UM MODELO ESPAÇO-TEMPORAL BASEADO EM PROCESSOS GAUSSIANOS PARA PREVISÃO DE INCIDÊNCIA DE DENGUE

JULIO ALBINATI CORTEZ

UM MODELO ESPAÇO-TEMPORAL BASEADO EM PROCESSOS GAUSSIANOS PARA PREVISÃO DE INCIDÊNCIA DE DENGUE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: GISELE LOBO PAPPA

Belo Horizonte, Minas Gerais

Janeiro de 2017

JULIO ALBINATI CORTEZ

A SPATIO-TEMPORAL GAUSSIAN PROCESS-BASED MODEL FOR FORECASTING DENGUE FEVER INCIDENCE

Dissertation presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

ADVISOR: GISELE LOBO PAPPA

Belo Horizonte, Minas Gerais

January 2017

.

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Cortez, Júlio Albinati.

C828s A spatio-temporal gaussian process-based model for forecasting dengue fever incidence. / Júlio Albinati Cortez. – Belo Horizonte, 2017.
xxvi, 99 f.: il.; 29 cm.

Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientadora: Gisele Lobo Pappa.

1. Computação – Teses. 2. Dengue. 3. Processos gaussianos. 4. Teoria da previsão. 5. Early warning system 6. Multi-task regression. I. Orientadora. II. Título.

CDU 519.6*61 (043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

A Spatio-temporal gaussian process-based model for forecasting dengue fever incidence

JÚLIO ALBINATI CORTEZ

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. GISELE LOBO PAPPA - Orientadora
Departamento de Ciência da Computação - UFMG

PROF. ANA PAULA COUTO DA SILVA
Departamento de Ciência da Computação - UFMG

PROF. FLAVIO VINICIUS DINIZ DE FIGUEIREDO
Departamento de Ciência da Computação - UFMG

PROF. WAGNER MEIRA JÚNIOR
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 16 de janeiro de 2017.

Agradecimentos

Primeiramente, gostaria de agradecer à minha esposa, que me acompanhou (literalmente) nesse percurso. Creio que podemos dizer que, em retrospectiva, os finais de semana em casa fazendo listas de exercícios ou TPs valeram a pena. Agradeço também à minha família, que me inspiraram desde muito cedo a me dedicar aos estudos. É o mal de ter pais professores...

Aos companheiros do laboratório *Speed*, em particular Vinícius, Paulo, Samuel, Derick, Denise, Walter, Osvaldo e Elverton (e também o Rubens, apesar da trairagem), deixo também meus agradecimentos. Tudo seria muito mais triste sem os momentos de descontração proporcionado por vocês. Podemos não ser o laboratório mais organizado ou o mais produtivo, mas duvido que exista outro mais acolhedor!

À minha orientadora, Gisele, agradeço por toda a confiança depositada na minha pessoa e por me acompanhar desde o início da minha graduação. Obrigado pelas portas abertas, e realmente espero que aproveite muito essa nova fase da sua vida! Ao professor Wagner Meira, também deixo os meus agradecimentos pelos conselhos e pelas oportunidades concedidas.

Não posso deixar de mencionar a secretaria do PPGCC, em particular Sônia, Linda e Sheila. Sem dúvidas, sem vocês o departamento estaria imerso no mais completo caos. Parabéns pelo excelente trabalho em manter tudo nos seus devidos eixos!

Por fim, agradeço ao PPGCC como um todo pela oportunidade de realizar meu mestrado. Sei que é um privilégio poder fazer um mestrado em um programa de prestígio e de forma gratuita. Espero um dia poder retornar a todo o mundo o que em mim foi investido.

Abstract

Dengue fever is a mosquito-borne disease present in all tropical zones of the world and affects almost 400 million people worldwide every year. Having no treatment nor vaccines available for public use, dengue fever can only be controlled by suppressing the vector population and quickly identifying outbreaks through the usage of accurate predictive models capable of estimating the number of dengue cases in a given area and period of time.

Brazil is responsible for the largest number of confirmed cases in the Americas, accounting for at least one fourth of the total number of cases in the continent. Motivated by this scenario, the main objective of this work is to develop a new model for predicting dengue fever incidence (DIR) in Brazilian cities. For that, we explored the non-parametric Bayesian framework of inference under Gaussian processes (GP), which lie in the intersection between interpretable and state-of-art machine learning modelling frameworks.

The proposed model is a GP-based model equipped with a spatio-temporal covariance function. The temporal component exploits local dependences and seasonality, expressed through the form of a quasi-periodic covariance function. The spatial component, in turn, is defined by an inter-cities covariance matrix learned from data, without requiring human intervention or specification. We also proposed an extension of the model that is capable of incorporating online data, such as data from Twitter, to account for the typical delay in the propagation of epidemiological data in a more realistic scenario, where online data acts as a proxy for epidemiological data.

We conducted an extensive experimental analysis to assess the accuracy of the proposed model and extension, verifying that they outperformed alternative approaches, including a model specifically designed for forecasting DIR in Brazil. Our results were particularly expressive in the scenario where DIR values can be categorized into three incidence levels – low, medium and high incidence – where the proposed model achieve a median area under the ROC curve of more than 0.90, compared to 0.74 obtained by the best alternative model.

Resumo

Dengue é uma doença presente em todas zonas tropicais do mundo, afetando quase 400 milhões de pessoas ao redor do mundo todos os anos. Como não há tratamento ou vacinas disponíveis para o público geral, a dengue só pode ser contida através do controle populacional do mosquito transmissor do vírus e identificando rapidamente novos focos da doença através de modelos preditivos capazes de estimar, de forma acurada, o número de casos de dengue em uma determinada área e período de tempo.

O Brasil é responsável pelo maior número de casos confirmados de dengue nas Américas, atingindo mais de um quarto do número total de casos no continente. Motivado por esse cenário, o principal objetivo desse trabalho é desenvolver um modelo para predição de número de casos de dengue em cidades brasileiras. Para tanto, exploramos o framework não-paramétrico e bayesiano de inferência utilizando processos gaussianos, um método que reside na interseção entre modelos interpretáveis e estado-da-arte.

O modelo proposto é equipado com uma função de covariância espaço-temporal. O componente temporal explora dependências locais e sazonalidade, sendo expresso através de uma função quasi-periódica. Já o componente espacial é definido por meio de uma matriz de covariância entre cidades, que é aprendida com base nos dados apenas, sem nenhuma intervenção humana. Além disso, propusemos uma metodologia para extender o modelo proposto de forma a utilizar dados de fontes *online*, como dados do Twitter, no cenário mais realista onde os dados epidemiológicos são fornecidos com atraso. Assim, os dados *online* atuam como *proxy* para os dados epidemiológicos.

Conduzimos uma análise experimental extensiva para analizar a acurácia do modelo proposto, bem como a sua extensão para o cenário descrito acima. Verificamos que as propostas obtiveram predições mais acuradas quando comparadas a formulações alternativas, incluindo um modelo previamente proposto para previsão de incidência de dengue no Brasil. Nossos resultados foram particularmente interessantes no cenário onde os valores de incidência são categorizados em níveis de incidência – baixa, média ou alta –, onde o modelo obteve uma área sob a curva ROC mediana superior a 0.90, comparada à area de 0.74 obtida pela melhor formulação alternativa.

List of Figures

List of Tables

List of Algorithms

Contents

Chapter 1

Introduction

Dengue fever is a mosquito-borne viral disease transmitted by females mosquitoes of the species *Aedes aegypti*, the same mosquito that carries the viruses of zika, yellow fever and chikungunya. As any vector-borne disease, dengue fever is not transmitted directly between humans, requiring the presence of an infected vector. In the specific case of dengue fever, a mosquito (vector) that bites an infected human may acquire the virus. After virus incubation from 4 to 10 days, an infected vector is capable of transmitting the virus for the rest of its life [World Health Organization Media Centre, 2016].

Dengue fever is a severe flu-like illness and may cause high fever, strong headache, pain behind the eyes, muscles and joints, nausea, vomiting, swollen glands and rash, which may last for one week. There is no specific treatment for dengue fever nor available vaccines and, therefore, control of the disease can only be made by suppressing the vector population, as well as identifying outbreaks as quickly as possible [World Health Organization Media Centre, 2016]. Although case fatality rate can be as low as 1% with proper treatment, the disease comes with economical and social burden [Samir et al., 2013].

According to Samir et al. [2013], dengue fever is ubiquitous throughout the tropics. The study estimates that about 390 million people worldwide contract dengue fever every year, with 96 million of these cases being symptomatic. About 70% of the symptomatic cases are from Asia, with India alone contributing with 34% of the global number of cases. The Americas contributed with 14% of the symptomatic infections, of which over half occurred in Brazil and Mexico. Finally, the study estimates a similar situation for Africa when compared to the Americas. Figure 1.1 summarizes the global distribution of dengue fever.

In Brazil, dengue fever was eradicated in the first half of the 20th century, but

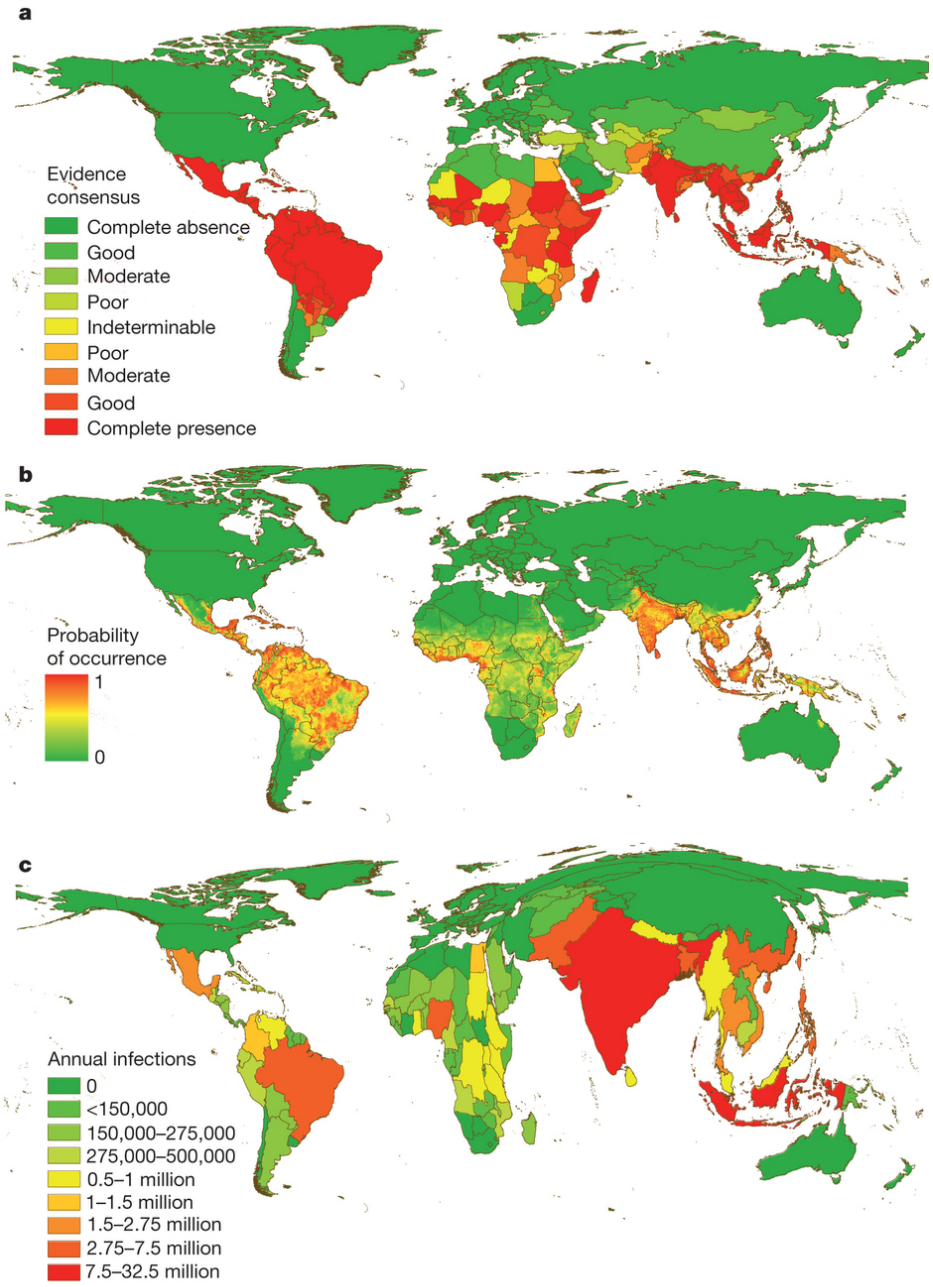

Figure 1.1. Global distribution of dengue fever, extracted from Samir et al. [2013]. The map on the top (a) shows the presence of dengue virus around the world. The map on the middle (b) shows the probability of occurence of dengue fever. The map on the bottom (c) indicates the number of dengue fever infections at country level, with areas proportional to the number of cases.

has been reintroduced in the 1970s, with the deactivation of surveillance and control of the disease's main vector, the mosquito *Aedes aegypti*. However, only after outbreaks in the city of Rio de Janeiro, in 1986, dengue fever became a major public health issue

[Teixeira et al., 2009]. Since then, the number of infections has increased significantly. The spread of the mosquito (and, consequently, of the disease) was facilitated by the country's appropriate climate [Teixeira et al., 2009] and population growth [Gubler, 2002]. Nowadays, dengue fever is endemic in most Brazilian regions and the country has the largest number of reported dengue fever cases in the Americas, accounting for at least one-fourth of the symptomatic cases in the continent [Samir et al., 2013]. In this context, providing tools capable of helping identifying the growth in the number of reported cases, allowing a quick response from the government, is essential to reduce the disease spread.

An *early warning system* (EWS) is a system capable of quickly identifying risks and plays a major role on disaster risk reduction by preventing loss of life and mitigating the economical and/or material impact [Wiltshire, 2006]. In the context of epidemiology, an EWS is a fundamental step on implementing effective interventions to control infectious diseases, thus reducing mortality and morbidity in human populations. However, identifying an epidemic is not an easy task, since it becomes evident only after a large number of infections have already happened and recorded. In this scenario, dengue fever is not an exception. In fact, the World Health Organization has indicated dengue fever as a disease for which better EWS are required [Kuhn et al., 2005].

According to Wiltshire [2006], an EWS is composed of four key elements: (i) risk knowledge, (ii) warning and monitoring services, (iii) dissemination and communication and (iv) response capability. The authors of Wiltshire [2006] also indicate three major questions that an appropriate warning system must answer affirmatively:

1. Are the right parameters being monitored?

2. Is there a sound scientific basis for making forecasts?

3. Can accurate and timely warning be generated?

Predictive models designed for forecasting dengue fever incidence typically use parametric models[1], such as generalized linear or ARIMA models, with covariates commonly associated to dengue fever, such as climate-related factors, urbanization and social-economical conditions (e.g., Martinez et al. [2011]; Lowe et al. [2013]; Shi et al. [2016]). In this sense, they answer affirmatively questions 1 and 2 listed above.

[1]Here, we define a parametric model as any model that is fully specified by a finite number of parameters *regardless* of the amount of data provided for training, as discussed by Rasmussen and Williams [2006]. On the other hand, nonparametric models are any model whose number of parameters grows with more training data.

However, these parametric models have limited complexity (defined by the number of parameters) and, consequently, may fail to capture complex patterns that could be exploited to enhance the accuracy of predictions. In order to assure accuracy, parametric models usually require careful analysis of data, leading to models that do not generalize to other contexts. Finally, epidemiological data typically takes time to be compiled and made available even for governmental authorities. Therefore, at time t, only epidemiological data associated up to time $t-\gamma$, $\gamma > 0$, is available. Most predictive models do not consider this fact, being forced to make predictions with much higher antecedence.

On the other hand, the machine learning community has proposed richer and more general non-parametric models which could be exploited to build epidemiological models. In the context of epidemiology, interpretability is an essential property, as indicated in question 2 above, making many state-of-art models infeasible. Motivated by this constraint, this work explores Gaussian processes [Rasmussen and Williams, 2006], a Bayesian non-parametric framework, to build predictive models for dengue fever incidence, as they lie in the intersection between state-of-art and interpretable models. Besides that, as previous works have already indicated associations between epidemiological data and online data sources [Gomide et al., 2011; Althouse et al., 2011; Souza et al., 2015], which can be obtained in real-time and used to fill the gap caused by delays on epidemiological data, we exploit Twitter data as a proxy for epidemiological data to build more accurate predictive dengue models.

1.1 Objectives

The main goal of this work is to develop a predictive model capable of generating warnings and working as a monitoring service within an EWS. We propose to explore Gaussian processes to generate predictive models capable of affirmatively answering all three questions raised above. We also propose a general framework for using online data to deal with delayed epidemiological data. In summary, our main objectives and associated contributions are:

- Characterization of weekly epidemiological dengue data with relation to temporal aspects, climate and weekly number of dengue-related tweets;

- Development of a spatio-temporal model based on Gaussian processes for forecasting the number of dengue cases on Brazilian cities;

- Development of a general framework for enhancing epidemiological models that mitigates the impact of delayed epidemiological data by using online data.

1.2 Thesis Organization

This thesis is organized as follows. In the next chapter, we give a more detail description on Gaussian processes and inference. In Chapter 3, we discuss some previous work on dengue fever modelling in Brazil and in other regions of the world. In Chapter 4 we present our dengue data collection and analyze temporal dependences, as well as climate and Twitter associations. Based on insights obtained through the characterization, we propose our model in Chapter 5. In Chapter 6, we conduct an extensive experimental evaluation to assess the accuracy of the proposed model and its variants. In Chapter 7, we propose our general framework for incorporating online data into epidemiological models. Finally, Chapter 8 concludes this work.

Chapter 2

Gaussian Processes

In this chapter, we introduce the Gaussian process modelling framework, as well as how it can be used for inference [Rasmussen and Williams, 2006]. Typically, there are two ways to interpret Gaussian processes (GPs): as a *distribution over functions* or as a *kernel machine*, that is, a *linear method applied over a set of points projected into a high-dimensional space*. Here we discuss both approaches, starting with the former and posteriorly the latter. We also discuss some covariance functions, hyperparameters optimization and provide a full pseudo-code for GP regression. Finally, we close this chapter with a discussion on extensions of the GP framework to the *multi-task learning* scenario, which will become useful later in this thesis.

2.1 Gaussian Process as a Distribution over Functions

We begin with the definition of a Gaussian process (GP) extracted from Rasmussen and Williams [2006]:

Definition 1 *A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

Under this interpretation, a GP can be seen as an extension from the multivariate Gaussian distribution to the space of functions. A GP is specified by a *mean function* $\mu(\boldsymbol{x})$ and a *covariance function* or (positive definite) *kernel* $k(\boldsymbol{x}, \boldsymbol{x'})$. Throughout this thesis, the term kernel will always refer to positive definite kernel. The terms kernel and covariance function will also be used interchangeably. We denote a GP as

$f \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'}))$, where

$$\begin{aligned} \mu(\boldsymbol{x}) &= \mathbb{E}[f(\boldsymbol{x})] \\ k(\boldsymbol{x}, \boldsymbol{x'}) &= \mathbb{E}[(f(\boldsymbol{x}) - \mu(\boldsymbol{x}))(f(\boldsymbol{x'}) - \mu(\boldsymbol{x'}))] \end{aligned} \tag{2.1}$$

and $\mathbb{E}$ denotes expectation.

In the particular scenario of regression, the random variables are defined as the evaluation of $f(\boldsymbol{x})$ for all $\boldsymbol{x}$ in the function's domain. Let X_{train} be a set of points for which we have observed evaluations of f, denoted as $\boldsymbol{f_{train}}$, and X_{test} be a set of points of interest for which we have not observed evaluations of f, denoted as the vector $\boldsymbol{f_{test}}$. Let us also define $\boldsymbol{f}$ as the vector resulting from concatenating $\boldsymbol{f_{train}}$ and $\boldsymbol{f_{test}}$. Using Definition 1, we have that

$$\boldsymbol{f}|X_{train}, X_{test} \sim \mathcal{N}\left(\begin{bmatrix} \mu(X_{train}) \\ \mu(X_{test}) \end{bmatrix}, \begin{bmatrix} K_{train} & K_{cross} \\ K_{cross}^T & K_{test} \end{bmatrix}\right) \tag{2.2}$$

where $\mathcal{N}$ denotes the multivariate Gaussian distribution, K_{train} indicates the covariance matrix between points in X_{train}, K_{test} indicates the covariance matrix between points in X_{test} and K_{cross} indicates the cross-covariance matrix between points in X_{train} and points in X_{test}.

Given that, we can calculate the conditional distribution of $\boldsymbol{f_{test}}$ given $\boldsymbol{f_{train}}$ using properties from the multivariate Gaussian distribution:

$$\begin{aligned} \boldsymbol{f_{test}}|\boldsymbol{f_{train}}, X_{train}, X_{test} &\sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}, \hat{K}\right) \\ \hat{\boldsymbol{\mu}} &= \mu(X_{test}) + K_{cross}^T K_{train}^{-1} \boldsymbol{f_{train}} \\ \hat{K} &= K_{test} - K_{cross}^T K_{train}^{-1} K_{cross} \end{aligned} \tag{2.3}$$

The equation above assumes that observations are *noise-free*. Usually, this is not the case. By assuming a constant and independent Gaussian noise, we define $y(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In this scenario, y is still a Gaussian process and the inference remains the same, although with a distinct covariance function:

$$y \sim \mathcal{GP}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'}) + \delta_{xx'}\sigma^2) \tag{2.4}$$

where $\delta_{xx'}$ is the Kronecker delta and is equal to 1 if and only if $\boldsymbol{x} = \boldsymbol{x'}$.

2.2 Gaussian Process as a Kernel Machine

Here we discuss an alternative view of GPs based on kernel machines instead of distributions over functions. We begin defining a linear model under a Bayesian treatment:

$$\begin{aligned} y(\boldsymbol{x}) &= \boldsymbol{x}^T \boldsymbol{w} + \epsilon \\ \boldsymbol{w} &\sim \mathcal{N}(\mathbf{0}, \Sigma_w) \\ \epsilon &\sim \mathcal{N}(0, \sigma_n^2) \end{aligned} \tag{2.5}$$

where $\boldsymbol{x} \in X_{train} \cup X_{test}$.

The linear model described above lacks expressiveness, since it can only model linear relationships. A strategy to improve this model is to map the input from a D_1-dimensional space to a D_2-dimensional space such that a linear model in this new space, commonly called *feature space*, is more appropriate than in its original space. Let us denote this mapping as $\phi(\boldsymbol{x})$. Then, the linear model is given by

$$\begin{aligned} y(\boldsymbol{x}) &= \phi(\boldsymbol{x})^T \boldsymbol{z} + \epsilon \\ \boldsymbol{z} &\sim \mathcal{N}(\mathbf{0}, \Sigma_z) \\ \epsilon &\sim \mathcal{N}(0, \sigma_n^2) \end{aligned} \tag{2.6}$$

Alternatively, let $N = |X_{train}|$, $M = |X_{test}|$ and Φ denote the $(N + M) \times D_2$ matrix obtained by applying the mapping ϕ to all $\boldsymbol{x} \in X_{train} \cup X_{test}$. Using properties from the multivariate Gaussian distribution [Eaton, 1983], we can state the following:

$$\boldsymbol{y} | X_{train}, X_{test} \sim \mathcal{N}(\mathbf{0}, \Phi \Sigma_w \Phi^T + \sigma_n^2 I) \tag{2.7}$$

For any finite $X_{train} \cup X_{test}$, we will have an associated vector $\boldsymbol{y}$ that will have joint Gaussian distribution. In other words, Equation 2.7 indicates that $y(\boldsymbol{x})$ is a zero-mean GP. However, to fully describe y as a GP, we need to define its covariance function. Since Σ_w is positive definite, we can define a new mapping function $\psi(\boldsymbol{x}) = \phi(\boldsymbol{x})^T \Sigma_w^{1/2}$ so that $\Phi \Sigma_w \Phi^T = \Psi \Psi^T$. This form is nothing more than an inner product in the new feature space induced by ψ. Therefore, the covariance function associated with y is $k(\boldsymbol{x}, \boldsymbol{x'}) = \psi(\boldsymbol{x})^T \psi(\boldsymbol{x'}) + \delta_{xy} \sigma_n^2$ and we have that

$$y \sim \mathcal{GP}(0, k(\boldsymbol{x}, \boldsymbol{x'}) = \psi(\boldsymbol{x})^T \psi(\boldsymbol{x'}) + \delta_{xx'} \sigma_n^2) \tag{2.8}$$

This reasoning shows that a GP is nothing more than a linear function in a feature space induced by the kernel used. The advantage of GPs comes from the fact that it

avoids explicitly mapping the inputs into a feature space, which may be time and space consuming or even impossible if the feature space has infinite dimensionality. It is very similar to other techniques such as SVM and kernel ridge regression, which are also linear methods that use kernels to deal with non-linearity. This concept is typically called *kernel trick* in the literature.

2.3 Covariance Functions

As indicated above, a GP is fully defined by its mean function and its covariance function. In practice, however, the mean function is typically assumed to be zero after centering the response variable [Rasmussen and Williams, 2006]. Therefore, the main decision to be made when modelling some phenomenon as a GP is to define an appropriate covariance function. This function indicates how data points interact between themselves and should reflect what we know (or expect) from data.

There are many *covariance function families* proposed in the literature. A covariance function family is a set of covariance functions defined by a parametric form, where its parameters are commonly called *hyperparameters*. In theory, any positive definite function could be used as a covariance function, but in practice a few set of functions (and their combinations) are commonly used. In this section, we define some covariance function families that are frequently used in the literature and that are used within this work. We also indicate how we can obtain an optimal set of hyperparameters in order to define a specific covariance function within its family.

2.3.1 Families of Covariance Functions

Squared Exponential Covariance Functions The squared exponential family, also known as the *radial basis function* (RBF) family, is defined as follows:

$$k_{SE}(\boldsymbol{x}, \boldsymbol{x'}) = \sigma^2 exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x'})^T M(\boldsymbol{x} - \boldsymbol{x'})\right) \tag{2.9}$$

where M is a diagonal matrix. The main assumption of this family is that data points separated by greater distances are less associated than data points located close by. In other words, the correlation between a pair of points decays monotonically with their distance. The hyperparameters associated to this family are σ^2, which indicates the strength of the covariance signal, and the diagonal entries of M, which defines the distance function used, that is, how each dimension contributes to the final distance

between pairs of points. Such elements are called the *characteristics length-scales* associated to each dimension, or simply length-scales.

Note that the squared exponential functions can be expressed in terms of $\boldsymbol{\tau} = \boldsymbol{x} - \boldsymbol{y}$. Functions of this kind are called *stationary*. If all length-scales are equal, then the squared exponential family can be expressed in terms of $|\boldsymbol{\tau}|$ and is called *isotropic*.

Figure 2.1 shows a 1-dimensional example of squared exponential covariance function obtained by fixing its hyperparameters as $\sigma = 1$ and $M = 10^{-2}$. Note how points separated by distances greater than 20 are nearly uncorrelated. The figure also show a sample of a zero-mean GP equipped with this covariance function, obtained using the transformation $\boldsymbol{y} = L\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, I)$ and $K = LL^T$ is the Cholesky decomposition of the covariance matrix K.

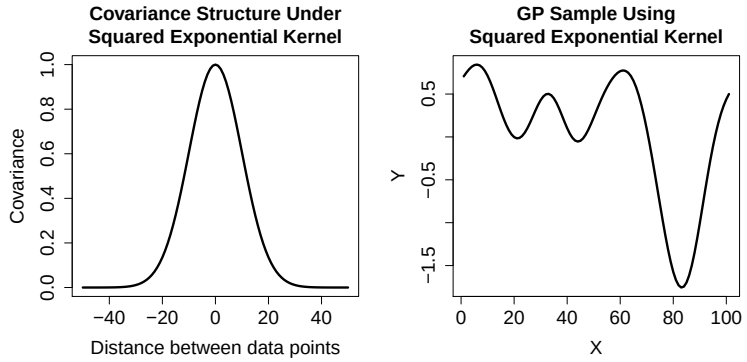

Figure 2.1. One-dimensional example of squared exponential covariance function with $\sigma = 1$ and $M = 10^{-2}$ (left figure) and a sample from a zero mean GP equipped with the illustrated covariance function (right figure).

Matérn Covariance Functions The Matérn family is similar to the squared exponential family in the sense that both are stationary and share the main assumption of monotonically decaying correlation with greater distances. However, while functions modelled using the squared exponential family are infinitely differentiable, functions modelled using the Matérn family are finitely differentiable and, therefore, are typically rougher.

In machine learning, this family is typically defined considering functions one or two times differentiable, where the functions are of the form:

$$k_{Mat}^{(\rho=1)}(\boldsymbol{x}, \boldsymbol{x'}) = \sigma^2 \left(1 + \sqrt{3}\boldsymbol{\tau}^T M \boldsymbol{\tau}\right) exp\left(-\sqrt{3}\boldsymbol{\tau}^T M \boldsymbol{\tau}\right) \tag{2.10}$$

$$k_{Mat}^{(\rho=2)}(\boldsymbol{x}, \boldsymbol{x'}) = \sigma^2 \left(1 + \sqrt{5}\boldsymbol{\tau}^T M \boldsymbol{\tau} + \frac{5(\boldsymbol{\tau}^T M \boldsymbol{\tau})^2}{3}\right) exp\left(-\sqrt{5}\boldsymbol{\tau}^T M \boldsymbol{\tau}\right) \tag{2.11}$$

where ρ denotes how many times modelled functions can be differentiated and $\boldsymbol{\tau} = \boldsymbol{x} - \boldsymbol{x'}$. Similarly to the squared exponential family, hyperparameters from Matérn family are σ^2, which is the strength of the covariance signal, and the diagonal entries of M, which also act as length-scales.

Figure 2.2 shows 1-dimensional examples of the Matérn covariance function obtained by fixing its hyperparameters as $\sigma = 1$ and $M = 10^{-2}$ with $\rho = 1$ or $\rho = 2$. Both samples were obtained using the same method as in Figure 2.1 and shared the same $\boldsymbol{\epsilon}$. Note that the sample obtained by using $\rho = 2$ is smoother than the one obtained using $\rho = 1$.

