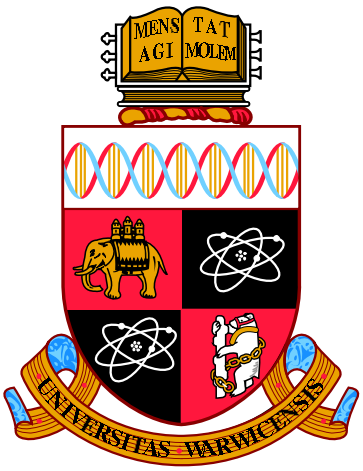

Improving dengue fever surveillance with online data

by

Giovanni Mizzi

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Mathematics for Real-World Systems

Centre for Doctoral Training

April 2019

THE UNIVERSITY OF
WARWICK

Contents

Acknowledgments

I would like to express my sincere gratitude to my advisors Prof. Suzy Moat and Prof. Tobias Preis for the continuous support of my PhD study and related research, for their patience, motivation, and teachings. Their guidance helped me in all the time of research and writing of this thesis.

My most sincere thanks also go to Dr. Claudia Codeço, Dr. Leonardo Bastos, Dr. Marcelo Gomes and the rest of the team in Fiocruz for the the close collaboration we established, for their help, their advice, and for allowing us access to the raw data which were essential for this thesis. I am also grateful for the support provided by the University of Warwick GRP Behavioural Science for our visit to Fiocruz in July 2016.

My thanks also to each and everyone who shared with me this experience at Warwick and to those who supported me from back home.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Abstract

Dengue is a major threat to public health in Brazil, the world's sixth largest country by population, with nearly 1.5 million cases reported in 2016 alone and case counts continuing to grow. However, official data on the current number of dengue cases can often be severely delayed, with incremental delivery of data and a wait of up to six months for full case count information. Previous studies have sought to exploit rapidly available data on dengue-related Google searches or Twitter messages to deliver improved estimates of dengue cases, but have not accounted for the true nature of the delays in dengue data across Brazil, rendering operational usage of these approaches unrealistic. Here, we develop a model which uses online data to deliver improved weekly estimates of dengue cases in Rio de Janeiro, whilst explicitly accounting for the structure of the delays in incoming dengue case count data. In contrast to previous approaches, we draw on data from Google Trends and Twitter in tandem, and demonstrate that this leads to better estimates compared to models using only one of these data streams alone. We also demonstrate how our model can be extended to forecast future dengue incidence. To underline the robustness of our approach, we apply our model to a range of Brazilian cities. Our results provide evidence that online data can be used to improve both estimates and predictions of disease incidence, even where the underlying case count data are severely delayed. Crucially, the model we present is operationally realistic, and can therefore be used in practice to support the decision-making processes of health authorities.

Abbreviations

AIC	Aikake Information Criterion;
ACF	Auto-Correlation Function;
ARIMA	Auto-Regressive Integrated Moving Average;
CDC	Centre for Disease Control;
Fiocruz	Oswaldo Cruz Foundation, Rio de Janeiro, Brazil;
GMRF	Gaussian Markov Random Field;
INLA	Integrated Nested Laplace Approximation;
LGM	Latent Gaussian Model;
LOG(Q)	Logarithmic Error;
MAE	Mean Absolute Error;
MAPE	Mean Absolute Percentage Error;
MEM	Moving Epidemic Method;
MPI	Mean Prediction Interval;
NB	Negative Binomial;
nMAE	normalised Mean Absolute Error;
nMPI	normalised Mean Prediction Interval;
rel LOG(Q)	relative Logarithmic Error;

relMAE relative Mean Absolute Error;

relMAPE relative Mean Absolute Percentage Error;

relMPI relative Mean Prediction Interval;

relRMSE relative Root Mean Squared Error;

RMSE Root Mean Squared Error;

SINAN Sistema de Informação de Agravos de Notificação, national notifiable diseases information system;

WAIC Watanabe-Akaike Information Criterion;

WHO World Health Organisation;

ZINB Zero Inflated Negative Binomial.

CHAPTER 1

Introduction

The primary goal of the present work is to find a way to improve estimates of the number of dengue cases in Brazil using online data sources such as Google searches and Twitter posts. We focus primarily on *nowcasting*, that is on estimating the current number of new dengue cases. When we say current, we refer to a short period of time, usually a week in our case, that just ended and for which complete official information is not yet available. This is generally why we need to make an estimate. Later in this thesis we also address short-term forecasting, by which we mean estimating the number of dengue cases in a future week.

Dengue is the most common mosquito-borne disease worldwide. It is widely spread in tropical and subtropical areas, with 50 to 100 million cases reported each year (Stanaway et al., 2016) and almost 4 billion people at risk (Brady et al., 2012). Its symptoms are usually similar to those of ordinary influenza but, unfortunately, in some cases it can be fatal (World Health Organization Regional Office for South-East Asia, 2009). Furthermore, there is currently no antiviral treatment to reduce severe illness (Endy, 2014) and considerable restrictions exist on the usage of Dengvaxia, the only vaccine licensed to date. In fact, it seems that, in patients that have not yet been infected, it might cause successive dengue infection to become severe (Aguiar and Stollenwerk, 2017a; Vogel, 2018; The Lancet Infectious Diseases, 2018; World Health Organization, 2018b).

Dengue is endemic in many countries, including Brazil. This means that the national health service, known as Unified Health System, has to deal with it every summer season. Thus dengue is not only a deadly disease but also a burden for the Brazilian economy.

Every day, we all make decisions. Some of them only affect ourselves, some of them may affect many people. Whether or not we are aware of it, whenever we make a decision, we use information, and we extract information from data. Information does not always drive our decisions, but we always look for and consider it. We use information when we decide what to wear, what to eat, when we decide what car we want to buy, when we choose a university, or a job, or a destination for our holidays. Physicians use information to decide therapies, businesses use information to decide what consumer sector to target, governments use information to create policies needed to run countries efficiently.

The higher the decision stakes are, the more important it is to consider information carefully. In order to do that, it is necessary to collect, organise, and interpret data correctly. Data collection can be dated back to ancient Egypt and even before, when first attempts were undertaken to estimate the population and catalogue trade activity and inventory. This kind of information was, at that time, a strategic advantage to grow and survive (Grajalez et al., 2013).

During the last couple of decades, we have been generating data at an unprecedented pace. Every two years the total amount of data produced by the whole population of the planet doubles in size, meaning that in the next two years we are going to generate as many data as we produced during the whole history of humanity (Gantz and Reinsel, 2011).

Scientists have been collecting incredible amounts of data through experiments at large scales. We can think of the data collected by radio astronomy experiments such as ASKAP[1], or by nuclear physics experiments such as those conducted at the CERN[2] facilities. We can also think of smaller scale data, such as those collected from surveys and experiments in psychology and related disciplines.

In today's world, countries and their administration are becoming more and more digitised and accessible. Many administrations make their data available for everyone to analyse and use. In many cities, such as New York[3], London[4] and Rome[5] just to name a few, data about transport, education, housing, health, crime and many other categories are made publicly available through open data portals. For some countries this is a faster process than for others.

[1] http://www.atnf.csiro.au/projects/askap/index.html
[2] https://home.cern/
[3] https://opendata.cityofnewyork.us/data/
[4] https://data.london.gov.uk/
[5] https://dati.comune.roma.it/

The commercial world is being digitally transformed as well, at an even faster pace. Traditional companies face increasing competition from digital disruptors in many sectors. Examples of these new disruptors include companies like Uber, Amazon or Airbnb. Even in the manufacturing sector, machine learning and artificial intelligence allow to improve production, reduce supply chain forecasting errors, automate quality testing. Such techniques are also being used to facilitate the recruiting process through automatic candidate screening, or in marketing to better tailor the offer to consumers (Manyika et al., 2017).

Finally, there are companies such as Google, Twitter or Facebook, that provide online services only and collect data on human interests, actions, communication and similar information. All these data are of particular interest to us to analyse, measure and predict human behaviour. We can think of all the data that we daily produce with our smart devices, our searches on the Internet, our social interactions, our payments, our travels. All this information is immensely valuable, and even though we release it for free scattered around our digital world as well as our real world, it is laboriously collected by such private companies that use them to make their own decisions, and only sometimes they release anonymised bits of these data publicly. As an example, Google collects information about what everybody around the world searches on the internet. This information might be used by services such as Google AdWords[6] to offer companies advertising tailored and targeted to potential clients. The same kind of information, in a more aggregated and anonymised form, can be publicly accessed through Google Trends[7] and analysed to extract information.

Looking back at Brazil and our specific problem, we see that while digital and online data grow at an increasingly fast pace, official data regarding disease spreading and the technology for their collection grow more slowly. Delays in the data retrieval process are still very long in Rio de Janeiro and other cities analysed in this thesis, and the process is not entirely digitised. In some places, health units still use handwritten notification forms and monitoring reports that then need to be sent to the municipality's epidemiological surveillance system section to be entered into the national notifiable diseases information system (SINAN) (Galvao et al., 2008), and because of this, it can take up to 6 months to successfully collect and classify all data about patients and dengue cases.

[6] https://ads.google.com/home/

[7] https://trends.google.com/trends/

For this reason, in the last decade researchers have explored the use of online data to better estimate the number of dengue cases instead of, or alongside more traditional statistical methods (Chan et al., 2011; Souza et al., 2015; Yang et al., 2017; Marques-Toledo et al., 2017). Most of this research analyses historical data and tries to find exploitable correlations between official and online data, but given the difficulty of accessing official data, it is not easy for researchers to build models that could be used in practice. Commonly used methods to nowcast the number of cases in the current week, such as auto-regressive models, are based on the assumption that complete official data are available up to the week before the one we want to make an estimate for. This assumption might not always hold, especially where the surveillance system works similarly to the Brazilian SINAN. In these cases, auto-regressive models might not be the most appropriate approach.

In the present work, we show how online data can be a useful source of information that could help practitioners of disease surveillance gain access to more timely and more accurate estimates, providing public health policymakers with more time and more information to make decisions. We describe methods that can be deployed in practice given the type of data that is available in reality in Brazil. In particular, we present methods that exploit the rapid availability of online data, whilst accounting for the complex structure of the delays in the official disease surveillance data and taking advantage of all released surveillance data.

