October 2018

Real-Time Dengue Forecasting In Thailand: A Comparison Of Penalized Regression Approaches Using Internet Search Data

Caroline Kusiak

REAL-TIME DENGUE FORECASTING IN THAILAND: A COMPARISON OF PENALIZED REGRESSION APPROACHES USING INTERNET SEARCH DATA

A Thesis Presented

by

CAROLINE KUSIAK

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

September 2018

Public Health
Biostatistics

REAL-TIME DENGUE FORECASTING IN THAILAND: A COMPARISON OF PENALIZED REGRESSION APPROACHES USING INTERNET SEARCH DATA

A Thesis Presented

by

CAROLINE KUSIAK

Approved as to style and content by:

Nicholas Reich, Chair

Laura Balzer, Member

Jing Qian, Member

Susan E. Hankinson, Department Chair
Department of Biostatistics and Epidemiology

ACKNOWLEDGMENTS

First, I would like to thank my advisor Prof. Nicholas Reich for introducing me to the Biostatistics world and bringing me into the ReichLab. He has been so kind, welcoming, and encouraging. His guidance has helped me to stay focused and taught me how to answer the important questions. Without him of course, I never would have had the opportunity to write this which has become near and dear to my heart.

Next, I would like to thank my academic advisor Prof. Laura Balzer. I am so happy I have been fortunate enough to take a course with her and learn from her. She has been so supportive in all things both within and outside of the academic world.

I would also like to thank Prof. Jing Qian for being such an approachable and knowledgeable teacher. I learned so much in just one semester with him and I am very grateful for him serving on my Thesis Committee.

Thank you to all of those in the Department of Epidemiology and Biostatistics who have taught me so much and made my time at UMass so enjoyable. A special thanks to Deborah Osowski for help coordinating my studies here and answering my many questions and Casey Gibson and Nutcha Wattanachit for their consultations.

Finally, I would like to thank our collaborators at the Thai Ministry of Public Health and Mauricio Santillana and his team. Their counsel and data have been invaluable and without both, this project would never have been possible.

ABSTRACT

REAL-TIME DENGUE FORECASTING IN THAILAND: A COMPARISON OF PENALIZED REGRESSION APPROACHES USING INTERNET SEARCH DATA

SEPTEMBER 2018

CAROLINE KUSIAK

B.A., AMHERST COLLEGE

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Nicholas Reich

Dengue fever affects over 390 million people annually worldwide and is of particular concern in Southeast Asia where it is one of the leading causes of hospitalization. Modeling trends in dengue occurrence can provide valuable information to Public Health officials, however many challenges arise depending on the data available. In Thailand, reporting of dengue cases is often delayed by more than 6 weeks, and a small fraction of cases may not be reported until over 11 months after they occurred. This study shows that incorporating data on Google Search trends can improve disease predictions in settings with severely underreported data. We compare penalized

regression approaches to seasonal baseline models and illustrate that incorporation of search data can improve prediction error. This builds on previous research showing that search data and recent surveillance data together can be used to create accurate forecasts for diseases such as influenza and dengue fever. This work shows that even in settings where timely surveillance data is not available, using search data in real-time can produce more accurate short-term forecasts than a seasonal baseline prediction. However, forecast accuracy degrades the further into the future the forecasts go. The relative accuracy of these forecasts compared to a seasonal average forecast varies depending on location. Overall, these data and models can improve short-term public health situational awareness and should be incorporated into larger real-time forecasting efforts.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

Dengue Fever is a mosquito-borne viral disease endemic in over 100 countries worldwide ([20]). Most commonly spread by the *Aedes aegypti* mosquito, there are over 390 million dengue infections resulting in about 96 million symptomatic cases each year [1]. It is estimated that 3.9 billion people are at risk of dengue virus infection and recently there has been a sharp increase in the number of reported cases [2]. An infection can develop into more severe forms such as dengue hemorrhagic fever (DHF) and eventually, dengue shock syndrome (DSS) which has a 44% fatality rate [17]. Because many dengue infections are asymptomatic, DHF is often used as a proxy for dengue incidence, since it is more severe and consequently more consistently reported.

In Thailand, dengue-related illness is the third leading cause of hospitalization and it is a problem endemic to most all of its provinces [5]. The majority of dengue infections are asymptomatic making it almost impossible to quantify the true number of people affected each year. The number of DHF cases varies across different locations and between seasons making it difficult to anticipate the impact it will have each year.

Making timely and accurate forecasts for future outbreaks can provide valuable information for Public Health officials. Predictions can be used to target times and