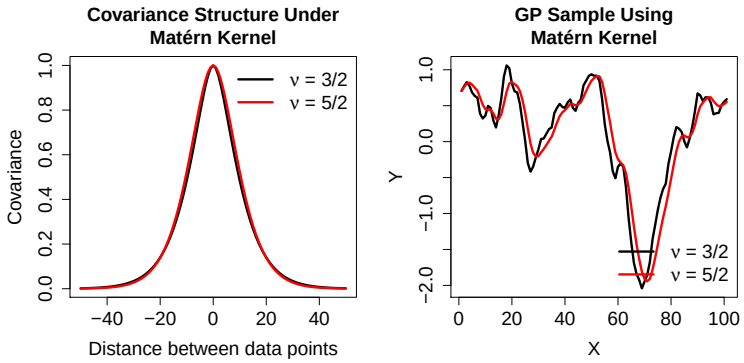

Figure 2.2. One-dimensional example of Matérn covariance function with $\sigma = 1$ and $M = 10^{-2}$ (left figure) and a sample from a zero mean GP equipped with the illustrated covariance function (right figure).

Periodic Covariance Functions Another stationary family is the periodic family. Different from the squared exponential and Matérn families, however, this family assumes a periodic behaviour instead of a monotonically decreasing one. It can be expressed as

$$k_{Per}(\boldsymbol{x}, \boldsymbol{x'}) = \sigma^2 exp\left(-2\frac{sin^2(\pi|\boldsymbol{x} - \boldsymbol{x'}|/p)}{\ell}\right) \tag{2.12}$$

where σ, p and ℓ are hyperparameters indicating the signal's strength, periodicity and decay, respectively. Note that this family is isotropic, as it can be expressed in terms of $|\boldsymbol{x} - \boldsymbol{x'}|$. More sophisticated non-isotropic formulations are available, but omitted here since they are not considered in this work.

Figure 2.3 shows an 1-dimensional example of periodic covariance function obtained by fixing its hyperparameters as $\sigma = 1$, $p = 30$ and $\ell = 1$.

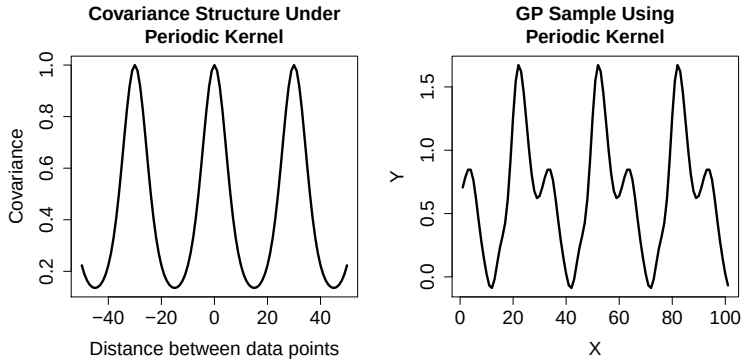

Figure 2.3. One-dimensional example of periodic covariance function with $\sigma = 1$, $p = 30$ and $\ell = 1$ (left figure) and a sample from a zero mean GP equipped with the illustrated covariance function (right figure).

Linear Covariance Functions The linear family assumes a linear relationship between covariates and the response variable. It can be expressed as

$$k_{Lin}(\boldsymbol{x}, \boldsymbol{x'}) = \boldsymbol{x}^T M \boldsymbol{x'} + \sigma^2 \tag{2.13}$$

where M is a diagonal matrix with length-scales indicating the effect of each dimension and σ allows for a bias term.

Figure 2.4 shows an 1-dimensional example of linear covariance function obtained by fixing its hyperparameters as $\sigma = 1$ and $M = 10^{-2}$.

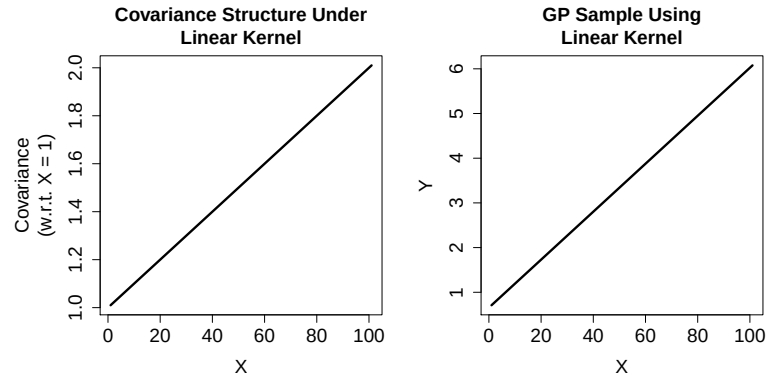

Figure 2.4. One-dimensional example of linear covariance function with $\sigma = 1$, $M = 10^{-2}$ (left figure) and a sample from a zero mean GP equipped with the illustrated covariance function (right figure).

Polynomial Covariance Functions The polynomial family is an extension of the linear family in the sense that is assumes a polynomial relationship between covariates

and the response variable. The homogeneous form of degree d is expressed as

$$k_{Poly}(\boldsymbol{x}, \boldsymbol{x'}) = \left(\boldsymbol{x}^T M \boldsymbol{x'}\right)^d \tag{2.14}$$

where M is a diagonal matrix with length-scales indicating the effect of each dimension. The inhomogeneous form of degree d, on the other hand, is expressed as

$$k_{Poly}(\boldsymbol{x}, \boldsymbol{x'}) = \sigma^2 + \sum_{i=1}^{d} \left(\boldsymbol{x}^T M \boldsymbol{x'}\right)^i \tag{2.15}$$

Figure 2.5 shows an 1-dimensional example of degree-2 polynomial covariance function obtained by fixing its hyperparameters as $\sigma = 1$, and $M = 10^{-2}$.

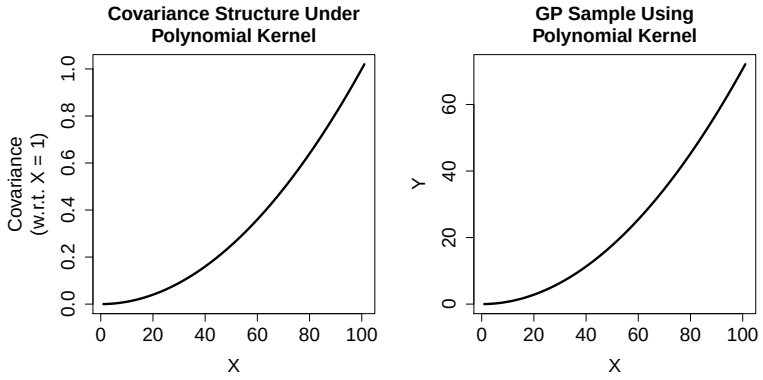

Figure 2.5. One-dimensional example of homogeneous polynomial (degree 2) covariance function with $\sigma = 1$ and $M = 10^{-2}$ (left figure) and a sample from a zero mean GP equipped with the illustrated covariance function (right figure).

Spectral Mixture Covariance Functions The spectral mixture family is a very general family of covariance functions recently proposed in Wilson and Adams [2013] capable of representing *any* stationary covariance function. It is based on Bochner's theorem, which shows that a stationary covariance function is completely specified by its spectral density. In other words, two stationary covariance functions are equal if and only if their spectral densities are equal. The spectral mixture family exploits this result by approximating the spectral density with a mixture of Gaussian functions, which turn into quasi-periodic functions when mapped from the frequency domain into the original domain. The larger the number of Gaussian components, the better is the approximation, but the larger the number of hyperparameters.

The spectral mixture family with D components can be expressed as

$$k_{SM}(\boldsymbol{x}, \boldsymbol{x'}) = \sum_{i=1}^{D} \sigma_i^2 \prod_{j=1}^{P} exp(-2\pi^2 \tau_j^2 v_{i,j}) cos(2\pi \tau_j \mu_{i,j}) \tag{2.16}$$

where P is the number of dimension, $\boldsymbol{\tau} = \boldsymbol{x} - \boldsymbol{x'}$, τ_j is the j-th component of $\boldsymbol{\tau}$ and σ_i, $v_{i,j}$ and $\mu_{i,j}$ are hyperparameters affecting the strength of the component i, the decay of component i on dimension j and the period of component i on dimension j, respectively.

Figure 2.6 shows a 1-dimensional example of a 3-component spectral mixture covariance function obtained by fixing its hyperparameters as $\boldsymbol{\sigma} = (1, 1, 1)^T$, $\boldsymbol{v} = (10^{-4}, 10^{-4}, 10^{-4})^T$ and $\boldsymbol{\mu} = (10^{-1}, 10^{-2}, 10^{-3})^T$. Note that spectral mixture kernel can exploit periodicity at multiple periods with distinct signal strength.

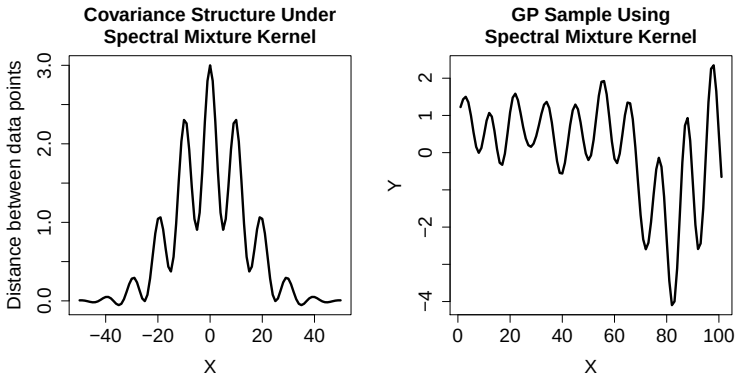

Figure 2.6. One-dimensional example of 3-component spectral mixture covariance function with $\boldsymbol{\sigma} = (1, 1, 1)^T$, $\boldsymbol{v} = (10^{-4}, 10^{-4}, 10^{-4})^T$ and $\boldsymbol{\mu} = (10^{-1}, 10^{-2}, 10^{-3})^T$ (left figure) and a sample from a zero mean GP equipped with the illustrated covariance function (right figure).

2.3.2 Hyperparameter Optimization

As shown above, each covariance function family is parametrized by hyperparameters. However, in order to define a GP, we need a specific covariance function, that is, we need to set a value for each hyperparameter. Defining these values manually may lead to misspecification of the model and may require a deeper knowledge of the data in hand. Fortunately, as a finite set of variables under a GP always has a joint Gaussian distribution, it is possible to find a closed-form expression for calculating the likelihood of the model given the data. In fact, it is also possible to find a closed-form expression for the derivatives of the likelihood with relation to the hyperparameters, enabling the

local optimization of the likelihood. This give us a principled way to find good estimates for hyperparameters, relieving the need of carefully finding appropriate values.

The log-likelihood under a GP model is given by

$$log\ p(\boldsymbol{y}|X,\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{y}^T K^{-1}\boldsymbol{y} - \frac{1}{2}log|K| - \frac{n}{2}log\ 2\pi \tag{2.17}$$

where K is the covariance matrix obtained by applying the covariance function parametrized by $\boldsymbol{\theta}$ on all pairs of points on X.

Derivatives with relation to hyperparameters can be found using the following expression:

$$\frac{\partial}{\partial\theta_i}log\ p(\boldsymbol{y}|X,\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{y}^T K^{-1}\frac{\partial K}{\partial\theta_i}K^{-1}\boldsymbol{y} - \frac{1}{2}trace\left(K^{-1}\frac{\partial K}{\partial\theta_i}\right) \tag{2.18}$$

Unfortunately, since entries of y are not conditionally independent given X and $\boldsymbol{\theta}$, optimization techniques based on *stochastic gradient descent* [Amari, 1993] are not directly applied. Rather, techniques based on conjugate gradients [Shewchuk et al., 1994] are more commonly used for hyperparameter optimization [Rasmussen and Nickisch, 2010].

2.4 Putting the Pieces Together

Using the results presented in the previous sections, we are now capable of defining a complete algorithm for inference under a GP model. Having defined the covariance function, which should reflect the relationships between variables in data, Equation 2.17 is locally maximized using any gradient-based method in order to obtain appropriate estimates for the covariance function hyperparameters. Then, these hyperparameters are used to compute the covariance matrices K_{train}, K_{cross} and K_{test} (as in Equation 2.2). We obtain the Cholesky decomposition of K_{train} to obtain the solution for the linear system $K_{train}^{-1}\boldsymbol{y}$, which is then used to compute the conditional predictive mean and covariance as shown in Equation 2.3. Finally, we calculate the log-likelihood using the expression in Equation 2.17. Algorithm 2.1 formalizes this reasoning.

The GP modelling framework allows for flexible (in the sense that any covariance function can be used, enabling the exploration of a wide range of patterns), exact inference with closed-form expressions. However, computational complexity may be a downside. Let N be the number of points in X_{train}. By inspecting Algorithm 2.1, we see that it requires the Cholesky decomposition of K_{train} (line 5). This operation requires $O(N^3)$ and dominates the complexity of Algorithm 2.1. In terms of spatial

2.1: Algorithm for inference under GP model

Input: training cases X_{train}, observed response variables $\boldsymbol{y}$, initial guess for hyperparameters of the covariance function k $\hat{\boldsymbol{\theta}}$, test cases X_{test}

Output: predictive mean $\hat{\boldsymbol{y}}$, predictive covariance matrix $\hat{K_{test}}$, optimal hyperparameters $\boldsymbol{\theta_*}$, log-likelihood ℓ

$\boldsymbol{\theta_*} \leftarrow$ gradient-based maximization of Equation 2.17 using $\hat{\boldsymbol{\theta}}$ as starting point
$K_{train} \leftarrow k(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta_*}) \; \forall \; \boldsymbol{x}, \boldsymbol{x'} \in X_{train}$
$K_{cross} \leftarrow k(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta_*}) \; \forall \; \boldsymbol{x} \in X_{train}, \boldsymbol{x'} \in X_{test}$
$K_{test} \leftarrow k(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta_*}) \; \forall \; \boldsymbol{x}, \boldsymbol{x'} \in X_{test}$
$L \leftarrow$ Cholesky decomposition of K_{train}
$\boldsymbol{\alpha} \leftarrow L^T \backslash (L \backslash \boldsymbol{y})$; `// calculating` $\alpha = K_{train}^{-1}\boldsymbol{y}$
$V \leftarrow L \backslash K_{cross}$; `// calculating` V `such that` $V^T V = K_{cross}^T K_{train}^{-1} K_{cross}$
$\hat{\boldsymbol{y}} \leftarrow K_{cross}^T \boldsymbol{\alpha}$
$\hat{K_{test}} \leftarrow K_{test} - V^T V$
$\ell \leftarrow -\frac{1}{2}\boldsymbol{y}^T\boldsymbol{\alpha} - \sum_i log\ L_{ii} - \frac{n}{2} log\ 2\pi$

complexity, Algorithm 2.1 requires $O(N^2)$, as it needs to store K_{train}. Therefore, a naïve implementation may struggle even with moderate datasets ($N \approx 10000$). Enabling inference under GP models for larger datasets is an active research topic [Williams and Seeger, 2000; Snelson and Ghahramani, 2005; Lázaro-Gredilla et al., 2010; Saatçi, 2012].

2.5 Multi-task Gaussian Process

In some occasions, one may need to solve a set of learning tasks which are associated between themselves. Although it is possible to deal with each task independently, it is also possible that tasks share information, so that learning them jointly would be beneficial. This scenario is known as *multi-task learning* [Caruana, 1993]. In the specific context of dengue fever in Brazil, we may define the learning task to be the prediction of new outbreaks in a given city. Therefore, our set of learning tasks would include a task per city and it would now be possible to use information from one city to predict new outbreaks in others. Multi-task GP (MTGP) is not a new topic in the literature, with different approaches being proposed since 2005. Here, we review some of these techniques.

MTGP using task descriptors Perhaps the simplest strategy for defining a MTGP model is to aggregate task descriptors into the original dataset [Bonilla et al., 2007a]. Task descriptors are attributes associated to tasks only, and not to data points. For instance, when tasks are associated to cities, task descriptors could be spatial coordi-

nates indicating the location of each city or the area extension occupied by each city. Importantly, these new attributes should impact the dependence between data points. These new attributes are then used in the covariance function as usual, inducing correlation between data points from distinct tasks, as well as data points from the same task.

This approach has, however, two major limitations. First, it requires defining attributes associated to tasks that can explain how tasks are related. Second, it suffers from a high computational cost in terms of number of operations required and space. Let N_i be the number of data points from task i, M be the number of tasks and $N_t = \sum_i N_i$, it requires storing and inverting a $N_t \times N_t$ matrix, leading to $O\left(\left(\sum_i N_i\right)^3\right)$ operations and $O\left(\left(\sum_i N_i\right)^2\right)$ bytes used. In contrast, independent inference would require $O\left(\sum_i N_i^3\right)$ operations and $O\left(\sum_i N_i^2\right)$ bytes. This is specially relevant if all tasks share approximately the same number of data points and M is large. In this scenario, MTGP can be $O(M^2)$ times slower and consume $O(M)$ times more memory.

MTGP using task-related covariance matrix In order to remove the need of defining task descriptors, Bonilla et al. [2007b] proposed to use a (symmetric positive definite) structure-free covariance matrix to model inter-task covariances. Let $\boldsymbol{x}^{(i)}$ and $\boldsymbol{y}^{(j)}$ be data points associated to tasks i and j, respectively. Then, the covariance between $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x'}^{(j)}$ is given by

$$k(\boldsymbol{x}^{(i)}, \boldsymbol{x'}^{(j)}) = K_f(i,j)k_x(\boldsymbol{x}, \boldsymbol{x'}) \tag{2.19}$$

where k_x is a covariance function whose outputs are independent of which tasks the data points come from and K_f is a task-related covariance matrix whose entries are considered hyperparameters and learned from data via likelihood maximization.

Although inefficient as it was originally proposed, MTGP using a task-related covariance matrix can become very efficient if all tasks share the same data points, that is, if for any $\boldsymbol{x}^{(i)}$ there exists $\boldsymbol{x'}^{(j)}$ such that $\boldsymbol{x}^{(i)} = \boldsymbol{x'}^{(j)}$, for all pairs i and j. This is a common property in time series domain, for instance, if multiple time series are observed simultaneously. In that case, data points will be time indices, which will be the same for all tasks. When this property holds, the final covariance matrix K obtained by applying Equation 2.19 to all pairs of data points from all tasks can be expressed as $K = K_f \otimes K_x$, where K_x is the matrix obtained by applying k_x (from Equation 2.19) to all pairs of data points from a single task and $\otimes$ denotes the Kronecker product. Recent results proposed in Saatçi [2012] show that inference under this scenario can be done in $O(N_i^3 + M^3)$ operations and using $O(N_i^2 + M^2)$ bytes. This improvement

in efficiency is due to two results. The first result is related to the eigendecomposition of K. Let $K = Q\Lambda Q^T$ be the eigendecomposition of K. Then,

$$K = (Q_f \otimes Q_x)(\Lambda_f \otimes \Lambda_x)(Q_f \otimes Q_x)^T \tag{2.20}$$

where $Q_f \Lambda_f Q_f^T$ and $Q_x \Lambda_x Q_x^T$ are the eigendecompositions of K_f and K_x, respectively. The second result is that, for any vector $\boldsymbol{x}$ of size $N_A N_B$ and matrices A and B of sizes $N_A \times N_A$ and $N_B \times N_B$, respectively, $\boldsymbol{z} = (A \otimes B)\boldsymbol{x}$ can be computed in $O(N_A^2 N_B + N_A N_B^2)$. This done by using the Algorithm 2.2, extracted from Saatçi [2012] and specialized to the context of MTGP.

2.2: Algorithm for calculating $\boldsymbol{z} = (A \otimes B)\boldsymbol{x}$

Input: matrices A and B of sizes $N_A \times N_A$ and $N_B \times N_B$, $(N_A N_B)$-length vector $\boldsymbol{x}$
Output: $(N_A N_B)$-length vector $\boldsymbol{z}$
1 $X \leftarrow \text{reshape}(\boldsymbol{x}, N_B, N_A)$
2 $X \leftarrow (BX)^T$
3 $X \leftarrow (AX)^T$
4 $\boldsymbol{z} \leftarrow \text{reshape}(X, N_A N_B, 1)$

Removing the need of task descriptors, however, does not come free. The major downside of this model is the number of hyperparameters. The task-related covariance matrix K_f is, except from being symmetric and positive definite, structure-free. Positive definiteness is guaranteed by using the Cholesky decomposition $K_f = L_f L_f^T$, where L_f is a lower triangular matrix. Non-null entries of L_f are then treated as hyperparameters and optimized via maximum likelihood. This leads to $O(M^2)$ hyperparameters, which may cause efficiency issues (due to the need of computing a large number of derivatives) and optimization issues (due to the high dimensionality).

Mixed-effect MTGP An alternative approach to explicitly using a task-related covariance matrix is to model each task as the sum of two independent components: an *average task* and a *individual shift*. This strategy is exploited in Pillonetto et al. [2010], where the authors model task i as

$$f_i(\boldsymbol{x}) = \bar{f}(\boldsymbol{x}) + \tilde{f}_i(\boldsymbol{x}) \tag{2.21}$$

where $\bar{f}$ and $\tilde{f}_i$ are GPs, namely the average task and the individual shift, respectively. The authors developed an algorithm for inference under this model which scales with $O(M\tilde{N}^3)$, where M is the number of tasks and $\tilde{N}$ is the number of unique data points

of all tasks. This technique can be rather efficient if tasks share a large amount of data points, but will be inefficient otherwise.

A more general model was proposed in Wang and Khardon [2012]. In contrast to the model in Equation 2.21, the model groups tasks into K disjunct clusters and uses K average tasks, one per cluster. The authors also developed an approximation scheme to speed up inference at the cost of exact inference by using a low rank approximation for the final covariance matrix, yielding a method efficient even when tasks do not share data points.

Convolution-based MTGP A similar but somewhat more sophisticated way to enable multi-output regression was originally proposed in Boyle and Frean [2004]. The main observation of the authors is that a GP can be expressed as the convolution between a Gaussian white noise process and a smoothing function. The authors then propose a model of N-output regression using a set of M independent Gaussian white noise processes and NM functions. By convolving the Gaussian white noise processes with the functions, NM GPs are produced. However, GPs that shared the same Gaussian white noise processes are no longer independent. For each output, M GPs are summed over, enabling dependences between each pair of outputs. Figure 2.7 shows an example with $M = 3$ independent Gaussian white noise processes and $N = 2$ outputs. Smoothing functions not shown explicitly in the figure are assumed to be zero. Note that, by making an appropriate choice of kernels, the convolution-based MTGP can simulate mixed-effect MTGP, thus being a more general framework.

Similarly to other approaches, this model is not very efficient, requiring $O\left(\left(\sum_i N_i\right)^3\right)$ operations, where N_i is the number of data points for task i, and $O\left(\left(\sum_i N_i\right)^2\right)$ bytes, since the model requires storing and inverting a full covariance matrix of size $\left(\sum_i N_i\right) \times \left(\sum_i N_i\right)$. In order to speed up inference under this model, the authors of Álvarez and Lawrence [2008, 2011] propose to approximate the full covariance matrix with a low rank one so that auto-covariance is kept intact and cross-covariances are approximated. The approximation scheme proposed is based on the notion that, given access to the latent Gaussian white processes, the output GPs are no longer dependent. The authors then extrapolate this reasoning by assuming conditional independence between outputs even when only a sample of size S obtained from the latent processes are observed. By doing so, inference can now be performed with $O(\sum_i N_i^3)$ operations and $O(\sum_i N_i^2)$ if $S \approx N$, the same complexity associated with inference of independent GPs.

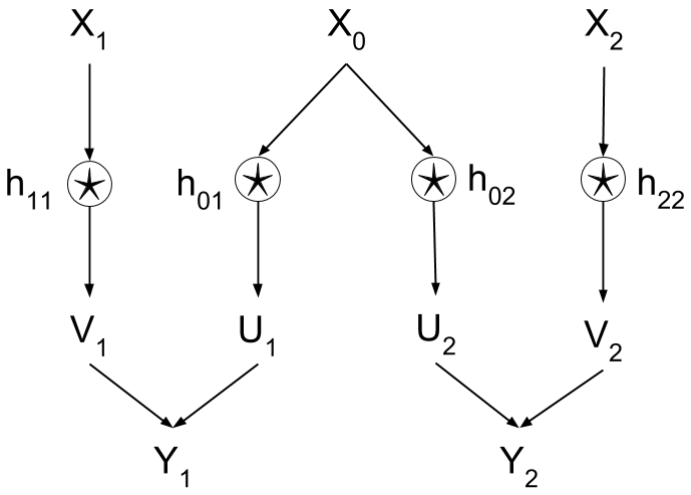

Figure 2.7. Diagram for convolution-based MTGP with three independent Gaussian white noise processes (X_0, X_1 and X_2), two outputs (Y_1 and Y_2) and four kernels (h_{01}, h_{02}, h_{11} and h_{22}). Kernels h_{12} and h_{21} are assumed to be zero and omitted. Dependence between Y_1 and Y_2 is done via U_1 and U_2, which share the same Gaussian white process.

Gaussian Process Regression Networks The authors of Wilson et al. [2012] proposed yet another model for MTGP regression similar to neural networks with a single hidden layer named Gaussian process regression networks (GPRNs). In this model, input feeds nodes from the hidden layer, which are modelled as GPs. Outputs from each node are then linearly combined in order to generate the outputs required. The combination weights, however, are also modelled as GPs, enabling the model to exploit varying dependences between outputs. Figure 2.8 shows a diagram of a GPRN with three hidden nodes and two outputs.

In contrast with independent GPs, inference under GPRN is not exact and requires the usage of more sophisticated inference techniques. In the original paper, the authors proposed two approaches. The first uses elliptical slice sampling [Murray et al., 2010], a recent sampling technique specifically designed for sampling from posteriors with strongly correlated Gaussian priors. The second approach approximates the posterior distribution performing a variational EM implementation [Jordan et al., 1999] by minimising the Kullback-Leibler divergence between real and approximate posterior distributions, and is also used for hyperparameter learning. Assuming that all tasks share the same data points, all hidden nodes and all weight functions share the same covariance function, inference using both approaches requires $O(N^3)$ operations, where N is the number of data points per task. Violating the precedent assumptions, the complexity can go as high as $O(N^3PQ)$, where P is the number of hidden nodes and

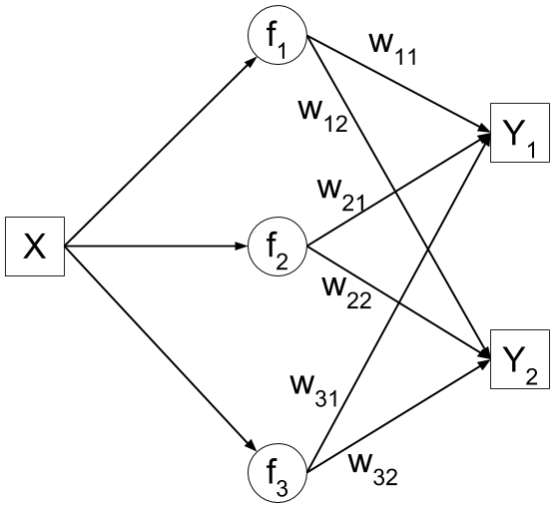

Figure 2.8. Diagram for GPRN with three hidden nodes (f_1, f_2 and f_3) and two outputs (Y_1 and Y_2). Combining weights are shown as $w_{..}$ and X act as input for the network.

Q is the number of outputs.

The number of variational parameters using variational Bayes, which is required for hyperparameter optimization, however, scales quadratically with N. Aiming to reduce the number of parameters, the authors of Nguyen and Bonilla [2013] proposed two new variational approaches for inference under the GPRN model. One uses full multivariate Gaussians distributions for each hidden node and each weight functions but with a parametrization that requires only $O(N)$ parameters. The second approach uses a recently proposed variational inference framework [Gershman et al., 2012] that approximates the posterior distribution with K isotropic multivariate Gaussian distributions, also requiring $O(N)$ variational parameters.

Chapter 3

Related Work

In this chapter, we discuss previous work on the development of predictive models for dengue fever around the world. Given the large volume of works on dengue modelling, we focused on works where authors intended to develop predictive models or assess the predictive capability of a set of covariates, ignoring the vast majority of works where the main objective is to establish relationships between dengue fever incidence and other variables. For a discussion on this matter, the reader is referred to Naish et al. [2014]; Louis et al. [2014].

Tables 3.1 and 3.2 summarize the 30 works found by looking at the 50 first results returned by Google Scholar using the queries *dengue forecast*, *dengue predict* and *dengue early warning* and manually filtered by inspecting their abstract. Works were summarized according to the following criteria:

- **Area under study:** Location from where dengue data was obtained.
- **Data resolution:** Spatial and temporal resolution of epidemiological data used for feeding the predictive model. Note that if data is obtained at a finer resolution but aggregated prior to fitting the model, the resolution considered is the one after the aggregation.
- **Model:** Type of model or technique used for modelling dengue data.
- **Covariates:** Set of covariates used by the model.
- **Spatial dependences:** Type of spatial structure imposed by the model, if any.
- **Predictive model:** We considered a model as a predictive model if the proposing work assessed its capabilities of issuing predictions, that is, if dengue incidence rate for time index t is predicted based only on information available at time

index $t - \beta$, where β is a positive integer. Note that this definition is valid for both the response variable and covariates. For each predictive model, this column indicates the largest possible value of β.

Area Under Study The vast majority of the works reviewed used data from locations in Asia or in the Americas, the continents with the largest number of dengue cases in the world. Out of 30 works, 14 are limited to a single city, 10 are limited to a state or region and only 6 are country-level studies, indicating a tendency of developing very specific dengue models, defined and evaluated for small areas. The models proposed within this work avoid that by using epidemiological data from cities of all Brazilian regions, thus covering the entire territory of the country.

Data Resolution The spatial resolution of epidemiological data used in the reviewed works varied drastically, ranging from country-level works to the work of Chan et al. [2015], which used data at the resolution of urban villages, subdivisions of the Kaohsuing City. With respect to temporal resolution, most models work at month or week level, with the notable exception of Chan et al. [2015], which uses daily data. Our models use data at city and week levels, thus being consistent with the works retrieved from the literature.