In Chapter 2, we present the highlights of the relevant research in the last few years concerning our particular topic. Being able to address a disease outbreak appropriately is very important in terms of resources for the economy of a country, but even more in terms of wellbeing of the population. In fact, since dengue can be fatal in its severe form, an appropriate response to an outbreak could potentially save many lives. Several approaches have been found to be successful at predicting the incidence of different diseases, and they complement traditional disease surveillance in suggesting what kind of actions should be taken against a disease outbreak. In the present thesis we address the case of dengue in Brazil. It is a particularly challenging case, especially because of the difficulty of having a quick and reliable diagnosis, and because of the slowness of the data collection process in Brazil that can produce delays in the availability of data of up to 6 months. We discuss what the main issues are and how they have been addressed in the past.

In Chapter 3, we present a thorough exploration of the data that are available for this study, highlighting the main issues that need to be addressed if we want to

improve on the state-of-the-art methods. Furthermore, we also provide a summary of the basic concepts behind the models and algorithms that we use in the following chapters. Finally, we address a critical aspect of predictive modelling, that is how to evaluate models, something that in previous research has not received the attention it deserves. As a case study, we first focus on the city of Rio de Janeiro.

Chapter 4 includes results regarding our attempt at using state-of-the-art auto-regressive models to estimate the weekly number of dengue cases in the city of Rio de Janeiro. We show that these models are not appropriate for estimating the current number of dengue cases in Brazilian cities, mainly due to severe delays in the data that such models cannot automatically take into account. In particular, we find that estimates are comparable to those currently produced by the InfoDengue nowcasting system and that using online data from Google and Twitter as external regressors improves these estimates. Unfortunately, we find that the prediction intervals generated by these models are not reliable because they do not capture the expected percentage of observed weekly dengue case counts. In other words, when considering 95% prediction intervals, we find that the observed weekly dengue case counts do fall within those intervals in less than 95% of the weeks.

In Chapter 5, we introduce the core model we use as a starting point for what we discuss in all later chapters. This new model is very different from those commonly used for disease nowcasting, and it is particularly suited for the kind of data we have. We also highlight how the addition of information from online data sources such as Google and Twitter does affect our estimates by making them more accurate and more precise. In particular, we show that we obtain the best results when we use these two data sources in tandem rather than separately, even though the improvement with respect to using only either one of them is relatively small.

We then build on this model to address one of the typical problems that we encounter in the data retrieval process of dengue cases in Rio de Janeiro. In Chapter 6 we address the problem of data not being released at the end of the week, which is when it is normally released. This might happen for several reasons, and when it does we are left with missing information in the data. Since our models do not use information which is not yet available, we explore how this delay affects the estimates and how online data can help at reducing the estimation error and the estimation intervals.

Building on these last results, in Chapter 7 we present a method for predicting the

number of dengue cases in the future, even before the first official data relative to the current week are released and when we only have partial online data to work with. To do so, we work with online data at a daily resolution, and we show that the models presented here can generate predictions even a few days earlier than our previous models could, before the week ends. Such models' prediction errors are, in the best cases, less than 20% higher than those of a model built knowing the official data that will be released at the end of the week.

Finally, in Chapter 8 we apply our methods to different Brazilian cities of variable size, and we try to identify the factors that would make our model usable in other cities.

Parts of this thesis have been presented to conferences or academic events. The results presented in Chapter 5 were presented at *Advances in Data Science 2018*, May 21st and 22nd, at the University of Manchester, and at *Data Natives 2017*, April 28th, at the City University of London. The results presented in Chapters 5, 6 and 7 have also been presented in several short talks between 2016 and 2018. A further paper is in preparation (Mizzi et al., in preparation).

The present research has the potential for impact on the way outbreak assessment is carried out. It also opens up multiple possibilities for further research in the field of disease surveillance, not just for dengue in Brazil, but also for different diseases in different countries where the data collection process is slow and the notification rate highly variable.

CHAPTER 2

Background

The present chapter aims to provide some context and motivation to what we discuss in the following chapters. We first describe what dengue is, and why it is so important that we improve current methods to monitor it. We then analyse the state of the art of disease nowcasting, and look at what further advantages modern technology gives us to build better models. In particular, we look at data coming from online sources and how they have been already used to produce more accurate estimates in different settings. Finally, we discuss InfoDengue, a Brazilian research project on dengue surveillance initially developed as a partnership between academia and the Rio de Janeiro health secretariat. InfoDengue has been successful in bringing a number of policymakers and other academic partners on board, and we discuss how the research carried out in this thesis work fits into it.

2.1 The burden of dengue

Mosquito-borne diseases are a common and important problem in many tropical and subtropical countries. Dengue is the most common mosquito-borne disease worldwide. It is caused by the dengue virus and transmitted by a mosquito known as *Aedes*, primarily *Aedes aegypti*, which also transmits other diseases such as Zika, chikungunya and yellow fever. Typical symptoms of dengue include high fever, rashes, muscle aches, joint pain and leucopenia, a disorder that causes a decrease in the number of white blood cells in the blood. A small proportion of patients develop a severe form of dengue, known as dengue haemorrhagic fever, which is potentially lethal and characterised by bleeding, low levels of blood platelets, blood plasma leakage and severe organ impairment (World Health Organization Regional Office

for South-East Asia, 2009). The virus has four different serotypes, all of which can cause the full spectrum of symptoms. Infection with one type gives lifelong immunity to that type but not to the others, providing only partial and temporary immunity to the other types. Furthermore, successive infections with different types are at higher risk of being severe (World Health Organization, 2018a). This effect is known as antibody-dependent enhancement of dengue virus replication and has to do with the fact that antibodies of a dengue type can attach to viruses of different types but cannot neutralise them, facilitating the replication of the virus (Whitehead et al., 2007).

It is challenging to quantify the global incidence of dengue because the actual number of dengue cases is generally underreported and cases are often misclassified. However, different independent estimates agree on a commonly cited interval of about 50-100 million cases reported each year (Bhatt et al., 2013; Stanaway et al., 2016). Another study about the prevalence of dengue estimates that almost 4 billion people, in 128 different countries, are at risk of infection with dengue viruses (Brady et al., 2012).

Dengue is also one of the fastest growing diseases in the world. The annual number of dengue infections continues to grow, having already grown by a factor of 30 worldwide over the last 50 years (World Health Organization, 2012). Unfortunately, there is currently no antiviral treatment to reduce severe illness (Endy, 2014), and considerable restrictions exist on the usage of Dengvaxia, the only vaccine licensed to date. The position paper by the World Health Organization (2016) provides guidelines for Dengvaxia administration, advising that it is only used in populations with a high percentage of people already infected by dengue of any type, and to administer it only to people more than 9 years old because of high risk of severe illness in patients between 2 and 5 years old. Later research on trial data has pointed out that usage of the vaccine in people who have not previously been infected appears to increase the chance of a subsequent dengue infection becoming severe (Aguiar and Stollenwerk, 2017a). After a reassessment of trial data, Dengvaxia's developer Sanofi Pasteur warned that the vaccine could increase the risk of severe dengue in particular circumstances, raising questions about the future of Dengvaxia (The Lancet Infectious Diseases, 2018). The WHO recently published an updated position paper recommending that the vaccine is only used in people who have previously been infected with dengue, and suggesting that a pre-vaccination screening strategy should be the preferred option (Vogel, 2018; World Health Organization, 2018b).

Dengue is endemic in more than 100 countries (World Health Organization, 2018a), including Brazil, where nearly 1.5 million cases were recorded in 2016 alone. Of these, 861 were confirmed cases of severe dengue and 8,402 were dengue cases supected to be severe. In 2016, there have been 642 confirmed deaths by dengue in Brazil, nearly 7% of the suspected and confirmed severe dengue cases (Brasil. Ministério da Saúde. Secretaria de Vigilância Epidemiológica, 2017). With such a high incidence rate, the disease is not only life-threatening but also a severe burden on the Brazilian economy. To maximise opportunities for the mitigation or avoidance of dengue outbreaks, policymakers would greatly benefit from accurate, rapidly available information on the number of citizens currently infected.

In Brazil, dengue fever transmission varies a lot within the year. There are periods of intense activity, typically during the summer, while in the rest of the year there is very low to no detectable activity. The variability in the magnitude of such activity, on the other hand, is very high because the complex interplay of several factors also influences dengue transmission. There are environmental factors such as temperature and humidity, human factors such as population immunity and mobility, and there are also different circulating strains. This complexity leads to high prediction uncertainties and makes it more difficult to prepare for an outbreak and allocate the right amount of resources to reduce the disease burden (Codeço et al., 2016). Furthermore, disease surveillance in Brazil is based on a passive system, relying on the cases reported by healthcare providers from patients seeking care. This also brings up numerous issues. In fact, cases are underreported, typically very delayed, there is high variability of the notification rate, and possible contamination of the data with similar diseases such as Zika and chikungunya (Galvao et al., 2008; Bastos et al., 2017).

2.2 Disease nowcasting

Disease nowcasting is a vast topic, and depending on the particular field it may have different meanings or interpretations. On top of being very broad, it is also widely studied and there exist many different approaches that developed independently. The results of this kind of study could help save lives as well as understand how to better allocate resources. It is a field where information is very precious, and for this reason, there has been much research around the topic of disease surveillance in order to provide public health policymakers and other similar stakeholders with

as much information as possible to inform their decisions. We review this body of research in the rest of this section.

Nowcasting diseases have many different purposes, because there are many different reasons why one might need more up-to-date data on a disease. For example, one could need to understand whether or not an outbreak is starting. One could need to estimate the strength of such an outbreak in terms of the number of cases or to estimate its geographical extension or its duration. The analysis could be made at the state level, or at the city level, on a monthly basis, or on a weekly basis. Of course, one would like to have the most accurate and precise information possible, but many factors intervene in how such problems can be addressed and limit what kind of estimates can be made. In particular, the main problem is usually that official information is not available in a timely fashion, making it difficult to assess the current situation. This is why we need a method to estimate information about an outbreak before official data become available.

In the context of the present thesis, we consider weekly data and the question we wish to answer is what the total number of new dengue cases in any given week is in the city of Rio de Janeiro.

To answer this question, in general, a mathematical model needs to be built. For many years the research on mathematical epidemiology was driven by models such as those proposed by Kermack and McKendrick (1927, 1932, 1933), which later led to the formulation of many other variations of the so-called *compartmental models*. These models assume that the population is divided in compartments, and that all individuals within the same compartment have the same characteristics. For example, one of the most widely known compartmental models is the SIR model (Kermack and McKendrick, 1927). Such model is composed of three compartments: the S compartment for susceptible individuals, i.e. those that can be infected, the I compartment for infected individuals, and the R compartment for recovered or immune individuals. The number of individuals in these three compartments is described by a set of differential equations, which provides a deterministic solution given the initial conditions. The SIR model is the simplest and more popular compartmental model, and many variations of it exist nowadays. In fact, compartmental models are still at the core of mathematical epidemiology research. Many steps have been taken in other directions, and many methods have been developed that effectively complement previous models and go beyond, providing new useful insights.

