vector-Borne and Zoonotic Diseases*, 6(3), September 2006.

[10] Nature Education. Dengue transmission. *https://www.nature.com/scitable/topicpage/dengue-transmission-22399758*, 2014.

[11] L. Eisen and S. Lozano-Fuentes. Use of mapping and spatial and space-time modeling approaches in operational control of aedes aegypti and dengue. *PLoS Neglected Tropical Diseases*, April 2009.

[12] P. L. Emerson. Numerical construction of orthogonal polynomials from a general recurrence formula. *Biometrics*, 24(3), September 1968.

[13] J. T. Fiusa Lima. Risco de urbanização da febre amarela no brasil. *Cadernos de Saúde Pública*, 1(3), July 1985.

[14] Centers for Disease Control and Prevention. Dengue. *https://www.cdc.gov/dengue/index.html*, 2016.

[15] D. O. Fuller, A. Troyo, and J. C. Beier. El niño southern oscillation and vegetation dynamics as predictors of dengue fever cases in costa rica. *Environmental Research Letters*, 4(1), March 2009.

[16] T. Getzen. Population aging and the growth of health expenditures. *Journal of Gerontology*, 47(3):S98–S104, May 1992.

[17] D. Gile. Explaining the almon distributed lag model. *Econometrics Beat: Dave Gile's Blog*, January 2017.

[18] R. T. Gluskin, M. A. Johansson, M. Santillana, and J. S. Brownstein. Evaluation of internet-based dengue query data: Google dengue trends. *PLoS Neglected Tropical Diseases*, February 2014.

[19] W. Griffiths and G. G. E. Judge. *Undergradute Econometrics*, chapter 15. John Wiley Sons Inc., 2 edition, 2001.

[20] https://support.sas.com/documentation/onlinedoc/ets/132/pdlreg.pdf. *SAS/ETS 13.2 User's Guide: The PDLREG Procedure*. SAS Institute Inc., Cary, NC, 2014.

[21] M. A. Johansson, F. Dominici, and G. E. Glass. Local and global effects of climate on dengue transmission in puerto rico. *Neglected Tropical Diseases*, February 2009.

[22] V. Lima, L. Figureiredo, H. Correa, O. Leite, O. Rangel, A. Vido, S. Oliveira, M. Owa, and R. Carlucci. Dengue: inquérito sorológico pós-epidêmico em zona urbana do estado de são paulo (brasil). *Revista de Saúde Pública*, 33(6), December 1999.

[23] C. Marques-Toledo, C. M. Degener, L. Vinhal, G. Coelho, W. Miera, C. T. Codeço, and M. Teixeira. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting dengue at country and city level. *PLoS Neglected Tropical Diseases*, July 2017.

[24] World Health Organization. *Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control*. World Health Organization and Special Programme for Research and Training in Tropical Diseases, Geneva, Switzerland, 2009.

[25] J. Parker. Distributed-lag models. 2013.

[26] A. T. Peterson, C. Martínez-Campos, Y. Nakazawa, and E. Martínez-Meyer. Time-specific ecological niche modeling predicts spatial dynamics of vector insects and human dengue cases. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 99(9):647–655, September 2005.

[27] S. Pollett, B. M. Althouse, B. Forshey, G. W. Rutherford, and R. G. Jarman. Internet-based biosurveillance methods for vector-borne diseases: Are they novel public health tools or just novelties? *PLoS Neglected Tropical Diseases*, November 2017.

[28] OECD Publishing. Oecd territorial reviews: Brazil 2013. *http://dx.doi.org/10.1787/9789264123229-en/content/book/9789264123229-en*, 2013.

[29] L. Ramadona, A. L. Lazuardi, L. Y. Hii, A Holmner, H. Kusnanto, and J Rocklöv. Prediction of dengue outbreaks based on disease surveillance and meteorological data. *Public Library of Science ONE*, 11(3), March 2016.

[30] World Health Organizaiton Fact Sheet. Vector-borne diseases. *http://www.who.int/mediacentre/factsheets/fs387/en/*.