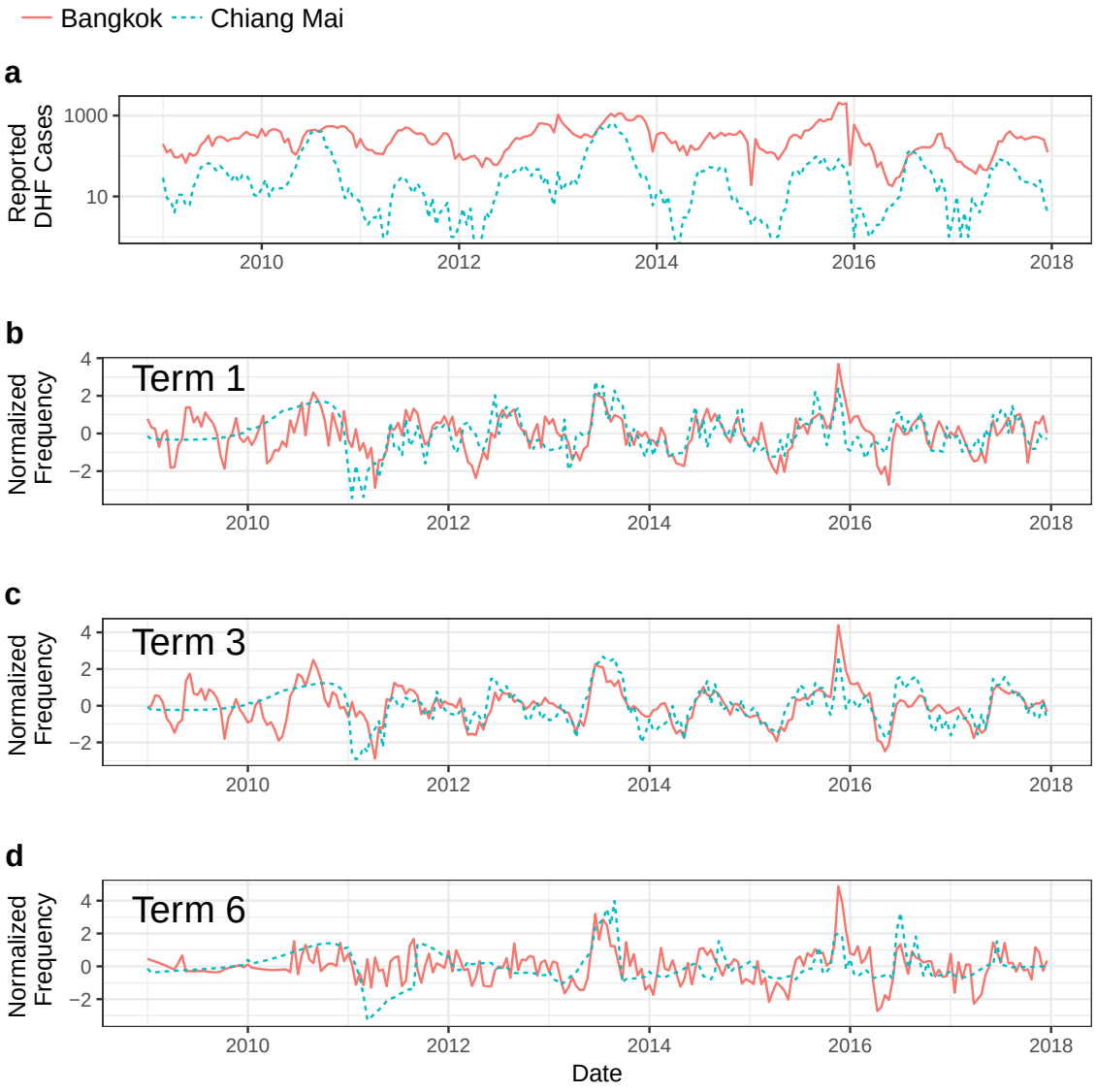

Figure 1.1. Dengue case counts and search term frequencies **a.** Logged DHF case count distributions since 2009. **b.-d.** Distributions of normalized search term frequency for Term 1 (Hemorrhagic fever disease), Term 3 (Hemorrhagic fever), and Term 6 (Symptoms of Hemorrhagic fever).

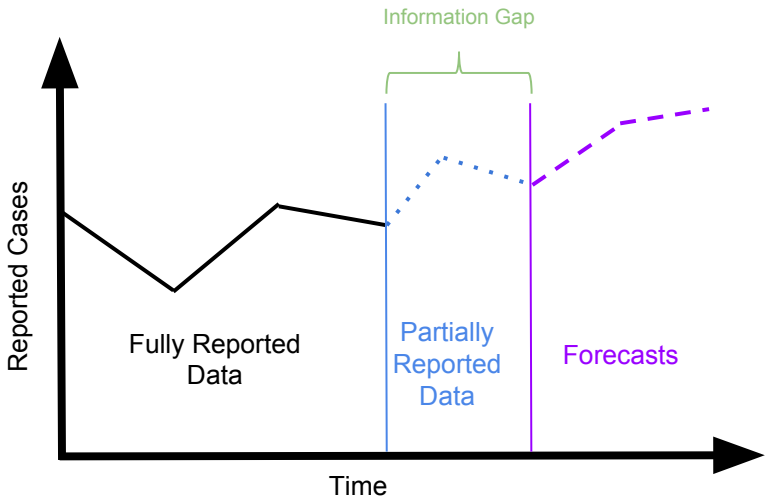

Figure 1.2. Illustration of modeling challenge. Because DHF case reports begin underreported, an information gap exists between our fully reported data and when we wish to begin making forecasts. We cannot trust that the data in the most recent past is representative of the eventually fully reported data and for this reason, we investigate using Google Search data in its place.

locations which could benefit from increased surveillance, treatment resources, and prevention education. Forecasts can be utilized for better planning in these areas and help to understand where to best allocate resources [14].

The Thai Ministry of Public Health provides our team with province-specific DHF case counts every two weeks (Figure 1.1). These estimates however, begin underreported and are revised in following months. Cases reports are often delayed by more than 6 weeks and some cases are not reported until over 11 months after they occur (Mr. Casey Gibson, personal communication). Therefore, accurate data for a given biweek may not be available until April of the following year when final revisions are made. Although we would like to make predictions into the future, we can not trust that the data in the most recent past will be representative of the eventually fully reported counts (Figure 1.2). Past forecasting efforts using this data have dealt with this information gap challenge by ignoring the most recent six weeks of data when making forecasts [16]. Ongoing efforts attempt to model the reporting delays themselves to allow for a estimate of current cases based on scaling up partially observed case counts (Mr. Casey Gibson, personal communication). Both of these approaches face serious challenges and obstacles.

It has been shown that internet search data can be helpful in predicting Influenza-Like-Ilness (ILI) epidemics, however limitations have also been identified [6, 8]. Determining trends in influenza-related search queries provides a valuable information source without expending many resources and while maintaining users' privacy. Google Flu Trends (GFT) was first proposed in 2008 to include information about this behavior in forecasting models [10]. At first, these models provided promising

Term	Thai	English Translation
1	โรคไข้เลือดออก	Hemorrhagic fever disease
2	อาการ โรค ไข้เลือดออก	Symptoms of hemorrhagic fever disease
3	ไข้เลือดออก	Hemorrhagic fever
4	โรค ไข้เลือดออก	Hemorrhagic fever disease
5	การ ป้องกัน ไข้เลือดออก	Prevention of hemorrhagic fever
6	อาการ ของ ไข้เลือดออก	Symptoms of hemorrhagic fever
7	สาเหตุ ไข้เลือดออก	Cause of hemorrhagic fever
8	สถานการณ์ โรค ไข้เลือดออก	Situation of hemorrhagic fever disease
9	สถานการณ์ ไข้เลือดออก	Situation of hemorrhagic fever
10	ไข้เด็งกี่	Dengue fever
11	ไข้เลือดออกช็อค	Hemorrhagic fever with shock (Dengue shock syndrome)
12	โรคไข้เลือดออกช็อค	Hemorrhagic fever disease with shock (Dengue shock syndrome)
13	เกล็ดเลือดต่ำ	Low platelet
14	ไข้เลือดออกระบาด	Hemorrhagic fever outbreak
15	โรคนำโดยยุงลาย	Aedes mosquito-borne disease

Table 1.1. Google Search terms and their English translations.

results in terms of prediction accuracy and excited a vision of using more creative "Big Data" sources in influenza forecasting. However, in 2013, these models failed dramatically, missing key incidence peaks by 140% and overestimating CDC estimates two-fold [12, 4]. These breakdowns resulted in Google abandoning the project altogether.

The ARGO model serves as a modification to GFT, and has been shown to make more robust and adaptable estimates in influenza predictions [23, 13, 25]. This work incorporates an autoregressive times series (AR) and Google search data (GO) into a regression model penalized with a least absolute shrinkage and selection operator (LASSO) constraint. This model addresses the limitations of GFT by (1) using a less static approach which instead evolves as new data is available, (2) not aggregating search terms into a single variable, and (3) including time series properties such as seasonality.

The main inspiration for this study is to adapt the methods proposed in the ARGO model to fit our modeling challenge. The two main differences arise from the need for (1) real-time predictions and (2) a way to accommodate situations without fully reported data. Past efforts have investigated using ARGO model in dengue prediction showing that using search data can help disease tracking in Mexico, Brazil, Thailand, Singapore, and Taiwan [22]. These analyses however, are not focused on making real-time predictions. The model as previously used, relies on the assumption of up-to-date and accurate case counts. This assumption is even a requirement of the software, which is meant to be refit to essentially complete data at every timestep. This functionality therefore is only equipped to make estimates for current levels

of infectious disease, with no obvious way to make forecasts further into the future [21]. In Thailand, we cannot trust that the most recent case count data is accurate and so, a model which needs to be refit to new data frequently is not practically useful. For our purposes, we need a model which can be fit once a year on only the most revised data and that is able to make predictions for multiple timesteps into the future. For these reasons, we set out to modify and extend this original ARGO model to accommodate situations when timely surveillance data is not available.