Model Linear and generalized linear models are the most commonly used models for dengue prediction, being used in 15 out of 30 works. Within the class of generalized linear models, Poisson and negative binomial models are standard practice, with the latter being preferred over the former due to its capacity of dealing with overdispersion, since Poisson models assume the mean value equals the variance. An alternative for this problem is the inclusion of an extra parameter responsible to increasing the variance, leading to the class of Quasi-Poisson models. Besides these models, ARIMA models and their seasonal variant, SARIMA, are also frequently used, specially in works prior to 2012, being explored in 12 works.

The previous analysis indicates a tendency in the literature to explore *parametric* models, that is, models that can be specified using a finite number of parameters. A major limitation of this class of models is the necessity of a strict specification of the expected relationships between covariates and the response variable, frequently requiring significant human intervention in the data analysis. Another drawback is the limited complexity modelled by these models, since their degrees of freedom are determined by the number of parameters. Thus, even with growing datasets, the complexity of the model remains the same. Non-parametric models such as Gaussian

Table 3.1. Selected previous work on dengue incidence rate prediction (2006-2011).

Reference	Area Under Study	Data Resolution		Model	Covariates	Spatial Dependences	Predictive Model
		Spatial	Temporal				
Promprou et al. [2006]	Southern Thailand	Province	Month	ARIMA	Temporal-only	N/A	1 year
Wu et al. [2007]	Kaohsuing City, Taiwan	City	Month	ARIMA	Temperature, humidity, rainfall, vector presence	N/A	No
Silawan et al. [2008]	Northeastern Thailand	Province	Month	SARIMA	Temporal-only	N/A	1 month
Choudhury et al. [2008]	Dhaka, Bangladesh	City	Month	SARIMA	Temporal-only	N/A	1 month
Luz et al. [2008]	Rio de Janeiro, Brazil	City	Month	ARIMA	Temperature, rainfall	N/A	1 and 12 months
Johansson et al. [2009]	Puerto Rico	City	Month	Poisson	Temperature, rainfall	No	No
Hu et al. [2010]	Queensland Australia	Local Gov. Area	Month	SARIMA	El Niño	N/A	No
Gharbi et al. [2011]	Guadeloupe, French Antilles	Region	Week	SARIMA	Temperature, humidity, rainfall	N/A	1, 3 and 12 months
Colón-González et al. [2011]	Mexico	Country	Month	Linear	Temperature, rainfall	N/A	No
Lowe et al. [2011]	Southeast Region, Brazil	Microregion	Month	Poisson	Temperature, rainfall, El Niño, altitude, urban population	Neighborhood	No
Yu et al. [2011]	Southern Taiwan	City	Week	Poisson	Temperature, rainfall, El Niño, vector presence	Distance	1 week
Martinez and Silva [2011]	Ribeirão Preto, Brazil	City	Month	SARIMA	Temporal-only	N/A	1 month
Martinez et al. [2011]	Campinas, Brazil	City	Month	SARIMA	Temporal-only	N/A	1 month
Yusof and Mustaffa [2011]	Selangor, Malaysia	District	Week	LS-SVM	Rainfall	Neighborhood	No
Gomide et al. [2011]	Brazil	City	Month	Linear	Dengue-related tweets	No	No
Althouse et al. [2011]	Singapore and Bangkok, Thailand	City	Week and Month	Linear, GBR, neg. binomial, SVM*, logistic regression*	Dengue-related queries on Google	N/A	No

* Models used for binary classification (epidemic vs non-epidemic)

Table 3.2. Selected previous work on dengue incidence rate prediction (2012-2015).

Reference	Area Under Study	Data Resolution		Model	Covariates	Spatial Dependences	Predictive Model
		Spatial	Temporal				
Earnest et al. [2012]	Singapore	City	Week	Poisson	Temperature, rainfall, sunshine, El Niño	N/A	No
Descloux et al. [2012]	Noumea, New Caledonia	City	Year	SVM*	Temperature, rainfall, humidity, wind force, evapotranspiration, El Niño, vector presence	N/A	1 month
Hu et al. [2012]	Queensland, Australia	Local Gov. Area	Month	Poisson	Temperature, rainfall, socialeconomical index	Neighborhood	No
Hii et al. [2012]	Singapore	City	Week	Quasi-Poisson	Temperature, rainfall	N/A	16 weeks
Buczak et al. [2012]	Peru	City	Week and Month	Fuzzy Association Rule Mining*	Temperature, rainfall, vegetation, El Niño, urbanization indices	No	3 to 7 weeks
Bhatnagar et al. [2012]	Rajasthan, India	State	Month	SARIMA	Temporal-only	N/A	1 month
Lowe et al. [2013]	Southeast Region, Brazil	Microregion	Month	Negative Binomial	Temperature, rainfall, El Niño, altitude, population density	Neighborhood	3 months
Dom et al. [2013]	Subang Jaya, Malaysia	City	Week	ARIMA	Temporal-only	N/A	4, 13 and 52 weeks
Lowe et al. [2014]	Brazil	Microregion	Month	Negative Binomial	Temperature, rainfall, population density, altitude	Neighborhood	3 months
Banu et al. [2014]	Dhaka, Bangladesh	City	Month	Quasi-Poisson	Temperature, rainfall, humidity	N/A	No
Eastin et al. [2014]	Cali, Colombia	City	2-weeks and Month	ARIMA	Temperature, rainfall, humidity, El Niño	N/A	2 and 4 weeks
Chan et al. [2015]	Kaohsuing City, Taiwan	Urban Village	Day	Logistic regression*	Temperature, rainfall population density	Neighborhood and co-occurence	1 day
Souza et al. [2015]	Brazil	City	Week	Poisson	Dengue-related tweets	No	No
Shi et al. [2016]	Singapore	City	Week	Linear	Population size, temperature, humidity, presence of vector	N/A	3 months

* Models used for binary classification (epidemic vs non-epidemic)

processes, on the other hand, have their degrees of freedom determined by the amount of data available, being capable of modelling more sophisticated behavior whenever more data is acquired.

In our literature review, we found only three works where dengue data is modelled using non-parametric models. Two of these works, however, are defined for a classification scenario (epidemic or non-epidemic period). Therefore, only one work used a non-parametric model to predict actual dengue fever incidence [Yusof and Mustaffa, 2011]. In this sense, our work is novel, as it helps extrapolating the usage of parametric models in favor of more sophisticated, state-of-art techniques capable of providing more accurate predictions while requiring less human supervision.

Covariates A wide set of covariates was used in the reviewed works. The most common are climate-related covariates, specially those related to temperature and rainfall. The major motivation is the known relationships between these covariates and the life cycle of *Aedes aegypti* [Kuhn et al., 2005]. Temperature interferes with the mosquito's life cycle by accelerating it. Thus, higher temperature leads to a faster transition between the mosquito's life stages and, consequently, to a higher presence of the vector in the environment. However, excessively high temperature may also lead to vector death, reducing dengue fever risk. Rainfall, on the other hand, helps creating still water reservoirs, which are then used by mosquitoes as breeding sites. Similarly to temperature, excessive rainfall can also reduce the vector presence by washing out breeding sites.

Another set of climate-related covariates used in the reviewed works are indices related to El Niño Southern Oscillation, which describes fluctuations in temperature between the ocean and the atmosphere, thus being directly related to temperature and rainfall. It is important to notice that most works using climate-related covariates do that by applying a *temporal lag* on them, in the sense that climate of week t is associated with DIR during week $t + \alpha$, where α is a positive number. This is due to the fact that climate is associated with the vector's life cycle, leading to an increase or decrease of vector presence. Hence, the proliferation of the vector in its early stage will only affect DIR after some time required by the vector to reach its final stage.

Urbanization and social-economical conditions are also exploited in some of the reviewed works through a variety of indices, such as population sizes and/or density and access to running water and hygienic services. The main intuition is that urbanization and social-economical conditions may facilitate dengue fever dissemination either by creating more infection possibilities, in the case of high density areas, or by increasing the usage of water containers, which may be used by the vector to proliferate.

More recently, works have begun using data coming from online sources, such

as search engines and online social networks, as covariates for dengue fever modelling. Although not directly related to the dissemination of the disease, this kind of data is a good sensor of people's reaction to dengue fever. Infected people may use online social media to indicate their health conditions and may use certain websites to find information about the disease. Therefore, by monitoring online sources, researchers have been able to identify current incidence of many diseases, including dengue fever.

In this work, we use climate-related covariates, namely average temperature, rainfall volume and average relative humidity in some of our proposed models. We have opted for not using covariates related to El Niño since its effect is captured in the other covariates. In Chapter 7, we also use data from Twitter to estimate current DIR. Urbanization and social-economical indices were not used due to the short period under study. These kind of indices typically vary slowly, being more relevant to works using data from a longer period of time.

Spatial Dependences As an infection disease, dengue fever outbreaks are known to be spatially dependent events. Despite that, 23 out of the 30 reviewed works proposed models to a single area or ignored this property, generating models that are completely unaware of dengue fever risk in other areas. The few works that explored some kind of spatial structure in dengue data enforce spatial dependences between neighboring or nearby areas. These structures, however, are very limiting in the sense that they required prior specification and ignore human mobility. In a highly connected world, human mobility may introduce dependences between areas that are not geographically close, but present significant human transit.

In this work, we used a multi-task learning approach, which enables the model to automatically specify spatial dependences based on data only, requiring much less human intervention, as indicated in Chapter 5. Being learned from data, it can automatically identify dependences regardless of it being related to human mobility or geographic proximity. The most similar approach found in the literature was proposed in Chan et al. [2015], which introduced spatial dependences between areas presenting co-occurrence of dengue fever cases. Our strategy, however, is broader, as it enforces spatial dependences between areas with similar DIR patterns.

Predictive Model In the reviewed works, we found 18 predictive models, with predictions made with a wide range of antecedence, from 1 day to 1 year. Antecedence and accuracy are two contradictory objectives. Higher antecedence is beneficial by allowing more time for health authorities to act, but leads to less accurate models as we cannot observed weeks that can be highly informative. Higher accuracy, on the other hand,

implies in more trustworthy predictions, but may require "fresher" data. In this work, we made predictions with 4 weeks of antecedence, a predictive window used by 10 of the reviewed works, long enough to for authorities to act while being short enough to allow accurate models.

Chapter 4

Dengue Data Characterization

As discussed in Chapter 2, a GP is mainly defined by its covariance function k. In order to use an appropriate covariance function, we must investigate useful relationships between variables in the data, so that the covariance function effectively exploits such relationships. In this chapter, we analyze our dengue data collection in order to obtain insights that will lead to the development of an adequate covariance function and, consequently, an accurate model. We first give an overview of our dengue dataset. Then, inspired by our literature review in Chapter 3, we analyze spatio-temporal patterns and relationship with other covariates, namely temperature, rainfall, relative humidity and volume of dengue-related tweets. Other covariates could be used, as suggested by some works reviewed in Chapter 3, but our goal is to define a model that could be simply instantiated and used, requiring only easily obtainable data.

4.1 Dengue Data Collection and Pre-Processing

We obtained the number of weekly confirmed dengue cases for 5303 Brazilian municipalities from January 2011 to December 2014, summing up to 209 weeks, provided by the Brazilian Ministry of Health. In order to account for different population sizes, we calculated the *dengue incidence rate* (DIR) as follows:

$$DIR_{s,t} = cases_{s,t} * \frac{100000}{pop_s} \tag{4.1}$$

where $DIR_{s,t}$ is the DIR at city s during week t, $cases_{s,t}$ is the number of confirmed cases at city s during week t and pop_s is the population size at city s.

The Brazilian Ministry of Health has defined a system of three *incidence levels* for dengue fever: high, medium and low. According to Lowe et al. [2013], high incidence

level at a given area occurs when there is more than 300 dengue cases per 100 thousands inhabitants during one month. Medium incidence, on the other hand, means more than 100 dengue cases and less than 300 dengue cases per 100 thousands inhabitants during one month. Finally, low incidence occurs when there is less than 100 cases per 100 thousands inhabitants at the same month. In this work, however, we deal with weekly dengue data. For that reason, we redefined these incidence levels by dividing the number of required cases by 4. Therefore, a high incidence week at a given area implies in a DIR of at least 75, while medium incidence requires DIR between 25 and 75 and low incidence requires DIR less than 25.

In order to reduce the effort of evaluating the proposed models, in this work we use only the cities with more than 100 thousands inhabitants, resulting in 298 cities. Although small in quantity, these cities account for more than 65% of the number of confirmed cases in Brazil. Therefore, cities for which an accurate and reliable EWS is fundamental in controlling dengue epidemics are included in the set of cities under study.

Figure 4.1 shows a map indicating the average DIR per state considering data from 2011 to 2014. The Southeast and Center-West regions were the most affected areas of the country, obtaining average DIR values equivalent to a high incidence level for almost all states. The Northeast region obtained average DIR values equivalent to medium incidence for almost all states, while the majority of states from the North and South regions obtained average DIR values equivalent to low incidence, being less affected by dengue fever.

In order to remove record errors, we pre-processed epidemiological data to remove additive outliers, that is, outliers that affect a single moment in time. We used the methodology proposed in Chen and Liu [1993], which identifies additive outliers (i.e., outliers that affect a single data point in time) by fitting auto-regressive models. This technique identified outliers happening most in the last and first week of each year, which are probably related to the reduced flux in health care posts due to holidays.

One issue when modelling DIR with GPs is dealing with count data using a Gaussian distribution, which may be inadequate. Here, we have two options: (i) change the model into a negative binomial or Poisson model or (ii) use an appropriate transformation on the response variable. An example of the former approach can be seen in Vanhatalo and Vehtari [2007], where the authors model the response variable as Poisson-distributed and the natural logarithm of its expected value as a GP. A similar strategy can be used with the negative binomial distribution, which has the advantage of handling over-dispersion. The downside of this approach is that it invalidates two interesting properties of GPs: mathematical tractability and exact predictive inference.

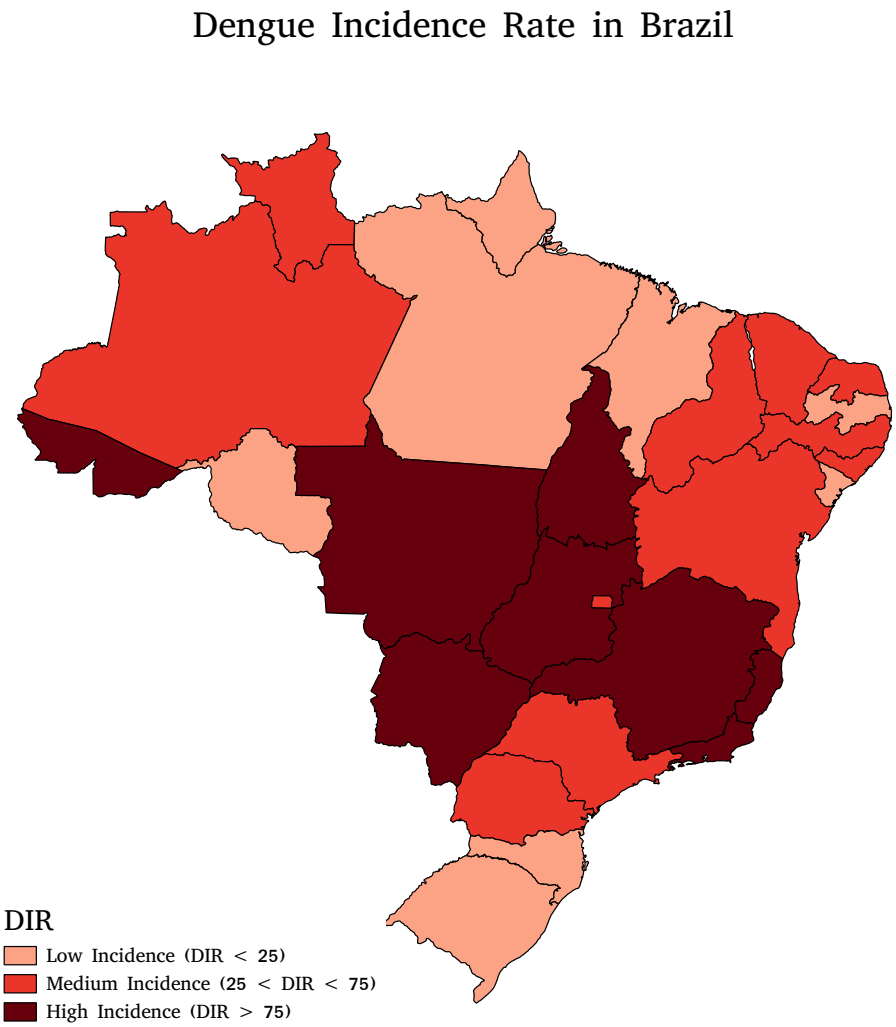

Figure 4.1. Average weekly DIR values per Brazilian state, with darker colors indicating higher incidence.

Inference under this kind of model requires more sophisticated techniques, typically Markov Chain Monte Carlo techniques as in Vanhatalo and Vehtari [2007], increasing the computational effort required.

For these reasons, we decided to apply a logarithmic transformation to the input data (after adding 1 to avoid logging 0) which, apart from reducing computational effort by allowing closed-form computations, also increases auto-correlation by smoothing unusually large dengue outbreaks. This approach, however, also has its downsides. Strictly speaking, it also does not allow exact inference, although it easily allows computation of exact confidence intervals. This is due to the fact that the model outputs the exact posterior distribution on the *logarithmic scale*, a Gaussian distribution with a given mean and covariance matrix. However, when exponentiating to return to the original scale, the predictive distribution is no longer a Gaussian distribution. Besides

that, it may present problems when used with cities with very low DIR. These cities, however, are the ones where dengue is not really a public health problem and for which an EWS would be less useful. Therefore, all following analysis were conducted using log-transformed DIR values.

4.2 Temporal Analysis

We first start by studying the temporal dependence within dengue data. From previous works on dengue modelling, we expect to see a local dependence, that is, dependence between successive or temporally nearby weeks, as well as a seasonal behavior (e.g., see Promprou et al. [2006]; Gharbi et al. [2011]; Martinez et al. [2011]). In order to visualize if this is true for our dataset, we computed the temporal auto-correlation for each city under study:

$$A_s(\tau) = cor(\boldsymbol{y}_{s,1:(N-\tau)}, \boldsymbol{y}_{s,(\tau+1):N}) \tag{4.2}$$

where $A_s(\tau)$ denotes the autocorrelation for city s and time lag τ, $\boldsymbol{y}$ are vectors containing log-transformed DIR values indexed by city and temporal range, N is the number of observations and *cor* denotes de Pearson correlation. Figure 4.2 shows the auto-correlation obtained for lags up to 156 weeks (approximately 3 years). It is possible to see a high correlation with small lags, meaning that nearby weeks are highly correlated in general. It is also possible to observe a weaker, yearly periodic correlation signal that slowly decays from the second year to the third, indicating the presence of seasonality.

4.3 Spatial Analysis

In our literature review, we observed that some models exploited spatial dependences, driven by the notion that dengue fever outbreaks are not geographically isolated events. To evaluate the spatial dependences in our dengue dataset, we compute Pearson correlation coefficient between each pair of cities.

Figure 4.3 shows the empirical cumulative distribution function of inter-cities correlations (left figure) and the distribution of these correlations with respect to inter-cities distances. These figures lead to two important conclusions. First, most pairs of cities exhibit low correlation, as only 8% of pairs of cities had a correlation coefficient greater than 0.7. This suggests that an useful spatial structure should not enforce dependences between each pair of cities, but only between a small subset of pairs. This reasoning will lead to some efficiency-related optimizations shown in Chapter

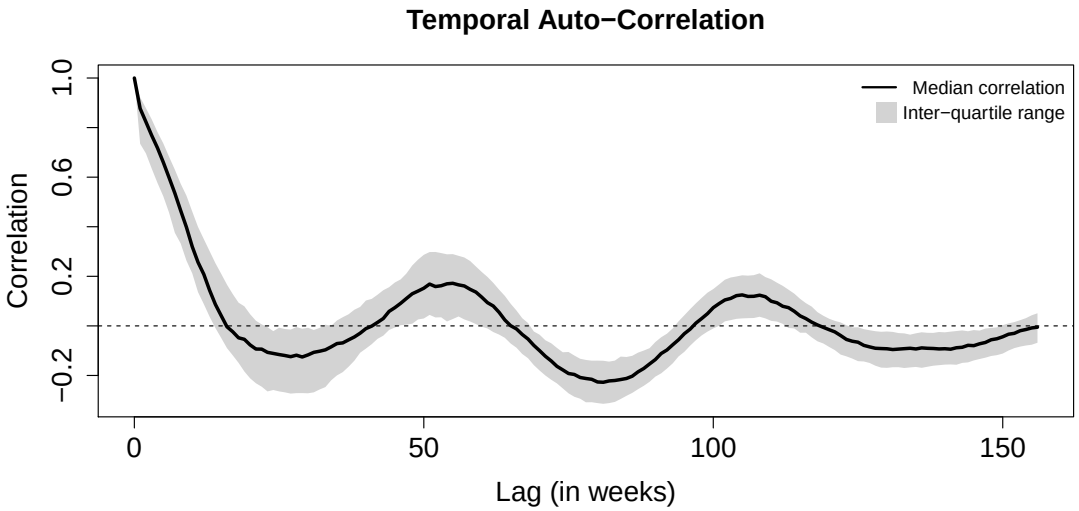

Figure 4.2. Median estimated temporal auto-correlation when considering all cities under study. Shaded area indicates inter-quartile range.

5. Second, although the distance between pair of cities seems to be associated with their correlation coefficient, the relationship between distance and spatial correlations is relatively weak, in the sense that it is not capable of drastically reducing the variability in data. Again, this fact will result in some of the conclusions discussed in Chapter 5.

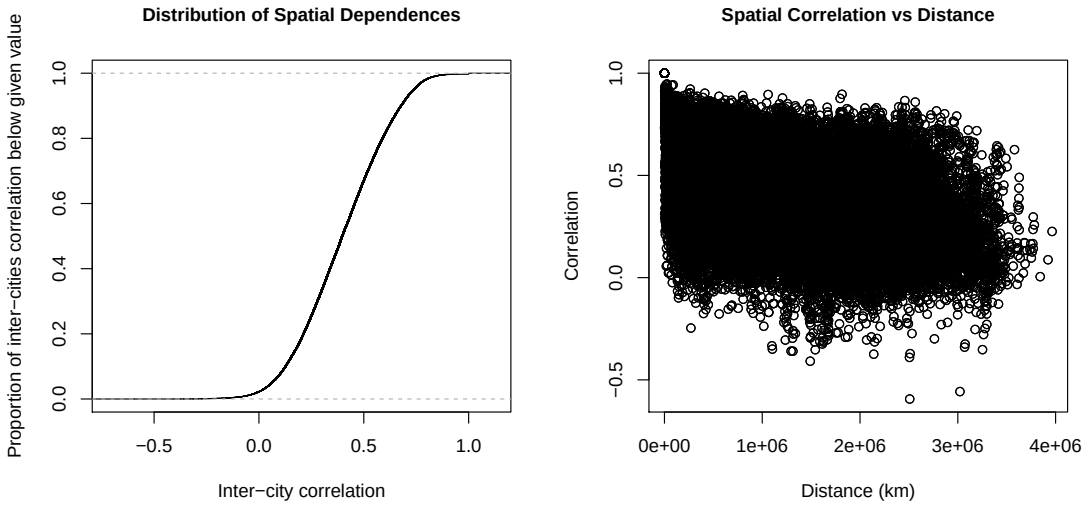

Figure 4.3. Empirical cumulative distribution function of spatial correlations (left figure) and the role of distance in spatial correlations (right figure).

4.4 Climate Dependences Analysis

From the literature, we also expect to see dependences between DIR and climate covariates, which are typically assumed to be linear (e.g., see Luz et al. [2008]; Lowe et al. [2013]; Shi et al. [2016]). Therefore, we collected daily average temperature, rainfall and relative humidity from 181 meteorological stations associated to INMET[1]. Since our epidemiological dataset provides weekly data, we transformed our daily climate data into weekly data by applying the average over each week. We associated each city with the closest meteorological station.

Climate covariates are typically used with some temporal lag, that is, DIR at a given week t is influenced by climate during week $t-\gamma$, $\gamma > 0$, as indicated in Chapter 3. We check for linear dependences and temporal lags by calculating the cross-correlations between epidemiological and climate-related time series while allowing shifting climate-related time series into the future, thus associating DIR at week t with climate during week $t - \gamma$:

$$C_s(\tau) = cor(\boldsymbol{y}_{s,(\tau+1):N}, \boldsymbol{c}_{s,1:(N-\tau)}) \tag{4.3}$$

where $C_s(\tau)$ denotes the cross-correlation for city s and time lag τ and $\boldsymbol{c}$ denotes a vector containing a climate-related covariate indexed by city and temporal range. Figure 4.4 shows the cross-correlation between DIR and climate covariates as a function of temporal lags. The three covariates, rainfall (top), average temperature (middle) and relative humidity (bottom), obtained peaking average correlations of 0.18, 0.26 and 0.18, respectively. Rainfall and temperature were better aligned with DIR time series using lags of 10 and 9 weeks, respectively, while higher cross-correlations between relative humidity and DIR were found when no lags are defined.

Figure 4.4 indicates relatively small cross-correlations between DIR and climate covariates, even if we define a lag per covariate in order to maximize correlation. However, it is also possible to allow lags specifically defined for each city. In this direction, Figure 4.5 shows that the optimal lag per city and covariate, that is, the lag that maximizes correlation between each covariate and DIR time series of a given city, can vary drastically. This motivated us to allow city-varying lags for all covariates, which increases cross-correlations between DIR and climate covariates. Using this approach, average peaking cross-correlations obtained are 0.28, 0.37 and 0.30 for rainfall, temperature and relative humidity, respectively, as indicated in Table 4.1.

[1]Brazilian Institute of Meteorology – Instituto Nacional de Meteorologia, in Portuguese: http://www.inmet.gov.br/portal/

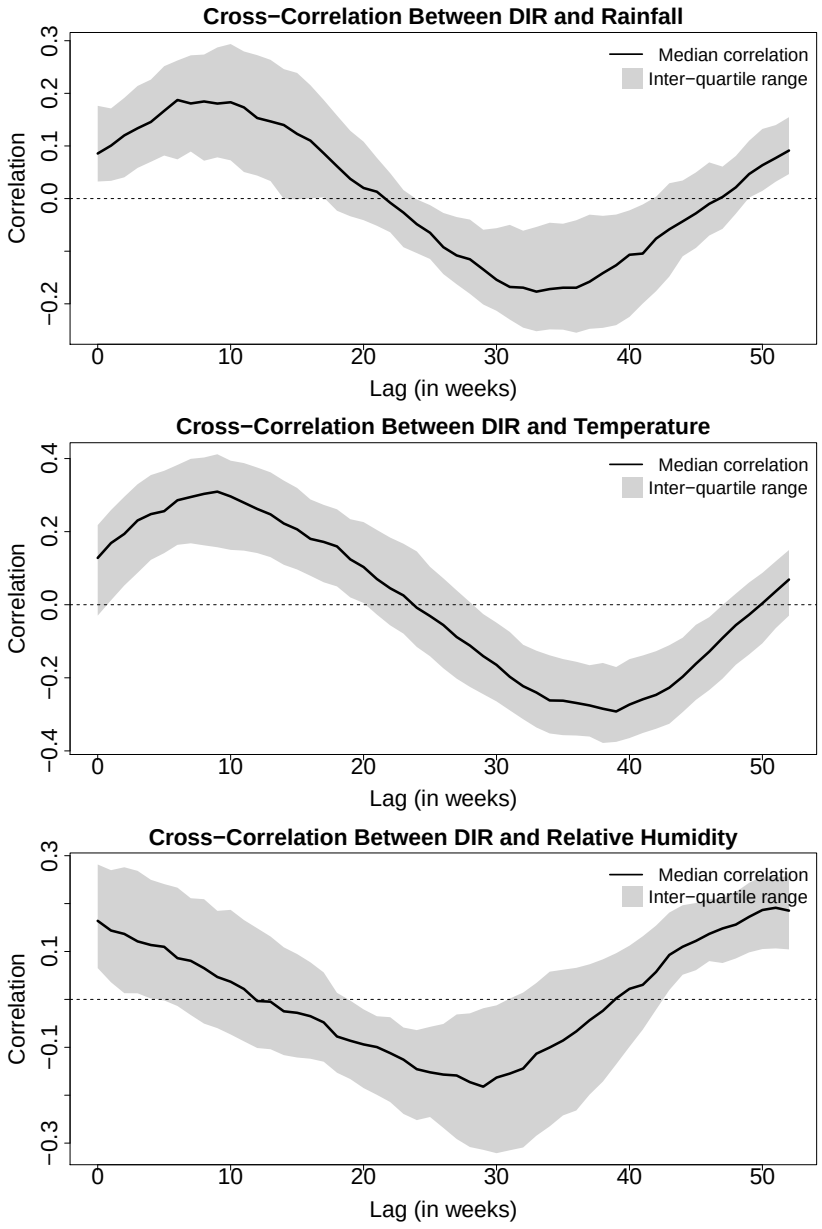

Figure 4.4. Average estimated cross-correlations between DIR and climate covariates considering all cities under study. Whiskers indicate 95% confidence intervals.

Table 4.1. Cross-correlation between DIR and climate covariates when lags are fixed and when lags are allowed to vary from one city to another.

Covariates	City-Varying Lags		Fixed Lags	
	Mean	Std. Dev.	Mean	Std. Dev.
Rainfall	0.28	0.10	0.18	0.15
Avg. Temperature	0.37	0.11	0.26	0.22
Rel. Humidity	0.30	0.12	0.18	0.17

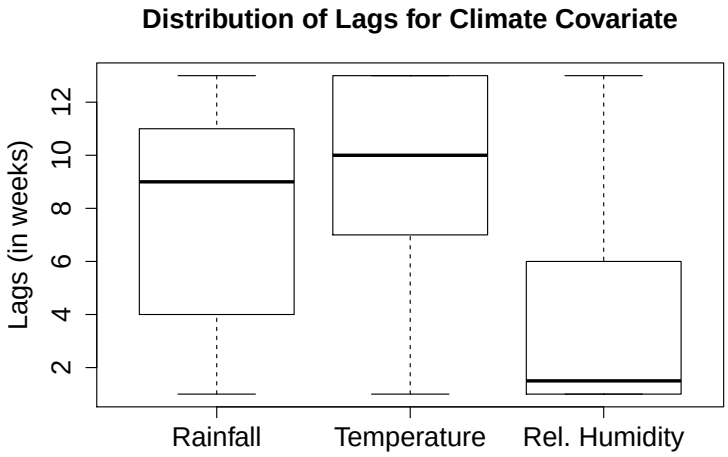

Figure 4.5. Distribution of the optimal lag per city for all three climate-related covariates.