In more recent years, network theory has played a prominent role in mathematical epidemiology, and the honing of these tools have made traditional compartmental models more efficient. Compartmental models work under the assumption that all individuals are in contact with each other, and they move from one compartment to another with certain rates. To match more closely the structure underlying human contact and disease transmission, more complex network models recently started to being used as the underlying structure of epidemiological models (Vespignani, 2009). To address such complexity it is not possible anymore to use compartmental models based on deterministic systems of differential equations (Kermack and McKendrick, 1927), but it is necessary to adopt a stochastic approach where every single individual changes its status with a certain probability, and where some of the network features may change with certain probabilities. Research has shown that the community structure of the social network and how this underlying structure changes and evolves in time also affect the spreading of an epidemics (Nadini et al., 2018). Mobility, i.e. how much and how fast individuals move on the network, has also a strong influence on the spreading of diseases, and long range mobility such as travelling with trains or air planes makes the model even more complex (Balcan et al., 2009). Furthermore, human behaviour also changes in response to large-scale spreading of infectious diseases, and people might try to avoid places where there is a higher chance of getting infected (Meloni et al., 2011).

Most of these models can either be used as descriptive exploratory models using simulations, or as predictive models using actual data. In the case of descriptive exploratory models, there is usually a real or synthetic underlying network which is fixed, and different parameters of the epidemic models can be varied to explore different scenarios. For example, in Balcan et al. (2009) the community structure and mobility features of the network are extracted from real data, and then the spreading of a disease is simulated with different sets of parameters. This allows to evaluate the effect of different features of the model and of the network such as human mobility or the epidemics' reproductive rate, i.e. the speed at which infected individuals infect non-infected individuals. When the initial values of the parameters are calculated on real data, this kind of models can also be used to make predictions. In general, once the model is trained, it is set for the whole duration of the epidemic being estimated. This means that if some parameters of the model change during an epidemic, these changes cannot be taken into account. A similar kind of models is used instead by Tizzoni et al. (2012) where the structure of the network is extracted from real data, but in this case the parameters of the model are

recalculated in real-time every time there is new information available. This allows for more reliable estimates that can be used in an operational setting to provide guidance to public health policymakers.

This kind of model can quickly become rather complex. They do not just try to predict a number, but they try to understand the underlying spreading dynamics that leads to some particular condition of the disease incidence, and in doing so they can be fed with many different layers of information. This approach is still prevalent, but it was a necessity until not many years ago because of the difficulty of retrieving data that could be used to train these models and make timely predictions. Futhermore, knowing the dynamics of the disease is also very important for prediction accuracy and precision because the prediction horizon might be far in the future.

Models such as those we just illustrated can describe diseases which require contact or proximity between two interacting individuals, such as influenza for example. This is a huge class of disease, but unfortunately not all disease can be easily described with these methods. For example, diseases such as dengue, Zika or chikungunya are transmitted by mosquitoes rather than by humans. But it is much more complicated to keep track of mosquitoes and of their interaction with humans in the same way we can keep track of interactions among humans, even though significant advances have been made in recent years on automatic tracking of mosquitoes (Spitzen and Takken, 2018).

Nevertheless, there is a vast and growing body of research addressing mosquito-borne diseases, and dengue in particular, that use models with roots on classic epidemiology models such as SIR. To be able to address dengue, these models need to take into account a further component, i.e. mosquitoes. These studies allow us to understand the underlying dynamic of such diseases, and therefore are fundamental in epidemiology research.

For example, Tennakone and De Silva (2018) show that there is a threshold of mosquitoes per person above which a population can become susceptible to a dengue outbreak, suggesting that vector control is necessary to limit such an occurrence. Oki et al. (2011) and Páez Chávez et al. (2017) also researched what should be the optimal time of insecticide fogging to minimise dengue cases. On the other hand, Carvalho et al. (2019) highlight that reducing mosquitoes is not enough to stop an outbreak. In fact, even after removing the infected population, subsequent infections

can still generate outbreaks, suggesting that the creation of an effective vaccine is fundamental to be able to really control the disease. Aguiar and Stollenwerk (2017b) produced a model that takes into account further biological aspects of the disease, such as the presence of multiple strains and disease severity, and study how a vaccine could affect all these aspects in the long term.

Several other studies focus on the simulation of the mechanism of dengue transmission and the effect of a vaccine on such mechanisms (Asmaidi and Sianturi, 2014; Chanprasopchai et al., 2018). Many of them also tried to capture the effect of multiple strains in dengue spreading. Nuraini et al. (2007) take into account the existence of two different strains as well as the presence of a more severe form of dengue, i.e. dengue hemorrhagic fever. The goal of their model was to reduce the number of patients with hemorrhagic fever. They find that, with multiple strains, their model has multiple equilibria, and only one of them is disease-free. Aguiar et al. (2011) studied the impact of seasonality and the differences between primary and secondary infections in a two-strain model. Kooi et al. (2014) show that when an asymmetry in the force of infection rates exists between different strains, the system ceases to have a finite set of endemic equilibria compared to the single strain case, and instead show periodic solution and possibly chaotic behaviour.

These studies offer several insights on what the challenges are of dealing with mosquito-borne disease such as dengue. For example, they show that it is important to produce an effective and safe vaccine as soon as possible. Vector control can only mitigate and control a dengue outbreak, but in order to eradicate the disease in a population a vaccine is necessary. Unfortunately, to validate these models highly detailed and complex data are needed, and often it is not possible to timely validate such models.

The last couple of decades have been characterised by an incredible growth of machines' computational power and amazing production and availability of data. Digitisation has made very easy to quickly collect, aggregate and transfer data. For example, in the United States, the Centre for Disease Control (CDC) can make influenza-like illness data available with a delay of just a week[1]. With such abundance, and most importantly with such timely availability of data, we can start to think of models that can estimate the spreading of a disease in real-time, recalibrating our models much more frequently. Unfortunately, while it might be easier to collect information like visits of patients to health units and their diagnoses, it

[1] https://www.cdc.gov/flu/index.htm

might be more complicated to collect data about patients mobility, contacts, immunity and similar information. Thus, recalibrating such kind of models might prove slower than expected even if some of the data becomes available very quickly.

The much more multidisciplinary approach of modern research allows us to use methods belonging to other fields to solve real-world problems such as this with completely different methods. An alternative approach to the problem of estimating and predicting diseases is that of time series analysis. This kind of model does not take into account the underlying dynamics of disease spreading, but works at a higher level by trying to predict future values of the disease incidence based on knowledge of previous values. Time series forecasting is a typical problem in econometrics, where one is interested in finding out if some variables influence the economy, and possibly how it will evolve in the future. There is a huge literature on the study of economic time series dating back more than thirty years (Beveridge and Nelson, 1981; Nelson and Plosser, 1982; Campbell and Mankiw, 1987b,a; Harvey, 1985; Watson, 1986; Clark, 1987; Engle et al., 1987; Hamilton, 1989), and time series methods have been in econometrics books for a long time (Johnston, 1963) proving to be some of the most indispensable tools of economists. Furthermore, Philip Howrey (1980) highlights that, differently from typical econometrics models where the specification of the parameters comes from the theory behind the phenomenon that one wants to describe, time series models have the further advantage of involving a slightly weaker set of assumptions, thus being more theory-independent. There are several reasons why these models can be interesting: they are easy and computationally cheap to produce; it might be expensive to retrieve more information to estimate a proper descriptive model; forecasts from such models can serve as a useful benchmark for comparison purposes and they are useful as a preliminary step for further modelling (Kennedy, 2008).

Time series methods have already been used in the field of epidemiology to forecast the incidence of diseases. For example, Allard (1998) uses time series models to estimate the number of Campylobacter infections in Montreal, Canada. He uses a particular type of time series models called ARIMA, which stands for Auto-Regressive Integrated Moving Average. This class of models is described thoroughly in Section 3.2. Allard highlights the need to update the model and the estimates whenever new data become available, and suggests that the usefulness of these methods consists not so much in the detection of an outbreak but more in giving policymakers a clearer idea of the variability that they can expect in the number of infections. He also suggests that time series models can be even more useful in smaller jurisdic-

tions, where it is more difficult to collect large samples of data. In these situations, being aware of the expected variability in the number of cases might help focus health efforts on suspect situations rather than on random fluctuations. Time series models can also be made more complicated by adding more parameters and external regressors. For example Imai et al. (2015) illustrate this by using cholera cases and rainfall from Bangladesh and influenza cases and temperature in Tokyo. Pan et al. (2016) also show that using an ARIMA model on data provided by CDC of Nanshan, China, they are able to outperform the predictions of simpler models used by practitioners in the CDC. With a focus on dengue, Dayama and Kameshwaran (2013) use univariate time series models to forecast dengue incidence in Singapore. Reich et al. (2016a) instead use official data only to estimate dengue case counts in various provinces of Thailand, and they explicitly address the problem of delays in official data reporting by using only official data that are available at the time they make the nowcasting.

When using time series models, the difficulty is often that official data are not available in due time. Depending on the circumstances, it might take weeks or months for all the official data about infections in one particular week to be collected. For this reason, it is necessary to forecast horizons far in the future. For example, if it takes two weeks to collect data about the current week, it means that in order to have an idea of what the situation is in this week it will be necessary to make a two-weeks forecast using data up to two weeks in the past. In these cases, the precision and accuracy of the estimation might not be enough, and it is necessary to look at different methods.

An approach that is becoming more and more popular is that of looking at other data sources to complement official data. For example, Imai et al. (2015) use rainfall and temperature data to estimate cohlera in Bangladesh and influenza in Tokyo. Roussel et al. (2016) try to quantify the role of climate on seasonal influenza in France, and they conclude that several factors have an impact on the spread of influenza, but since many of them are correlated, it is difficult to clearly identify those that have a real importance. Deyle et al. (2016) confirm that temperature and absolute humidity are drivers of influenza outbreaks at a global level. For dengue too, there are several studies that consider the use of weather data to improve estimates. Hii et al. (2012) use temperature and rainfall to forecast dengue in Singapore. They find an association between the number of dengue cases and lagged meteorological data, which is thought to represent a connection with the biological development of mosquitoes life cycles and long hatching times. Luz et al. (2008) carry out a similar

analysis in Rio de Janeiro, Brazil, while Ramadona et al. (2016) use meteorological data to predict dengue outbreaks in the Yogyakarta province, Indonesia.

More recently, internet data has become increasingly popular among researchers. In the next section we examine how they have changed the way we can study people's behaviour, and how we can use them to monitor dengue outbreaks.