This study focuses on the impact of using such internet data to enhance forecasts when surveillance data is imperfect. Our search data consists of frequencies of 15 terms related to dengue fever each potentially related to DHF trends (Table 1.1). Our analyses are focused on two provinces in Thailand, Bangkok and Chiang Mai, both which exhibit different disease dynamics. We incorporate province-level search data with seasonality components to investigate whether short-term forecasts can be improved.

CHAPTER 2

METHODS

2.1 The Data

Dengue hemorrhagic fever (DHF) case count data comes from the Thai Ministry of Public Health. This includes the reported number of DHF cases in the provinces of Bangkok and Chiang Mai from 1968 to the present. Because disease dynamics are different in each province, a separate model is fit for Bangkok and for Chiang Mai. Trends in this data from 2009 onward can be seen in Figure 1.1. These counts begin underreported and are updated in the following months, with a final report of all cases in one calendar year delivered in April of the subsequent year. This means that accurate data for a biweek may not be available until April of the following year when final revisions are made. We fit models to make biweekly predictions for DHF on the log scale.

Google Search data was collected with help from collaborators at Harvard University with special access to the Google Trends API . This data is available in realtime. Based on random samples of the all of the world's Google searches, the site provides province-level "interest" estimates on specific topics. Here interest is defined as the "proportion of all searches on all topics in Google in that same place and time" [18]. We began with information on 15 commonly queried Thai terms relating to

DHF sampled from each province, Bangkok and Chiang Mai. These terms were chosen based on previous modeling efforts and suggestions from our collaborators [22]. Their translations can seen in Table 1.1. In Bangkok, 5 of these terms contained all 0s resulting in 10 terms used in our analysis. In Chiang Mai, only 4 terms were used also due to too many 0s. Some of their distributions are shown in Figure 1.1.

2.2 Prediction Time Unit

Dengue has a generation time (or time between two consecutive generations) of two weeks. For this reason, we mapped all data into biweekly intervals. The first biweek of every year begins on January 1st, at 00h00m00s and the last biweek ends on December 31st, at 11h59m59s. A more explicit definition of this time scale can be found in previous work with this same data [16].

2.3 The Baseline Models

2.3.1 Seasonal Average

The primary model we use as a reference is a **Seasonal** model. This model has a separate fixed effect for each biweek in the season combined in a simple linear model. Using this model as a reference allows us to compare each predicted value to the historical mean at that timestep. These comparisons will help us to determine if our more complicated models make predictions better than the seasonal mean. This model will serve as our baseline for all comparisons and is defined as

$$y_t = \text{Log-transformed DHF case counts in biweek t}$$

t = biweekly timestep of interest

$$y_t = \mu + \sum_{b=1}^{B} \phi_b I(biweek = b) + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

with $B = 25$ biweek indicators. The 26th biweek is included in the intercept term, μ which is the average logged DHF cases for biweek 26.

2.3.2 ARGO Model

As explained previously, the original ARGO model is only designed to make 1-step ahead predictions, refitting the model each week when additional data is received. However, because the data from Thailand is substantially underreported, we do not consider the data from the most recent biweeks as reliable. For this reason, we instead only use the revised data from the previous year, and discard data from the most recent past. Instead of considering the incoming data for new prediction models, we want a model which can be fit once a year to fully reported data. This model would then be used to create forecasts at each timepoint for the next year, until new fully reported data is available.

The original ARGO model uses a combination of past case counts and search data as regressors. This model is given by

$$y_t = \mu + \sum_{j=1}^{L} \alpha_j y_{t-j} + \sum_{k=1}^{K} \beta_k X_{k,t} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

with y_t the timeseries of case counts with L lags and X_t a vector of K exogenous variables, or in this case, search terms at time t.

2.3.3 Baseline SeaGO

In Thailand, delays in reported DHF case data often mean than an autoregressive model has limited practical utility for forecasting into the future. More explicitly, at certain timesteps in our challenge we do not have complete information. Therefore rather than including an "AR" term, we instead include a "Sea" component of seasonal indicators. We propose the following **SeaGO** model

$$y_t = \mu + \sum_{k=1}^{K} \beta_k X_{k,t} + \sum_{b=1}^{B} \phi_b I(biweek = b) + \epsilon_t \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

with X_t a vector of K search terms and $I(biweek = b)$ indicator variables for the biweek at time t. In our specifications in Bangkok, we consider $K = 15$ Google search terms and $B = 25$ biweeks, with the 26th biweek being absorbed into the intercept term. Because this is a relatively large number of predictors, ordinary least squares maximum likelihood estimation may lead to overfitting. Thus, to optimize this equation, we constrain our coefficient estimates using an L_1, or LASSO penalty [19]. The LASSO finds $\beta = \{\beta_k\}$ and $\phi = \{\phi_b\}$ to minimize

$$\sum_{i=1}^{N} \left(y_i - \sum_{k} x_{i,k}\beta_k - \sum_{b} \phi_b I_{i,b} \right)^2 + \lambda \left[\sum_{k=1}^{K} |\beta_k| + \sum_{b=1}^{B} |\phi_b| \right]. \tag{4}$$

with $N =$ the number of training observations. This is equivalent to minimizing the residual sums of squares with the constraint $\sum |\beta| + \sum |\phi| \leq s$. This penalty shrinks regression coefficients towards and sometimes exactly to 0.

2.4 Extensions of the SeaGO

2.4.1 Lagged SeaGO

These models also consider multiple past lag values of these search terms, meaning the $\sum_{k=1}^{K} \beta_i X_{k,t}$ in Equation 3 can be replaced with $\sum_{k=1}^{K} \sum_{j=0}^{L} \beta_{k,j} X_{k,t-j}$. The full model can be written as

$$y_t = \mu + \sum_{k=1}^{K} \sum_{j=0}^{L} \beta_{k,j} X_{k,t-j} + \sum_{b=1}^{B} \phi_b I(biweek = b) + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{5}$$

After consulting autocorrelation plots for each search term, we determined that each term is autocorrelated with up to 4 lags. Therefore, we decided to use $L = 4$ as the possible number of lags to be considered in this model. This is the maximum number of lags that are *possibly* correlated with case counts allowing the LASSO algorithm to decide which should remain.

2.4.2 Group SeaGO

One downfall of using a LASSO penalty is that it may not perform well when many predictors are highly correlated with each other. In cases of such multicollinearity the algorithm arbitrarily places all of the weight on 1 of the predictors from this group [7]. All other correlated predictors are shrunk to 0.

In our data, many of the predictor trends arise from similar signals. For example, our biweek indicators are directly related with each other. **Term 1** and **Term 3** can be translated to "hemorrhagic fever disease" and "hemorrhagic fever" which would reasonably be queried either at the same time or with similar trends. They

have a correlation coefficient of 0.75. These terms are also, by a function of our choosing, correlated with each of their lag terms. For example, correlation coefficients between search terms and first lagged values range between 0.46 and 0.96. In order to accommodate these correlations, we investigate the use of the Group LASSO [24].

The Group LASSO is very similar to the optimization problem seen in Equation 4, but it instead applies the penalty

$$\lambda \sum_{h=1}^{G} m_h ||\theta_h||_2 \tag{6}$$

with $G =$ the number of groups considered and $||\theta_h||_2$ the L_2 norm. Here, θ_h is a vector of coefficients from group h where $\bigcup_h \theta_h = \{\beta, \phi\}$ and θ_h is disjoint. $k_h =$ the number of predictors in group h and m_h serves as a scalar to account for differences in group size. $m_h = \sqrt{k_h}$ and $\sum_h k_h = p =$ the total number of predictors considered in the model. This specification of m_h means that the same amount of penalty is applied to groups with different numbers of predictors.