4.5 Twitter Dependences Analysis

Recent works have already explored dependences between dengue and online data, more specifically the number of dengue-related tweets posted during a given period of time [Gomide et al., 2011; Souza et al., 2014, 2015]. The main intuition is that the online behavior of people may provide hints on incidence of dengue fever, as infected people can use online social networks to update friends about their health conditions, as well as use search engines and other online sources of information to search for symptoms or treatments. The main advantage of this kind of data when contrasted with epidemiological data is that online data can usually be obtained in real-time, while epidemiological data is usually published with a delay, since it requires, more often than not, confirmation tests and time to propagate through many levels of governmental hierarchy.

In this work, we use a dataset kindly provided by the authors of Souza et al. [2014] containing the weekly number of dengue-related tweets for the 298 Brazilian cities with more than 100,000 inhabitants, from January 2011 to December 2014. Based on this dataset and our dengue dataset, we calculated the cross-correlation for each city between the two data sources, obtaining the results exposed in Figure 4.6. For computing the cross-correlation, we used both Twitter data in the original scale, as well as in log-scale, since we have already computed the log-transform of DIR values. Note that the lag for the cross-correlation can now assume negative values. In that case, it would mean that Twitter data is delayed with relation to epidemiological data.

In other words, peaks on Twitter data would appear *after* peaks on epidemiological data. Similarly to the climate analysis, we first tried fixing a single lag value for all cities, obtaining maximum correlation for lags around zero or small negative values. We then allowed city-varying lags, but we did not observe a significant increase in average correlation, as shown in Table 4.2.

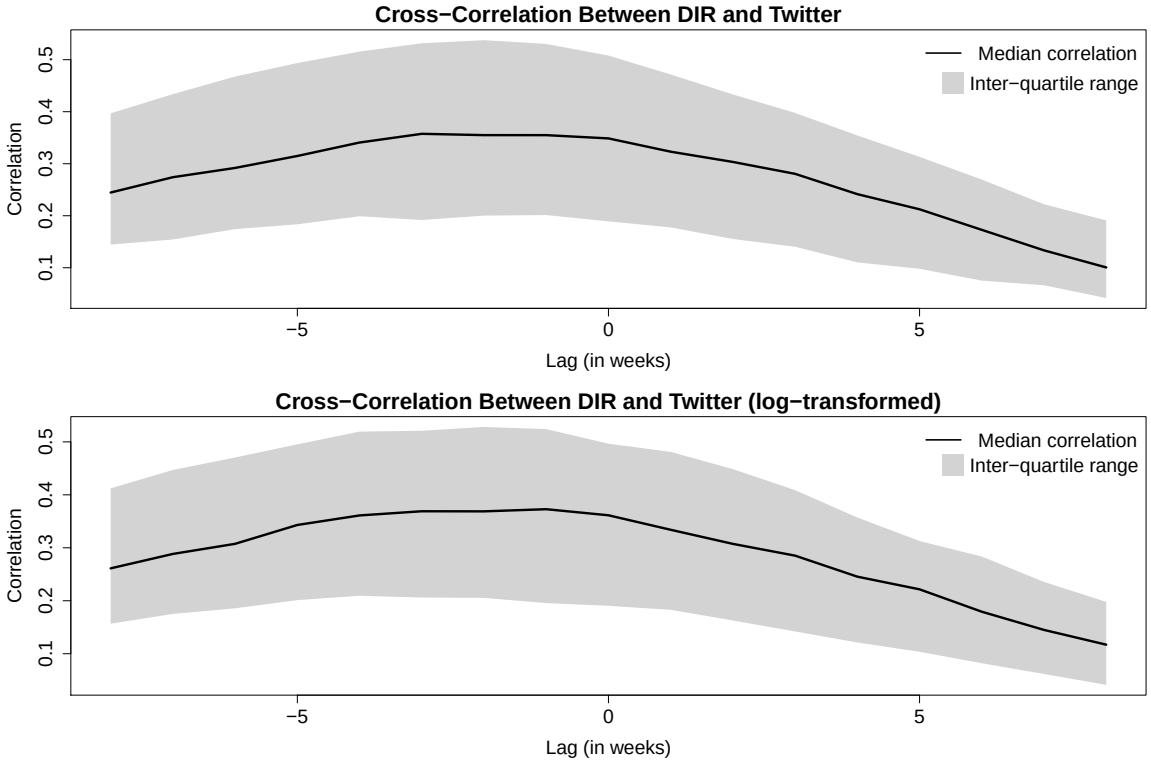

Figure 4.6. Average estimated cross-correlations between DIR and volume of dengue-related tweets considering all cities under study. Whiskers indicate 95% confidence intervals. Negative values indicate that Twitter data is delayed with relation to epidemiological data, while positive values indicate the opposite scenario.

Table 4.2. Cross-correlation between DIR and volume of dengue-related tweets when lags are fixed and when lags are allowed to vary from one city to another.

Covariates	City-Varying Lags		Fixed Lags	
	Mean	Std. Dev.	Mean	Std. Dev.
Twitter	0.36	0.21	0.40	0.18
Twitter (log scale)	0.37	0.21	0.41	0.19

From the standard deviation in Table 4.2, we observed that correlations between DIR values and Twitter data may vary drastically from one city to another. In order to better understand this variability and considering that the cross-correlation did not

pointed out that applying any kind of lag would lead to statistically better correlations, we investigated the impact of the total number of dengue-related tweets on the correlations between DIR values and Twitter data when no lag is applied, resulting in Figure 4.7. By looking at this figure, we can see that cities from which we have more Twitter data typically present higher correlation. This is expected, since a small number of dengue-related tweets will hardly be representative of the whole population. Low incidence cities also presented smaller correlations. For this kind of cities, dengue fever records exhibit patterns similar to white noise, making any useful correlation impossible.

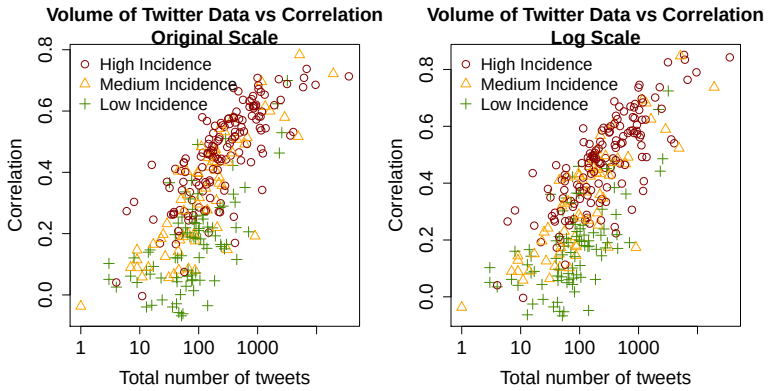

Figure 4.7. Impact of the total number of dengue-related tweets on the correlation between DIR values and Twitter data per city. Each symbol denotes a city, with color and shape indicating its highest incidence level achieved in the period under study.

4.6 Summary and Discussion

In this chapter, we characterized our dengue data collection in order to verify if patterns observed previously in the dengue literature were also present in our scenario. More specifically, we checked for five patterns: (i) local dependences, (ii) seasonality, (iii) spatial dependences, (iv) linear relationship with climate-based features and (v) linear relationship with the number of dengue-related tweets posted during a given period of time. This step is fundamental to propose an appropriate DIR forecast GP-based model, since we need to specify a covariance function that is capable of exploiting patterns present in data.

The temporal analysis was based on computing the auto-correlation for each city under study and revealed a strong local dependence, as well as a seasonality effect, verifying that the first two patterns indicated above are indeed present in our data.

The spatial analysis was performed based on the correlation between DIR time series of each pair of cities. We verified that spatial dependences exist, but that only a few pairs presented relevant correlation. We also verified that spatial dependences cannot be simply described by the distance between cities, requiring a more sophisticated formulation.

Dependences between DIR and climate-based features were studied through the cross-correlation of DIR and climate time series. For this analysis, we used weekly average temperature, rainfall and relative humidity. We observed that climate-based features were correlated to DIR values when considering a time lag. This is necessary due to the fact that climate affects the life cycle of the dengue's main vector, potentially leading to an increase of vector population. Therefore, time is required for new mosquitoes to get infected and start to infect human beings. Besides that, we observed that the time lag for each city seems to vary drastically. Therefore, in order to better model DIR, we need to use a specific time lag for each city.

Finally, we also confirmed the relationship between DIR and Twitter data previously indicated in the literature [Gomide et al., 2011]. For this analysis, we calculated the cross-correlation between DIR and Twitter data for each city, considering both Twitter data in its original and log scale, since DIR values were previously log-transformed. Both versions obtained similar results and could be used indistinguishably. Different from climate-based features, however, we noted that maximum correlation was obtained when a null or negative time lag was applied, indicating that Twitter data is synchronized or delayed with relation to epidemiological data. This indicates that Twitter data is not very useful when *predicting* DIR values, as it would require knowledge of *future* Twitter data. Besides that, correlations seem to be strongly associated to the number of dengue-related tweets on each city, making it useless for cities where this kind of data is not abundant. Therefore, Twitter data is not going to be used in our proposed model as a covariate, but as proxy to epidemiological data in an extension discussed in Chapter 7.

Chapter 5

Dengue Fever Incidence Modelling

In this chapter, we formalize and discuss our proposed model, named *dengue GP* model (DGP). In order to arrive at the proposed model, we will need to empirically compare candidate formulations. Therefore, we first introduce general aspects of our experimental design. Then, we discuss possible temporal and climate-related covariance functions and compare them empirically. After, we discuss the spatial component of our model, as well as efficiency issues that arise by using it and optimizations to mitigate them. Additional experiments are then performed to evaluate the impact of optimizations in both computational effort required for inference and accuracy of predictions.

5.1 Experimental Setup

As indicated in Chapter 1, our goal is to develop a predictive model capable of issuing accurate predictions with some antecedence, so that health authorities have time to act in order to minimize the impact of a new outbreak. Therefore, we evaluate all models with predictions made with 4 weeks of antecedence. That is, when predicting week t, we use only data available at week $t-4$. The choice of antecedence for predictions is a tricky one, as more antecedence allows for better planning on how to act to minimize the impact of dengue fever, but leads to less accurate models, as it would imply in not having access to weeks nearby the week of interest, which can be highly informative. Therefore, we chose 4 weeks because it would allow some time for planning and acting, but still allow the design of accurate models. The available data is then used for hyperparameter learning and computing the conditional predictive distribution for week t (Equation 2.3). We also defined that the first two years (weeks 1 to 104) would be used for training only. Predictions are then emitted for third and fourth years, with a constantly growing training set, as indicated previously.

In order to evaluate models, we defined three evaluation metrics, which are calculated per city under study:

1. **Pearson correlation coefficient:** we computed the correlation coefficient between real and predicted values for third and fourth years. Although not frequently used for measuring accuracy, this metric has the advantage of being bounded, easily interpretable and comparable between cities.

2. **Normalized mean absolute error (NMAE):** mean absolute error is a commonly use metric for evaluating regression models. However, it does not allow for comparison between cities, since the scale of DIR may change drastically from one city to another. In order to minimize this issue, we first normalize the response variable and predictions so that all cities had unit variance, and then calculated the mean absolute error between real and predicted values.

3. **Area under receiver operating characteristic curve (AUC):** we used the three incidence levels indicated in Chapter 4 (high, medium and low incidence) to transform the regression problem of predicting DIR into a classification problem. For each week and each city, we assigned real and predicted incidence levels, which were then interpreted as class labels. For each incidence level, we calculated the receiver operating characteristic (ROC) curve, which indicates the trade-off between true positive rate (sensitivity or recall) and false positive rate (1 - sensibility), that is, how much the error rate increases as we increase sensitivity. The area under a ROC curve is a metric used for summarizing this relationship and lies in the interval between 0 and 1. An area close to 1 indicates a good trade-off between true positive and false positive rate, that is, we can increase sensitivity incurring in only a few classification errors. On the other hand, an area of 0.5 indicates performance similar to a random classifier, while areas below 0.5 indicate performance worse than a random classifier. Since AUC is a metric defined for binary classification, we computed three AUC values, where one incidence level is assumed to be positive and the remaining to be negative. The final AUC is calculated as the mean of the three values. We highlight that cities that never reached medium and/or high incidence levels are not evaluated by this metric, since this city would have a single class label and AUC is not defined in this scenario.

After computing the evaluation metrics, we have to compare between competing models. In order to do so, we applied the Wilcoxon signed-rank test, using the scores

obtained for all cities per metric and a confidence level of 95%. Whenever multiple tests are required, we applied a Bonferroni correction to ensure statistical significance.

Experiments were conducted using GNU Octave[1] 4.0.0, GPML toolbox [Rasmussen and Nickisch, 2010] and R[2] 3.0.2. Statistical tests were conducted using R 3.0.2. Time measurements were conducted using a single core of a machine equipped with a octa-core Intel Xeon E5620 2.40GHz processor and 72 GB of RAM.

5.2 Temporal and Climate-Related Covariance Function

Recall that a GP is defined by a mean and a covariance function. While the mean function is typically assumed to be zero after centering the response variable, the covariance function is fundamental to capture patterns present in data.

The analysis exposed in Chapter 4 shows that our data collection exhibit three commonly exploit patterns in dengue literature: local dependences (smoothness), seasonality and linear relationship with climate. This reasoning motivates a three-component covariance function:

- The first component enforces local dependences by correlating temporally nearby weeks;
- The second component enforces quasi-periodicity to exploit seasonality, allowing usage of information from past years while giving less relevance to more distant years;
- The third component allows for linear dependences between climate covariates and DIR.

In summary, the proposed covariance function has the following general formulation:

$$k_x(\boldsymbol{x}_t, \boldsymbol{y}_{t'}) = k_{loc}(t, t') + k_{qp}(t, t') + k_{weather}(\boldsymbol{x}, \boldsymbol{y}) \tag{5.1}$$

where $\boldsymbol{x}_t$ indicates a data point associated to time index t, $\boldsymbol{y}_{t'}$ indicates a data point associated to time index t', k_{loc} is the local component, k_{qp} is the quasi-periodic component and $k_{weather}$ is the weather-related component. Note that the first two components are temporal-only, while the third does not take time indices into consideration.

[1] http://www.gnu.org/software/octave/
[2] http://cran.r-project.org/

5.2.1 Evaluation of Candidate Covariance Functions

The proposed covariance function (Equation 5.1) is composed of three components: a local, a quasi-periodic and a weather-related component. While the linear weather-related component is formulated using a linear kernel, the local and quasi-periodic components can be formulated in multiple ways. The goal of this section is to verify which candidate formulation issues predictions with the highest accuracy.

The local component requires a function that decays with temporal distance, while the quasi-periodic component is expressed as the product of a periodic function with a function that also decays with temporal distance. The GP literature provides some covariance functions that have the desired behavior of monotonic decay, as indicated in Chapter 2, leading to the following candidate formulations: (i) squared exponential kernel, (ii) Matérn kernel with $\nu = 3/2$ and (iii) Matérn kernel with $\nu = 5/2$. All three formulations hold similar assumptions, but vary on the smoothness of the obtained function.

Figure 5.1 shows the comparison between three candidate covariance functions obtained by equipping the local and quasi-periodic components with the Matérn kernel with $\nu = 5/2$, the Matérn kernel with $\nu = 3/2$ and the squared exponential kernel. The weather-related component was fixed with a linear kernel. Since this is the first set of comparative graphs and similar graphs will be presented later, we explain here how this kind of graphs should be read. Each symbol in the graphs denote a city, with color and shape indicating the highest incidence level reached by the city during the period under study. Each graph is associated to an evaluation metric (indicated in the title) and is used to compare between two models, which are indicated in the x-axis and y-axis. Given a point, its coordinates indicate the value obtained for the respective city according to the respective evaluation metric for both models being compared. Therefore, the solid diagonal line indicates equal performance, with points not in the diagonal line denoting differences in accuracy. Points above the diagonal line indicate higher values achieved for the model in the y-axis, while points below the diagonal indicate higher values achieved for the model in x-axis. Note that correlation coefficient and AUC are intended to be maximized, so higher values are preferred. NMAE, on the other hand, is intended to be minimized, making smaller values preferred. Considering Figure 5.1, we observed similar accuracy for all three formulations when considering correlation coefficient and NMAE. However, according to AUC, the Matérn covariance function with $\nu = 5/2$ obtained statistically better performance than alternative formulations.

The formulation in Equation 5.1 is based on the data characterization and on the previous literature, but it is also important to verify if each component is really neces-

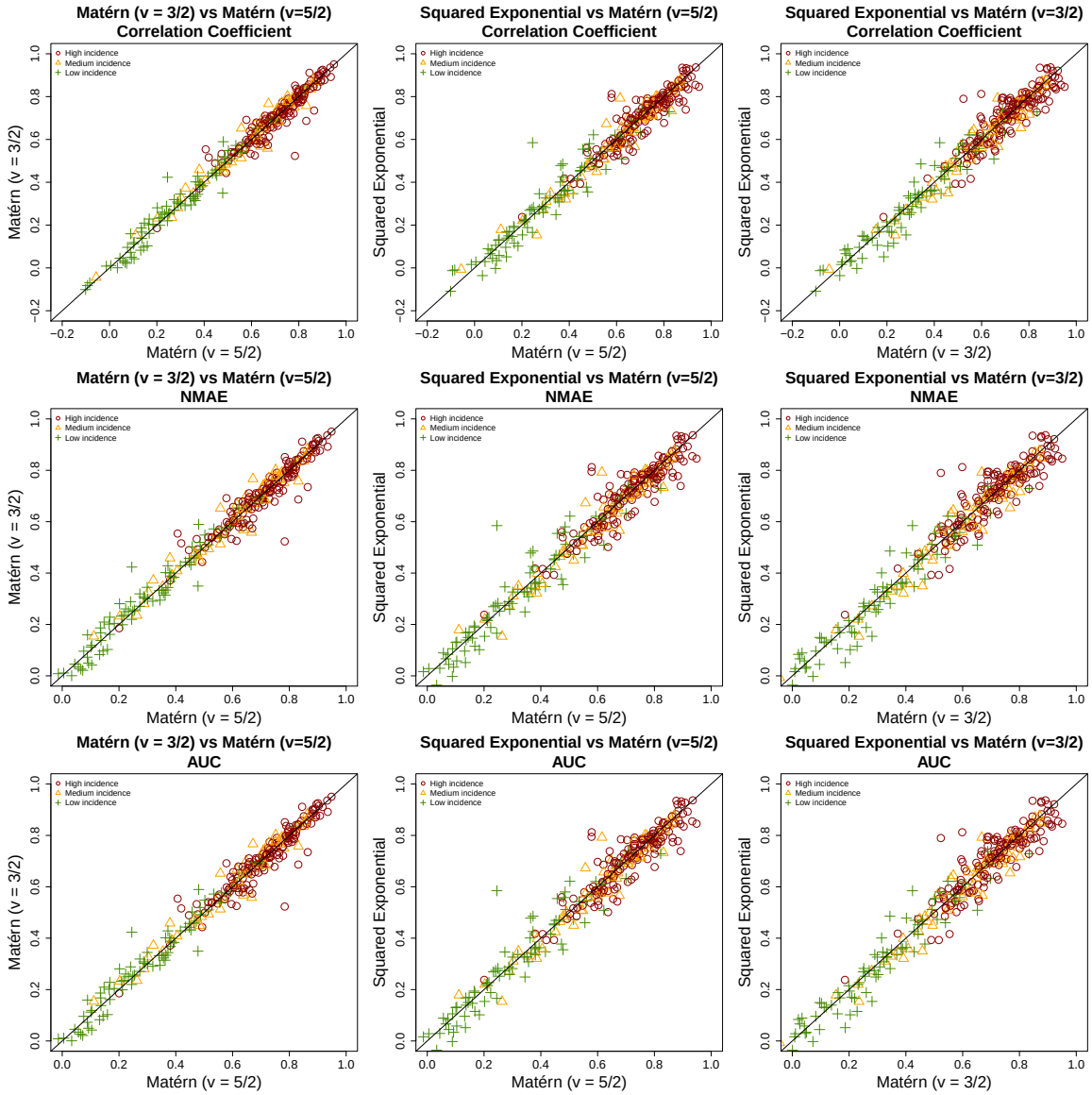

Figure 5.1. Comparison between alternative formulations of quasi-periodic covariance functions. Each symbol represents a city and color and shape indicate the highest incidence level achieved by the corresponding city. The solid black line indicates equal performance between models. Cities above the solid line indicate higher values obtained by the y-axis formulation, while cities below the line indicate higher values obtained by x-axis formulation.

sary. In order to do so, we defined three new alternative formulations of the proposed function, each ignoring one of the three components. Figure 5.2 shows the results according to all three evaluation metrics. According to correlation, all formulations obtained similar performance. When considering NMAE, on the other hand, excluding the local component led to statistically worse results, although differences were marginal in absolute values. Finally, according to AUC, excluding any component led

to statistically worse results. Since for at least one evaluation metric excluding a component was detrimental, we conclude that the all components contribute to improving accuracy of predictions.

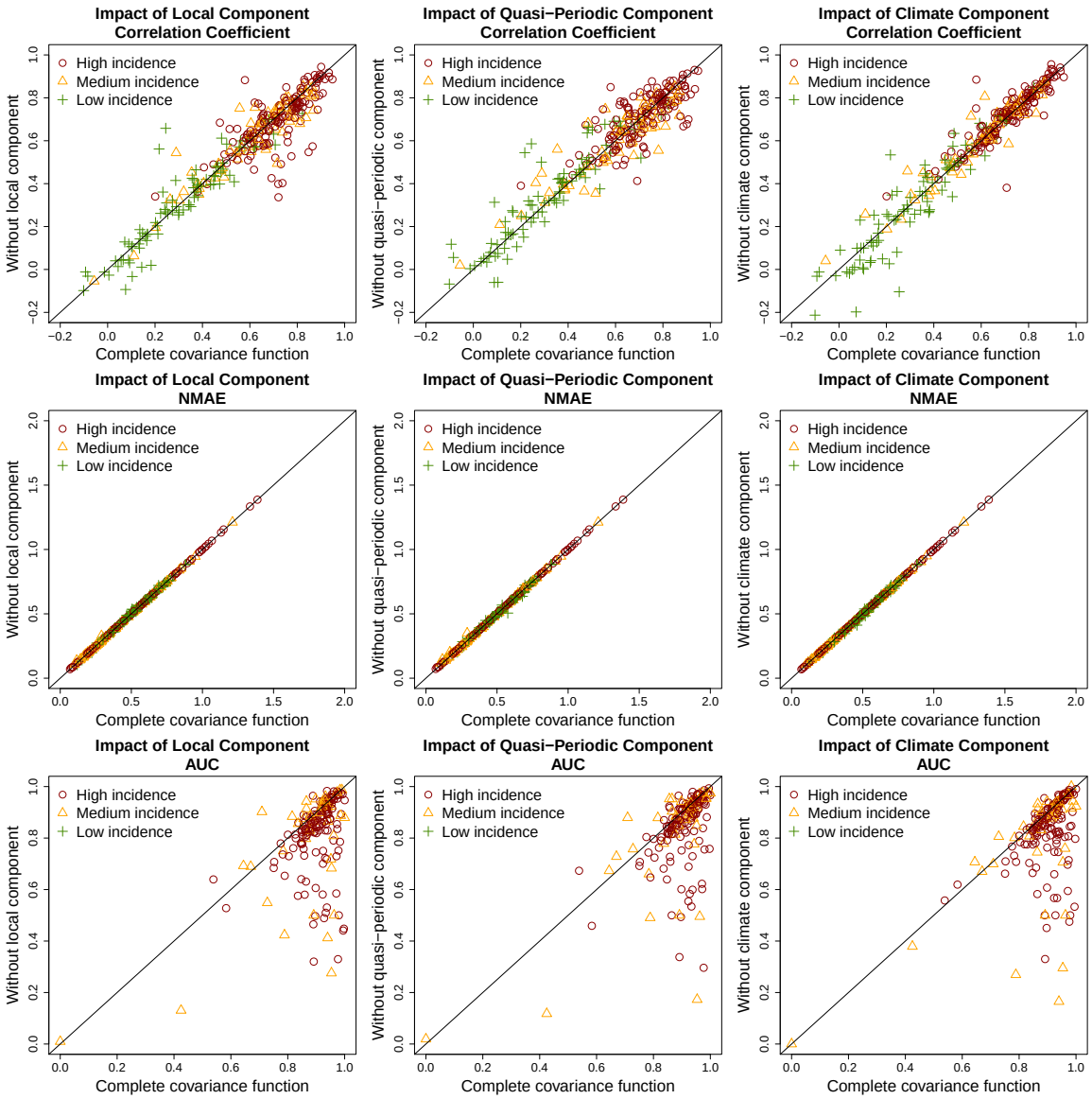

Figure 5.2. Comparison between alternative formulations of the proposed covariance function obtained by removing one of its three components. Each symbol represents a city and color and shape indicate the highest incidence level achieved by the corresponding city. The solid black line indicates equal performance between models. Cities above the solid line indicate higher values obtained by the y-axis formulation, while cities below the line indicate higher values obtained by x-axis formulation.

An alternative formulation to the quasi-periodic component, as discussed in Chapter 5, is to use the spectral mixture kernel (Equation 2.16), which is defined as a sum-

mation of quasi-periodic functions, thus being able to discover other patterns besides annual seasonality. Figure 5.3 shows the comparison between results obtained by DGP equipped with the covariance function in Equation 5.1 with a single quasi-periodic function and results obtained with the spectral mixture kernel with 5 components. For all evaluation metrics, the single function obtained statistically better results. This suggests that being able to explore other periodicities besides annual periodicity is not necessarily beneficial, and may be even detrimental. A possible explanation is the increased number of hyperparameters, which may lead to overfitting or optimization issues.

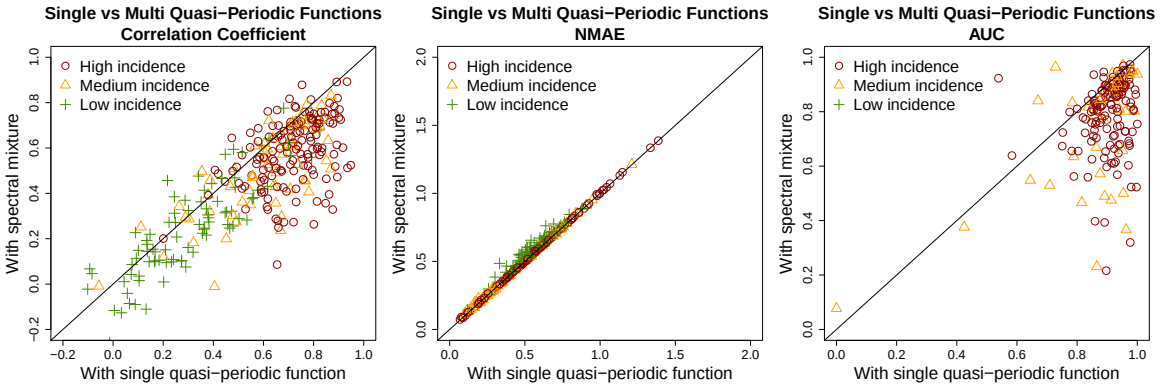

Figure 5.3. Comparison between the covariance function with a single quasi-periodic function and the spectral mixture kernel. Each symbol represents a city and color and shape indicate the highest incidence level achieved by the corresponding city. The solid black line indicates equal performance between models. Cities above the solid line indicate higher values obtained by the y-axis formulation, while cities below the line indicate higher values obtained by x-axis formulation.

Based on the results above, we arrived at the following covariance function

$$k_x(\boldsymbol{x}_t, \boldsymbol{y}_{t'}) = k_{Mat_1}(t, t') + k_{Mat_2}(t, t')k_{Per}(t, t') + k_{Lin}(\boldsymbol{x}, \boldsymbol{y}) \tag{5.2}$$

where k_{Mat_1} and k_{Mat_2} indicates Matérn kernels with $\nu = 5/2$ and distinct hyperparameters (Equation 2.11), k_{Per} indicates a periodic kernel (Equation 2.12) and k_{Lin} indicates a linear kernel (Equation 2.13). Table 5.1 summarizes numerically the comparison between the covariance function above and the alternative formulations evaluated. As indicated previously, all formulations performed similarly according to correlation coefficient and NMAE, with no statistically significant differences or marginal differences. However, the proposed covariance function obtained significantly better results according to AUC when compared to all other formulations.

Table 5.1. Difference according to each evaluation metric between DGP equipped with proposed covariance function and with alternative formulations.

Candidate Covariance Function	Evaluation Metric	Difference Between Proposed and Alternative Formulations		Number of Cities Where Proposal Wins
		95% Conf. Interval	Median	
Matérn $\nu = 3/2$	Correlation	[-0.005, 0.001]	-0.002	139 (47%)
	NMAE	[0.000, 0.000]	0.000	152 (51%)
	AUC	[0.032, 0.062]	0.046	135 (75%)
Squared Exponential	Correlation	[0.000, 0.010]	0.005	158 (53%)
	NMAE	[0.000, 0.000]	0.000	162 (54%)
	AUC	[0.031, 0.065]	0.046	137 (76%)
DGP (without local component)	Correlation	[0.002, 0.015]	0.009	162 (54%)
	NMAE	[-0.001, 0.000]	0.000	178 (60%)
	AUC	[0.034, 0.066]	0.046	136 (75%)
DGP (without quasi-periodic component)	Correlation	[-0.001, 0.006]	-0.002	144 (48%)
	NMAE	[0.000, 0.000]	0.000	164 (55%)
	AUC	[0.022, 0.041]	0.030	136 (75%)
DGP (without climate component)	Correlation	[-0.007, 0.003]	-0.002	140 (47%)
	NMAE	[0.000, 0.000]	0.000	140 (47%)
	AUC	[0.026, 0.056]	0.038	121 (67%)
Spectral Mixture	Correlation	[0.083, 0.115]	0.099	231 (78%)
	NMAE	[-0.007, -0.003]	-0.005	217 (73%)
	AUC	[0.073, 0.116]	0.093	161 (89%)

5.3 Including Spatial Dependences to the Covariance Function

The covariance function expressed in Equation 5.2 treats dengue fever outbreaks as geographically independent events, since it does not include any spatial dependences. In order to introduce spatial dependences in the covariance function, we view the problem of estimating future DIR values at Brazilian cities as a multi-task learning problem. Under this formalism, each city is treated as a task and, therefore, models for each city are learned jointly, enabling knowledge transfer between models.