[31] M. Teixeira, M. Barreto, M. Costa, L. Ferreira, P. Vasconcelos, and S. Cairncross. Dynamics of dengue virus circulation: a silent epidemic in a complex urban area. *Tropical Medicine and International Health*, 7(9):757–762, September 2002.

[32] C. J. Villabona-Arenas, J. L. de Oliveira, C. S. Capra, K. Balarini, M. Loureiro, C. R. T. Fonseca, S. D. Passos, and P. M. A. Zanotto. Detection of four dengue serotypes suggests rise in hyperendemicity in urban centers of brazil. *PLoS Neglected Tropical Diseases*, 8(2), February 2014.

[33] H. J. Wearing and P. Rohani. Ecological and immunological determinants of dengue epidemics. *Proceedings of the National Academy of Sciences*, 103(31):11802–11807, August 2006.

[34] L. J. Welty, R. D. Peng, S.L. Zeger, and F. Dominici. Bayesian distributed lag models: Estimating effects of particulate matter air pollution on daily mortality. *Biometrics*, 65:282–291, March 2009.

[35] C. Yung, K. Lee, T. Thein, L. Tan, V. Gan, J. G. X. Wong, L. Ng, and Y. Leo. Dengue serotype-specific differences in clinical manifestation, laboratory parameters and risk of severe disease in adults, singapore. *American Journal of Tropical Medicine and Hygiene*, 92(5):999–1005, May 2015.