This optimization allows members of a group to share a LASSO penalty. This means that either all members of each group are 0 or all are non-0. By comparing the coefficients from these group LASSOs, we can determine if a *group* of predictors is helpful in our predictions.

In total, we consider $P = 75$ predictor variables in our modeling in Bangkok. These consist of $B = 25$ seasonal indicators, $K = 10$ search terms, and $L \cdot K = 40$ lagged search terms. Due to the complex collinearity scheme existing between these independent variables, there are a few different grouping approaches which could reasonably be employed as described below.

Option 1: Seasonal Grouping (SeaGO-GrSeas)

This first scheme groups together all $B = 25$ seasonal indicator variables and leaves the remaining variables alone, each in a group of 1. In Bangkok, G = 1 + 50 = 51. This finds the optimal values of β and ϕ to

$$minimize\left(||y - X\beta - I\phi||^2 + \lambda \sum_{h=1}^{G=51} m_h||\theta_h||_2\right). \tag{7}$$

Option 2: Seasonal and Search Grouping (SeaGO-GrSearch)

This scheme groups together all $B = 25$ seasonal indicators and groups together each of the $K = 10$ search terms with their lag terms. For this scheme in Bangkok, G = 1 + 10 = 11.

Option 3: Seasonal and Lag Grouping (SeaGO-GrLag)

This scheme groups together all $B = 25$ biweek indicators and groups together all search terms with terms of the same lag. Here, G = 1 + 5 = 6.

Option 4: Seasonal and Correlated Grouping (SeaGO-GrCorr)

This approach groups together the $B = 25$ binary biweek variables and creates 3 groups based on which predictors are most highly correlated. Here, G = 1 + 3 = 4.

2.4.3 Adaptive SeaGO

Another extension of the LASSO often considered is the Adaptive LASSO [26]. This algorithm further reduces LASSO's bias by penalizing large coefficients and thus further prevents overfitting. This method applies the penalty $\lambda \sum_{h=1}^{P} \hat{w}_h|\theta_h|$

with P = the number of predictors considered and $\hat{w}_h = \frac{1}{\hat{\theta_h}^{\text{initial}}}$ where $\hat{\theta_h}^{\text{initial}}$ are *pilot estimates* obtained from a preliminary LASSO run. This finds the argument to

$$minimize\left(||y - X\beta - I\phi||^2 + \lambda \sum_{h=1}^{P} \hat{w}_h|\theta_h|\right) \tag{11}$$

The Adaptive LASSO also has the benefit of the oracle properties [26]. These include (1) the ability to correctly identify the best subset of predictors and (2) that it provides the optimal estimation rate.

2.5 Principal Components Regression

The final approach we investigate is the popular dimension reduction method of Principal Components Regression (PCR) [9]. First, Principal Components Analysis is performed to identify low-dimensional projections of the data in the form of linear combinations of the available variables. The optimal number of components to be included is determined through cross-validation. Our best models chose 43 and 45 components on average. Dimension reduction is then performed by only using the principal components identified through PCA. These linear combinations are then used as regressors in least squares regression. This method avoids the many problems associated with multicollinearity and decreases the amount of overfitting. However, because this model considers linear combinations of the total number of P predictors as regressors, it performs dimension reduction, but not feature selection. Therefore, this method can be very challenging to interpret, and conclusions about specific predictors are much more difficult to draw than from LASSO methods.

2.6 Multiple Prediction Horizons

Here, we define a *horizon* as the number of time steps into the future we hope to predict. To make predictions for multiple time steps into the future, models are trained on search term data shifted back in time based on the horizon of interest. An illustration of this process is shown in Equation 12, for a two-step ahead forecast.

$$y_{t+2} = \mu + \sum_{i=1}^{K}\sum_{j=0}^{L}\beta_{i,j}X_{i,t-j} + \sum_{l=1}^{B}\phi_{l-2}I(biweek = l) + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0,\sigma^2) \tag{12}$$

This way, a new model is fit with different coefficients for the specific modeling challenges of making predictions in each horizon and each province.

2.7 Cross Validation

Each of the methods outline previously have built-in cross validation approaches in order to determine which parameterizations are optimal. For each LASSO method, the best value of λ is determined by 10-fold cross validation over a sequence of possible values. The optimal value of λ corresponds to the value with the lowest mean-squared out-of-sample prediction error [3]. For principal components methods, the optimal number of components is determined through 10-fold cross validation considering 1 through P possible components. The optimal value corresponds to the lowest bias corrected mean-squared prediction error [15].

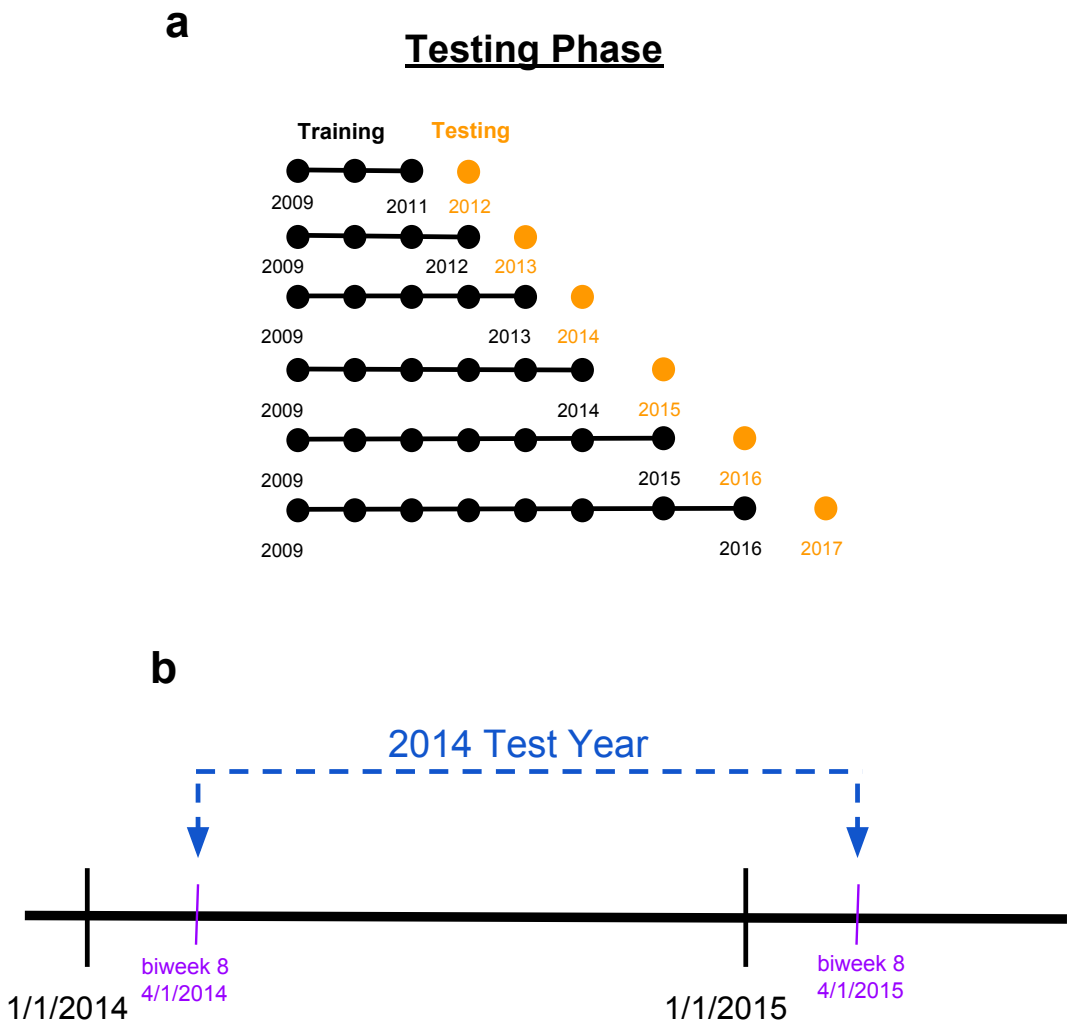

Figure 2.1. Illustration of testing scheme. **a.** Yellow dots represents years of data used for our test sets. Back dots represent years of data included in the training sets. **b.** As an example, this illustrates our definition of the test year for 2014.