As indicated in Chapter 2, multi-task approaches for GP modelling have already been proposed in the literature. Note that extending a GP model to a multi-task scenario for dengue data in Brazil is not an easy task due to the large number of cities. In this work, we use up to 298 cities, and multi-task GP models were not designed for such a large number of tasks, specially if tasks do not share the same data points. We opted for using the multi-task GP model proposed in Bonilla et al. [2007b], which simply uses a inter-task covariance matrix to induce covariance between data

points from distinct tasks. Entries of the inter-task covariance matrix are learned via likelihood maximization and the matrix is required only to be symmetric and positive definite. Although the complexity for inference under this model is cubic on the number of tasks, we chose this model for three major reasons: (i) it is highly interpretable as the inter-task covariance matrix can be seem as a measure of dependence between tasks, (ii) it does not require any sophisticate inference technique and (iii) it enables simple modifications that can drastically reduce the computational effort required for inference.

Therefore, after including the spatial component, the covariance function used by Dengue GP model (DGP) can be defined as follows. Let $\boldsymbol{x}_t^{(i)}$ and $\boldsymbol{y}_{t'}^{(j)}$ be data points associated to time t and t' and cities i and j, respectively, K_C be the inter-cities (inter-task) covariance matrix and $K_C(i,j)$ be the value of K_C for cities i and j. Then, the covariance function $k\left(\boldsymbol{x}_t^{(i)}, \boldsymbol{y}_{t'}^{(j)}\right)$ is given by

$$k\left(\boldsymbol{x}_t^{(i)}, \boldsymbol{y}_{t'}^{(j)}\right) = K_C(i,j)k_x(\boldsymbol{x}_t, \boldsymbol{x}_{t'}) \tag{5.3}$$

where $k_x(\boldsymbol{x}_t, \boldsymbol{x}_{t'})$ is defined in Equation 5.2 and does not take into account the tasks from which data points come from.

In summary, we can define the proposed model as follows:

$$\begin{aligned} DIR_{s,t} &= exp(z_{s,t} + \bar{z}_s) - 1 \\ z_{\cdot,\cdot} &\sim \mathcal{GP}\left(0, k\left(\boldsymbol{x}_t^{(i)}, \boldsymbol{x}_{t'}^{(j)}\right) + \delta_{ij}\delta_{tt'}\sigma_n^2\right) \end{aligned} \tag{5.4}$$

where $DIR_{s,t}$ is the DIR at city s during week t, $\bar{z}_s$ is the log-transformed average DIR for city s and δ_{ij} denotes the Kronecker delta, which is equal to 1 if and only if $i = j$ and 0 otherwise.

5.3.1 Improving the Performance of DGP

A problem with DGP as proposed above is the computational effort required for inference. Assuming that all cities have the same number of data points, inference requires $O(N^3M^3)$ operations, where N is the number of data points *per city* and M is the number of cities. This is caused by the need of inverting the full $(NM) \times (NM)$ covariance matrix obtained by applying the covariance function in Equation 5.3 to all pairs of data points, as previously discussed in Chapter 2. In this work, we explore two strategies for reducing the computational effort required for inference: turning K_C into a block-diagonal matrix and exploiting Kronecker structure using a temporal-only

version of DGP.

Another issue with DGP is the large number of hyperparameters to be optimized. We dealt with this issue by fixing K_C based on an empirical approximation using available training data, instead of allowing it to be learned via likelihood maximization.

All three strategies are presented in more detailed in the following sections.

5.3.1.1 Block-Diagonal Inter-Cities Covariance Matrix

The covariance function indicated in Equation 5.3 is capable of inducing non-zero covariance between data points from all pairs of cities. However, it is not expected for all pairs of cities to present significant dependence. In fact, we expect only a subset of the cities to be strongly associated with a given city. This motivates sparsifying the inter-cities covariance matrix K_C. However, turning entries of K_C to null arbitrarily may violate the positive definiteness of the matrix.

A safe strategy to sparsify K_C while still keeping it positive definite is to turn it into a block-diagonal matrix. Each block is the covariance matrix for a subset of cities and, consequently, is positive definite by construction. Since each block is positive definite, the whole matrix will also be positive definite. K_C being block-diagonal, the full covariance matrix K obtained by applying Equation 5.3 to all pairs of data points will also be block-diagonal (possibly after an appropriate ordering of rows and collumns). In order to calculate K^{-1}, we can now deal with each block *individually*, which reduces the computational complexity. In fact, assuming all blocks have size S, inference can be done in $O(N^3S^2M)$, which can lead to drastic reduce in computational effort if $S << M$.

For forming appropriate blocks of cities, we applied an agglomerative complete-link hierarchical clustering algorithm [Maimon and Rokach, 2005]. Cities from the same cluster will be on the same block and may have non-null covariance, while cities from distinct clusters will have null covariance, thus being independent. We opted for the hierarchical clustering because it allowed us to easily introduce an extra constraint: the maximum allowed number of cities within each cluster. This constraint is relevant as it is directly associated to the computational effort required for inference and should be defined by the user. Other clustering algorithms, such as k-means, do not allow for an easy control of clusters' sizes, and use extra parameters.

The clustering strategy used is described in Algorithm 5.1. At each iteration, it computes the complete-link distance between each pair of clusters and merges the two closest clusters whose union would not violate the maximum size constraint. For computing the distance between each pair of cities, we used the correlation between

their log-transformed DIR time series:

$$dist(i,j) = 1 - cor\left(log(DIR_i + 1), log(DIR_j + 1)\right) \tag{5.5}$$

where DIR_i is the time series of DIR values for city i and $cor(\cdot,\cdot)$ denotes the Pearson correlation coefficient. This is a natural choice for measuring distance as K_C is a covariance matrix. Thus, by applying the clustering strategy, we expect to maintain high covariance values present in the original inter-cities matrix.

5.1: Agglomerative Complete-Link Hierarchical Clustering Algorithm

Input: log-transformed centered DIR values for all cities $\boldsymbol{z}$, number of cities M, number of weeks N, maximum cluster size MAX
Output: clusters obtained C

$D_{ct}(i,j) \leftarrow$ Equation 5.5, $i,j = 1,2,...,M$, $i \neq j$
$C(i) =\leftarrow \{i\}$, $i = 1,2,...,M$
$D_{cl}(i,j) = D_{ct}(i,j)$
$S_1 \leftarrow 1$
$S_2 \leftarrow 1$
while $S_1 + S_2 \leq MAX$ **do**
    $(i^*, j^*) \leftarrow \arg\min_{i \neq j} D_{cl}(i,j)$ subject to $|C(i) \cup C(j)| \leq MAX$
    Add $C(i^*) \cup C(j^*)$ to C
    Remove $C(i^*)$ and $C(j^*)$ from C
    $D_{cl}(k,l) = max_{i \in C(k), j \in C(l)} D_{ct}(i,j)$, $k,l = 1,2,...,|C|$
    $S_1 \leftarrow$ size of smallest cluster in C
    $S_2 \leftarrow$ size of second smallest cluster in C
end
Return C

Figure 5.4 shows a didactic example of how block-diagonalization is performed. First, we use Algorithm 5.1 to clusters cities, obtaining the clusters indicated by the colors in the left figure. Then, rows and collumns of the covariance matrix K obtained by applying Equation 5.3 to all pairs of data points are orded in a way that K becomes block-diagonal, with each cluster associated to a block, as indicated by the red, green and blue blocks in the right figure. The white area denotes the covariances between cities from distinct clusters, which are assumed to be zero.

5.3.1.2 Exploiting Kronecker Structure on Temporal-Only DGP

As indicated in Chapter 2, the multi-task approach proposed in Bonilla et al. [2007b] and adopted by DGP can save computational effort if all tasks share the same data points. This is due to the fact that the full covariance matrix K can be expressed

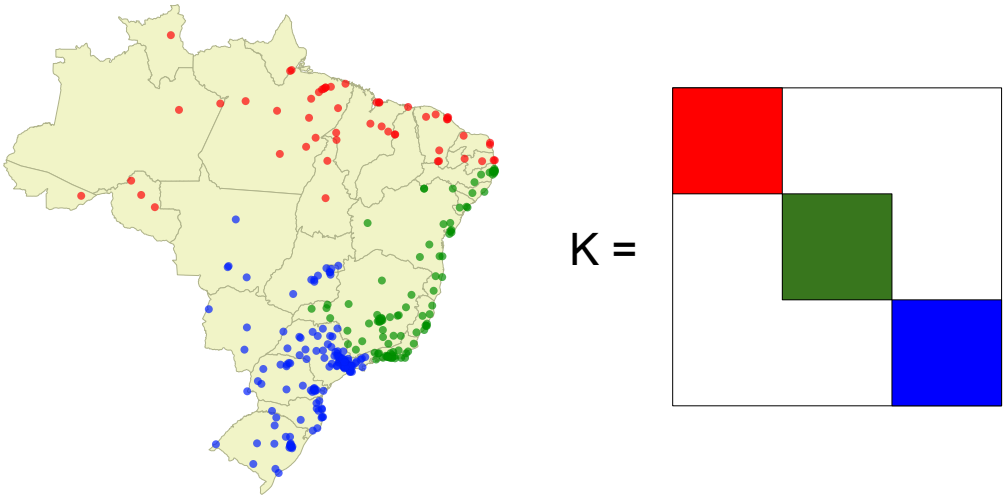

Figure 5.4. Illustrative example of block-diagonalization of the covariance matrix K. In the left figure, each symbol denotes a city, with color indicating the cluster it belongs. The right figure shows the resulting covariance matrix, with white space indicating null values and non-white space indicating covariances between cities within the same cluster.

Table 5.2. Computational complexity when using proposed strategies

	Regular K_C	Block-Diagonal K_C
Full Model	$O(N^3M^3)$	$O(N^3S^2M)$
Temporal-Only Model	$O(N^3 + M^3)$	$O(N^3 + S^2M)$

in terms of a Kronecker product between the task-related and data-related covariance matrices. This structure allows us to compute K^{-1} by dealing with each matrix individually and to perform fast matrix-vector multiplications.

In the context of this work, the Kronecker structure will be present whenever we do not use climate-related covariates. Then, the model will use only temporal indices, which are the same for all cities. By exploiting this structure, inference requires only $O(N^3 + M^3)$ operations.

Note that both strategies presented so far are orthogonal and can be used in conjunction, leading to a even more drastic reduction in computational effort. Table 5.2 summarizes the computational complexity of each variant, where N is the number of data points per city, M is the number of cities and S is the maximum size allowed for each block for the block-diagonal K_C variants.

5.3.1.3 Empirical Approximation of Inter-Cities Matrix

The block-diagonalization of the inter-cities matrix K_C described above, besides reducing computational effort required for inference, also helps reducing the number of hyperparameters to be optimized under the DGP model. Originally, DGP used $O(M^2)$ parameters, as K_C was structure-free. After block-diagonalization, the number of parameters drops to $O(SM)$.

Another effective way to reduce to the number of hyperparameters is to approximate K_C using available training data. That is, at week t, when it is required to provide forecasts for week $t+4$, we compute the Pearson correlation coefficient between each pair of cities (or each pair of cities within the same block) using dengue data up to week t. The main intuition behind this approximation is that the covariance between a pair of cities should not change drastically from week t to week $t+4$, specially when t is large. Since it is based on training data, we call this strategy *empirical approximation* of K_C. The inter-cities matrix is then fixed for optimizing the remaining hyperparameters and inference.

5.3.2 Evaluation of the Spatial Component

Having proposed three optimizations for reducing the computational effort required for inference, we now present results from experiments designed to measure the impact of each optimization, from both efficiency and accuracy points of view.

5.3.2.1 Clustering Analysis

In order to analyze the impact of block-diagonalization of the inter-cities covariance matrix K_C, we conducted experiments by applying Algorithm 5.1 with distinct maximum cluster sizes, ranging from 10 to 50. For comparison, we also conducted experiments by grouping cities by state and by region. To reduce computational effort, experiments conducted here used the temporal-only version of DGP, using empirical approximation of the K_C or letting it be learned via likelihood maximization. We then extrapolate conclusions for the full version of the model.

We first consider the clusters obtained through Algorithm 5.1 when considering distinct maximum size values. Figure 5.5 shows the number of clusters obtained when the maximum size of clusters grows from 10 to 50 cities. For comparison, the black dotted line indicates the number of clusters obtained when grouping cities per state, while the red dotted lines indicates the number of clusters obtained when grouping cities per administrative region. As expected, the larger the cluster sizes allowed,

the smaller the number of clusters. Note, however, that grouping cities by state led to a comparable number of clusters obtained by Algorithm 5.1 with maximum size of 10 cities per cluster. In contrast, grouping per region leads to a smaller number of clusters than allowing clusters composed by as much as 50 cities. Figure 5.1 also indicates how close cities within the same cluster are when clustering by state, by region or by correlation. Clustering by state or by region leads to clusters limited by distance between pairs of cities, while clustering by correlation does not. Even when considering relatively small clusters, the clusters obtained are composed of cities separated by greater distances. In fact, approximately 30% of the cities within the same clusters were separated by more than 1000 km when clustering by correlation. When considering clustering by region, less than 10% of cities within the same clusters were separated by more than 1000 km, while this percentage decreases to less than 1% when considering clustering by states.

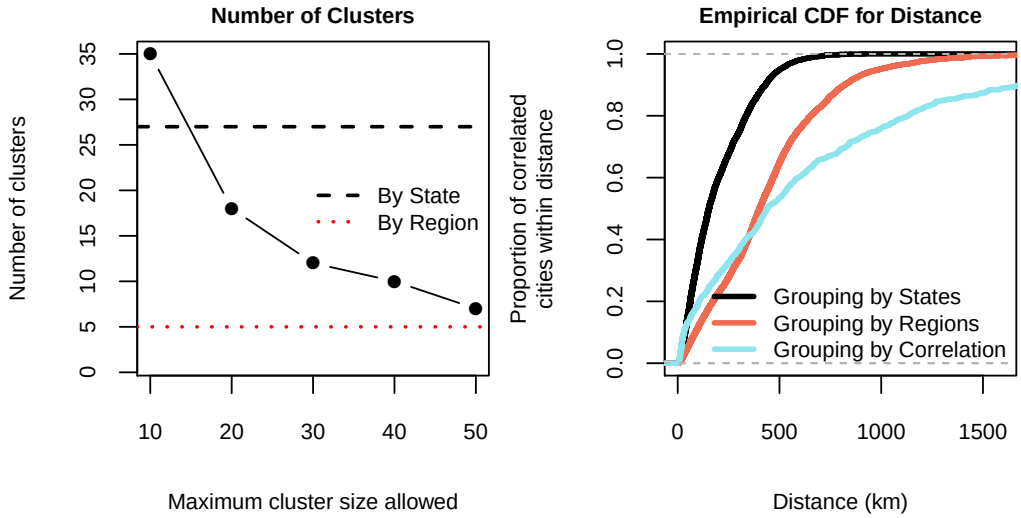

Figure 5.5. The left figure shows the number of clusters obtained according to the maximum size allowed in Algorithm 5.1. Dotted lines indicate number of clusters when clustering by state and region for comparison. The right figure shows the proportion of cities within the same clusters separated by a given distance when clustering by states, by region and by correlation with maximum allowed size of 10 cities.

Figure 5.6 shows the results obtained by DGP for clusters formed with distinct maximum size constraints, as well as clusters obtained by grouping cities by state or region, for both versions using fixed or optimized inter-cities covariance matrices. For both versions and all three evaluation metrics, the best results were obtained when using small clusters defined by correlation, with the impact of maximum cluster size and clustering strategy are more noticeable when using optimized matrices. We believe this

difference is due to the high number of hyperparameters to be optimized when larger clusters are used. Besides that, the fact that DGP using correlation-based clustering outperformed the same model using distance-based clustering suggests that distance is not a major factor when considering the spatial dependences of dengue outbreaks. This is problematic, since most works on dengue modelling that enforce some kind of spatial dependence use topological or distance features only. In contrast, DGP is able to automatically identify appropriate dependences in a more flexible fashion.

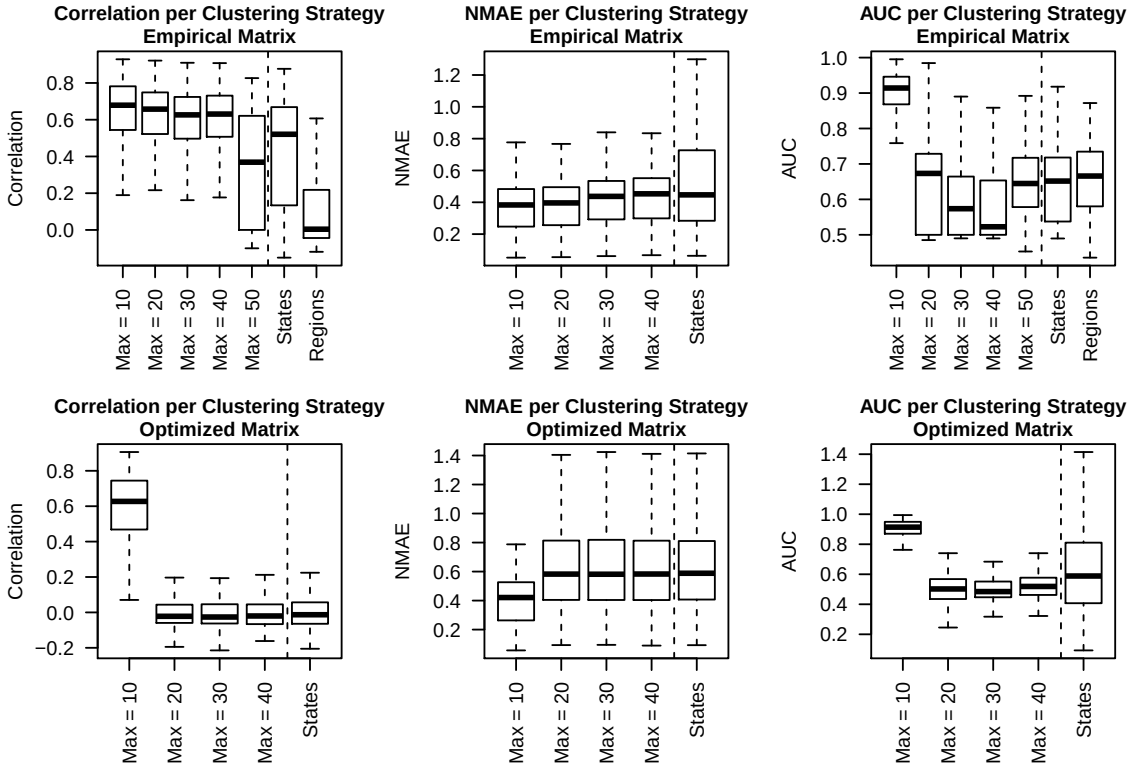

Figure 5.6. Results obtained by DGP for distinct clustering strategies according to all three evaluation metrics. Experiments with optimized covariance matrix and clusters of size up to 50 cities or grouped by states were not performed due to the large computational effort required.

Figure 5.7 shows the time required for inference with $N = 209$ weeks, $M = 298$ cities and maximum cluster sizes ranging from 10 to 50 for both versions using optimized or fixed inter-cities covariance matrices. Given the complexity shown in Table 5.2, we would expect an increase in required time for inference when S increases. This is observed when we used optimized inter-cities covariance matrix. However, when fixing it, we observed an inverse phenomenon, with time required for inference decreasing as S increases. We believe it to be associated with implementation details, since using smaller cluster leads to a larger number of clusters and, consequently, more

loop iterations, which are inefficient when implemented in Octave.

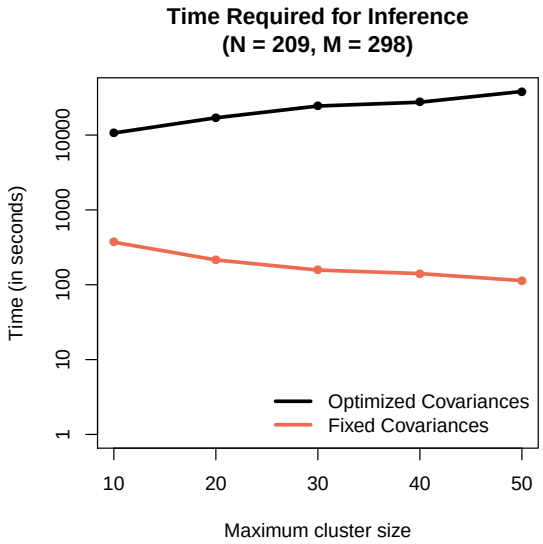

Figure 5.7. Time required for inference with $N = 209$ weeks, $M = 298$ cities and maximum cluster sizes ranging from 10 to 50.

5.3.2.2 Comparison Between Learned and Empirical Matrices

We now focus on understanding the impact of using an empirical approximation for the inter-cities covariance matrix K_C. Based on the previous results, we compare both approach using block-diagonalization with clusters of up to 10 cities. We also perform the comparison using the temporal-only version of DGP and extrapolate the conclusions to the full version of the model, as we did in the last section.

Figure 5.8 shows the comparison between the two approaches according to the three evaluation metrics. Both strategies led to similar results, with empirically learned covariances slightly outperforming optimized covariances according to correlation coefficient and NMAE. This suggests that assuming spatial dependences as hyperparameters can be even detrimental, possibly due to difficulties solving an optimization problem in high dimensionality. On the other hand, the computational time required for inference differs drastically: for $N = 209$ weeks, inference using empirical approximations required 370 seconds on average, while letting covariance to be optimized required 10658 seconds on average, as shown in Table 5.4.

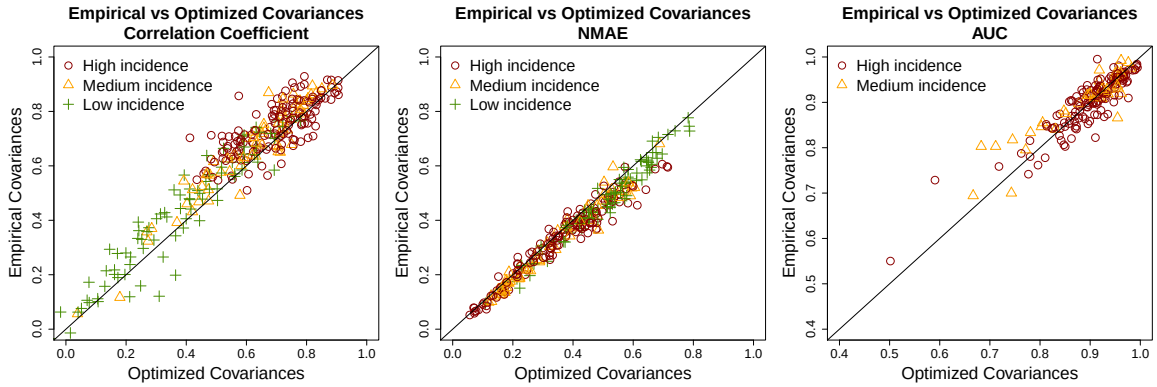

Figure 5.8. Comparison between optimizing covariances and using empirical approximations according to all three evaluation metrics. Each symbol represents a city and color and shape indicate the highest incidence level achieved by the corresponding city. The solid black line indicates equal performance between models. Cities above the solid line indicate higher values obtained by empirically approximating K_C, while cities below the line indicate higher values obtained by letting it be optimized via likelihood maximization.

5.3.2.3 Comparison Between Full and Temporal-Only Model

Finally, we evaluate our last strategy for reducing computational effort required for inference under DGP by comparing results obtained by the full DGP model and the temporal-only DGP, which does not use climate data. Both models were evaluated using a block-diagonal inter-cities matrix K_C formed based on clusters of up to 10 cities and with entries estimated empirically and fixed throughout hyperparameter optimization.

Figure 5.9 shows the comparison graphically according to all three evaluation metrics, clearly indicating that the temporal-only version outperforms full DGP. Although this result may seem to be unexpected, it is due to the high *spatial heterogeneity* within dengue data from Brazilian cities. That is, due to the large area of the country, the relationship between climate and DIR is not uniform for all Brazilian cities. Thus, attempting to model it without taking this fact into consideration may lead to worse results. As indicated in Equation 5.3, an uniform linear relationship between climate and DIR is expected within each cluster, forcing some spatial homogeneity that is not always present. In the next experiments, we present a more clear example of how expecting spatial homogeneity may fail within the context of dengue data from Brazilian cities.

Considering the computational effort required for inference, Table 5.4 also indicates that temporal-only DGP is to be preferred over full DGP.

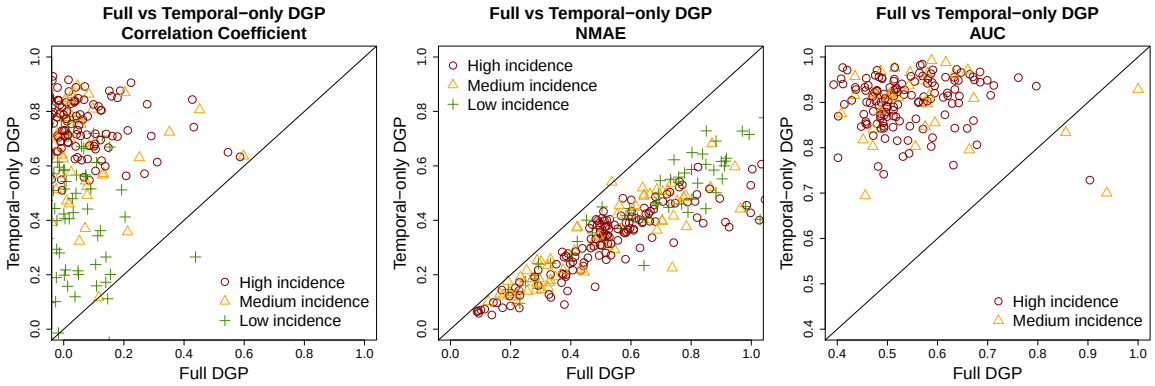

Figure 5.9. Comparison between temporal-only and full (using climate) DGP models according to all three evaluation metrics. Each symbol represents a city and color and shape indicate the highest incidence level achieved by the corresponding city. The solid black line indicates equal performance between models. Cities above the solid line indicate higher values obtained by temporal-only DGP, while cities below the line indicate higher values obtained by full DGP.

5.3.2.4 Impact of Spatial Dependences in the Model

Having verified the impact of each strategy proposed to reduce computational effort and noted that DGP achieve better accuracy using only temporal information, with K_C being estimated empirically and block-diagonalized with blocks of up to 10 cities, we can now check if the inclusion of the spatial component is indeed beneficial.

Figure 5.10 shows the results obtained by DGP when spatial dependences are enforced and when they are ignored (with $K_C = I$). The usage of spatial dependences led to more accurate predictions according to correlation and NMAE, while both models are indistinguishable according to AUC. Given that, we see that spatial dependences are beneficial to the model, as it increases the amount of information available for predictions.

5.4 Summary and Discussion

In this chapter, we propose many candidate GP-based models for DIR forecasting. All models were based on the characterization performed on Chapter 4, which indicated that an appropriate covariance function for dengue data would exploit local dependences, seasonality and association with climate. Additionally, we also introduced spatial dependences, exploiting the notion that dengue fever outbreaks are not geographically isolated events.

We began by proposing a general formulation, which led to a three-part covariance function composed of a local, a quasi-periodic and a climate-related components.

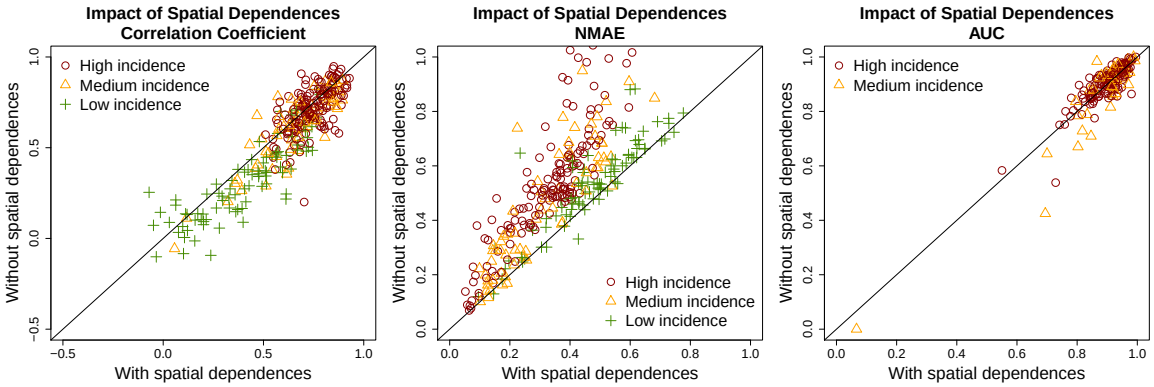

Figure 5.10. Comparison between DGP with and without spatial dependences according to all three evaluation metrics. Each symbol represents a city and color and shape indicate the highest incidence level achieved by the corresponding city. The solid black line indicates equal performance between models. Cities above the solid line indicate higher values obtained by ignoring spatial dependences, while cities below the line indicate higher values obtained by enforcing spatial dependences.

Table 5.3. Difference according to each evaluation metric between DGP in its best configuration (temporal-only, $S = 10$ and empirical approximation of K_C) and alternative formulations.