2.3 Nowcasting with online data

One of the most popular alternative data sources researchers started to consistently look at in the last decade is the Internet. In the modern digital era, every single person produces an incredible amount of data on a daily basis, leaving digital traces on the Internet or around themselves. We leave traces whenever we connect with our smartphones or computers to the Internet, whenever we search something on Google, whenever we write something on our favourite social network, or make a purchase on Amazon. But we also leave traces in the real world, whenever we make a phone call, whenever we buy something with our credit card, whenever we swipe our pass to get on the tube, or when we use our loyalty card at our favourite store. A new field of computational social science is emerging that studies this kind of data thanks to unprecedented depth, breadth and scale (Lazer et al., 2009).

All these data are digitised, quickly and automatically collected, transferred to data storing facilities and made available for analysis. This means that, for example, Google can predict what we are looking for, Amazon can predict what we would like to buy, Facebook can predict what we would like to read right now and respond in real time to our actions. While in certain cases the information from such data is used by these private companies to improve their services, these much faster data collection methods and the data sharing speed offered by modern technologies could help us to rapidly collect and analyse data to predict future collective behaviour and provide strategic insights to policymakers (Moat et al., 2014).

Today, humanity is facing important social and political challenges including financial and economical instability; social, economical and political divide; threats against health, such as the spreading of epidemics; organised crime and unethical use of communication and information systems. The rapidly developing field of computational social science aims to address such and similar problems of real-world societies with data driven quantitative and qualitative methods based on the

abundance and timely availability of data (Conte et al., 2012). Unfortunately, from the point of view of academic research, several obstacles still need to be overcome. In fact, while individual-level information can be collected to provide services, it is very difficult to make it available to researchers. It is necessary to appropriately anonymise these data to comply with regulations in terms of data protection and guarantee the privacy of users. To do so, it is necessary to develop a new paradigm for data sharing which can simplify the data retrieval process for researchers and protect those whose data are about (King, 2011).

Research in other fields has illustrated the opportunities that already exist to use the massive data sets generated by society's everyday actions to support better forecasting of future behaviour. In particular, data about what people search on Google have been consistently used in the past few years for many different purposes (Moat et al., 2016). For example, Google data has been widely used to predict the behaviour of the stock market. Preis et al. (2010) found evidence that there is a clear correlation between the weekly transaction volumes of companies listed in the S&P 500 American stock market index, and search volumes of the corresponding company names. Preis et al. (2013b) and Preis and Moat (2015) show that search volumes of specific terms related to the financial field could have been used to guide a strategy that would have generated much higher profits with respect to buying and holding the same stocks in the period they analysed. Curme et al. (2014) found evidence that increases in Google searches about topics in politics or business precede stock market falls. These investigations of the relationship between searches for information and stock market movements complement analyses of the relationship between the distribution of information in the news and stock market moves (Alanyali et al., 2013; Curme et al., 2017).

Researchers have also explored a variety of other fields beyond the stock market. Goel et al. (2010) show that Google search data could be used to predict the opening weekend box office revenues and the rank of songs in the Billboard Hot 100 chart, and that including online data can improve the performance of models based solely on official, publicly available data. McLaren and Shanbhogue (2011) study how Google search data can be used to nowcast the unemployment rate as well as house prices in the United Kingdom, showing that some search terms can outperform some existing indicators that are normally collected through surveys. Kristoufek et al. (2016) use Google search data to improve estimates of suicide occurrence in England before official data are made available.

Google search data can be retrieved from Google Trends[2] in the form of time series. As a result, Google data have been widely used in time series analysis problems, as the examples above illustrate. However, a different kind of exploration of such data can provide insights of a different nature. Letchford et al. (2016) studied how the most searched terms on Google in different US states correlates with some demographic features. For example, they find that people in states where there is a higher birth rate tend to look more for information about pregnancy, while people in states with a lower birth rate look for information concerning cats. Preis et al. (2012) look at how much people search about past or future years on Google, showing that the propensity to search about future years correlates with per-capita gross domestic product (GDP), a strong economic indicator. Noguchi et al. (2014) build on this work by constructing measures of time-perspective of nations and show that, in fact, nations with higher per-capita GDP are more interested in the future and less in the past, showing a link between a psychological characteristic and the economical activity of nations.

But Google is not the only source of online data that has been and could be used to get insights on human behaviour. Another of such sources, for example, is Wikipedia. Moat et al. (2013) show that data on changes in the number of visits to Wikipedia pages about companies or financially related topics might have been used as indicators of stock market falls in the period they analysed. Alis et al. (2015a) investigate whether Wikipedia page views can be used to estimate tourism statistics. Data about users' posts and interactions on social media has also been used in several works. Alis et al. (2015b) explore how the median length of Twitter messages changes across the United Kingdom, but they find only minor deviations. Bollen et al. (2011) analyse the sentiment of Twitter messages and show that the collective mood states of Twitter feeds are correlated with the Dow Jones Industrial Average (DJIA); they show that this association could be used to improve predictions of the DJIA closing prices. Kramer et al. (2014) show that emotional contagion is possible also on a virtual social network such as Facebook, suggesting that contrarily to what is normally assumed, it is not necessary do have direct, in-person interaction with someone to influence their mood, and that observing someone's positive experiences is a positive experience for people. Twitter data and mobile phone data can also help estimate the size of crowds such as those attending political events or those attending football matches or music concerts in stadiums, as shown by Botta et al. (2015).

[2]https://trends.google.com/trends/

Data from photographs have been used increasingly in computational social science studies as the use of online photo sharing services continues to grow, alongside machine learning techniques to analyse their contents. Many social networks and internet portals offer the possibility to share photos and pictures. Flickr is one of those portals, and it has been used extensively for research purposes. Alanyali et al. (2016) use Flickr photos across the world to track protests, and find that where more photos related to protests are available there is a higher number of protests reported on newspapers, suggesting that photos published online could be used as a cheap monitoring tool not only for protests, but for human behaviour in general across the world. In fact, for example, Barchiesi et al. (2015a) try to quantify international travel flows by tracking geotagged photos published by users that were moving over time, and thus inferring their trajectories. They report that their estimates of UK visitors from different countries correlate significantly with the estimates of UK authorities regarding the number of visitors from such countries. Barchiesi et al. (2015b) conduct a similar study at the national level, tracking the mobility of users within the UK, and find that their estimates of the probability of users travelling between major cities are in agreement with official data about travel flows within the UK. Preis et al. (2013a) find a strong correlation between the number of Flickr photos uploaded relating to the Hurricane Sandy and the value of atmospheric pressure, with the landfall of Hurricane Sandy corresponding to the moment with the higher number of published photos. This suggests that photos uploaded on Flickr could be used as a real-time sensor of events attracting human attention. Finally, Seresinhe et al. (2016) find an association between the number of Flickr photos tagged as *art* and the prices of houses in the neighbourhood, showing that where there is a higher density of photos concerning art property prices tend to be higher.

Through a crowdsourcing website called *Scenic-Or-Not*, Seresinhe et al. (2015) obtain data about how much geotagged photos showing different parts of the UK are considered scenic. Then, they combine this information with UK census data on citizen-reported health and observe that people living in more scenic environments report better health, suggesting a potential effect of the environment on wellbeing. Using more sophisticated machine learning techniques, Seresinhe et al. (2017) try to understand what are the features of outdoor locations that make people find them beautiful, observing that not only natural features but also man-made ones lead people to consider places more scenic. This kind of information could provide insights to policymakers in charge with designing and protecting built and natural

environments. Based on the research just described, Seresinhe et al. (2018) show that by combining data such as pictures from Flickr and crowdsourced geographic data from *OpenStreetMap* it is possbile to quantitatively estimate the scenicness of areas and extend this type of measurement to areas where crowsourced opinion data about scenicness are not available.

We have seen that online data have had a prominent role in the research of the last decade, focusing on a huge variety of fields and topics. A lot of research has also been developed with a focus on health and epidemiology, and in particular on using online data to monitor disease spreading. One of the most popular products of such research is that of Google's on Google Flu (Ginsberg et al., 2009). They automatically select queries in the Google search data that correlate very well with the average number of physicians visits in the US where the patients presents symptoms of influenza. Because the CDC official data are released with a 1-2 weeks delay, they use such search data to estimate the current percentage of patients affected by influenza, before official data are released. Initially, these findings became a well-known example of the advantages of using online data in the generation of rapid indicators of society. However, in later years, the project became perhaps equally well-known for errors in the estimates generated (Lazer et al., 2014). Specifically, in 2013 Google Flu estimated more than double the proportion of doctor visits for influenza than the CDC subsequently reported. While it is difficult to pin down the exact reason for this mismatch, Google itself suggests that a new stem of the flu spreading at that time is likely to be the main reason, as people started to search for flu symptoms without being ill, triggered by media coverage about this upcoming flu wave (Copeland et al., 2013). For such kind of algorithms to work in practice, it is necessary to account for this kind of changes and adapt the model accordingly (Copeland et al., 2013; Preis and Moat, 2014).

As demonstrated by the Google Flu episode, using online data only may not be a reliable method to make predictions (Lazer et al., 2014; Preis and Moat, 2014). Different research, again from Google employees, shows that Google Trends data could similarly be used to help when it is necessary to estimate present behaviour that would not otherwise be known, if the search data are added to models using official data (Choi and Varian, 2012). Choi and Varian (2012) illustrate this by estimating motor vehicles and parts sales in the US, claims for unemployment benefits in the US, and monthly visitor arrival statistics in Hong Kong, and show that Google search data contributes a significant improvement in the estimates over a model that only uses official data.

The importance of nowcasting using online data such as Google searches lies in the fact that it takes some time for data about a certain time period, for example the current week, to be cleaned, aggregated, and collated in a way that is usable. This is the case when we deal, for example, with public services which have a distributed structure with many offices in the city, in the region, or in the state. It takes one week for the CDC in the US, which we mentioned before, to collect, aggregate and publish data about influenza cases[3]. Google search data[4], instead, are immediately available, and they can be used to estimate information about the current week that is not yet known, and that will only be available in one week's time. Using official data alone to estimate current influenza cases would mean using past official data to effectively forecast the current number of cases. Using Google data only, instead, is prone to errors due to sudden changes in the search behaviour caused by external factors such as media coverage, as the Google Flu Trends example illustrates. A third option is to use information from Google search data together with previous official data, and to retrain the model each time new official data arrive, in a way such that delayed official data can be used to recalibrate the online data in the training window to the correct size and counter situations similar to what happened in the Google Flu example, making the model more robust to sudden changes in the search behaviour (Preis and Moat, 2014).