8 Appendix

8.1 R Code for Cleaning Dataset

All R code for creating the proper work environment.

```
# This script is edited for analysis of the dengue data of Ceara,
   Brazil
# It assumes that you have the following data:
#   Shapefiles for Brazil
#   Case Data
#   Predictor Data:
#     (max/mean/min)ndvi, (max/mean/min)green_ndwi,(max/[mean]/min)swir
   _ndwi,
#     (max/mean/min)nbr, percent_cloudy_pixels,relhum, (max/mean/min)
   temp, google_trends
#
# This data is called by the files:
```

```
# rawDeng.RData, rawClustDeng.RData, rawStateDeng.RData,
    datastreamsNov20.RData, datastreamsJan20.RData, datastreamsJan20.
    RData
#  **stored in a folder called "RData" in your working directory (wd)
# municipios_2010 shapefiles
#  **stored in a folder called "municipios_2010"
# *Epidemic_Weeks_BRA.csv is currently not being used. User can comment
     lines 126-132 if they do not have this.
# Google Trends csv files
#  **stored in a folder called "raw-data"
#
# Edit the wd lines before running. Check that you have all the
    necessary libraries installed.

#### Set up the Work Environment
    ##############################################################################

##Clear the work environment
rm(list=ls())

#Set working directory
setwd("/Volumes/Jessie's Hard Drive/LANL_Items/Research") #/Volumes/
    LIFELINE/LANL_Items/Research
home = setwd("/Volumes/Jessie's Hard Drive/LANL_Items/Research") #/
    Volumes/LIFELINE/LANL_Items/Research

library(xtable)
library(matrixStats)
library(sp)
library(rgdal)
library(raster)
library(ggplot2)
library(plyr)
library(lattice)
library(RColorBrewer)
library(classInt)
library(reshape2)
library(maptools)
```

```
library(gridExtra)
library(mgcv)
library(dlnm)

#Call in shape files
  BRAadm2 = readOGR("municipios_2010","municipios_2010")
  names(BRAadm2)[names(BRAadm2)=='codigo_ibg']<-'Mun_Number'
  names(BRAadm2)[names(BRAadm2)=='nome'] <- 'Mun_Resid_BR'
  names(BRAadm2)[names(BRAadm2)=='estado_id'] <- 'State_Number'
  names(BRAadm2)[names(BRAadm2)=='uf'] <- 'State_ID'
  BRAadm2$Mun_Number = substr(BRAadm2$Mun_Number,1,nchar(as.character(
      BRAadm2$Mun_Number))-1) #Chop off extra digit of UID
  BRAadm2$Mun_Number <- as.numeric(BRAadm2$Mun_Number)
  assign("BRAadm2",BRAadm2,.GlobalEnv)
  BRAadm1 = readOGR("estados_2010","estados_2010")
  names(BRAadm1)[names(BRAadm1)=='id']<-'State_Number'
  names(BRAadm1)[names(BRAadm1)=='nome'] <- 'State_Name'
  names(BRAadm1)[names(BRAadm1)=='sigla'] <- 'State_ID'
  assign("BRAadm1",BRAadm1,.GlobalEnv)

##Call in satellite files
fpath = file.path(home,"RData","datastreamsJan20.RData")
load(fpath)
fpath = file.path(home,"RData","datastreamsJan20_meso.RData")
load(fpath)
fpath = file.path(home,"RData","datastreamsJan20_state.RData")
load(fpath)

datastreams <- c("max_ndvi", "mean_ndvi", "min_ndvi", "max_green_ndwi",
    "mean_green_ndwi", "min_green_ndwi",
                 "max_swir_ndwi", "min_swir_ndwi", "max_nbr", "mean_nbr
                    ", "min_nbr", "percent_cloudy_pixels",
                 "relhum", "tmax", "tmean", "tmin")


#Call in case data
fpath = file.path(home,"RData","rawDeng.RData")    #Raw case data for
   dengue by municipality, mesolevel, and state
```

```
load(fpath)
fpath = file.path(home,"RData","rawClustDeng.RData")
load(fpath)
fpath = file.path(home,"RData","rawStateDeng.RData")
load(fpath)

Deng.all <- cbind(dlist1[[1]][,1:(ncol(dlist1[[1]])-2)],dlist1
    [[2]][,3:(ncol(dlist1[[2]])-2)], dlist1[[3]][,3:(ncol(dlist1[[3]])
    -2)],
                  dlist1[[4]][,3:(ncol(dlist1[[4]])-2)], dlist1
                     [[5]][,3:(ncol(dlist1[[5]])-2)], dlist1[[6]][,3:(
                     ncol(dlist1[[6]])-2)],
                  dlist1[[7]][,3:(ncol(dlist1[[7]])-2)])
Deng.all.state <- cbind(dlist4[[1]][,1:(ncol(dlist4[[1]])-2)],dlist4
    [[2]][,2:(ncol(dlist4[[2]])-2)], dlist4[[3]][,2:(ncol(dlist4[[3]])