Candidate Formulation	Evaluation Metric	Difference Between Proposal and Alternative Formulations		Number of Cities Where Proposal Wins
		95% Conf. Interval	Median	
Temporal-only, $S = 1$, Emp. Matrix	Correlation	[0.037, 0.061]	0.049	200 (67%)
	NMAE	[-0.148, -0.120]	-0.133	280 (94%)
	AUC	[-0.008, 0.001]	-0.004	76 (42%)
Temporal-only, $S = 10$, Opt. Matrix	Correlation	[0.039, 0.054]	0.047	231 (78%)
	NMAE	[-0.032, -0.025]	-0.028	250 (84%)
	AUC	[-0.003, 0.005]	0.001	90 (50%)
Complete model, $S = 10$, Emp. Matrix	Correlation	[0.600, 0.662]	0.634	294 (99%)
	NMAE	[-0.238, -0.198]	-0.217	297 (99%)
	AUC	[0.357, 0.389]	0.373	176 (97%)

We then investigated if each component in the general formulation proposed was indeed necessary, as well as compare it with another similar covariance function proposed in the literature, known as spectral mixture kernel, which has the advantage of being able to exploit multiple seasonality effects. Our experiments indicated that all three components are beneficial and increased the accuracy of the model on at least one evaluation metric, confirming hypothesis usually exploited in the dengue literature. Besides that, our experiments also indicated that a single seasonality effect was sufficient to model our dengue data, and including more seasonality effects led to worse results, possibly

Table 5.4. Time required for hyperparameter optimization and inference for variants studied.

Configuration / Model	Time (s)	
	mean	std
Temporal-only, $S = 1$, Identity Matrix	3218	79
Temporal-only, $S = 10$, Emp. Matrix	370	9
Temporal-only, $S = 10$, Opt. Matrix	10,658	62
Complete model, $S = 10$, Opt. Matrix	52,576	1,698

due to overfitting. Considering the relatively short period of time under study, this result was not surprising. However, using spectral mixture may be beneficial if a longer period of time is modelled, as it may be useful in modelling seasonality effects due to variation of immunological properties of the population.

We then moved to the study of the spatial component, represented by a inter-cities covariance matrix K_C, which explicitly indicates the covariance between cities. Our first finding was that allowing spatial dependences between cities according to the correlation between their DIR values led to better results than simply enforcing spatial dependences according to distance. This is an interesting finding since distance-based spatial structures are very common in dengue literature. However, in a world that becomes more and more connected, with more accessible fast means of transportation, distance is no longer a good proxy for the interaction between two areas. In this context, the fact that our methodology is capable of automatically identifying relevant spatial associations is very interesting, as these associations tend to become more complex and, consequently, harder to analyze. Another important insight obtained through the experiments is that DGP with small clusters outperformed DGP with larger clusters, indicating that using a small, but carefully selected, set of spatial dependences leads to higher accuracy. Finally, we observed that the proposed strategies for reducing computational effort required for inference not only reduced the time used for hyperparameter optimization and issuing predictions, but also lead to more accurate predictions. The main reasons behind these results were the reduction of dimensionality for hyperparamenter optimization and the high spatial heterogeneity, which does not allow enforcing an uniform effect for climate features for all Brazilian cities under study.

After analyzing all experiments shown in this chapter, we concluded that the best formulation for DGP would be temporal-only and follow Equation 5.4, where

the inter-cities covariance matrix K_C is block-diagonalized according to Algorithm 5.1 with maximum cluster size $S = 10$ cities and its entries are fixed using an empirical approximation based on available training data. Having arrived at the final model, the next chapter will introduce additional experiments to assess the accuracy of DGP when compared to other models previously proposed for DIR prediction. A more detailed study of the predictions and the hyperparameters obtained by DGP will also be available in the next chapter.

Chapter 6

Experimental Analysis

In Chapter 5, we discussed our proposed model for DIR modelling, named DGP. In this chapter, we assess the accuracy of the proposed model by comparing it with three previously proposed models for DIR forecasting. Then, we provide a deeper analysis of predictions and hyperparameters obtained by DGP.

6.1 Comparison Between DGP and Previous Models

The experiments in Chapter 5 were intended to define the covariance function used in DGP, as well as tune some parameters, such as the maximum size for clusters of cities, required for the proposed methodology. However, it is still necessary to compare the accuracy of predictions provided by DGP with those provided by previously proposed approaches. For that, we selected three models based on the works we found in the dengue literature:

- **Linear Model (LM):** the first approach is a simple linear model based on climate-related covariates, namely weekly average temperature, rainfall and relative humidity. To avoid predictions of negative values, DIR values are previously log-transformed.
- **Autoregressive Model (AR):** the second approach is a first order autoregressive model $y_{s,t} = \beta_0 + \beta_1 y_{s,t-1} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $y_{s,t}$ denotes the log-transformed DIR value at city s during week t.
- **Negative Binomial Model (NB):** the third approach is a negative binomial model based on Lowe et al. [2013], which proposes a model specifically designed

for predicting DIR in Brazilian cities:

$$
\begin{aligned}
DIR_{s,t} &\sim NegBin(\mu_{s,t} = e_s\rho_{s,t}, \kappa) \qquad (6.1)\\
log(\rho_{s,t}) &= \alpha + \sum_{j=1}^{3}\beta_j x_{jst} + \sum_{j=1}^{2}\gamma_j w_{js} + \delta z_{st} + \omega_{Month(t)} + \phi_s + \nu_s\\
\alpha &\sim U(-\infty, +\infty)\\
\beta_j &\sim \mathcal{N}(0, 10^6)\\
\gamma_j &\sim \mathcal{N}(0, 10^6)\\
\delta &\sim \mathcal{N}(0, 10^6)\\
\omega_1 &= 0,\ \omega_{Month(t)}|\omega_{Month(t)-1} \sim \mathcal{N}(\omega_{Month(t)-1}, \sigma^2_\omega),\ Month(t) = 2, ..., 12\\
\phi_s &\sim \mathcal{N}(0, \sigma^2_\phi)\\
\nu_s|\nu_{j\neq s} &\sim CAR(\sigma^2_\nu)\\
\tau_\omega &= 1/\sigma^2_\omega \sim Gamma(0.5, 0.0005)\\
\tau_\phi &= 1/\sigma^2_\phi \sim Gamma(0.5, 0.0005)\\
\tau_\nu &= 1/\sigma^2_\nu \sim Gamma(0.5, 0.0005)\\
\kappa &\sim Gamma(0.5, 0.0005)
\end{aligned}
$$

Since it was originally proposed to work with monthly data at micro-region level, we adapted it to work on weekly data at municipality-level by transforming averages over 3 and 4 months into averages over 13 and 17 weeks, respectively. In Equation 6.1, e_s stands for the expected DIR at city s, calculated as the multiplication between the global dengue risk and the population size of city s. The variables x_{jst} are climate covariates for city s and week t: rainfall ($j = 1$) and temperature ($j = 2$) averaged over 13 weeks, and Oceanic Niño Index (ONI) ($j = 3$) averaged over 17 weeks. The variables w_{js} are altitude ($j = 1$) and population density ($j = 2$), while z_{st} is given by $log\left(\frac{DIR_{s,t}}{e_s}\right)$, $\omega_{Month(t)}$ is an auto-regressive term and ϕ_s allows for unstructured spatial variance. Structured spatial variance is provided by the conditional auto-regressive term ν_s, where the neighborhood function was modified to consider cities within 500 km of each other as neighbors. Finally, κ allows for overdispersion.

Note that we have not included here more sophisticated non-parametric regression techniques, such as *support vector regression* (SVR). This decision was motivated by two major facts. First, we would like to compare the proposed model with commonly

used techniques in dengue literature, as indicated in our literature review. Second, these methods tend to generate *black-box* models, that is, models whose predictions are very hard to understand, leading to non-interpretable models. These models are not appropriate for EWS, since we cannot be sure if predictions are based in scientific knowledge, violating the principles for designing appropriate warning services.

Figure 6.1 shows the comparison between DGP and the three previous models according to the three evaluation metrics used in this study. When compared to all models, DGP obtained statistically better results on all metrics, with better predictions for at least 83% of the cities under study. Table 6.1 shows the numeric values obtained by each model.

Table 6.1. Difference according to each evaluation metric between DGP and alternative models.

Model	Evaluation Metric	Difference Between DGP and Baseline		Number of Cities Where DGP Wins
		95% Conf. Interval	Median	
LM	Correlation	[0.300, 0.346]	0.323	283 (95%)
	NMAE	[-0.161, -0.135]	-0.147	291 (98%)
	AUC	[0.157, 0.192]	0.174	173 (96%)
AR	Correlation	[0.103, 0.131]	0.116	270 (91%)
	NMAE	[-0.159, -0.132]	-0.145	289 (97%)
	AUC	[0.028, 0.046]	0.036	151 (83%)
NB	Correlation	[0.523, 0.591]	0.557	290 (97%)
	NMAE	[-1.918, -1.007]	-1.456	298 (100%)
	AUC	[0.394, 0.413]	0.404	180 (99%)

Previous experiments showed that the introduction of spatial dependences improved our model. However, this does not seem to be the case for NB. In order to better understand why DGP was better on virtually every city, we analyzed the effects of climate in spatially-aware NB and spatially-unaware NB. When using spatial effects, we fit a single model instead of a model per city, while that spatially-unaware NB is fitted for each city individually. When we compare coefficients for climate-related variables, we noticed a huge variation when cities are modelled independently. This strengths our hypothesis that climate effects suffer from high spatial heterogeneity. Figure 6.2 shows an example of the city of São José dos Campos, in São Paulo, where NB obtained a correlation of 0.86 when fitted independently and a correlation of 0.00 when jointly fitting all cities. Note that distributions vary drastically, even changing signs. This is a very illustrative example of when parametric models may fail: by assuming a parametric form, NB forced all cities to have similar effects on climate.

In contrast, temporal-only DGP assumes the same *prior distribution* over all cities, a much looser assumption. Thus, it is capable of dealing with spatial hetero-

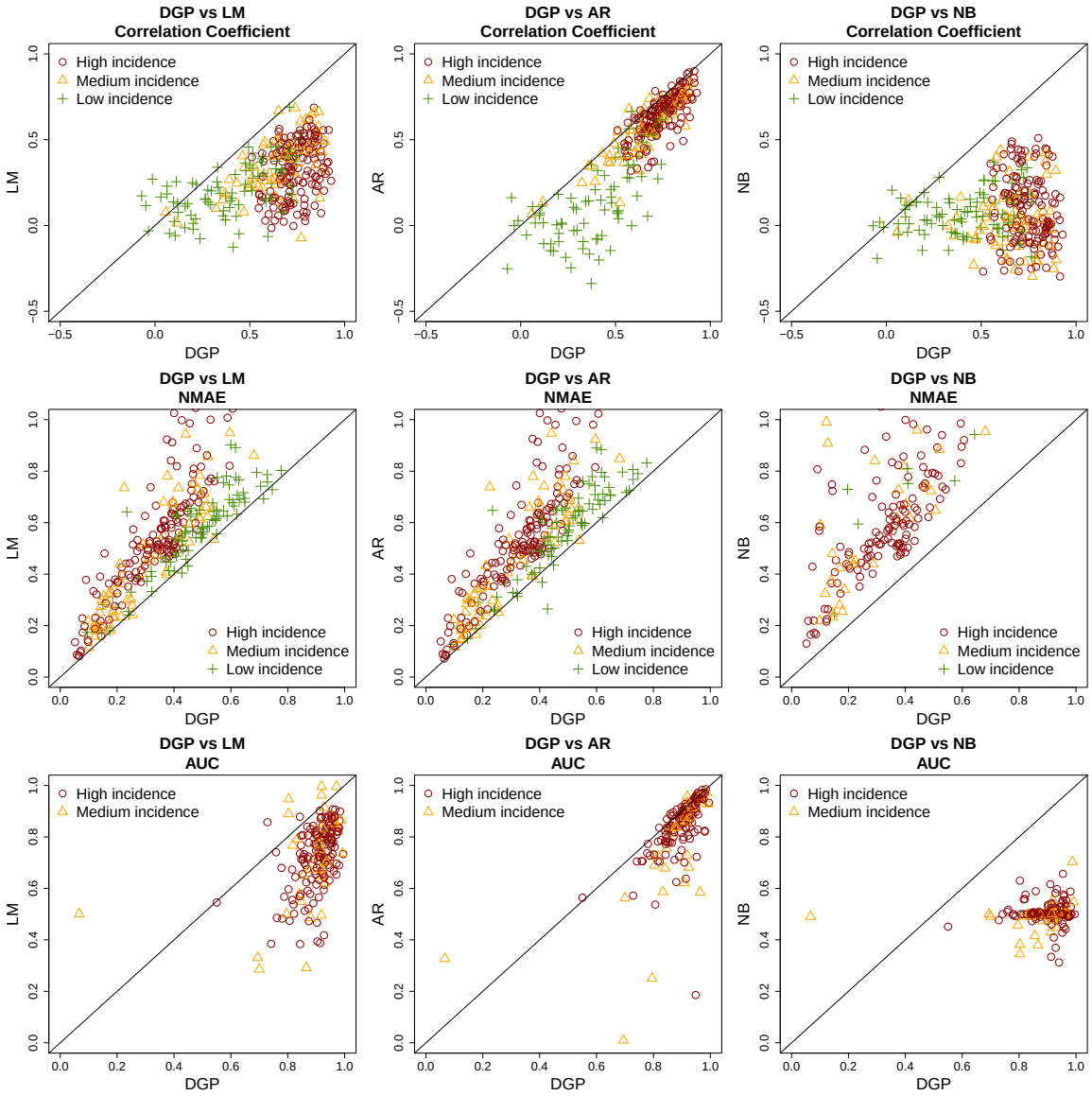

Figure 6.1. Comparison between DGP, LM, AR and NB according to all three evaluation metrics. Each symbol represents a city and color and shape indicate the highest incidence level achieved by the corresponding city. The solid black line indicates equal performance between models. Cities above the solid line indicate higher values obtained by LM, AR or NB , while cities below the line indicate higher values obtained by DGP.

geneity in a much more effective way than NB. Alternatives for improving NB would be fitting independent coefficients per city, or clustering coefficients that should be similar. The former approach leads to a potentially large increase in parameters, while the latter requires a careful analysis of data, specially when the number of effects included in the model is large.

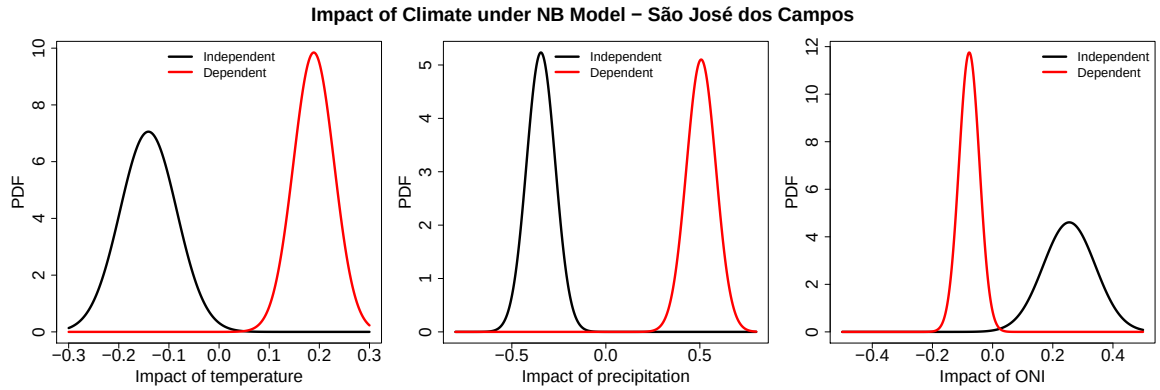

Figure 6.2. Posterior distribution of climate-related effects for the city of São José dos Campos, São Paulo, when NB is fitted per city (black line) or jointly with all cities (red line).

6.2 Analysis of Hyperparameters and Predictions

We now provide an analysis of hyperparameters obtained after likelihood maximization and predictions obtained under temporal-only DGP model.

Table 6.2 shows the optimized hyperparameters obtained. The hyperparameters σ_{loc} and σ_{qp} indicate the strength of local and quasi-periodic components, respectively. By comparing both values, we observed that the quasi-periodic signal is stronger than the local signal, reinforcing the hypothesis that seasonality provides useful information for DIR forecast. The hyperparameters ℓ_{loc} and ℓ_{qp}, on the other hand, are related to the decay of local and quasi-periodic signal, that is, how fast each signal vanishes as a function of time. The low value obtained for ℓ_{loc} shows that local component signal decays quickly, with only weeks within 2-weeks distance of each other obtaining high correlation (above 0.5). This shows how challenging forecasting DIR with 4 weeks in advance can be, as we are not provided with the most informative data. The value of ℓ_{qp} is approximately of half a year, indicating that, although the quasi-periodic component is responsible for the major part of the total covariance signal, it decays fast enough that weeks separated by one year will still have relatively low covariance when compared to nearby weeks. Finally, the periodicity of quasi-periodic signal is approximately of 1 year, exploiting the annual seasonality previously observed in dengue data.

Table 6.2. Hyperparameters obtained by DGP.

Hyperparameter	ℓ_{loc}	σ^2_{loc}	ℓ_{qp}	σ^2_{qp}	ℓ_{per}	p
Value	2.3572	0.12244	24.323	0.42781	0.77978	56.993

Figure 6.3 shows the cumulative distribution of each evaluation metric, considering the 298 cities under study. It indicates the proportion of cities that obtained

a result according to a given metric lower or equal to a given value. For correlation coefficient and AUC, which are intended to be maximized, the more to the right the curve is, the better. For NMAE, the reasoning is the opposite. It is possible to see that approximately 80% of the cities obtained correlation above 0.5, while approximately the same proportion obtained NMAE below 0.5. When considering AUC, more than 90% of the cities obtained value above 0.8.

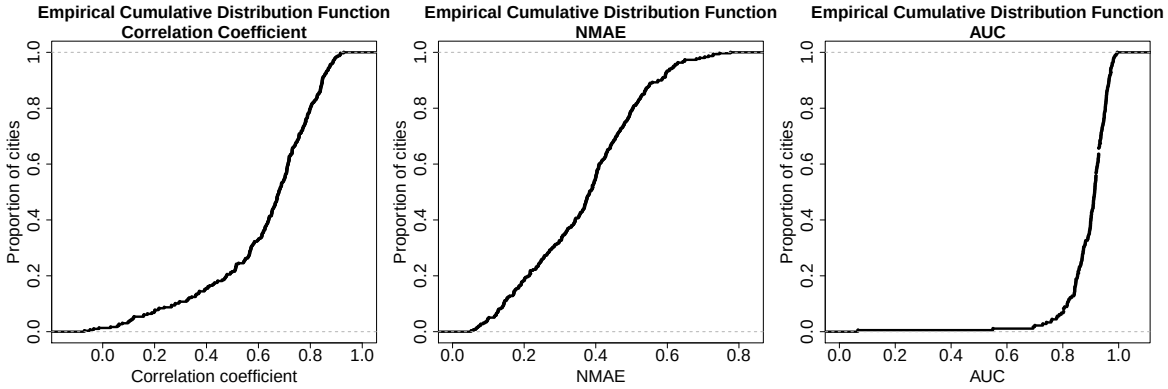

Figure 6.3. Empirical cumulative distribution function for each evaluation metric.

Figure 6.4 shows the spatial distribution of results obtained by DGP according to all three evaluation metrics. Considering both correlation coefficient and AUC, we observed an approximately uniform spread, with good results obtained all over the country. The only exception would be the South region, specially the states of Santa Catarina and Rio Grande do Sul, where no city obtained correlations above 0.75. This region is the least affected region of the country considering DIR, leading a large proportion of cities that never reached medium or high incidence. According to NMAE, we observed that the best results were concentrated in North and Northeast region. Figure 6.5 shows a quantitative analysis of the spatial distribution of results, confirming the analysis obtained through Figure 6.4.

Finally, Figure 6.6 shows a more qualitative view of predictions obtained by DGP for the six Brazilian capital cities most affect by dengue fever. Note that predictions follow a similar pattern to real values, which are almost always within the 95% confidence interval. Exceptions occurred in abrupt increases in DIR, such as in the beginning of 2013 in Belo Horizonte, where DGP may issue underestimated predictions.

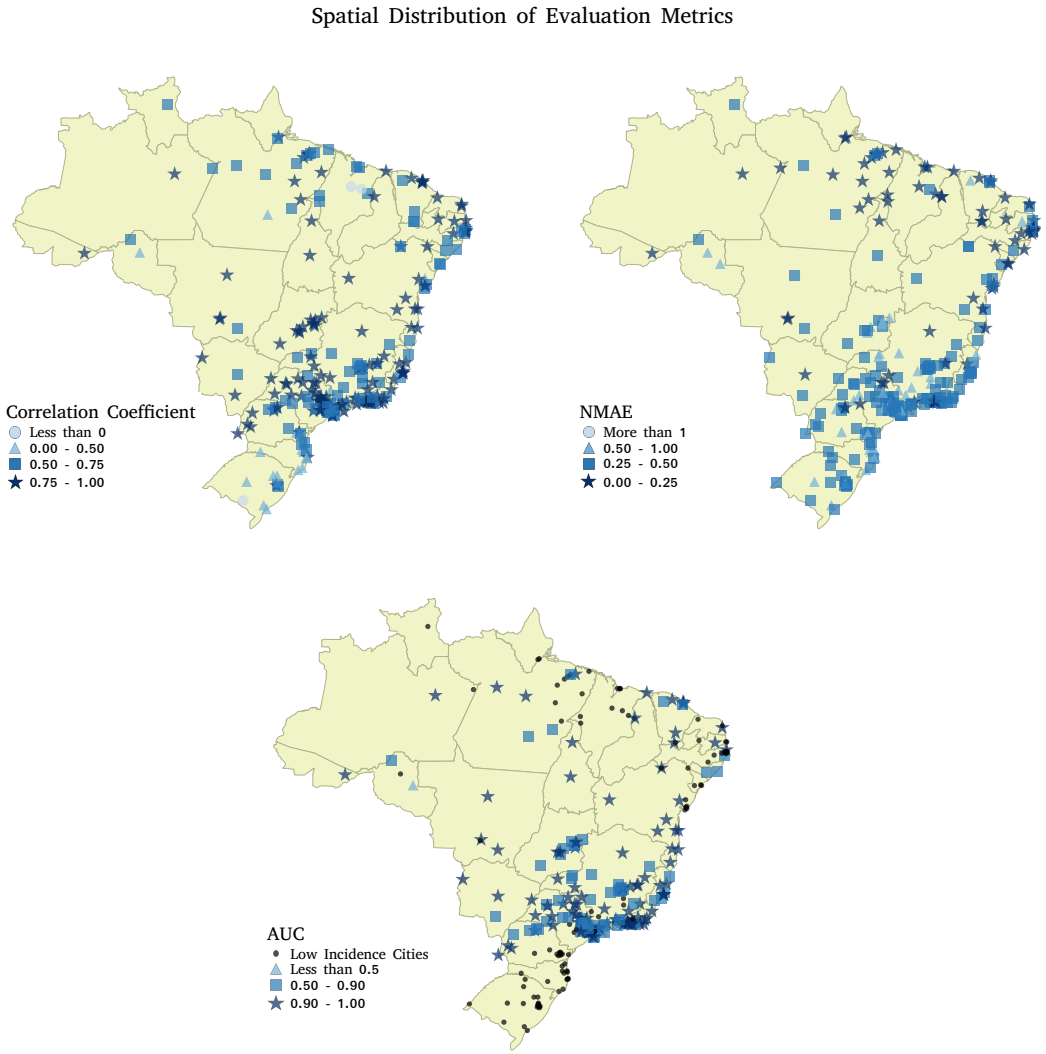

Figure 6.4. Spatial distribution of evaluation metrics per city obtained by DGP. Each symbol denotes a city, with color and shape associated to the value obtained in each metric on the corresponding city.

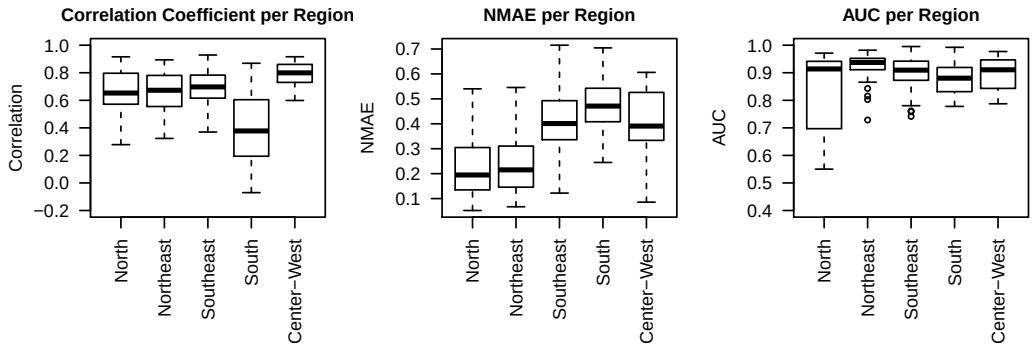

Figure 6.5. Evaluation metrics stratified per Brazilian region.

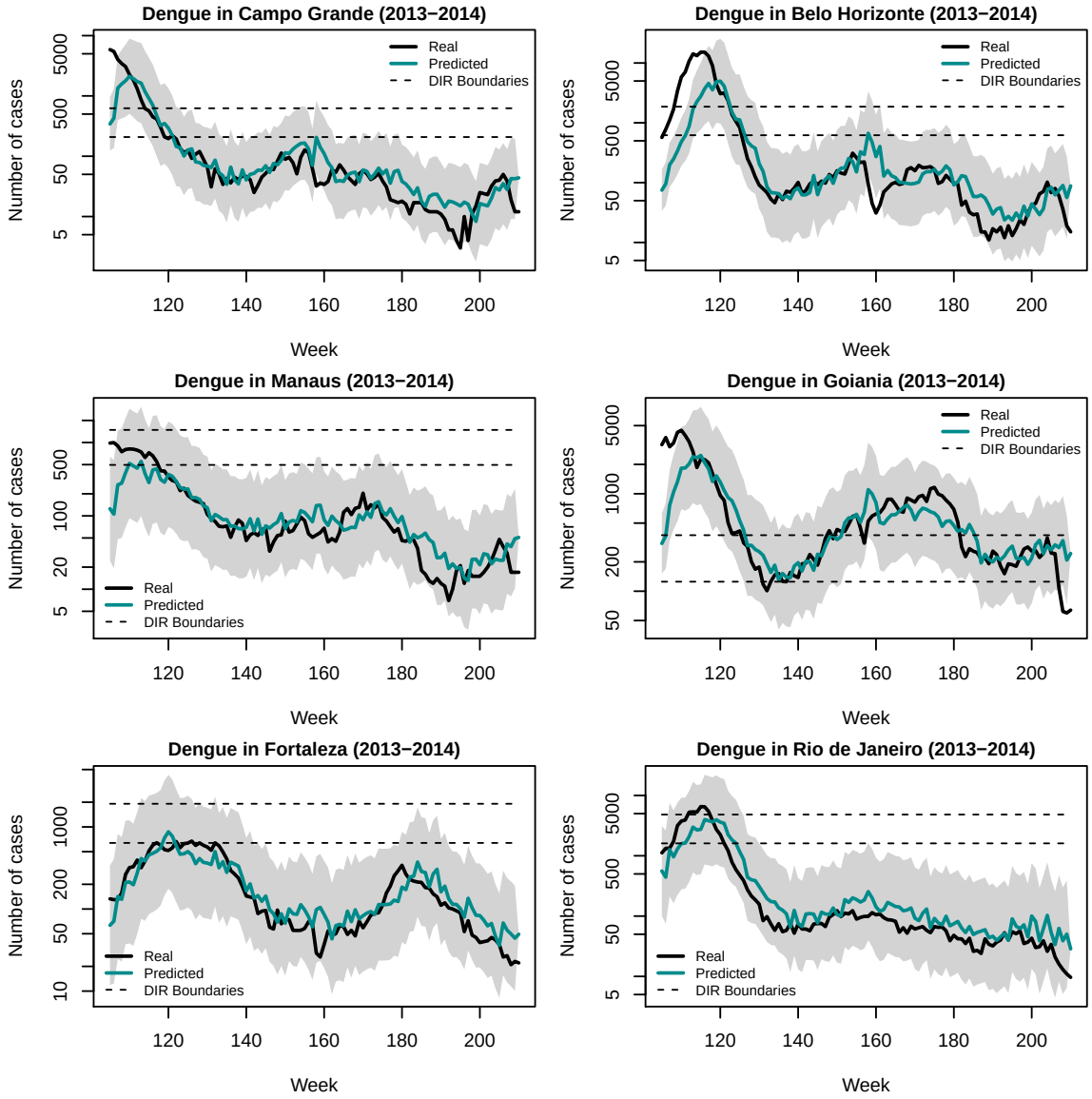

Figure 6.6. Predictions issued for the six Brazilian capital cities with highest DIR values. The black line indicates real values, while the blue line show the predictions obtained. The blue shaded area indicates the 95% confidence interval.

6.3 Summary and Discussion

In this chapter, we exposed and discussed some experiments designed to evaluate the accuracy of the proposed model. Besides that, we performed an analysis on hyperparameters and predictions obtained by DGP.

When assessing the accuracy of DGP, we compared it with three previously proposed approaches for DIR modelling: a linear, an autoregressive and a negative binomal model, the last one being specifically proposed for DIR modelling in Brazilian cities. We then verified that DGP outperformed all alternative models on all evaluation met-

rics. NB obtained results particularly bad, showing again the result of ignoring spatial heterogeneity, on a very illustrative example of how parametric models may fail if not applied carefully.

In the end, we took a deeper look on hyperparameters and predictions obtained by DGP. The set of hyperparameters obtained showed the relevance of quasi-periodic component, which was the main contributor the total covariance signal. It also revealed the difficulty of predicting DIR with 4 weeks in advance: we are not allowed access to the most informative data, which would be weeks nearby to the week for which prediction is being made. Our proposed model obtained good overall results, specially when considering AUC. For this evaluation metric, more than 90% obtained values above 0.8, which indicates a very good trade-off between true positive and false positive rate. By looking at the spatial distribution of accuracy, we observed that, for correlation and AUC, spatial location is not significantly associated to accuracy. However, when considering NMAE, cities from the North and Northeast obtained the best results. Finally, we showed predictions obtained by the six Brazilian capital cities most affected by dengue fever, which demonstrated that predictions follow the expected behavior. A limitation seem to be very abrupt DIR peaks, where the model captures the growing trend, but has difficulties in estimating the correct value, leading to underestimated predictions.