This is a technique that has been increasingly used in recent years. Yang et al. (2015) use a sophisticated time series model they call ARGO (Auto-Regression with GOogle search data) to again nowcast influenza cases using official data from the CDC together with Google searches. Lampos et al. (2015) also make a thorough comparison of different models, showing that their best performing model is again one using both official data from the CDC and Google search data. Davidson et al. (2015) instead use a different approach and build a network model to extract correlations between the number of influenza cases in different US states, as reported by the CDC. They then use this information to mitigate possibly inflated Google search volumes during influenza epidemics to improve Google Flu estimates. Similarly to what we have seen before for other research fields, Google search data are not the only possibility when looking for online data sources. In fact, Paul et al. (2014) find that also Twitter data produce a significant improvement on the nowcasting of influenza cases with respect to models using official data from the CDC only.

Finally, a an emergent approach in digital epidemiology is that of model fusion.

[3]https://www.cdc.gov/flu/index.htm
[4]https://trends.google.com/trends/

Model selection in standard statistical analysis is carried out by minimizing some information criterion, and in general one picks the best model among an ensamble of models (Akaike, 1974). More advanced techniques allow to combine such models in the ensamble to generate predictions averaged on all the models within the ensamble. In fact, it has been shown that model averaging over an ensamble of models generally leads to better performance than model selection among the ensamble (Burnham and Anderson, 2004). In a recent paper, Xu et al. (2017) show how a model fusion approach can be used to forecast Influenza in Hong Kong by using Google search queries coupled with official data. This is done by means of Bayesian Model Averaging (BMA), a coherent framework that allows to combine models that might be very different from each other. BMA allows to build a meta-model that assigning weights to the original models to calculate an average prediction. BMA is slightly different from standard model averaging in that it does not reqire all models in the ensamble to be the same type of model. In their paper, Xu et al. (2017) combine a Generalised Linear Model, a LASSO model, an ARIMA model and a Neural Network, and they demonstrate how the model built with BMA generally outperforms all the other models according to the metric they use. A detailed review of BMA works can be found in Hoeting et al. (1999) and Raftery et al. (2005).

BMA approaches have also been considered in several other fields and are becoming increasingly popular because of their accuracy and flexibility. For example Faust et al. (2013) use BMA to predict economic activity with credit spreads, Sloughter et al. (2010) to produce wind forecasts, and Wöhling et al. (2015) for soil-plant model selection and prediction.

In this thesis, we want to explore how these existing nowcasting methods and the rapid availability of online data can help provide fast monitoring of dengue spreading in Brazil to provide guidance to public health policymakers. Dealing with dengue in Brazil instead of influenza in the US poses some additional difficulties, and approaches from the influenza literature would not transfer straightforwardly. The main reason for this is that, while the CDC in the US makes data about patients visits to physicians available with one week delay, in Brazil the delay with which the SINAN makes data available is much more variable. For example, in Rio de Janeiro some visits might be reported at the end of the same week in which they occurred, with no delay, but others continue to be reported in the following weeks, with delays of up to six months in the worst cases. The rate at which these data are made available also varies from city to city and from state to state. This makes it difficult to use auto-regressive models such as those used in many of the works cited

before in this context, since one of the assumptions those models are based upon is that data are fully available with a short, fixed delay. In the case of the CDC it is one week, in our case it is variable, and much longer. This aspect is analysed in much more detail in Section 3.1.

Nevertheless, there has been a lot of research on how to use online data such as Google searches and Twitter posts to produce insights on the spreading of dengue. Gomide et al. (2011) produced a monitoring system[5] based on the screening of Twitter posts in Brazil, where they extract only tweets that express personal experience of dengue filtering out tweets that are just broadly related to dengue. They find a significant correlation between the number of tweets expressing personal experience of dengue and the actual number of dengue cases reported in many locations in Brazil. This system is also described more in detail in Section 3.1 because it is one of the data sources we use in this thesis.

Chan et al. (2011) use Google search data instead, and follow an approach very similar to that of Google Flu (Ginsberg et al., 2009) by producing a linear model to predict the number of dengue cases in different countries based on the search volumes of dengue-related queries. They also find a good association between dengue-related web searches and the number of dengue cases in all the countries they consider, and that their model can correctly estimate peaks in such countries. Althouse et al. (2011) also use Google data for their estimates exploring many other models.

All these works precede 2013, when Google Flu predicted more than double the actual number of influenza cases and the research community started to understand the need to use official data together with online data. They share with that version of Google Flu the problem of only relying on online data. In fact, Souza et al. (2015) produce a Bayesian model that is based both on official dengue case count and the number of Twitter posts. They produce estimates of the number of dengue cases using official data with a delay of four weeks, and Twitter data available at the moment of nowcast. For this reason, and also because they work at the city level, the model they develop could be considered closer to an operational nowcasting model. However, as we see in 3.1, this assumption is not always true in the case that we consider in this thesis, and in many of the cities we analyse it takes more time than that to collect all dengue cases.

Similarly, Marques-Toledo et al. (2017) use Twitter data and official data with a

[5]http://www.observatorio.inweb.org.br/dengue/

delay of three weeks to estimate the current number of dengue cases. They make a very detailed analysis of the advantages of using Twitter as an external data source to estimate the number of dengue cases, and they end up preferring a model using only Twitter data and the time structure as explanatory variable because of ease of implementation and because the results were not too different from those of a model with lagged official data.

Yang et al. (2017) build on the ARGO model they developed in Yang et al. (2015) and apply the same methodology to study dengue in Brazil and other countries. Instead of Twitter, they use Google search data, and they work at a monthly resolution. As in the case of Souza et al. (2015), even if they explicitly state that they use only official information that is already available when they produce the nowcast, they make the assumption that data up to the previous month is available. Again, as they highlight themselves, these assumptions about data availability might not hold in practice, affecting the potential practical applicability of the methodology they describe.

Each of these works provides solid evidence of the fact that online data from Google searches and Twitter posts could be used to improve estimates of the number of dengue cases. However, when considering disease surveillance in Brazil, it becomes clear that none of these approaches would be appropriate for practical application, as they do not deal with the true nature of the delays in the official data provided by the SINAN. This is the core issue that we address in the present thesis. Automatically dealing with the delay structure of the dengue surveillance data provided by the SINAN is the first step towards producing a nowcasting system which is operationally realistic.

We also note that all of the studies presented in this section, focusing on influenza, dengue or a different disease, either use Twitter or Google search data to enhance surveillance or to nowcast the number of current cases before official data are made available, rather than using both Twitter and Google data together. In this thesis, we also explore the advantage of using Google and Twitter data in tandem within the same nowcasting model.

2.4 InfoDengue

InfoDengue[6] is a nowcasting system for the surveillance of dengue fever transmission in Brazil, developed by a team of researchers at the Oswaldo Cruz Foundation (Fiocruz) in Rio de Janeiro (Codeço et al., 2016). The system currently operates in 790 cities across Brazil in the regions of Ceará, Espírito Santo, Minas Gerais, Paraná and Rio de Janeiro, and the number of monitored cities keeps growing. Currently, the InfoDengue system monitors not only the number of dengue cases but also Zika and chikungunya. In the present work, we focus solely on dengue.

Every week, when new official data become available from all the monitored cities through the SINAN, a nowcasting model is run to estimate dengue incidence in each of them. The nowcasting model currently used, as described by Codeço et al. (2016) is only based on historical official data, and in particular on the historical notification rate. To estimate the number of dengue cases in one week, the current algorithm considers the number of reported cases in that week, and it uses a function to correct this number. The function it uses, as detailed in Codeço et al. (2016), is a Poisson distribution where the parameter depends on the delay and it represents the proportion of unknown dengue cases as a function of the number of days after the case is notified. It is fitted on historical data and used to correct the number of reported dengue cases every week. The parameters are fit again periodically to account for possible changes in the notification rates over time. With this corrected number of reported dengue cases, the system assesses the risk or the gravity of a dengue outbreak, and then this information is passed on to public health policymakers to guide them in addressing such problems.

Information about climate and about Twitter posts expressing personal experience of dengue (Gomide et al., 2011) is also used in the InfoDengue system to complement the nowcasting model at detecting a possible outbreak, but at the moment they are not used to produce more accurate estimates of the number of dengue cases.

Recently Bastos et al. (2017) introduced a Bayesian approach to help estimate the number of missing dengue cases using the dengue case count data alone, based on Integrated Nested Laplace Approximation (INLA) introduced in Rue et al. (2009). Bastos et al.'s model is detailed later in Section 5.1 as it is the foundation of the nowcasting model we present in this thesis. Estimates made using Bastos et al.'s

[6] https://info.dengue.mat.br

method largely outperform those of the model currently used in InfoDengue, making this model the natural starting point for further improvement.

The research presented in this thesis is intended to produce a model that is operationally realistic, which could become part of the InfoDengue pipeline. Our focus is on functionality that relates to part of the InfoDengue system, i.e. on a model to produce estimates of the weekly number of dengue cases. We concentrate our attention on the city of Rio de Janeiro for most of the chapters. In Chapter 8, we extend our analysis to other Brazilian cities. In Chapters 5 to 8, we build on the Bastos et al. (2017) model mentioned above to produce our nowcasting models that use online data from Google searches and Twitter posts alongside official data, and we also explore other potential applications based on Bastos et al.'s model. In particular, with a focus on the use of online data, in Chapter 6 we assess the robustness of behaviour of our nowcasting models to the situation in which official data fail to be delivered at the end of the week. This is a situation that rarely occurs, but that nevertheless causes high uncertainties in the estimation of the nowcasting models. Finally, in Chapter 7, we address the possibility of producing short term forecasts using online data from Google and Twitter.

CHAPTER 3

Data and methods

In the present chapter we first describe the data that are available to us for this study, and then we explore the basic concepts behind the different approaches we use to nowcast the number of dengue cases in Rio de Janeiro in the period between 2012 to 2016.

The central focus of the present thesis is to create a model that is operationally realistic. Creating an operationally realistic model means that the assumptions we make on the data availability do actually hold in the real world operation of the model, at least in the particular setting we consider. We want to create a model that could produce estimates of the current number of dengue cases using only data that are available when the estimates are made.

Thus, we need to understand what data are available to train our models at each time step. The major problems with these data are the lack of timeliness in the collection process along with the possibility of data being changed retroactively. We explore the data and discuss the challenges associated with them in Section 3.1

In Section 3.2, we give a summary of the methods commonly used in time series analysis and we look more thoroughly at ARIMA models. These are at the basis of *Adaptive Nowcasting* models we use in Chapter 4 for a first attempt at solving our problem.