    -2)],
                        dlist4[[4]][,2:(ncol(dlist4[[4]])-2)], dlist4
                           [[5]][,2:(ncol(dlist4[[5]])-2)], dlist4
                           [[6]][,2:(ncol(dlist4[[6]])-2)],
                        dlist4[[7]][,2:(ncol(dlist4[[7]])-2)])
Deng.all.meso <- as.data.frame(cbind(dlist7[[1]][,1:(ncol(dlist7[[1]])
    -2)],dlist7[[2]][,2:(ncol(dlist7[[2]])-2)], dlist7[[3]][,2:(ncol(
    dlist7[[3]])-2)],
                                     dlist7[[4]][,2:(ncol(dlist7[[4]])
                                        -2)], dlist7[[5]][,2:(ncol(
                                        dlist7[[5]])-2)], dlist7
                                        [[6]][,2:(ncol(dlist7[[6]])-2)
                                        ],
                                     dlist7[[7]][,2:(ncol(dlist7[[7]])
                                        -2)])))

municipality_list <- dlist1
state_list <- dlist4
mesolevel_list <- dlist7
rm(list = c('dlist1','dlist4','dlist7'))


# Reorder Case Data to match satellite
```

```
idx <- sapply(datastreamsLIST[[1]]$Mun_Number, function(x) {
  which(municipality_list[[1]]$Mun_Number == x) })
for (i in 1:7){
  municipality_list[[i]] <- municipality_list[[i]][unlist(idx),]
}
Deng.all <- Deng.all[unlist(idx),]

idx <- sapply(datastreamsLIST_state[[1]]$State_Number, function(x) {
  which(state_list[[1]]$State_Number == x) })
for (i in 1:7){
  state_list[[i]] <- state_list[[i]][unlist(idx),]
}
Deng.all.state <- Deng.all.state[unlist(idx),]

idx <- sapply(datastreamsLIST_meso[[1]]$Meso_Number, function(x) {
  which(mesolevel_list[[1]]$cluster == x) })
for (i in 1:7){
  mesolevel_list[[i]] <- mesolevel_list[[i]][unlist(idx),]
}
Deng.all.meso <- Deng.all.meso[unlist(idx),]


# Grab in true dates
fpath = file.path(home,"Dengue_data","Epidemic_Weeks_BRA.csv")
epi.weeks <- read.csv(fpath, header = TRUE)
epi.weeks1 <- c(as.character(epi.weeks[1:52,2]),as.character(epi.weeks
   [1:52,3]),as.character(epi.weeks[1:52,4]),
                as.character(epi.weeks[1:52,5]),as.character(epi.weeks
                   [1:53,6]),as.character(epi.weeks[1:52,7]),
                as.character(epi.weeks[1:52,8]))
epi.weeks1 <- as.Date(epi.weeks1, "%m/%d/%y")
epi.weeks1 <- epi.weeks1[1:ncol(Deng.all)]

# Unlist State level data
datastreams <- c("max_ndvi", "mean_ndvi", "min_ndvi", "max_green_ndwi",
    "mean_green_ndwi", "min_green_ndwi",
                 "max_swir_ndwi", "min_swir_ndwi", "max_nbr", "mean_nbr
                    ", "min_nbr", "percent_cloudy_pixels",
```

```
                "relhum", "tmax", "tmean", "tmin")
for (i in 1:length(datastreams)){  #This loop unzips the list into its
    separate elements
  assign(paste(datastreams[i]), datastreamsLIST_state[[i]], .GlobalEnv)
}

state <- c('AC', 'AL','AM', 'AP','BA', 'CE', 'DF', 'ES', 'GO', 'MA', '
    MG','MS','MT',
           'PA', 'PB',  'PE', 'PI', 'PR','RJ', 'RN','RO', 'RR','RS', '
              SC', 'SE', 'SP','TO')

# Create functions to be used
grab.google.data <- function(state_num=6){
  ## Function to grab Google Health data
  fpath = file.path(home,"raw-data",paste("BR-", state[state_num],".csv
      ", sep = ""))
  health.trends.data.mun <- read.csv(fpath, header = TRUE, sep = ",",
      colClasses = c("Date",rep("numeric",20)))
  fpath = file.path(home,"raw-data","BR.csv")
  health.trends.data <- read.csv(fpath, header = TRUE, sep = ",",
      colClasses = c("Date",rep("numeric",20)))

  assign("health.trends.data.state", health.trends.data.mun, .GlobalEnv
      )
  assign("health.trends.data", health.trends.data, .GlobalEnv)
}

select.season <- function(season,mydat,flag="off"){
  #Decide on season subset. Seaons are set up such that they follow
      this:
  #     Dengue Season: Jan - April Wk 1:18 of a year (18 weeks)
  #     Pre-Dengue Season: May-Aug Wk 19:35 of a year (17 weeks)
  #     Post-Dengue Season: Sept-Dec Wk 36:52 of a year (17 weeks [18
      weeks for 2014])