Another limitation of the proposed model is the fact that it assumes that epidemiological data is provided on real-time. In other words, at week t predictions relative to week $t+4$ are issued using information associated to weeks 1 to t. A more realistic scenario would consider that, at week t, only epidemiological data up to week $t-\beta$, $\beta > 0$, would be available, due to the time necessary for epidemiological data to be ready for public use. In the next chapter, we propose a general framework to deal with delayed epidemiological data using online data sources aimed to mitigate this issue.

Chapter 7

Using Proxies for Epidemiological Data

In Chapter 5, we proposed a model for forecasting DIR at Brazilian cities. However, a major assumption of this model is that epidemiological data is provided in real-time. Therefore, at week t, we use data available up to week t to forecast DIR during week $t+\beta$, where β is the antecedence for which predictions are required. In particular, the experiments conducted in Chapters 5 and 6 assumed $\beta = 4$ weeks. Epidemiological data, however, requires time for being available even for governmental authorities, since cases need to be confirmed and information needs to propagate from local health care units to federal authorities. Hence, data from time t is not always available to make predictions with β weeks in advance.

A simple approach for dealing with this issue is to use only delayed epidemiological data available at time t. Thus, if epidemiological data requires γ weeks to be ready to use, $\gamma > 0$, predictions would be made with $\beta + \gamma$ weeks in antecedence. This approach, however, leads to a potential decrease in accuracy, specially if γ is large, since predictions would be issued in practice with a much higher antecedence. Consequently, the model would not have access to up-to-date data, which is typically considered as highly informative data [Dom et al., 2013; Eastin et al., 2014].

An alternative approach would be to use a real-time related data source as a proxy for epidemiological data. As indicated in Chapter 3, some works have already used online data, such as Twitter or Google data, to model dengue data [Gomide et al., 2011; Althouse et al., 2011; Souza et al., 2015]. However, the focus of these works were to provide real-time estimates of dengue data, and not to use it in a *predictive model*. In this chapter, we present two approaches for building predictive models using online data without assuming that epidemiological data is provided in real-time. First, we

discuss the two proposed approaches. Then, we conduct experiments to find parameters associated to the approaches and to compare between the two proposals. Finally, we compare them to the simple approach discussed above, which simply makes predictions with higher antecedence.

For discussing our approaches for incorporating online data into predictive epidemiological models, we assume that t denotes the current moment in time. We also define *delayed epidemiological data* as data associated to time between $t-\gamma$ and t, that is, data that is not available at time t because of the delay associated to the difusion of epidemiological data. We define *future epidemiological* or *future online data* as any epidemiological or online data associated to time $t' > t$. Finally, we call a epidemiological data collection as up-to-date if it contains epidemiological data up to week t.

7.1 Two Approaches for Incorporating Online Data

We propose two approaches for mitigating the problem of the lack of up-to-date epidemiological data by exploiting online data, such as data from Twitter. The first approach *estimates delayed epidemiological data* up to the current moment in time using a relationship between epidemiological and online data (such as the one presented in Chapter 4) and then uses a traditional epidemiological predictive model to *forecast future epidemiological data*, which now has access to up-to-date noisy epidemiological data. The second approach, in turn, first *estimates future online data*, and then uses these estimates to *forecast future epidemiological data*, as detailed below.

7.1.1 Hybrid Approach

The first approach, which we call a *Hybrid Approach* (HA), deals with the problem of delayed epidemiological data by (i) estimating delayed epidemiological data up to current time t using a relationship, such as a linear or polynomial dependence, between epidemiological and online data; and (ii) deciding whether to use estimated delayed epidemiological data or not to forecast future epidemiological data. The second step is relevant since the former predictions can have high levels of uncertainty, which may be detrimental for the accuracy of the predictive model. We highlight that, although it would be possible to use uncertainty associated to estimates instead of applying a binary decision of using or not estimates, incorporating distinct uncertainty for training points would lead to a heteroskedastic model, violating the structure necessary for better exploiting Kronecker structure. The approach is called hybrid because it can

simply ignore online data if it is too uncertain or use a hybrid model that considers both types of information simultaneously when appropriate.

This approach has three main components, the first and third being predictive models, as presented in Figure 7.1. The predictive model in component (i) takes online data as a covariate and epidemiological data as the response variable, outputting delayed epidemiological data. The second component decides whether estimates of delayed epidemiological data from online data are safe to use, i.e., if the uncertainty associated to estimates is low. This module is interesting because uncertain estimates may introduce extra noise to the training set of the traditional EWS, which can be detrimental. The simplest way to define when to use information from which model is to use an uncertainty estimation threshold. Finally, the predictive model in component (iii) takes information associated to epidemiological data (e.g., temperature, rainfall, time of the year) as covariates and outputs future epidemiological data as the response variable, using estimated delayed epidemiological data if they are safe to use.

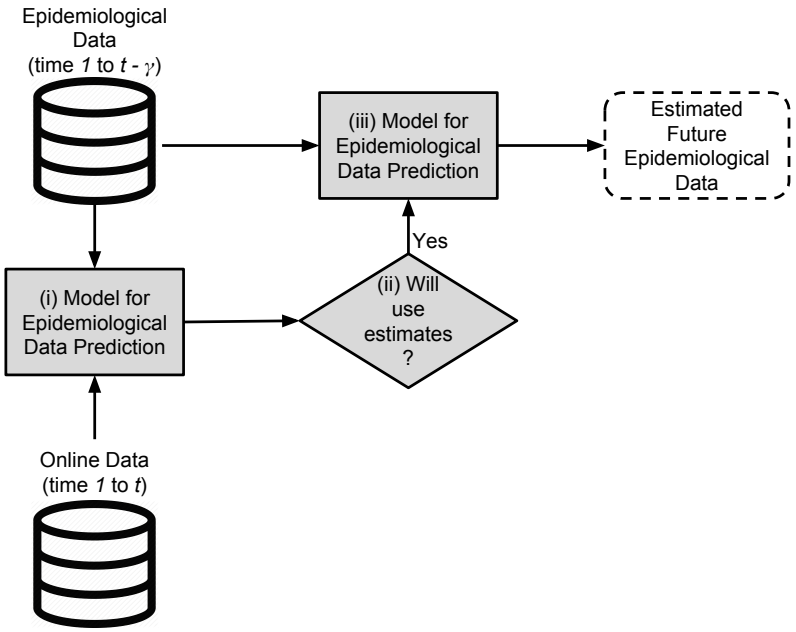

Figure 7.1. Three-step hybrid approach proposed for using Twitter data to improve epidemiological predictive models for EWSs. Gray boxes represent the approach components.

This approach has two major advantages. First, it employs traditional models used by EWSs without requiring any modification for the final prediction. Second, it offers flexibility for us to easily introduce any module responsible for defining when estimates of epidemiological data are going to be used, even if other online data sources are added to the framework.

7.1.2 Online-only Approach

Assuming there exists a strong relationship between online and epidemiological data, an alternative approach is to first *forecast future online data*. Based on these estimates, we can estimate epidemiological data not only with delayed data, but with "future" data. This is the main reasoning behind the second proposed approach, named *Online-only Approach* (OA), described in Figure 7.2.

This approach is composed of two components, both predictive models. Component (i) is a model for *online data prediction* that takes a online data time series as input and outputs as the response variable *future online data*. Component (ii) is a model for *epidemiological data prediction* that takes current and future online data as covariates and epidemiological data as the response variable, and outputs *future epidemiological data*. The rationale behind this approach is that, instead of estimating delayed epidemiological data which is then used to feed EWSs, we could directly exploit the dependencies between online and epidemiological data.

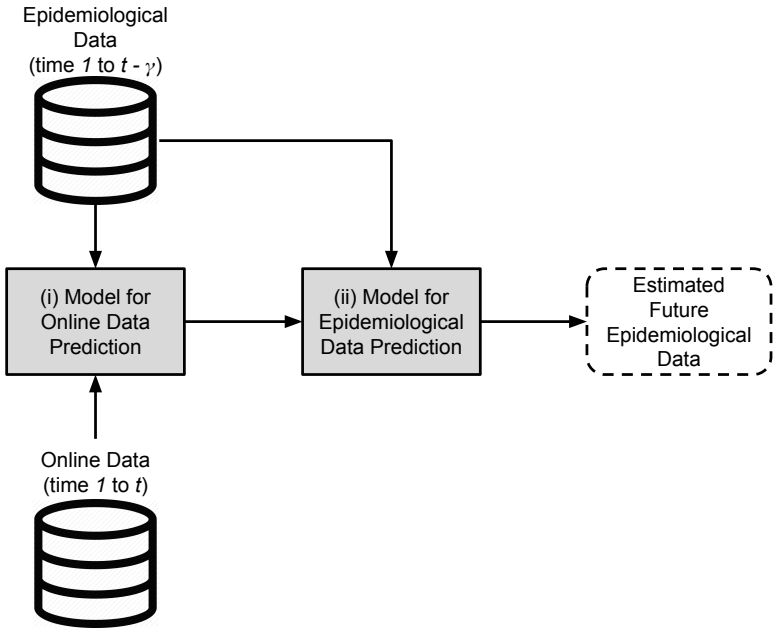

Figure 7.2. Two-step online-only approach proposed for using Twitter data to improve EWSs for dengue fever. Gray boxes represent the approach components.

This approach is more objective than the first one in the sense that it directly exploits the association between the two data sources. Besides that, dependencies between epidemiological and online data are typically simpler than dependencies exploited in traditional EWSs. For instance, Gomide et al. [2011] identified a simple linear relationship between the amount of dengue-related tweets and dengue fever inci-

dence in Brazil. Thus, if future online data can be predicted accurately, this approach can lead to better results.

7.1.3 Components Specification

The two approaches previously presented have components that may be specified in very different ways and that are problem-dependent, being able to be instantiated in epidemiological scenarios different from those of dengue, which is the case study in this thesis. For this work, we define that epidemiological data consists on weekly DIR values and online data consists on the weekly number of dengue-related tweets. For HA, we need to define a traditional EWS prediction model, a model for DIR prediction and a strategy for deciding when to use estimated DIR. For OA, we need a model for online data prediction and a model for predicting future DIR. Since we are going to use the same model for DIR predictions in both approaches, this leaves us to define four components, which are described below.

Traditional EWS DIR Prediction Model Based on the results presented in the previous chapters, we used temporal-only DGP described in Chapter 5 as the traditional prediction DIR model for EWS.

DIR prediction models using online data The main goal of this component is to provide DIR prediction models with estimates of delayed DIR data by exploiting Twitter data, a real-time data source. Based on the linear relationship observed between Twitter and dengue data in Brazil in previous works [Gomide et al., 2011; Souza et al., 2015] and on the predictive power of GPs for dengue data, we opted for modelling dengue data as a GP equipped with a linear kernel over the number of dengue-related tweets observed at a given city during a given week. We followed the decision in Chapter 4 and applied a logarithmic transformation on the epidemiological data. This was done to avoid modelling count data with a Gaussian distribution, which may be inadequate, requiring an appropriate transformation.

More formally, let $DIR_{s,t}$ denote the DIR at city s during week t, $\overline{y_s}$ denote the mean of log-transformed DIR values for city s and $x_{s,t}$ denote the number of dengue-related tweets observed at city s during week t. Then, the proposed model for inference of missing epidemiological data is given by

$$\begin{aligned} DIR_{s,t} &= exp(y_{s,t} + \overline{y_s}) - 1 \\ y_{s,\cdot} &\sim \mathcal{GP}\left(0, k(x_{s,i}, x_{s,j}) = \sigma_f^2 + \frac{x_{s,i} * x_{s,j}}{\ell^2} + \delta_{ij}\sigma_n^2\right) \end{aligned} \tag{7.1}$$

where σ_f, ℓ and σ_n are hyperparameters learned from data via likelihood maximization, σ_f allows for a bias term in the linear relationship between Twitter and epidemiological data, ℓ controls the impact of Twitter data and σ_n allows for noise-corrupted observations.

Deciding whether to use estimated DIR In order to avoid using poorly estimated DIR values from online data, which could introduce noise into epidemiological data, we defined an extra module to HA responsible for deciding whether estimated epidemiological data is going to be useful or not. Since we are assuming a linear relationship between Twitter and dengue data, we opted for using a threshold based on the correlation between epidemiological and Twitter data. Whenever the correlation exceeds a threshold, estimated data from the model trained with online data is considered. Otherwise, only real dengue data is used.

A possible strategy for defining appropriate values for the threshold is to perform a regression of the differences in accuracy obtained when using a model that accounts only for real epidemiological data and a model that always considers estimates of epidemiological data as a function of correlation between epidemiological and Twitter data. Given the regressed function, we can estimate for which correlation values it is better to use estimates of DIR values and for which it is not. A more rigid formulation would be to use estimates only when a city has a 0.95 probability of being higher than zero according to the regressed function.

Model for online-data prediction The work of Gomide et al. [2011] indicated that epidemiological and number of dengue-related tweets presented similar temporal patterns. Based on this, we propose to use the same model for epidemiological dengue data, as indicated in Chapter 5.

7.2 Experimental Results

In this section, we empirically evaluate the accuracy obtained by both approaches and compare it to simply enlarging the antecedence which predictions are issued. Before that, however, we have to find an appropriate threshold for HA to decide whether to use estimated DIR values or not. For all these experiments, we assume a delay for epidemiological data to be available of approximately 2 months (8 weeks). In the end, we included a study of impact of this delay.

7.2.1 Determining the Threshold for Using the Hybrid Approach

Our first experiment was conducted in order to obtain estimates of the threshold to be used for deciding whether to use estimates of epidemiological data under the HA. For that, we defined two versions of HA representing two extreme behaviors. The first, named as *Hybrid Approach - Never Estimates* (HAnE), makes predictions using only real epidemiological data, and Twitter data is omitted. This implies that predictions are made not with 4 weeks in advance, but with 12 weeks in advance, since epidemiological data is delayed by 8 weeks. The second model, named as *Hybrid Approach - Always Estimates* (HAaE), uses Twitter data to estimate missing DIR values for all cities and then uses these estimates to feed DGP to estimate DIR with 4 weeks in advance. Note that these versions differ only on the module for deciding whether to use estimated DIR values: the former never uses them, while the latter always uses them.

Figure 7.3 shows the difference observed for each city between HAaE and HAnE as a function of the total number of dengue-related tweets observed for the corresponding city and the correlation between epidemiological and Twitter data. Positive values indicate a higher value obtained by the former, while negative values indicate the opposite. Note that NMAE is intended to be minimized, while correlation and AUC are intended to be maximized. The figure leads to some conclusions. First, we observed that for some cities it is better to use HAaE, while for others it is better to use HAnE. This seems to be associated with the total number of tweets and the correlation between the two data sources in the sense that, the higher the volume of tweets and the correlation, the more beneficial it is to use estimated DIR values according to the evaluation metrics. Second, we observed that for both volume of tweets and correlation between Twitter and dengue data we obtained similar patterns, indicating that the two measures are highly associated. However, patterns obtained as a function of correlation were more well-behaved and exhibited less variability. Finally, we observed that differences in AUC followed a linear pattern, while differences in correlation (when considering number of tweets) and NMAE followed a less clear, more sophisticated pattern.

Given these observations, we opted for a linear regression to model the relationship between the differences in AUC and the correlation between dengue and Twitter data, obtaining the values in Table 7.1. Using these estimates, we calculated the two threshold values indicated previously: the point where the regressed function crosses the x-axis and the point where the regressed function is greater than zero with 0.95 probability. We obtained the threshold values of 0.42 and 0.49, respectively.

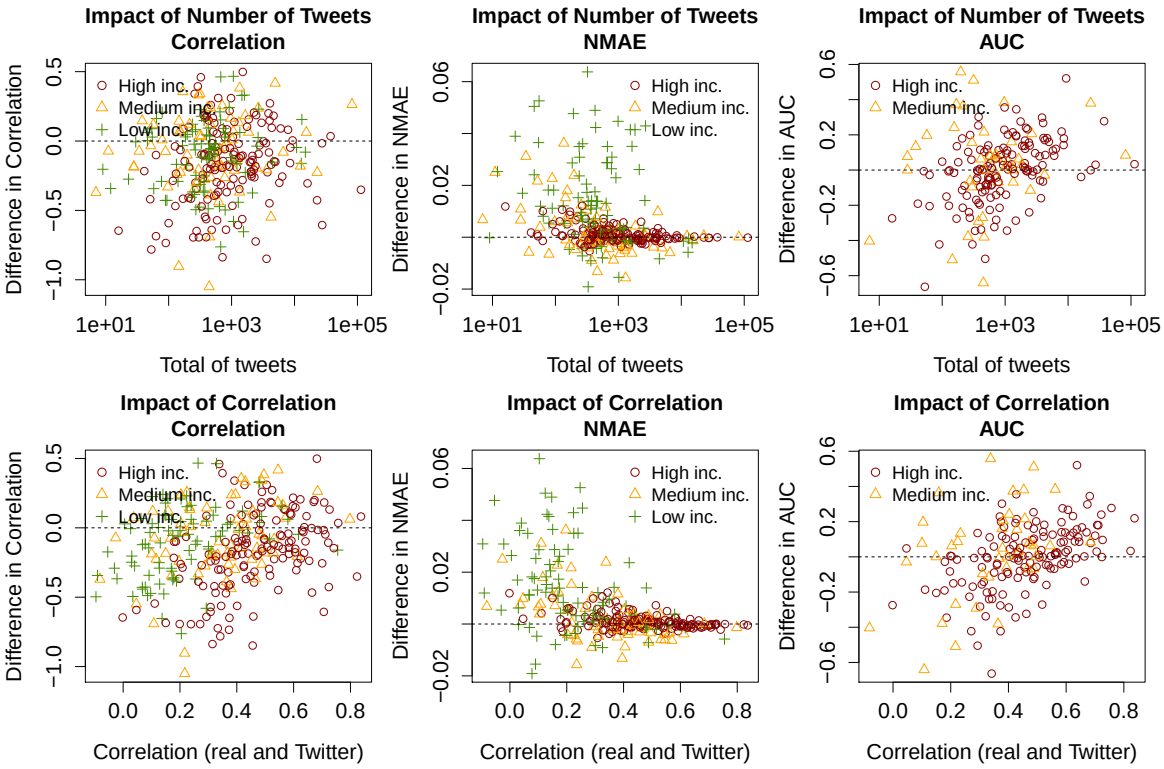

Figure 7.3. Impact of the number of dengue-related tweets and correlation between epidemiological and online data for each city on the difference of accuracy between HAaE and HAnE. Each symbol represents a city and color and shape indicate the highest incidence level achieved by the corresponding city. The dashed line indicates equal performance. Points above this line indicate higher values obtained by HAaE, while points below the line indicate higher values obtained by HAnE.

Table 7.1. Estimated coefficients for linear regression of difference in AUC over correlation between Twitter and dengue data.

	Estimated Value	Standard Error
Bias	-0.219	0.038
Correlation	0.515	0.082

7.2.2 Comparison Between Online-only and Hybrid Approaches

After obtaining estimates for the threshold value, we can now compare the results obtained by the two proposed approaches. We evaluate two versions of HA, using the two threshold values. To differentiate between these two versions, we denote HA using the larger threshold value as *Confident Hybrid Approach* (cHA). We also name the OA using Twitter data as OAT.

Figure 7.4 shows a comparison between OAT, HA and cHA. For the top figures,

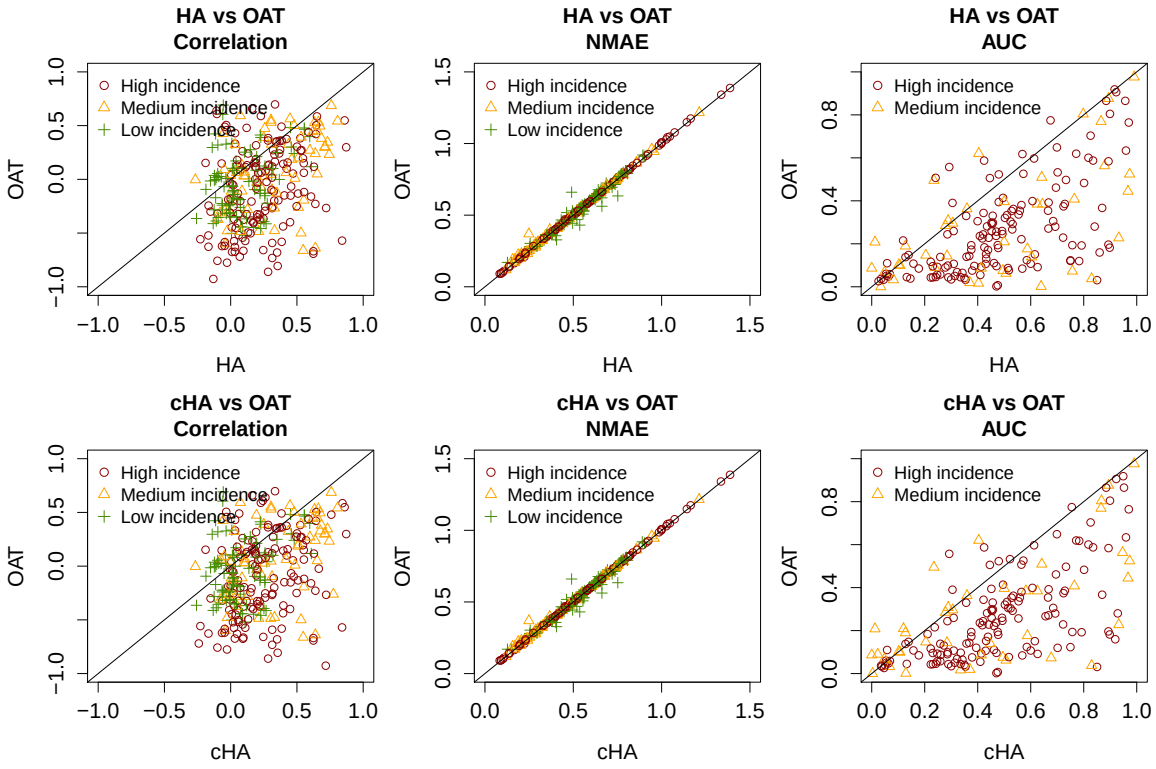

Figure 7.4. Comparison between the two proposed approaches. The x-axis and y-axis are related to the value achieved using a given evaluation metric and a given approach. Each symbol represents a city and color and shape indicate the highest incidence level achieved by the corresponding city. The diagonal line indicates equal performance. Points above this line indicate higher values obtained by OAT, while points below the line indicate higher values obtained by HA or cHA.

points below the diagonal line indicates higher accuracy. For the bottom figures, points above the diagonal line indicates higher accuracy. The figure shows that, specially considering AUC, HA and cHA obtained better results, indicating that HA outperforms OAT. We believe there are two major reasons behind this result. First, HA is more flexible in the sense that it can ignore estimated DIR values using Twitter data when they are poorly estimated. OAT, on the other hand, always uses Twitter data, even when its relationship with epidemiological data is not very clear. Second, Twitter data can be very hard to predict, as it is typically much noisier than epidemiological data.

7.2.3 Evaluation of the Proposed Approaches

We now focus on the accuracy of the proposed approaches when compared with HAnE and HAaE, called the *stand-alone* models. We use the two versions of the proposed approaches that obtained better results in the previous analysis, HA and cHA. Figure 7.5 shows graphically the comparison between stand-alone and hybrid models according

Table 7.2. Difference according to each evaluation metric between HA and OAT.

		Difference w.r.t OAT		Number of Cities
		95% Conf. Interval	Median	Where OAT Loses
HA	Correlation	[0.147, 0.218]	0.183	220 (74%)
	NMAE	[-0.003, -0.002]	-0.002	230 (77%)
	AUC	[0.136, 0.195]	0.164	154 (85%)
cHA	Correlation	[0.150, 0.225]	0.187	217 (73%)
	NMAE	[-0.003, -0.002]	-0.003	231 (78%)
	AUC	[0.125, 0.186]	0.153	149 (82%)

to evaluation metrics while Table 7.3 shows the output of a paired Wilcoxon test used to compare between models. When compared to HAnE, HA and cHA were not statistically superior than simply extending predictive antecedence. However, both versions outperformed HAnE when considering AUC. Differences were more noticeable when comparing HAnE with HA, due to the fact that the latter uses Twitter data more frequently than cHA. When considering HAaE, differences were noticeable when considering correlation and NMAE. HA and cHA obtained similar accuracy, so for the following analysis we are going to use cHA.

Table 7.3. Difference according to each evaluation metric between HA and stand-alone models.

		Difference w.r.t. HAnE		Difference w.r.t. HAaE	
		95% Conf. Interval	Median	95% Conf. Interval	Median
HA	Correlation	[-0.102, 0.021]	-0.039	[0.119, 0.247]	0.182
	NMAE	[-0.001, 0.000]	0.000	[-0.010, -0.004]	-0.006
	AUC	[0.027, 0.112]	0.068	[-0.001, 0.150]	0.074
cHA	Correlation	[-0.111, 0.035]	-0.035	[0.101, 0.216]	0.157
	NMAE	[-0.001, 0.000]	0.000	[-0.007 -0.003]	-0.005
	AUC	[0.037, 0.136]	0.084	[-0.023, 0.094]	0.034

For the following analysis, we classify each city as a *win* or a *loss*, where a win happens whenever cHA correctly chooses to use estimated DIR values or not. Figures 7.6 shows the spatial distribution of wins and losses, while Figure 7.7 shows the proportion of wins stratified by Brazilian regions. We could not observe any clear spatial pattern or dramatic differences between regions, indicating that the proposed methodology is safe to be applied to any region of Brazil.

Figure 7.8 shows the distribution of number of dengue-related tweets and its correlation with epidemiological data stratified by wins and losses. Again, we do not observe any clear pattern, indicating that the proposed methodology is robust to the amount and quality of online data.

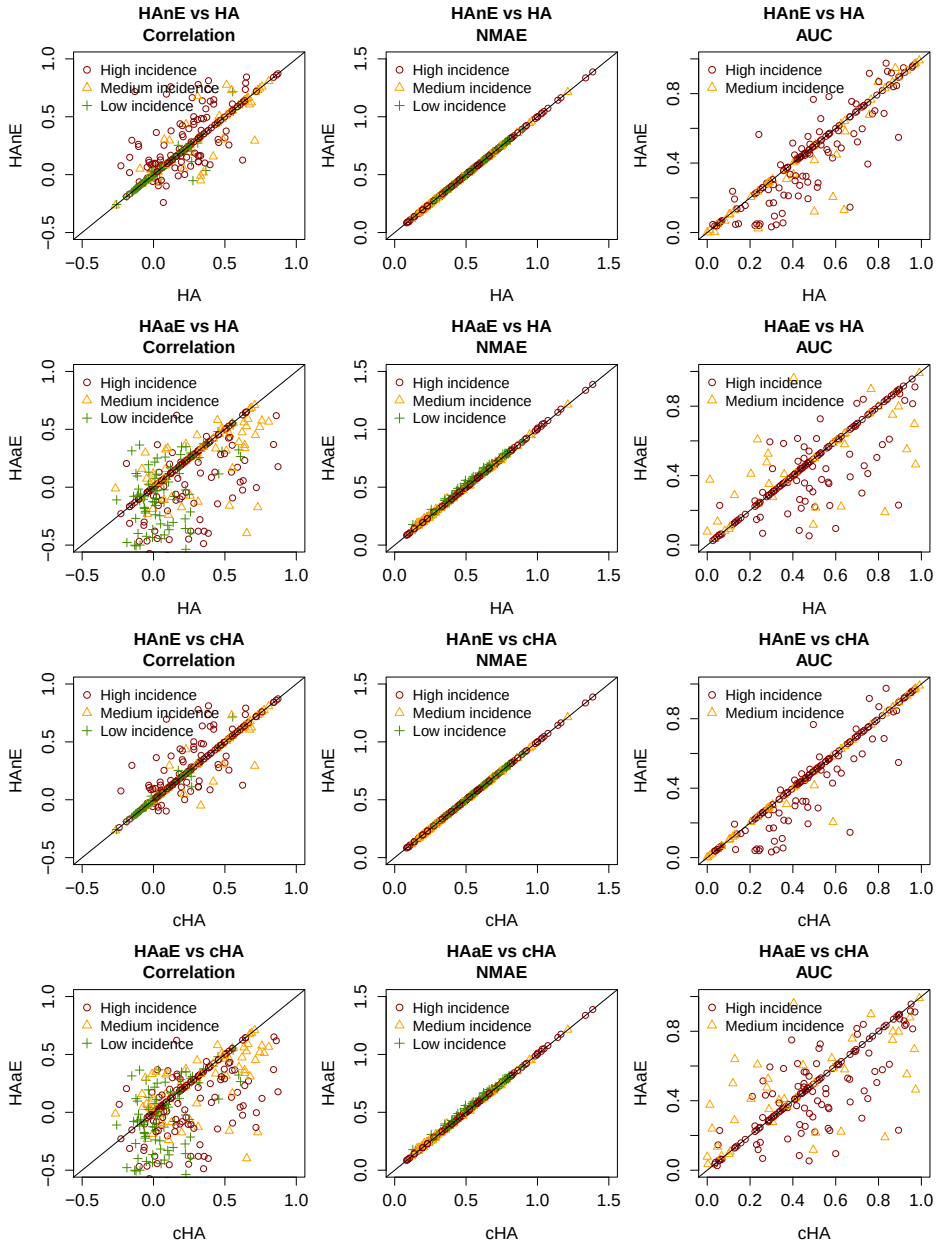

Figure 7.5. Comparison between hybrid approaches and stand-alone models. Each symbol represents a city and color and shape indicate the highest incidence level achieved by the corresponding city. The solid black line indicates equal performance between models. Cities above the solid line indicate higher values obtained by stand-alone models, while cities below the solid line indicate higher values obtained by hybrid approaches.