2.8 Validation

To test the performance of each of these models, we compare their errors in prospective out-of-sample predictions. Each model was fit on a set of training data and used to make predictions on a subsequent set of test data (Figure 2.1 a.). We chose 6 test years between 2012 and 2017. Our training data consisted of all data through December of the year preceding the test year. The test sets began with data in April of the test year through March of the following year (Figure 2.1 b.). For each test year, prediction error was calculated using root-mean-square error (RMSE).

Previous studies have suggested that models perform better with a 2-year sliding window of training data [23]. This is because internet search behaviors change frequently and trends from 3 years in the past make not be indicative of behaviors today. For these reasons, we included 2 additional models with this 2-year sliding window. These models employ the window on the **SeaGO Lag** and **Adaptive SeaGO** and are referred to as the **SeaGO Sliding** and **A. SeaGO Sliding**, respectively.

CHAPTER 3

RESULTS

3.1 Including Search Data to Improve Short-Term Forecasts

We found that a simple model including Google Search data (**SeaGO**) improves case count predictions over a seasonal average model (**Seasonal**). Predictions for the years 2014 and 2015 are shown in (Figure 3.1). As can be seen in (Table 3.2) neither model outperformed the other in each location and in each year. However in both locations, the **SeaGo** model had better average prediction accuracy. In Bangkok, the **Seasonal** model performed best in 2013, the year with highest total annual incidence. However, the **SeaGO** model performed best in 2015, another year with high incidence. In Chiang Mai, the **SeaGO** model performed better in 2013, the year with the highest number of reported cases. We can conclude that including search term data in our models improved forecasts in 8 of 12 province-years and did better on average than a seasonal average model.

3.2 Extensions of SeaGO

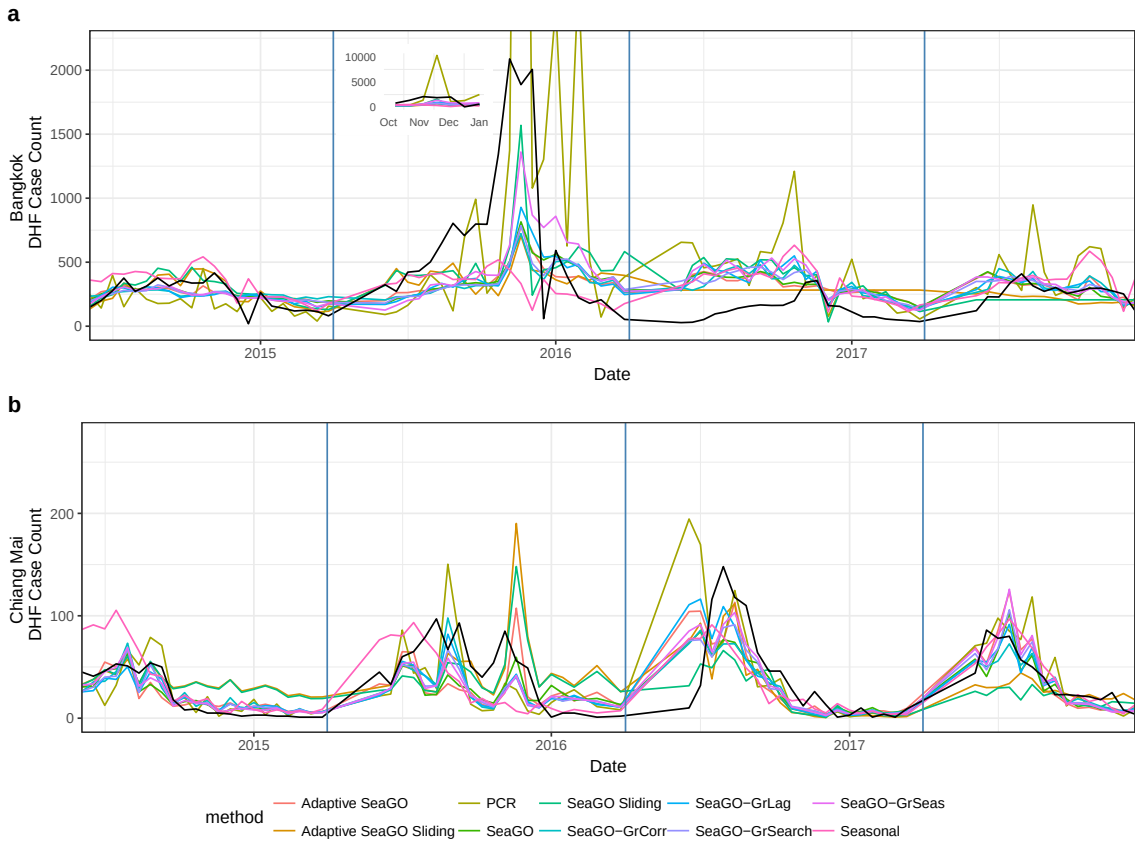

Figure 3.1. All model predictions across 4 test years. This figure shows 1-biweek ahead predictions across the 4 test years. The black line represents the fully-reported case counts from the MoPH. Blue lines delineate breaks in test data which occur on April 1st of each year. At this point, we expect that data through December of the previous year is fully reported.

Bangkok

	Seasonal	SeaGO	Lagged SeaGO	Adaptive SeaGO	SeaGO GrSeas	SeaGO GrSearch	SeaGO GrLag	SeaGO GrCorr	SeaGO Sliding	A. SeaGO Sliding	PCR
2012	276.78	268.23	261.22	260.17	259.19	261.10	261.60	260.69	**253.93**	253.99	4942.49
2013	**441.71**	485.75	486.01	455.19	487.23	500.51	489.27	477.20	576.95	593.10	859.37
2014	126.78	92.03	85.15	79.48	77.62	88.50	86.80	95.66	71.63	**62.67**	133.83
2015	683.34	591.13	625.63	600.77	**544.91**	623.79	581.10	625.56	604.02	649.61	1972.93
2016	244.12	224.75	219.62	223.26	235.44	213.89	231.27	225.26	276.16	**179.57**	396.03
2017	146.73	99.28	88.06	93.29	94.21	**78.32**	95.45	95.74	93.12	92.18	243.45
Mean	319.91	293.53	294.28	**285.36**	**283.10**	294.35	**290.91**	296.69	312.64	305.19	1424.68