In Section 3.3, we then give a summary of the foundations of the INLA framework, which is at the core of a more complex Bayesian approach. This constitutes the basis for all other models that we present, and in particular, it is the basis for the model that allows us to obtain the most interesting results.

Finally, in Section 3.4, we dedicate particular attention to how these models can be evaluated. Based on recent research, we expand on the commonly used methods and also propose to put more emphasis on the importance of prediction intervals, which are kept in high consideration by practitioners. In fact, as suggested by Allard (1998), the usefulness of these methods consists not so much in the detection of an outbreak but more in giving policymakers a clearer idea of the variability that they can expect in the number of infections.

The intended end consumer of such a model are surveillance practitioners and policymakers. They should be able to monitor the situation frequently, produce alerts when there is risk of an outbreak, and use these estimates to inform strategies to address or prevent outbreaks.

An operationally realistic nowcasting model should be able to produce estimates timely and frequently. We would like to produce estimates in a short amount of time once we obtain new data, and we should be able to run the model frequently, possibly at the same rate we obtain new data. These estimates can help disease surveillance practicioners to detect an outbreak in its early phases, providing time to policymakers to address the problem. For example, the population could be alerted via an information campaign through different media, or areas close to water could be reclaimed to prevent mosquitos proliferation.

We would like to produce estimates that are close to the notified dengue case counts we observe, but we would also like to know the variability around these estimates, to infer what the worst and best case scenario are in terms of expected dengue case counts. This information can be used by policymakers to adequately prepare for an outbreak, for example by allocating enough resources to hospitals and clinics. On the other hand, having an idea of the worst-case scenario can help policymakers to avoid allocating an excess of resources to the hospitals, precious resources that could better be used otherwise.

To summarise, our core aim is to produce an operationally realistic model that could actually be used in practice. This constraint drives all our choices.

3.1 Data

Familiarisation with the dataset is a fundamental step of the work pipeline whenever any kind of data are involved. In this section, a thorough description of the data sources available for this study will be provided, and then some features of the datasets will be explored to provide insight on how to best use them for this study.

There are two different types of data sources that we use during the present study. On one side there are *official* data. These are obtained from the national notifiable diseases information system (SINAN) and constitute the principal source of data needed for the operation of the InfoDengue[1] system (Codeço et al., 2016) and of the models we discuss in this thesis. The second type of data comes from sources related to the Internet and will be referred to as *online* data.

3.1.1 Dengue cases official data

Official data are the backbone of all the nowcasting models we consider. We want to estimate the number of dengue cases in the current week, and these data represent what we already know about the number of dengue cases in this and previolus weeks. As we show later in Section 3.1.3, we only have partial data for the current week and for many past weeks, and this is the very reason why we need methods to estimate the total number of dengue cases in all these weeks, and in the current one in particular.

For this study, official data are in the form of a list of dengue cases for the city of Rio de Janeiro during the period from 1st January 2012 to 23rd July 2016. There is high variability in the yearly number of dengue cases in Rio de Janeiro. In the period we consider we see more than 18,000 cases in 2012, but less than 4,000 in 2014. Each case has a *date of notification* and a *date of system entry*. The *date of notification* is the date on which the patient visits the doctor and dengue is diagnosed. The diagnosis is often made on the sole basis of clinical symptoms before laboratory analysis confirms that it is, in fact, a dengue case. For this reason, it is only a suspected dengue case, and it can be removed from the list if it is later confirmed to be of a different disease. The vast majority of the cases are notified in the first part of the year, between January and June, due to the seasonality of dengue incidence.

[1] https://info.dengue.mat.br/

The *date of system entry* is the date on which the information about this case is inserted into the official database and becomes available for analysis, for example in nowcasting models such as those described here. If laboratory analysis finds that a suspected dengue case is a case of a different disease, the case is retroactively removed from the list even after it is entered in the system. The effect of this is a higher uncertainty on the weekly estimates of the number of cases. This and other issues will be analysed with more detail in Section 3.1.3.

The official data were obtained from the Health Secretariat of Rio de Janeiro, via the InfoDengue[2] project (Codeço et al., 2016).

3.1.2 Online data

Two different sources of online data are used in the present study to enhance the nowcasting models we consider. The advantage of these online sources is that, while it takes many weeks for all the official data related to a given week to be collected and digitised, when it comes to online data there is no such kind of delay. At the end of any given week, all the online data relating to activity in that week and all the previous weeks are fully available. This means that if there is any correlation between these online data and the complete notified dengue cases data, then it may be possible to use this correlation to improve the nowcasting of the true data by employing online data.

The online data that are used in the models presented in this work are the following:

Google Health Trends. Data on search behaviour was obtained via the *Google Extended Trends API for Health*, for which Google granted us access for the objectives of this project. The *Google Extended Trends API for Health* is an enhanced version of the Google Trends API that allows for more refined queries with higher resolution on more extended time periods. It also provides more accurate data, especially in the case of low search volumes. We obtained daily data for the whole period of analysis from 1st January 2012 to 23rd July 2016. To identify searches relating to the topic of *dengue*, we searched for the topic using *Wikidata*[3], and then used the identified topic's Freebase identifier to query the *Google Extended Trends API for Health*. For the topic of *dengue*

[2]https://info.dengue.mat.br/
[3]https://www.wikidata.org/

fever (referred to as *dengue* from now on), the Freebase ID is */m/09wsg*. We chose the topic *dengue fever* rather than *dengue virus* as the search volume for the latter was much lower. In Brazil, the finest geographical resolution for data retrieved from the *Google Extended Trends API for Health* is state level. We therefore requested data on searches made in the state of Rio de Janeiro only. The data returned by the API represents the probability of a few consecutive searches relating to dengue, including typos and indirect descriptions of the disease, within the state of Rio de Janeiro on each day in the period of analysis.

Since 2015, the *Zika* arbovirus has presented an additional risk in Rio de Janeiro, with considerable media coverage. This disease is spread by the same mosquito as dengue, and also shares some symptoms. The same is true of a further arbovirus, *chikungunya*, which has also been present in Rio de Janeiro since 2015, although with lower case counts. To allow us to investigate whether data on *Google* searches relating to these two arboviruses might act as an additional potential signal for dengue incidence, we also retrieve searches relating to the topics of *Zika virus* (referred to as *Zika* from now on, Freebase ID */m/080m_5j*; chosen instead of *Zika fever* due to higher search volume) and *chikungunya* (Freebase ID */m/01___7l*).

Twitter. We also analyse data on the volume of tweets relating to dengue that were posted to Twitter during each week between 1st January 2012 and 23rd July 2016, for which the user location was determined to be in the city of Rio de Janeiro. Location was inferred from the user location specified in the Twitter user's user information, as described in more detail by Gomide et al. (2011). The data reflects the volume of tweets that meet both the criteria of containing the word 'dengue' and expressing personal experience of dengue (e.g., in English, "You know I have had dengue?") rather than other sentiment categories such as ironic/sarcastic tweets, tweets expressing the opinion about some fact related to dengue and informative tweets or tweets echoing public campaigns (Gomide et al., 2011). This dataset was made available to us by the *Observatorio da Dengue*[4] via the InfoDengue[5] project (Codeço et al., 2016).

We depict all the time series described above together with official data in Figure 3.1. It is possible to see that there is a correlation between the number of dengue cases notified to doctors in a given week (Figure 3.1A, black) and both the volume

[4]http://www.observatorio.inweb.org.br/dengue/
[5]https://info.dengue.mat.br/

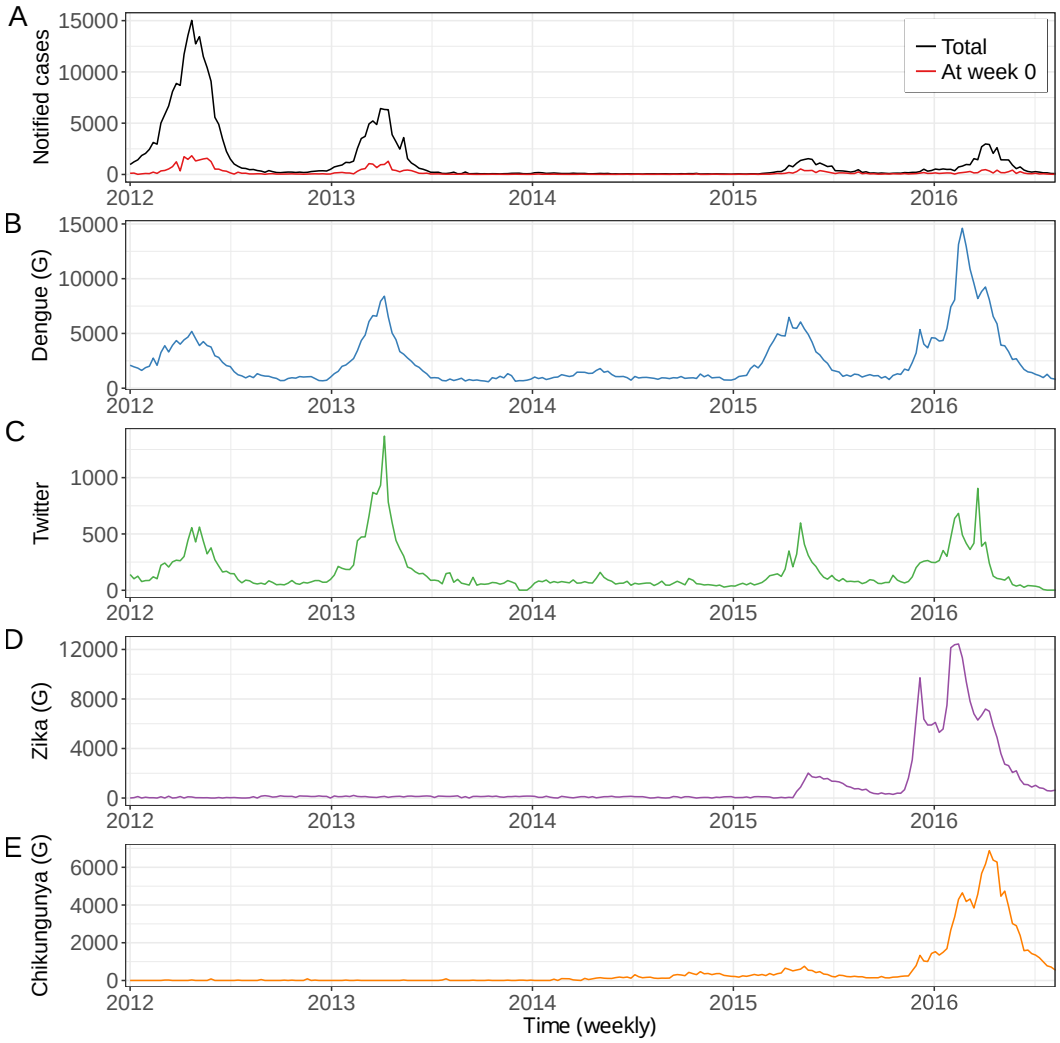

Figure 3.1: Dengue case count data compared to data from *Google* and *Twitter*. (A) In black, we depict official data on the total number of dengue cases recorded in official data for each week in Rio de Janeiro, from January 2012 until July 2016. The city frequently experiences dengue seasons during which thousands of people are infected. In red, we depict the total number of dengue cases known to the authorities by the end of each week. It is clear that only a small fraction of dengue cases are entered into the database by the end of each week. We examine the nature of these delays further in Figure 3.2, where we illustrate that there is a mean delay of 9 weeks before 95% of the final number of notified cases for a given week are entered into the system. (B) We therefore investigate whether rapidly available data on *Google* searches relating to dengue can help improve our understanding of the number of dengue cases in the previous week. We can see that peaks in dengue-related searches occur at roughly the same time as peaks in dengue cases. *(continues on the following page)*