  #The data we have is Jan 2010 through July 2016
  mydata2 <- mydat
  if (season=="Dengue") {
    seas <- c(1:18,53:70,105:122,157:174,209:226,262:279,314:331)
```

```
    print("You are calling Dengue season. 2016 data is available for
        forecasting this season.")
  } else if (season=="Post-Dengue") {
    seas <- c(19:35,71:87,123:139,175:191,227:243,280:296,332:344)
    print("You are calling Post-dengue season. 2016 data is missing 4
        weeks of this season. Forecasting is limited.")
  } else{#season=="Pre-Dengue"
    seas <- c(36:52,88:104,140:156,192:208,244:261,297:313)
    print("You are calling Pre-dengue season. 2016 data is missing for
        this season. 2015 will be used for forecasing.")
  }
  if (flag=="off"){
    epi.weeks2 <- epi.weeks1[seas]
    assign("epi.weeks2",epi.weeks2,.GlobalEnv)
  }
  mydata2 <- mydata2[which(mydata2$season==season),]
  return(mydata2)
}

set.data <- function(state_num=6,season1) {
  grab.google.data(state_num)
  caseofinterest = Deng.all.state[which(Deng.all.state$State_Number ==
      state_num),]
  caseofinterest1 <- caseofinterest[2:length(caseofinterest)]

  #Build data matrix with info of interest
  idx <- which(mean_ndvi$State_Number == state_num)
  mydata <- data.frame(time = 1:344, dengue_cases = t(caseofinterest1),
      #GoogleHealthTrends = health.trends.data.mun[53:396,3], #Calling
      the search term "dengue" only
                       max_ndvi = t(max_ndvi[idx,3:346]), mean_ndvi= t(
                           mean_ndvi[idx,3:346]), min_ndvi= t(min_ndvi[
                           idx,3:346]),
                       max_green_ndwi = t(max_green_ndwi[idx,3:346]),
                           mean_green_ndwi = t(mean_green_ndwi[idx
                           ,3:346]), min_green_ndwi = t(min_green_ndwi[
                           idx,3:346]),
```

```
                        max_swir_ndwi = t(max_swir_ndwi[idx,3:346]), min
                            _swir_ndwi = t(min_swir_ndwi[idx,3:346]),
                        max_nbr = t(max_nbr[idx,3:346]),mean_nbr = t(
                            mean_nbr[idx,3:346]),min_nbr = t(min_nbr[idx
                            ,3:346]),
                        percent_cloudy_pixels = t(percent_cloudy_pixels[
                            idx,3:346]), relhum = t(relhum[idx,41:384]),
                        max_temp = t(tmax[idx,41:384]),mean_temp = t(
                            tmean[idx,41:384]), min_temp = t(tmin[idx
                            ,41:384]),
                        google_trends = health.trends.data.state
                            [1:344,3])
  mydata.names <- c("time", "dengue_cases", "max_ndvi", "mean_ndvi", "
      min_ndvi", "max_green_ndwi", "mean_green_ndwi", "min_green_ndwi",
                    "max_swir_ndwi", "min_swir_ndwi", "max_nbr", "mean_
                        nbr", "min_nbr", "percent_cloudy_pixels",
                    "relhum", "max_temp", "mean_temp", "min_temp", "
                        google_trends")
  colnames(mydata) <- mydata.names

  #Subset for seasons
  mydata$season <- c(1:344)
  mydata$season[c(1:18,53:70,105:122,157:174,209:226,262:279,314:331)]
      <- "Dengue"
  mydata$season[c(19:35,71:87,123:139,175:191,227:243,280:296,332:344)]
       <- "Pre-Dengue"
  mydata$season[c(36:52,88:104,140:156,192:208,244:261,297:313)] <- "
      Post-Dengue"
  mydata$year <- c(rep(2010,52),rep(2011,52),rep(2012,52),rep(2013,52),
      rep(2014,53),rep(2015,52),rep(2016,31))

  assign("mydata",mydata,.GlobalEnv)
  assign("mydata.names",mydata.names,.GlobalEnv)

  mydata.season.raw <- select.season(season1, mydat = mydata)
  assign("mydata.season.raw",mydata.season.raw,.GlobalEnv)
}
```

```
state_num <- 6
state[state_num] #Check which state you are calling
```