Chiang Mai

	Seasonal	SeaGO	Lagged SeaGO	Adaptive SeaGO	SeaGO GrSeas	SeaGO GrSearch	SeaGO GrLag	SeaGO GrCorr	SeaGO Sliding	A. SeaGO Sliding	PCR
2012	25.78	33.79	28.60	39.97	28.34	25.79	54.14	32.14	**22.79**	32.43	132.57
2013	259.59	219.12	220.65	**207.15**	219.06	228.92	240.43	243.27	284.20	283.56	216.30
2014	22.79	13.00	11.15	12.75	8.71	10.63	**10.57**	10.90	20.21	18.30	13.96
2015	33.38	32.49	33.96	35.66	31.23	30.44	**29.80**	30.02	35.41	37.72	35.69
2016	32.98	38.63	32.70	36.65	30.15	**29.41**	35.07	34.84	39.28	33.63	57.01
2017	**12.93**	17.14	16.27	16.88	18.81	16.46	13.30	13.97	28.14	23.10	23.67
Mean	64.58	59.03	**57.22**	58.18	**56.05**	**56.94**	63.88	60.86	71.67	71.45	79.87

Table 3.1. Root-mean-square prediction errors across all models and years.

To find the method for best forecasting dengue cases we further compared this **SeaGO** model against extensions of LASSO approaches and a model using principal components regression. In Bangkok the **Adaptive SeaGO**, **SeaGO-GrSeas**, and **SeaGO-GrLag** models did better on average than other models (Table 3.2). In Chiang Mai, the **Lagged SeaGO**, **SeaGO-GrSeas**, and **SeaGO-GrSearch** performed the best on average. In both provinces, **PCR** methods did not seem to provide an improvement over the **Seasonal** model on average. All SeaGO extension methods showed an improvement over the **Seasonal** baseline in at least one year and one province. In both provinces, at least 1 SeaGO model outperformed the **Seasonal** model in 5 of the 6 test years.

As shown in (Figure 3.1) for Bangkok in 2015, the LASSO extension models were able to anticipate the late season spike in case counts when the **Seasonal** model was not. In this same year in Chiang Mai, these models incorrectly predicted a similar peak (Figure 3.1). In 2016 in Bangkok, all models seem to over-predict the number of DHF in early months. This makes sense because the last data used to train those models was through December of 2015, when the number of reported cases rocketed to 2,000.

No model achieved the lowest error in every season, highlighting the impact of season-to-season variation on model performance. In Bangkok, all 9 **SeaGO** models outperformed the **Seasonal** model in 5 of the 6 test years and also on average. Sliding-window models had the lowest prediction error in 3 of the 6 test years. In Chiang Mai, all SeaGO extensions, except for the sliding-window models, outperformed the **Seasonal** baseline model on average. Sliding-window models performed

better than the **Seasonal** model in only 3 of 12 province-years. In both Bangkok and Chiang Mai the sliding SeaGO and Adaptive SeaGO models did worse than their non-sliding counterparts.

We can conclude that in Bangkok the **Adaptive SeaGO**, **SeaGO-GrSeas**, and **SeaGO GrLag** models had the lowest prediction error in our testing phase and we recommend them for future forecasting efforts. We recommend the **Adaptive SeaGO**, **SeaGO-GrSeas**, and **SeaGO-GrSearch** models for future efforts.

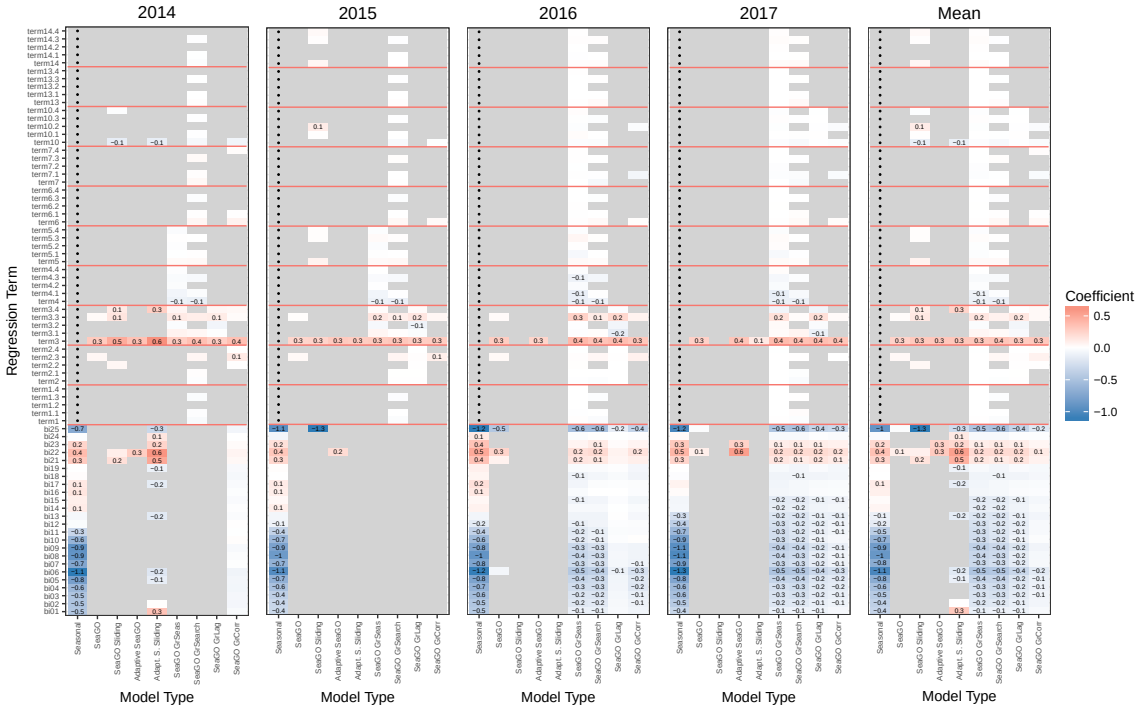

Figure 3.2. Regression coefficients in Bangkok. This plot shows regression term coefficients for each model in each test year in Bangkok. Labels are provided for coefficients with magnitude of at least 0.1. White cells represent coefficients close to, but not exactly 0. Gray cells represent the regression terms which were deselected out of the model during cross validation. The exception for this is the Seasonal baseline model, where search terms were not considered for modeling efforts.