Figure 3.1: *(continues from previous page)* However, we note that the size of the peak in searches often does not directly correspond to the size of the peak in dengue cases. (C) We also examine the relationship between dengue case counts and the number of tweets in the city of Rio de Janeiro that express personal experience of dengue. Again, we see that peaks in tweets occur at roughly the same time as peaks in cases, but the relative size of the peaks does not always correspond. (D) Since 2015, the *Zika* arbovirus has presented an additional risk in Rio de Janeiro, with considerable media coverage. This disease is spread by the same mosquito as dengue, and also shares some symptoms. We therefore also investigate whether data on *Google* searches relating to Zika might act as an additional potential signal for dengue incidence. (E) For similar reasons, we also consider data on *Google* searches relating to the arbovirus *chikungunya*. In Brazil, *Google* data are made available via the *Google Extended Trends API for Health* at the state level and therefore relates to searches in the state of Rio de Janeiro.

of Google searches (Figure 3.1B; Kendall's $\tau = 0.506$, $N = 238$, $p < 0.001$) and tweets (Figure 3.1C; Kendall's $\tau = 0.557$, $N = 238$, $p < 0.001$) relating to the topic of *dengue*. Here we used a Kendall's correlation test because some of the assumptions required for a standard Pearson's test do not hold: our data are not normally distributed or homoskedastic, and outliers are present. For this reason we prefer to use a non-parametric test that can also keep into account the fact that our data is ordinal.

For example, let us consider the time series of the number of notified dengue cases n_t and the volume of Google searches G_t. For every two weeks t and u, where $t < u$, the pairs (n_t, G_t) and (n_u, G_u) are concordant if $(n_u - n_t) \cdot (G_u - G_t) > 0$ and discordant if $(n_u - n_t) \cdot (G_u - G_t) > 0$. We then take the difference between the number of concordant and discordant pairs, and we divide this difference by the number of pairs. The Kendall's τ does not keep into account the actual difference between two values of the same time series at different times, but it only takes into account the information that one time series goes up or down in the same or opposite way than the other time series. Then, it can potentially also be used to compare time series with very different distributions.

Whereas data on Google searches and tweets are available almost immediately, only a small fraction of dengue cases are entered into the surveillance system and therefore known to policymakers and analysts in the same week in which the patient visits the doctor (Figure 3.1A, red). Indeed, there is a mean delay of 9 weeks before 95% of the cases notified to doctors in a given week are entered into the system

(Figure 3.2C). This means that in any given week, the official data on dengue cases in previous weeks is also notably incomplete. This presents clear obstacles for autoregressive models that seek to infer the number of cases in a given week by drawing on complete knowledge about previous weeks. It can also be seen that the number of cases entered into the system in the same week in which the patient visited the doctor cannot simply be multiplied by a constant to determine the total number of cases notified to doctors in that week (Figure 3.1A).

For the reasons outlined when introducing the Google Trends data above, we also examine the volume of Google searches for the topics of *Zika* (Figure 3.1D) and *chikungunya* (Figure 3.1E). We find a correlation between the number of dengue cases notified to doctors in a given week and Google searches for both *Zika* and *chikungunya* in the same week, both when considering the whole period of analysis (*Zika* searches: Kendall's $\tau = 0.127$, $N = 238$, $p < 0.01$; *chikungunya* searches: Kendall's $\tau = -0.09$, $N = 238$, $p < 0.05$) and the period beginning in the 1st epidemiological week in 2015, the year in which Zika and chikungunya became present in Rio de Janeiro (*Zika* searches: Kendall's $\tau = 0.499$, $N = 81$, $p < 0.001$; *chikungunya* searches: Kendall's $\tau = 0.526$, $N = 81$, $p < 0.001$).

3.1.3 Issues

There are several issues that affect the data sources we consider here. While we might use many different sources of information to help generate a prediction, the importance of such sources is based on their correlation with the data we want to estimate, the weekly number of dengue cases. Official data, instead, is the observed data. If we had all this data as soon as it is generated, there would be no reason to estimate the current number of dengue cases. Very often, instead, as in the case of the dataset we use here, the official data can be delayed, and this effect might be severe. This is the very reason why we need to develop methods that can accurately estimate the current number of dengue cases.

The lack of timeliness in the reporting process is an effect that could be accounted for, but in doing so there is the additional problem of high variability of the notification rate over time.

Figure 3.2 provides an overview of the situation. If we focus our attention on a particular week, we see that we know only about 25% of the cases at the end of such

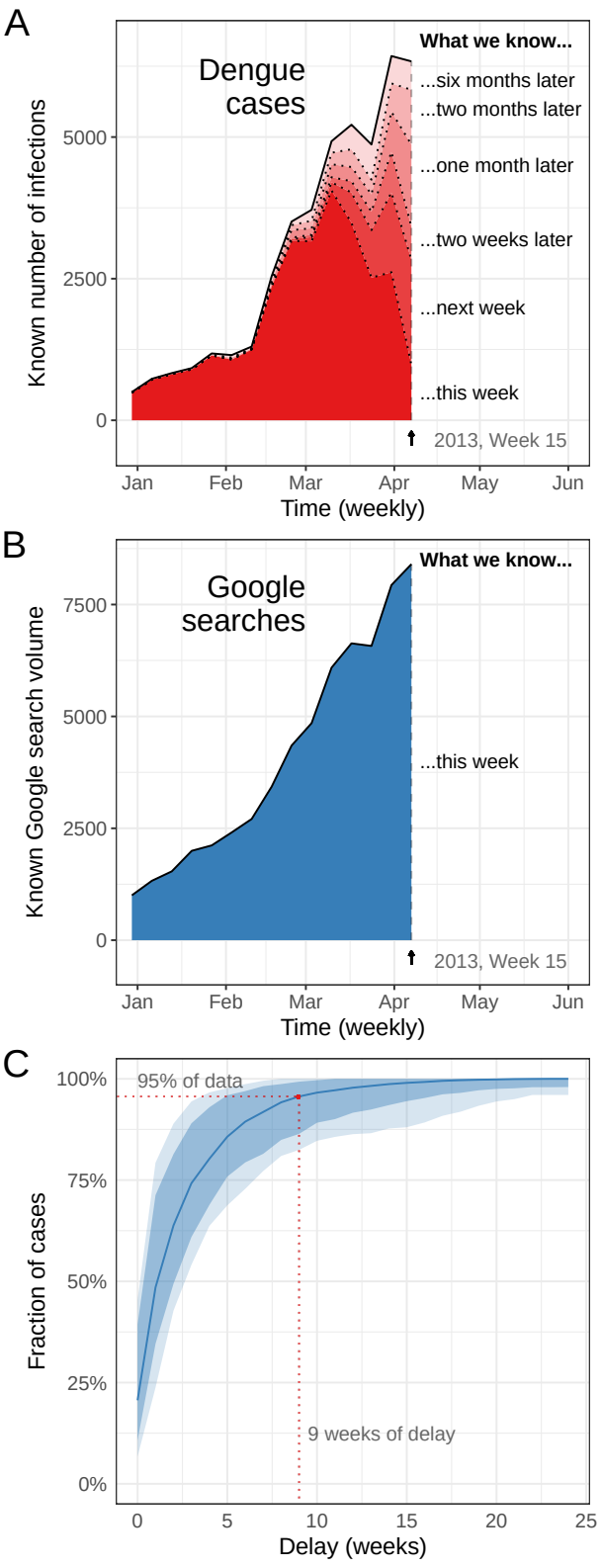

Figure 3.2: Delays in official data on dengue case counts. (A) We examine the true nature of the delays in the availability of official data on cases of dengue in Rio de Janeiro. We consider data from the 15$^{\text{th}}$ epidemiological week of 2013 as an example. We can see that only a small fraction of dengue cases have been entered into the surveillance system by the end of the week. Indeed, data relating to this week continues to arrive over a period of six months. Furthermore, by the end of the 15$^{\text{th}}$ epidemiological week of 2013, data on dengue cases in the previous weeks is also severely incomplete. This creates problems for auto-regressive methods that seek to use complete knowledge about previous weeks to compensate for delays in the arrival of data relating to the current week. (B) In contrast to official data on dengue cases, data on *Google* searches in the 15$^{\text{th}}$ epidemiological week of 2013 is available in full by the end of the week. The same applies to data on tweets posted on *Twitter*. This opens up possibilities to use data on *Google* searches and tweets relating to dengue to improve estimates of the number of dengue cases in a given week. (C) We further examine the rate with which dengue cases for a given week are added into the system. Here, we depict the empirical distribution of the delays in dengue case count entry over the whole time series. The blue line depicts the mean fraction of cases entered into the system after a given delay. *(continues on the following page)*

Figure 3.2: *(continues from previous page)* The dark shading indicates 80% of the empirical distribution of the fraction of cases notified after a given delay, and the light shading indicates 95% of the empirical distribution. We see that there is a mean delay of 9 weeks before 95% of dengue cases for a given week are entered into the system.

week. After one week, we know only 50% of the cases. It is necessary to wait for more than two months to get 95% of the data, and up to 6 months to have complete data about that week.

Furthermore, it is important to stress that the official dataset is a list of *suspected* cases of dengue. Cases that are later confirmed to be something different from dengue infection are retroactively removed from the list. This is, of course, yet another source of uncertainty in the data that makes it difficult to generate an estimate. It means that when considering the data available at a given week, one must be aware that there is a large component of missing data that has not yet been entered into the system, but also a component of cases that have been classified as dengue but that will be removed in the future. Unfortunately, this is a quantity which is even more variable. This is the operational condition in which we find ourselves when we need to estimate the weekly number of dengue cases. For this reason, it is difficult to have reliable auto-regressive models, since they rely on complete data about previous weeks, but in our case data about the current week and several past weeks are incomplete.