8.2 SAS Code for Proc PDLREG

All SAS code for Proc PDLREG procedure. Example code shown here is for Maximum NDVI with truncation lag $L = 5$ and degrees of freedom $D = 8$ for the polynomial constructing the coefficients.

```
proc import out= mydat datafile= 'C:/Users/jconrad4/Documents/ceara_
   data.xlsx'
dbms=xlsx replace;
getnames=yes;
run;
proc pdlreg data=mydat;
model dengue_cases = max_ndvi(5,4);
ods output FitSummary=testset;
run;
```

8.3 SAS Macro Code for Marginal β_L

All SAS code for finding when the marginal coefficient of the polynomial distributed lag model is equivalently zero. This was made in reference to the SAS manual for Proc Pdlreg. [20]

```
%macro marginalloop(xvar);
proc import out= mydat datafile= 'C:/Users/jconrad4/Documents/ceara_
   data.xlsx'
dbms=xlsx replace;
getnames=yes;
run;
%let n=9; /* first lag case to test */
proc pdlreg data=mydat;
model dengue_cases = &xvar(&n,8);
ods output LagDist = testset;
run;
/* first lag case to test */
data lastrow;
if 0 then set testset nobs=nobs end=eof;
set testset point = nobs;
output;
```

```
stop;
run;
data _null_;
set lastrow;
call symputx("BTerm", Probt);
run;
%put BTerm = &BTerm.;
/* Do loop until terminal coeff insignificant  */
%do %until (&BTerm < 0.10);
%put n=&n.;
%let n=%eval(&n. + 1); /* &n holds the value */
ods exclude all;              /* suspend all open destinations */
proc pdlreg data=mydat;
model dengue_cases = &xvar(&n,8);
ods output LagDist = testset;
run;
/* find minimum prob(t) */
proc sort data=testset;
by Probt; run;
proc sort data=testset (obs=1);
by Probt; run;
data _null_;
set testset;
call symputx("BTerm", Probt);
run;
%put BTerm = &BTerm.;
%end;
/* find minimum marginal t */
%do %until (&BTerm > 0.10);
%put n=&n.;
%let n=%eval(&n. + 1); /* &n holds the value */
ods exclude all;              /* suspend all open destinations */
proc pdlreg data=mydat;
model dengue_cases = &xvar(&n,8);
ods output LagDist = testset;
run;
data lastrow;
if 0 then set testset nobs=nobs end=eof;
```

```
set testset point = nobs;
output;
stop;
run;
data _null_;
set lastrow;
call symputx("BTerm", Probt);
run;
%put BTerm = &BTerm.;
%end;
/* Print final result  */
ods exclude none;
proc pdlreg data=mydat;
model dengue_cases = &xvar(&n,8);
run;
%mend marginalloop;


%marginalloop(max_ndvi);
```

8.4 SAS Macro Code for Minimizing AIC Score

All SAS code for minimizing the AIC score of the polynomial distributed lag model. This was made in reference to the SAS manual for Proc Pdlreg. [20]

```
%macro AICloop(xvar);
proc import out= mydat datafile= 'C:/Users/jconrad4/Documents/ceara_
   data.xlsx'
dbms=xlsx replace;
getnames=yes;
run;
%let n=9; /* first lag case to test */
proc pdlreg data=mydat;
model dengue_cases = &xvar(&n,8);
ods output FitSummary=testset;
run;
data _null_;
set testset;
if Label2="AIC" then call symputx("AICnew", nValue2);
```

```
run;
%put AICnew = &AICnew.;
%put AICold = 1000000000;/* Arbitrarily large to initialize loop  */
/* Do loop while AIC decreases  */
%do %until (&AICnew > &AICold);
%put n=&n.;
%let n=%eval(&n. +1); /* &n holds the value */
%let AICold = &AICnew; /* Replace AICold with previous AICnew value */
%put AICold = &AICold;
ods exclude all;              /* suspend all open destinations */
proc pdlreg data=mydat;
model dengue_cases = &xvar(&n,8);
ods output FitSummary=testset;
run;
data _null_;
set testset;
if Label2="AIC" then call symputx("AICnew", nValue2);
run;
%put AICnew = &AICnew;
%end;
/* Print final result  */
ods exclude none;
proc pdlreg data=mydat;
model dengue_cases = &xvar(&n,8);
run;
%mend AICloop;

%AICloop(max_ndvi);
```