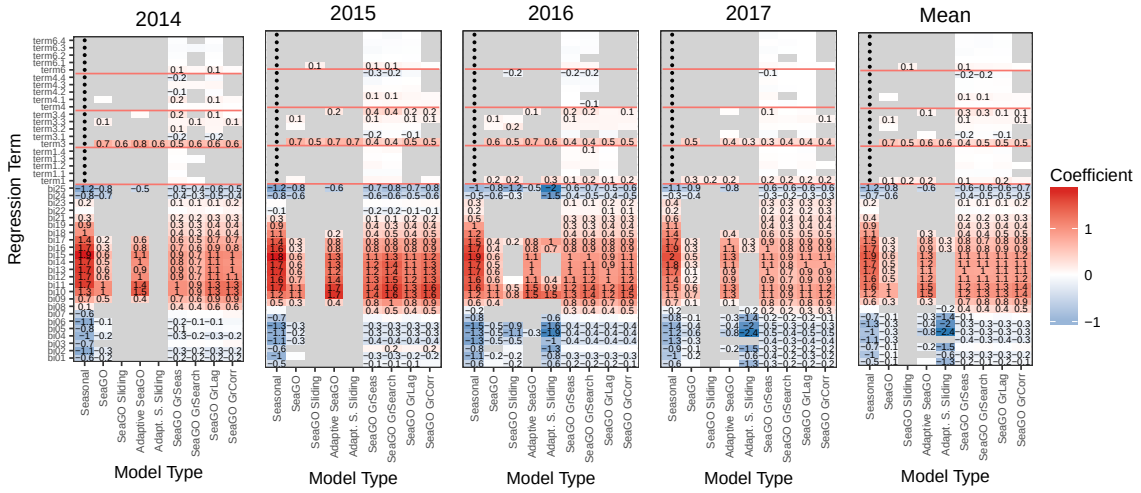

Figure 3.3. Regression coefficients in Chiang Mai. This plot shows regression term coefficients for each model in each test year in Chiang Mai. Labels are provided for coefficients with magnitude of at least 0.1. White cells represent coefficients close to, but not exactly 0. Gray cells represent the regression terms which were deselected out of the model during cross validation. The exception for this is the Seasonal baseline model, where search terms were not considered for modeling efforts.

Figure 3.2 shows the coefficients fit for each model in each year. Labels from these plots are the coefficients for each regression term and can be interpreted as the term's relationship with expected log-DHF cases. For example, in 2014 the **A. Sliding SeaGO** model found after that holding all else equal, being in the 22nd biweek is associated $e^{0.6} = 1.82$, or 82% more expected DHF cases than being in the 1st biweek of that year. This model also found that a one unit increase in the standardized frequency of Term 3 is associated with $e^{0.3} = 1.35$, or 35% more expected DHF cases, after controlling for other terms.

In 2014 and 2015, very few of the SeaGO extension models selected the seasonal predictors in their fits. Here, it is evident that the SeaGO extension models spread weight onto more regression terms than the **SeaGO** model. For example in 3.3 , the baseline **SeaGO** model put all weight on Term 3. However, SeaGO extension models (which account for multicollinearity) spread weight onto multiple search terms.

In Chiang Mai, much more weight was put on the seasonality terms than in Bangkok (Figure 3.3). Intuitively, this makes sense because Chiang Mai has stronger seasonal patterns. Bangkok is nearly "a-seasonal" meaning that cases are harder to predict and more weight is put on the Google data. Chiang Mai models also had a smaller subset of search terms to penalize because less search data was available in this province. Only search terms 1, 3, 4, and 6 had reasonable frequencies for our modeling.

"Hemorrhagic fever" (Term 3) was consistently found to be predictive of DHF case counts in both provinces. "Hemorrhagic fever disease" (Term 4), a close variant of Term 3, is similarly weighted into the models across method types. "Dengue fever"

(Term 10) was not found to be predictive of DHF in any model. This is reasonable, because the term "dengue fever" is often not used in Thailand, whereas "hemorrhagic fever" is much more common for the same illness. More technical terms such as "Low Platelet" (Term 13) and "Aedes mosquito-borne disease" (Term 15) were also not included in any models. These terms were not selected in Bangkok and were not searched with enough frequency to be used in our models in Chiang Mai.

3.3 Multiple Prediction Horizons

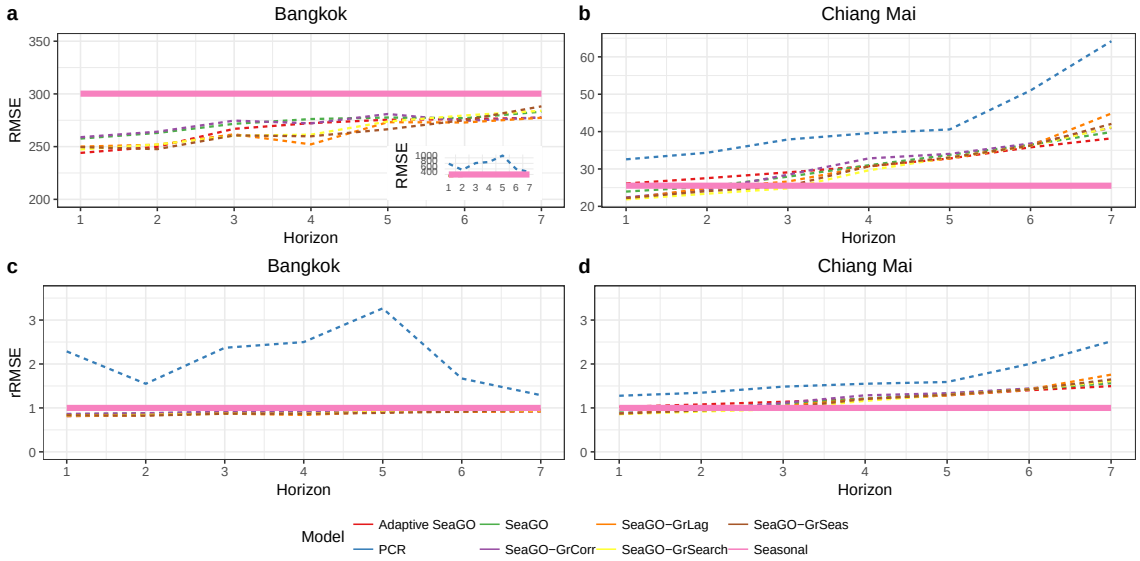

Figure 3.4. Prediction error for all models across multiple horizons. **a.** Average RMSE for each model prediction up to 7 horizons into the future. **b.** RMSE relative to the **Seasonal** model. All models below the pink line at 1 have lower error than the **Seasonal** baseline model.

These analyses were replicated for making predictions multiple timesteps into the future. As can be seen in Figure 3.4 **a.**, Bangkok models perform similarly across multiple prediction horizons. **SeaGO** models continue to outperform the **Seasonal** model for all horizons. For example, the **Adaptive SeaGO** model had 18.3%, 9.5%, and 6.2% less error than the **Seasonal** model at 1, 3, and 7 horizons into the future, respectively. Even at 7 horizons into the future, all **SeaGO** models have less prediction error that the **Seasonal** model.

In Chiang Mai, **SeaGO** models perform better than the **Seasonal** model for the 1st horizon, however by the 3rd horizon, their predictions are worse than a seasonal mean. For example, the **SeaGO-GrCorr** model has 13% less error than the **Seasonal** model at 1 horizon, but has 8.9% more error for predictions made for 3 horizons into the future. By 7 horizons into the future, the **SeaGO-GrCorr** model has 60.7% more error than the **Seasonal** model. From this we can conclude that **SeaGO** models are useful in predicting 1 horizon into the future, but for further targets in Chiang Mai, a **Seasonal** model may be more accurate.

CHAPTER 4

DISCUSSION

We have shown that real-time internet data can be used in the place of underreported case count time series. On average our SeaGO models which incorporate Google Search data outperform seasonal average models in both provinces. In Bangkok, these models have lower average prediction error than a seasonal model for up to at least 7 horizons into the future. In Chiang Mai, these models perform better for short-term forecasts, but for making predictions 3 or more horizons into the future, a seasonal model is preferable.

It is evident that there is not one modeling approach which uniformly outperforms all others. Certain models are better able to adapt to a province's unique disease dynamics. This further substantiates the need for fitting different models for different provinces in Thailand. Different methods work better depending on the provinces and years they are fit for. This means it is important to investigate which approach works best in each location and to not generalize findings from one location to another.