7.2.4 Impact of Epidemiological Data Delay

All previous experiments assumed that epidemiological data would be available with a 8-week delay. However, this delay is expected to affect the proposed frameworks, as well as HAnE, which corresponds to the simplistic approach that only extends antecedence

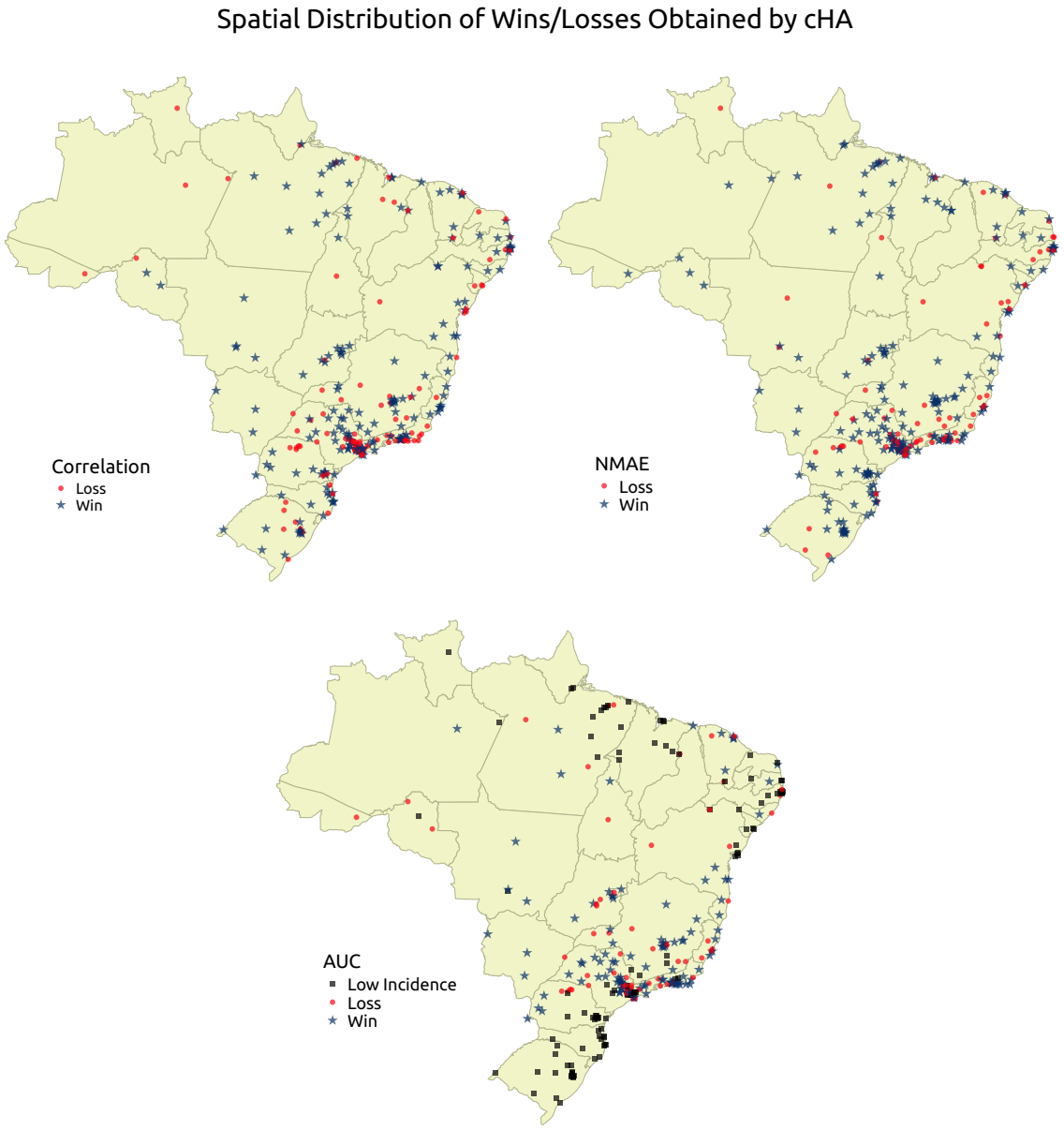

Figure 7.6. Spatial distribution of wins and losses obtained by cHA. Each point denote a city, with color and shape indicating whether cHA chose correctly to use estimated DIR values or not.

of predictions. In this section, we study the impact of this delay.

Figure 7.9 shows the difference in accuracy according to all three evaluation metrics between cHA and HAnE. Correlation and AUC exhibited a similar pattern, with differences growing with larger delays. This is expected, since larger delays imply in more outdated data available for HAnE. According to correlation, HAnE is better for delays up to 4 weeks, both methods are similar for delays of 6 and 8 weeks, and cHA is more appropriate for delays greater than 8 weeks. In contrast, according to AUC, cHA

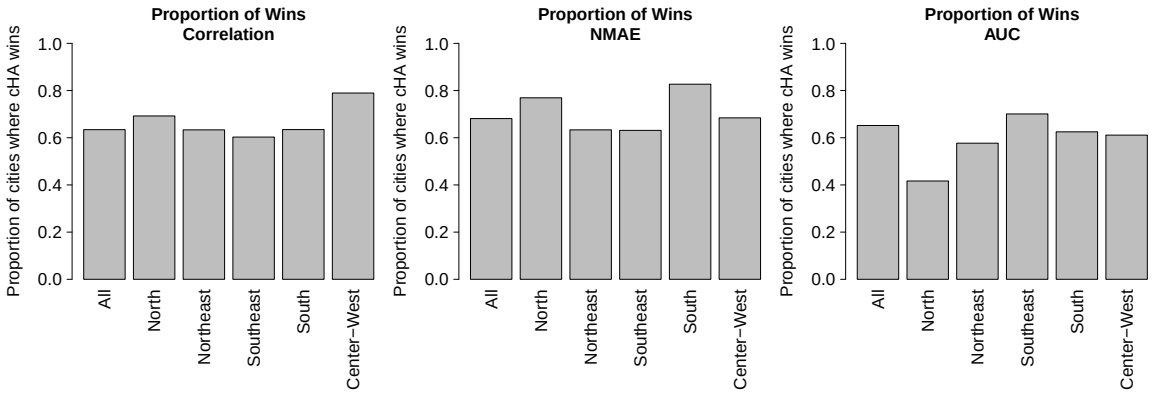

Figure 7.7. Proportion of wins/losses obtained by the cHA stratified by Brazilian administrative region.

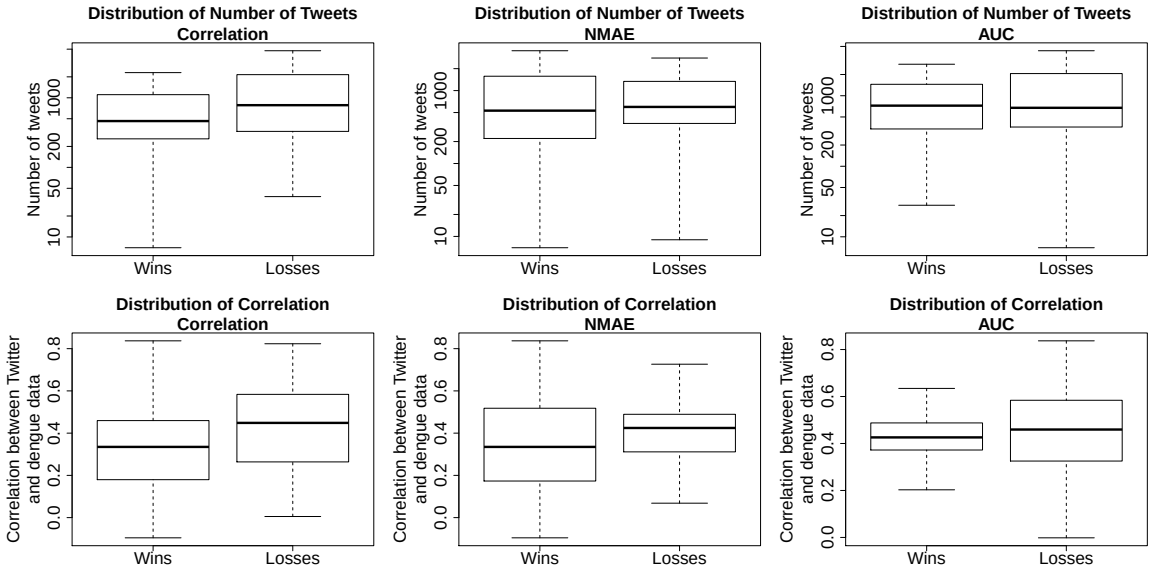

Figure 7.8. Distribution of number of tweets and correlation between Twitter and dengue data for cities where the proposed approach obtained better results and where it obtained worse results.

is to be preferred for delays greater than or equal to 6 weeks. The remaining metric, NMAE, exhibited only marginal differences between both approaches regardless of how much epidemiological data is delayed.

7.3 Summary and Discussion

Many predictive epidemiological models require up-to-date data for making accurate predictions. However, epidemiological data usually takes time to be publicly available, as it may require time-consuming confirmation tests or time to propagate through the

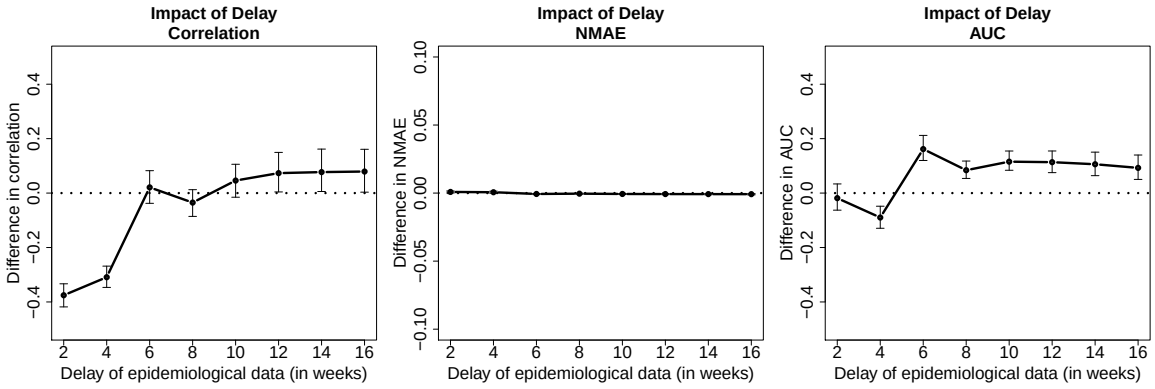

Figure 7.9. Difference in each evaluation metric for predictions issued by cHA and HAnE for epidemiological data delay ranging from 2 to 16 weeks. Positive values indicate higher values achieved by cHA, while negative values indicate the opposite scenario.

hierarchy of health authorities. On the other hand, online data sources provide data in real-time and may be associated to the incidence of a given disease. Motivate by this scenario, in this chapter we proposed to enrich the model proposed in Chapter 5, DGP, to use Twitter data as proxy for epidemiological data. By doing so, we aimed to maintain the predictive capability of DGP, while using it in a more realistic scenario where epidemiological data is provided in a delayed fashion.

Two approaches were proposed in this chapter. The first approach, names Hybrid Approach (HA), uses the number of online data to estimate missing epidemiological data, which can be used posteriorly by DGP. The second approach, names Online-only Approach (OA), forecasts *future* online data and exploits a linear relationship between online and epidemiological data to predict future epidemiological data.

Assuming that epidemiological data is provided with delay of 8 weeks, we conducted experiments to compare both proposed approaches and found that cHA obtains more accurate predictions. We then compared cHA with a simple strategy that simply extends the antecedence with predictions are made, finding that HA obtains statistically better predictions when considering AUC, while obtaining similar results when considering correlation and NMAE.

By evaluating the accuracy of cHA for varying delays associated to epidemiological data, we verified that cHA is indeed superior to other approaches for sufficiently large delays. AUC were again the most benefited evaluation metric: cHA was superior to not using online data for delays greater than or equal to 6 weeks. This reinforces the validity of the proposal, since delays of more than one month are reasonable to be expected.

Chapter 8

Conclusions and Future Work

Dengue fever is a mosquito-borne disease transmitted by females *Aedes aegypti* mosquitoes that affects hundreds of millions of human beings worldwide. Although case fatality rate is typically low with proper treatment, it causes social and economical burden in almost all tropical countries of the world. Particularly, Brazil contributes to a significant proportion of the global number of cases, being the most affected country in the Americas. Since vaccines or treatments for dengue fever are not yet available for public use, control of the disease can only be done through suppression of vector population and quick identification of new outbreaks. For the latter, appropriate early warning systems are required. Therefore, the main goal of this work was the development of a predictive model capable of effectively integrating an EWS.

According to Wiltshire [2006], an appropriate predictive model for EWS should be able to affirmatively answer the following three questions:

1. Are the right parameters being monitored?
2. Is there a sound scientific basis for making forecasts?
3. Can accurate and timely warning be generated?

Traditional models for DIR usually can obtain affirmative answers for the first and second questions, since they exploit parametric models based on known relationships between dengue fever and other covariates. However, models used in this paradigm have their complexity constrained by the number of parameters, being unable to model more complex models even when more data is available.

Based on this reasoning and on the three questions above, we defined a spatio-temporal model based on Gaussian processes equipped with a quasi-periodic covariance function. Our model is built based on previous dengue fever incidence. In this sense,

we answer affirmatively to the first question. Forecasts made by the proposed models explore seasonality and spatial dependences, which are deeply studied in the literature and commonly used for dengue modelling (see surveys Naish et al. [2014]; Louis et al. [2014]). Besides that, the Gaussian process modelling framework leads to interpretable models, since covariance functions can be seen as a measure of similarity and/or dependence between data points. Therefore, predictions made by the proposed models are trustworthy, leading to an affirmative answer to the second question. Finally, higher accuracy is provided by the non-parametric nature of the proposed models, which allows for higher spatial heterogeneity and exploitation of more complex patterns than traditional parametric approaches. Predictions are always made with four weeks in advance, thus allowing time for health authorities to act in the direction of minimizing the impact of future dengue outbreaks. Given that, we believe the proposed models are capable of successfully integrating an real EWS for dengue fever in Brazil.

Another advantage of the proposed model is automatically identifying spatial dependences. Most previous models for dengue fever that exploit spatial dependences enforce relationships based on distance or neighborhood functions. We observed that dependences not constrained by distance led to more accurate models. To some extent, this result is expected, as distance only is not capable of correctly identifying relationships between areas in a highly connected world, where fast means of transportation are available. In this context, more flexible spatial structures are necessary, and the proposed model does not impose almost any prior structure, learning spatial dependences from data.

In this work, we also showed that the proposed model can use data from online sources in a more realistic scenario where epidemiological data is not provided in real-time. We proposed frameworks that safely use Twitter data to enhance the accuracy of traditional epidemiological models by estimating missing dengue data and selecting cities where Twitter and epidemiological data are associated. Although we were capable of improving accuracy, there is still room for improvement in this area, specially when considering smaller cities, where Twitter data is not abundant.

An additional limitation of traditional models for DIR that use parametric models is the necessity of a careful analysis of data in order to define an appropriate model. This may lead to very specific models, which have trouble when applied in other scenarios. For instance, we were unable to reproduce results from a previous model for dengue fever in Brazil by simply changing data resolution and period under study. The model proposed in this work, however, is much more general, requiring only spatial dependences and seasonality, besides being non-parametric. In this sense, we believe it could be applied to other diseases in other regions of the world.

We highlight that, despite positive results in general, the usage of the proposed predictive model in a real-life EWS requires careful analysis. Considering the results obtained in this thesis, we observed that, if only incidence levels are required and up-to-date epidemiological data is provided, the proposed model obtained good results for the vast majority of cities, independent of where they are located. On the other hand, if the epidemiological data is provided with significant delays, then our model is more adequate to regions with facilitated access to Internet, such as the Southeast region of Brazil, where people will publish more often information that could be used to infer the real incidence of dengue fever. However, even for this region, predictions issued by the model are to be treated only as alarms, indicating that, perhaps, human specialists should analyze the current situation in a given area at risk. In this sense, the main advantage in using such tools is to help specialists in better navigating the huge amount of data available, facilitating the perception of new outbreaks. A blind belief in any EWS could lead to an erroneous comprehension of the real incidence of dengue fever, in a similar fashion to the Google Flu case [Lazer et al., 2014].

Based on these conclusions, we indicate the following future work:

Dealing with online data scarcity We intend to integrate other online data sources, such as volume of dengue-related queries on search engines and Wikipedia, to our methodology. With more data, we expect to obtain more clear relationships between online and epidemiological data even for less populated cities, consequently improving the accuracy of the proposed framework.

Evaluation the generality of the proposed model We intend to apply the proposed model in other scenarios, particularly for other mosquito-borne diseases. In the case that accurate predictions are obtained, this would imply in a step towards building general epidemiological models. For doing so, we first would need to collect epidemiological data (and possible online data) associated to the respective disease. Then, it is necessary to check if the same spatio-temporal patterns are present, as well as dependences between epidemiological and online data sources. Finally, we could apply the same methodology described in this thesis, with possible small adjustments in the covariance function used.

Integrate the model to a real EWS Finally, we aim to integrate the proposed model to a real-life EWS, capable of issuing alerts based on incidence levels. For that, we would need to better identify for which cities the proposed model is appropriate

and, therefore, could be used safely. It is also necessary to build a system capable of continuously integrating new epidemiological and online data.

Bibliography

Althouse, B. M., Ng, Y. Y., and Cummings, D. A. T. (2011). Prediction of dengue incidence using search query surveillance. *PLoS Negl Trop Dis*, 5(8):1–7.

Álvarez, M. A. and Lawrence, N. D. (2008). Sparse convolved gaussian processes for multi-output regression. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pages 57--64. Curran Associates, Inc.

Álvarez, M. A. and Lawrence, N. D. (2011). Computationally efficient convolved multiple output gaussian processes. *Journal of Machine Learning Research*, 12:1459--1500.

Amari, S. (1993). Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(3):185--196.

Banu, S., Hu, W., Guo, Y., Hurst, C., and Tong, S. (2014). Projecting the impact of climate change on dengue transmission in dhaka, bangladesh. *Environment International*, 63:137 – 142. ISSN 0160-4120.

Bhatnagar, S., Lal, V., Gupta, S. D., Gupta, O. P., et al. (2012). Forecasting incidence of dengue in rajasthan, using time series analyses. *Indian journal of public health*, 56(4):281--285.

Bonilla, E. V., Agakov, F. V., and Williams, C. K. I. (2007a). Kernel multi-task learning using task-specific features. In Meila, M. and Shen, X., editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, AISTATS 2007, San Juan, Puerto Rico, March 21-24, 2007*, volume 2 of *JMLR Proceedings*, pages 43--50. JMLR.org.

Bonilla, E. V., Chai, K. M. A., and Williams, C. K. I. (2007b). Multi-task gaussian process prediction. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Ad-*

vances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007, pages 153--160. Curran Associates, Inc.

Boyle, P. and Frean, M. R. (2004). Dependent gaussian processes. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 217--224.

Buczak, A. L., Koshute, P. T., Babin, S. M., Feighner, B. H., and Lewis, S. H. (2012). A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. *BMC medical informatics and decision making*, 12(124).

Caruana, R. (1993). Multitask learning: A knowledge-based source of inductive bias. In *Machine Learning, Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, MA, USA, June 27-29, 1993*, pages 41--48. Morgan Kaufmann.

Chan, T.-C., Hu, T.-H., and Hwang, J.-S. (2015). Daily forecast of dengue fever incidents for urban villages in a city. *International journal of health geographics*, 14(9).

Chen, C. and Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88(421):284–297. ISSN 01621459.

Choudhury, Z. M., Banu, S., and Islam, A. M. (2008). Forecasting dengue incidence in dhaka, bangladesh: A time series analysis. *Dengue Bulletin*, 32:29–37.

Colón-González, F. J., Lake, I. R., and Bentham, G. (2011). Climate variability and dengue fever in warm and humid mexico. *The American journal of tropical medicine and hygiene*, 84(5):757--763.

Descloux, E., Mangeas, M., Menkes, C. E., Lengaigne, M., Leroy, A., Tehei, T., Guillaumot, L., Teurlai, M., Gourinat, A.-C., Benzler, J., et al. (2012). Climate-based models for understanding and forecasting dengue epidemics. *PLoS Negl Trop Dis*, 6(2):e1470.

Dom, N. C., Hassan, A. A., Latif, Z. A., and Ismail, R. (2013). Generating temporal model using climate variables for the prediction of dengue cases in subang jaya, malaysia. *Asian Pacific Journal of Tropical Disease*, 3(5):352 – 361. ISSN 2222-1808.

Earnest, A., Tan, S., and Wilder-Smith, A. (2012). Meteorological factors and el nino southern oscillation are independently associated with dengue infections. *Epidemiology and infection*, 140(07):1244--1251.

Eastin, M. D., Delmelle, E., Casas, I., Wexler, J., and Self, C. (2014). Intra-and interseasonal autoregressive prediction of dengue outbreaks using local weather and regional climate for a tropical environment in colombia. *The American journal of tropical medicine and hygiene*, 91(3):598--610.

Eaton, M. L. (1983). *Multivariate statistics: A vector space approach*. John Wiley & Sons. ISBN 0-471-02776-6.

Gershman, S., Hoffman, M. D., and Blei, D. M. (2012). Nonparametric variational inference.

Gharbi, M., Quenel, P., Gustave, J., Cassadou, S., La Ruche, G., Girdary, L., and Marrama, L. (2011). Time series analysis of dengue incidence in guadeloupe, french west indies: forecasting models using climate variables as predictors. *BMC infectious diseases*, 11.

Gomide, J., Veloso, A., Meira, Jr., W., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, pages 3:1--3:8, New York, NY, USA. ACM.

Gubler, D. J. (2002). Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. *Trends in Microbiology*, 10(2):100 – 103. ISSN 0966-842X.

Hii, Y. L., Zhu, H., Ng, N., Ng, L. C., and Rocklöv, J. (2012). Forecast of dengue incidence using temperature and rainfall. *PLoS Negl Trop Dis*, 6(11):1–9.

Hu, W., Clements, A., Williams, G., and Tong, S. (2010). Dengue fever and el nino/southern oscillation in queensland, australia: a time series predictive model. *Occupational and environmental medicine*, 67(5):307--311.

Hu, W., Clements, A., Williams, G., Tong, S., and Mengersen, K. (2012). Spatial patterns and socioecological drivers of dengue fever transmission in queensland, australia. *Environmental health perspectives*, 120(2):260--266.

Johansson, M. A., Dominici, F., and Glass, G. E. (2009). Local and global effects of climate on dengue transmission in puerto rico. *PLoS Negl Trop Dis*, 3(2):1–5.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183--233.

Kuhn, K., Campbell-Lendrum, D., Haines, A., Cox, J., Corvalán, C., Anker, M., et al. (2005). Using climate to predict infectious disease epidemics. *Geneva: WHO*.

Lázaro-Gredilla, M., Quiñonero Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. (2010). Sparse spectrum gaussian process regression. *Journal of Machine Learning Research*, 11:1865--1881. ISSN 1532-4435.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203--1205.

Louis, V. R., Phalkey, R., Horstick, O., Ratanawong, P., Wilder-Smith, A., Tozan, Y., and Dambach, P. (2014). Modeling tools for dengue risk mapping - a systematic review. *International Journal of Health Geographics*, 13(1):1--15. ISSN 1476-072X.

Lowe, R., Bailey, T. C., Stephenson, D. B., Graham, R. J., Coelho, C. A., Carvalho, M. S., and Barcellos, C. (2011). Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in brazil. *Computers & Geosciences*, 37(3):371--381.

Lowe, R., Bailey, T. C., Stephenson, D. B., Jupp, T. E., Graham, R. J., Barcellos, C., and Carvalho, M. S. (2013). The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in southeast brazil. *Statistics in medicine*, 32(5):864--883.

Lowe, R., Barcellos, C., Coelho, C. A., Bailey, T. C., Coelho, G. E., Graham, R., Jupp, T., Ramalho, W. M., Carvalho, M. S., Stephenson, D. B., et al. (2014). Dengue outlook for the world cup in brazil: an early warning model framework driven by real-time seasonal climate forecasts. *The Lancet infectious diseases*, 14(7):619--626.

Luz, P. M., Mendes, B. V. M., Codeço, C. T., Struchiner, C. J., and Galvani, A. P. (2008). Time series analysis of dengue incidence in rio de. *American Journal of Tropical Medicine and Hygiene*, 79(6):933--939.

Maimon, O. and Rokach, L. (2005). *Data mining and knowledge discovery handbook*, volume 2. Springer.

Martinez, E. Z. and Silva, E. A. S. d. (2011). Predicting the number of cases of dengue infection in ribeirão preto, são paulo state, brazil, using a sarima model. *Cadernos de saude publica*, 27(9):1809--1818.

Martinez, E. Z., Silva, E. A. S. d., and Fabbro, A. L. D. (2011). A sarima forecasting model to predict the number of cases of dengue in campinas, state of são paulo, brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, 44(4):436--440.

Murray, I., Adams, R. P., and MacKay, D. J. C. (2010). Elliptical slice sampling. In Teh, Y. W. and Titterington, D. M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, pages 541--548. JMLR.org.

Naish, S., Dale, P., Mackenzie, J. S., McBride, J., Mengersen, K., and Tong, S. (2014). Climate change and dengue: a critical and systematic review of quantitative modelling approaches. *BMC infectious diseases*, 14(1):167.

Nguyen, T. V. and Bonilla, E. V. (2013). Efficient variational inference for gaussian process regression networks. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, volume 31 of *JMLR Workshop and Conference Proceedings*, pages 472--480. JMLR.org.

Pillonetto, G., Dinuzzo, F., and Nicolao, G. D. (2010). Bayesian online multitask learning of gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):193–205. ISSN 0162-8828.

Promprou, S., Jaroensutasinee, M., and Jaroensutasinee, K. (2006). Forecasting dengue haemorrhagic fever cases in southern thailand using arima models. *Dengue Bulletin*, 30:99.

Rasmussen, C. E. and Nickisch, H. (2010). Gaussian processes for machine learning (gpml) toolbox. *Journal of Machine Learning Research*, 11:3011--3015. ISSN 1532-4435.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142. ISBN 0-262-18253-X.

Saatçi, Y. (2012). *Scalable inference for structured Gaussian process models*. PhD thesis, University of Cambridge.

Samir, B., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., Drake, J. M., Brownstein, J. S., Hoen, A. G., Sankoh, O., Myers, M. F.,

George, D. B., Jaenisch, T., Wint, G. R. W., Simmons, C. P., Scott, T. W., Farrar, J. J., and Hay, S. I. (2013). The global distribution and burden of dengue. *Nature*, 496(7446):504–507. ISSN 0028-0836.

Shewchuk, J. R. et al. (1994). An introduction to the conjugate gradient method without the agonizing pain.

Shi, Y., Liu, X., Kok, S.-Y., Rajarethinam, J., Liang, S., Yap, G., Chong, C.-S., Lee, K.-S., Tan, S., Chin, C., et al. (2016). Three-month real-time dengue forecast models: An early warning system for outbreak alerts and policy decision support in singapore. *Environ Health Perspect*, 124(9). ISSN 1552-9924.

Silawan, T., Singhasivanon, P., Kaewkungwal, J., Nimmanitya, S., and Suwonkerd, W. (2008). Temporal patterns and forecast of dengue infection in northeastern thailand. *Southeast Asian J Trop Med Public Health*, 39(1):90--98. ISSN 0125-1562.

Snelson, E. and Ghahramani, Z. (2005). Sparse gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*, pages 1257--1264.

Souza, R. C., de Brito, D. E., Assunção, R. M., and Meira Jr, W. (2015). A latent shared-component generative model for real-time disease surveillance using twitter data. *arXiv preprint arXiv:1510.05981*.

Souza, R. C. S. N. P., de Brito, D. E. F., Cardoso, R. L., de Oliveira, D. M., Jr., W. M., and Pappa, G. L. (2014). An evolutionary methodology for handling data scarcity and noise in monitoring real events from social media data. In *IBERAMIA*, volume 8864 of *Lecture Notes in Computer Science*, pages 295--306. Springer.

Teixeira, M. G., Costa, M. d. C. N., Barreto, F., and Barreto, M. L. (2009). Dengue: twenty-five years since reemergence in Brazil. *Cadernos de Saúde Pública*, 25:S7 – S18. ISSN 0102-311X.

Vanhatalo, J. and Vehtari, A. (2007). Sparse log gaussian processes via MCMC for spatial epidemiology. In Lawrence, N. D., Schwaighofer, A., and Candela, J. Q., editors, *Gaussian Processes in Practice, Bletchley Park, Bletchley, UK, June 12-13, 2006*, volume 1 of *JMLR Proceedings*, pages 73--89. JMLR.org.

Wang, Y. and Khardon, R. (2012). Sparse gaussian processes for multi-task learning. In Flach, P. A., Bie, T. D., and Cristianini, N., editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol,*

UK, September 24-28, 2012. Proceedings, Part I, volume 7523 of *Lecture Notes in Computer Science*, pages 711--727. Springer.

Williams, C. K. I. and Seeger, M. W. (2000). Using the nyström method to speed up kernel machines. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 682--688. MIT Press.

Wilson, A. G. and Adams, R. P. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1067--1075. JMLR.org.

Wilson, A. G., Knowles, D. A., and Ghahramani, Z. (2012). Gaussian process regression networks. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.

Wiltshire, A. (2006). 'developing early warning systems: A checklist. In *Proc. 3rd Int. Conf. Early Warning (EWC)*.

World Health Organization Media Centre (2016). Dengue and severe dengue. http://www.who.int/mediacentre/factsheets/fs117/en/. [Online; accessed 25-April-2016].

Wu, P.-C., Guo, H.-R., Lung, S.-C., Lin, C.-Y., and Su, H.-J. (2007). Weather as an effective predictor for occurrence of dengue fever in taiwan. *Acta Tropica*, 103(1):50 – 57. ISSN 0001-706X.

Yu, H.-L., Yang, S.-J., Yen, H.-J., and Christakos, G. (2011). A spatio-temporal climate-based model of early dengue fever warning in southern taiwan. *Stochastic Environmental Research and Risk Assessment*, 25(4):485--494.

Yusof, Y. and Mustaffa, Z. (2011). Dengue outbreak prediction: A least squares support vector machines approach. *International Journal of Computer Theory and Engineering*, 3(4):489.