Predicting similar behaviors between provinces is unsurprising because our search term behavior is often indicative of national trends. Therefore, when a there is a large influx in search term frequency our models react similarly but, in different

magnitudes. In December of 2015, a beloved Thai actor Por Sahawong, died in Bangkok after developing dengue shock syndrome. His illness sparked national fear in the disease and resulted in surges in searches related to both him and dengue [11]. Although his illness was centered in the Bangkok outbreak, people all over Thailand were interested in his story, meaning search data at this time did not mirror province-specific behavior. Thus it is unsurprising, that in 2015 SeaGO models in Chiang Mai made predictions more suited for activity in Bangkok.

Results from 2015 show that these models still have susceptibility to spurious search traffic that is not necessarily related to symptomatic illness. Unfortunately, it is difficult to reproduce the GFT models in order to make a comparison with our results. It is reasonable to believe though that our models may do better in times with misleading amounts of search traffic. This may be because GFT uses a single variable for the fraction of ILI-related search queries at a given time while the SeaGO models do not aggregate all search information into one predictor [8]. Instead, the SeaGO models treat each search term separately, perhaps allowing for a more nuanced accommodation of internet trends. These models also include a seasonality component. More work is needed to evaluate how these models are sensitive to unrelated search behaviors and what methods can be used to combat those issues.

This study leaves many doors open for future work. Successes with the moving-window models in Bangkok suggest further investigation is promising. An ensemble of the SeaGO models may provide smoother and more accurate predictions. Another area of interest would be investigating probabilistic predictions in addition to our point estimates.

The algorithms utilized for both the LASSO and PCR methods used mean-squared errors to choose optimal parameterizations. For this reason, we chose out-of-sample prediction RMSE as our main metric for model performance. Mean absolute error is a similar metric, however it is though to be is less sensitive to outliers. We replicated these same procedures instead using MAE and our results were much the same. Future work may benefit from further investigating this relationship between parameter optimization and prediction error metrics.

Our work has been able to extend the ARGO model to make real-time predictions in settings with imperfect surveillance. We have shown that internet trends can be a powerful alternative source of data. With careful modeling approaches this data can improve predictions otherwise limited by underreporting. These approaches can likely be applied to many different model challenges, infectious disease or otherwise.

BIBLIOGRAPHY

[1] Bhatt, Samir, Gething, Peter, Brady, Oliver, Messina, Jane, Farlow, Andrew, Moyes, Catherine, Drake, John, Brownstein, John, Hoen, Anne, Sankoh, Osman, MYers, Monica, George, Dylan, Jaenisch, Thomas, Wint, William, Simmons, Cameron, Scott, Thomas, Farrar, Jeremy, and siman Hay. The global distribution and burden of dengue. Nature 496 (2013), 504–7.

[2] Brady, Oliver, Gething, Peter, Bhatt, Samir, Messina, Jane, Brownstein, John, Hoen, Anne, Moyes, Catherine, Farlow, Andrew, Scott, Thomas, and Hay, Simon. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. Public Library of Science Neglected Tropical Diseases 6 (2012), 1–15.

[3] Breheny, Patrick, and Huang, Jian. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Statistics and Computing 25 (2015), 173–187.

[4] Butler, Declan. When google got flu wrong. Nautre 494 (2 2013).

[5] Chareonsook, O, Foy, HM, Teeraratkul, A, and Silarug, N. Changing epidemiology of dengue hemorrhagic fever in thailand. Epidemiology and Infection 122 (2 1999), 161–6.

[6] Cook, Samantha, Conrad, Corrie, Fowlkes, Ashley, and Mohebbi, Matthew. Assessing google flu trends performance in the united states during the 2009 influenza virus a (h1n1) pandemic. Public Library of Science One 6 (8 2011).

[7] Friedman, Jerome, Hastie, Trevor, and Tibshirani, Rob. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33 (2010), 1–22.

[8] Ginsberg, Jeremy, Mohebbi, Matthew, Patel, Rajan, Brammer, Lynnette, Smolinksi, Mark, and Brilliant, Larry. Detecting influenza epidemics using search engine query data. Nature 457 (2 2009), 1012–14.

[9] Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2 ed. Springer, 2017, ch. Methods Using Derived Input Directions.

[10] Helft, Miguel. Google uses searches to track flu's spread. New York Times (11 2008).

[11] Herriman, Robert. Dengue takes the life of por sahawong, outpouring of tributes on twitter. Outbreak News Today (1 2016).

[12] Lazer, David, Kennedy, Ryan, King, Gary, and Vespignani, Alessandro. The parable of google flu: Traps in big data analysis. Science 343 (2014), 1203–1205.

[13] Lu, FS, Hou, S, Baltrusaitis, K, Shah, M, Leskovec, J, Sosic, R, Hawkins, J, Brownstein, J, Conidi, G, Gunn, J, Gray, J, Zink, A, and Santillana, M. Accurate influenza monitoring and forecasting using novel internet data streams: A case study in the boston metropolis. Journal of Medical Internet Research Public Health and Surveillance 4 (January 2018).

[14] Manheim, David, Chamberlin, Margaret, Osoba, Osonde A, Vardavas, Raffaele, and Moore, Melinda. Improving Decision Support for Infectious Disease Prevention and Control. RAND Corportation, 2016.

[15] Mevik, Bjrn-Helge, Wehrens, Ron, and Liland, Kristian Hovde. pls: Partial Least Squares and Principal Component Regression, 2016. R package version 2.6-0.

[16] Reich, Nicholas G, Lauer, Stephen A, Sakrejda, Krzysztof, Iamsirithaworn, Sopon, Soawapak, Hinjoy, Suangtho, Paphanij, Suthachana, Suthanun, Clapham, Hannah E, Salje, Henrik, Cummings, Derek AT, and Lessler, Justin. Challenges in real-time prediction of infectious disease: A case study of dengue in thailand. Public Library of Science Neglected Tropical Diseases (June 2016).

[17] Rigau-Perez, JG, Clark, GG, Gubler, DJ, Reiter, P, Sanders, EJ, and Vomdam, AV. Dengue and dengue haemorrhagic fever. Lancet 352 (9 1998), 971 – 7.

[18] Rogers, Simon. What is google trends data - and what does it mean? Google News Lab (July 2016).

[19] Tibrshirani, Robert. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B 58 (1996), 267–288.

[20] World Health Organization. Dengue and severe dengue, 2 2018.

[21] Yang, Shihao. argo: (ARGO) AutoRegression with GOogle search data - accurate estimation of influenza epidemics, 2015. R package version 1.0.

[22] Yang, Shihao, Kou, Samuel C., Lu, Fred, Brownstein, John S., Brooke, Nicholas, and Santillana, Mauricio. Advances in using internet searches to track dengue. Public Library of Science Computational Biology (July 2017).

[23] Yang, Shihao, Santillana, Mauricio, and Kou, S.C. Accurate estimation of influenza epidemics using google search data via argo. Proceedings of the National Academy of Sciences of the United States of America 112, 47 (11 2015), 14476–78.

[24] Yuan, Ming, and Lin, Yi. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society B 68 (2006), 29–67.

[25] Zhang, Y, Bambrick, H, Mengersen, K, Tong, S, and Hu, W. Using google trends and ambient temperature to predict seasonal influenza outbreaks. Environment International 117 (August 2018), 284–291.

[26] Zou, Hui. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (2006), 1418–1429.