On the other hand, online data about the same week are entirely available at the end of the week, and because of this they offer a significant advantage, but there is a cost to pay when dealing with them. This cost is related to the fact that they are not fully controllable by the end user, they belong to private companies such as Google and Twitter that make them available at no cost but with some restrictions. For what concerns Google, the *Google Extended Trends API for Health* provides search volumes based on a sample of the data. Furthermore, since *topics* are used in the model we consider, it is not possible to be 100% sure about the procedures of elaboration or aggregation these data have gone through. Instead, the Twitter streaming API[6] which is used by Gomide et al. (2011) should cover approximately 1% of the public tweet volumes at any time. However, according to some clarification by the Twitter staff[7] the actual volume was sometimes more than 1% and there have

[6] https://developer.twitter.com/

[7] https://twittercommunity.com/t/potential-adjustments-to-streaming-api-sample-volumes/

been some changes in the sampling rate in early 2015 to rebalance it to 1%.

The data and algorithms these companies use are under continuous evolution and improvement. Slight changes in data format or sampling algorithms are likely to occur in long time windows such as those we consider in this study. For this reason, when we decide to use online data sources such as Google Trends and Twitter, we need to necessarily accept to compromise with this and continuously monitor the data to be aware of any change that could affect our models.

Finally, let us note that the spatial resolution of the data that we consider is different. While we have Twitter data and official data at the city level for Rio de Janeiro, we only have Google Trends data at the state level, i.e. for the state of Rio de Janeiro, not the city. This is something that should be generally kept into consideration while examining the results. In the particular case of the present study, we find that that there is a good correlation between the number of notified dengue cases in the city of Rio de Janeiro and the Google search volume about the topic *dengue* in the state of Rio de Janeiro. This might be due to the fact that the city of Rio de Janeiro accounts for more than a third of the population of the entire state. Being much larger than all other cities, Rio de Janeiro might also account for an even higher proportion of the population of Internet users in the state. This implicit bias might help explain why there is such a high correlation with the notified dengue case count in Rio de Janeiro.

3.2 Auto-regressive models

In Chapter 4 we use a particular type of auto-regressive model that we refer to as an *Adaptive Nowcasting model* (Preis and Moat, 2014), which is built on an ARIMA model. ARIMA stands for Auto-Regressive Integrated Moving Average, and is one of the most common and widely used time series forecasting tools. In this section, a quick review of time series analysis is provided, with the aim of introducing concepts useful to understand auto-regressive models better.

The summary we provide below draws on the overview provided by Hyndman and Athanasopoulos (2013).

3.2.1 Time series decomposition

A time series can generally be decomposed into three components:

Trend. A trend component is present in a time series when there is a long-term increase or decrease in the data. It could be linear, but it does not need to be. The trend could change in time, and it accounts for any variation which does not have a fixed and known period but is of a time scale much longer than fluctuations.

Seasonal. There is a seasonal component in a time series when the data show a cyclic behaviour of fixed and known period.

Error. The error component is always assumed to be present. It is an irregular component that accounts for anything else not accounted for by the trend and seasonal components.

If an additive model is assumed, then the time series y_t will look like

$$y_t = T_t + S_t + E_t$$

where the T_t is the trend component, S_t is the seasonal component and E_t is the error component. The index t represents time. This kind of decomposition is most suitable if the error magnitude is independent of the scale of the time series. In many real-world applications, very often in economy and finance, and also in the problem of nowcasting dengue incidence, this is not the case. Usually, the error becomes higher when the scale of the time series gets higher. In these situations a more appropriate assumption is that of a multiplicative model:

$$y_t = T_t \times S_t \times E_t$$

This is equivalent to an additive model when a log transformation has been applied to both data and components:

$$\log y_t = \log T_t + \log S_t + \log E_t$$

In fact, an effect of applying a log transformation is to stabilise the variation in the series, making it more suitable to be represented by an additive model.

3.2.2 Stationarity

A time series is said to be stationary if its properties do not depend on the time at which it is observed. For example, a time series with a trend or seasonality is not stationary. The value of these components will affect the time series differently at different times. White noise, instead, is stationary. A time series with an unpredictable cyclic pattern may be stationary. In general, we can also say that a time series is stationary if it has no predictable patterns in the long-term.

3.2.2.1 Differencing

Even if stationary time series are the objects one usually wants to work with, real-world time series are often non-stationary, including those that we analyse in this thesis. For this reason, it is necessary to have a method to transform a non-stationary time series in a stationary one. One of such methods is called *differencing*, and it consists of considering the differences between consecutive points of the time series. The first-order differenced time series y'_t is given by:

$$y'_t = y_t - y_{t-1}$$

This time series has one data point less than the original time series, because it is impossible to use the previous formula to calculate y'_1 if y_1 is the first point in the time series. This is an effective method for eliminating trend and seasonality at the cost of one data point (typically the first one).

The Auto-Correlation Function (ACF) is a useful tool to check if a time series is stationary. For stationary time series, the ACF drops to zero relatively quickly. It might be possible that some time series have an ACF that drops more slowly even if the process is stationary. This happens when there are cyclic components, or for example in the case of time series generated through some random process. Differencing in general removes this kind of autocorrelation, but if it does not, it is possible to differentiate the time series further to remove this effect. The second-order differenced time series y''_t is given by:

$$y''_t = y'_t - y'_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$$

This time series has two data points less than the original time series, because it is

not possible to use the previous formula to calculate y_1'' and y_2'' if y_1 and y_2 are the first two points in the time series. It is possible, but usually not necessary, to go beyond second-order differences.

A more rigorous method to establish if a time series is stationary is to use a *unit root test*. There exist different unit root tests, and since they are based on different assumptions, they may lead to conflicting results. One of the most popular tests is the *Augmented Dickey-Fuller (ADF) test*. In this test, the following regression model is estimated:

$$y_t' = \phi y_t + \beta_1 y_{t-1}' + \beta_2 y_{t-2}' + ... + \beta_k y_{t-k}'$$

where y' is the time series of first-order differences, and k is the number of lags to include in the regression. If the original time series needs differencing, then it is expected that $\hat{\phi} \simeq 0$, while if it is already stationary, it is expected $\hat{\phi} < 0$. The null-hypothesis for an ADF test is that the data are non-stationary.

Finally, note that not all time series can be made stationary by differencing multiple times. We can imagine a time series shaped like a sine function, or a time series representing exponential growth. In both these cases, the time series will never become stationary, no matter how many times we differentiate. Furthermore, the more differences we take, the more data points we lose.

3.2.3 Auto-regressive models

A model is said to be auto-regressive when the variable of interest is expressed as a linear combination of past values of the same variable. An auto-regressive model of order p, $\mathbf{AR(}\boldsymbol{p}\mathbf{)}$, is defined as follows:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + e_t \tag{3.1}$$

where c and $\phi_{1,...,p}$ are parameters and e_t is white noise.

This kind of prediction is based on the autocorrelation of the time series. Changing the parameters ϕ_i results in different time series patterns, while changing the variance of e_t results in a different scale of the time series, but not different patterns. There are some constraints that the parameters must satisfy. For example, for an AR(1) model, $-1 < \phi_1 < 1$. For an AR(2) model, $-1 < \phi_2 < 1$, $\phi_1 + \phi_2 < 1$,

$\phi_2 - \phi_1 < 1$. The constraints became more complicated as p grows. Usually, software that estimates this kind of models automatically take them into account.

3.2.4 Moving average models

A model is said to be moving average when the variable of interest is expressed as a linear combination of past values of the forecast error. A moving average model of order q, $\mathbf{MA(}\boldsymbol{q}\mathbf{)}$, is defined as follows:

$$y_t = c + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q} \tag{3.2}$$

where c and $\theta_{1,...,p}$ are parameters. This kind of regression is a little bit more complicated to think of. We do not observe the value of e_t, so it is not the usual kind of regression. As with the case of auto-regressive models, changing the parameters results in different patterns, but changing the variance of the error distribution will only change the scale of the patterns.

In general, it is possible to write an AR(p) model as an MA(∞) model. This is easily proven by repeated substitution. The reverse can happen if the MA model satisfies some requirements on the parameters, and then it is possible to write an MA(q) model as an AR(∞). The constraints are very similar to those for the parameters of the auto-regressive models. For MA(1) model, $-1 < \theta_1 < 1$. For an MA(2) model, $-1 < \theta_2 < 1$, $\theta_1 + \theta_2 < 1$, $\theta_2 - \theta_1 < 1$.

3.2.5 ARIMA

ARIMA stands for Auto-Regressive Integrated Moving Average. This class of models is a combination of an auto-regressive model, a moving average model, and an appropriate choice of differencing to make the time series stationary (*integration* is the inverse process of differencing). So given the time series y_t the full model $\mathbf{ARIMA(}\boldsymbol{p,d,q}\mathbf{)}$ can be written as:

$$y'_t = c + \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + ... + \phi_p y'_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q} \tag{3.3}$$

where y'_t is the d_{th}-order differenced time series, p is the order of the auto-regressive component, and q is the order of the moving average component.

Auto-regressive, moving average, and many other common models can be derived as particular cases of ARIMA models. White noise is an ARIMA(0,0,0), a random walk is an ARIMA(0,1,0) without a constant, while a random walk with a drift effect is the same with a constant. An auto-regressive model is merely an ARIMA(p,0,0) while a moving average model is an ARIMA(0,0,q).

Given a time series y_t the task of fitting it with an ARIMA model, thus choosing the order p, d, q of the model, can be difficult. Fortunately, it is something that can be done automatically. In R, this can be done with the `auto.arima()` function provided by the *forecast* package (Hyndman et al., 2018; Hyndman and Athanasopoulos, 2013).

3.2.5.1 Order selection and model comparison

Once the order of the model p, d, q is fixed, it can be estimated. R does this by Maximum Likelihood Estimation, i.e. it tries to find the parameters that maximise the probability of the observed data coming from the estimated model. At the same time, though, we are also trying to find the order of the model that maximises the likelihood overall; it is then necessary to use more sophisticated information criteria that penalise models with more parameters. With the same likelihood, a model with fewer parameters is preferable to one with more parameters. This is a common principle in the philosophy of science stating that if there are two models explaining a phenomenon, the model which requires less assumptions is to be considered more powerful than the model which requires more assumptions. Furthermore, a model with fewer parameters is less prone to overfitting, and it is more accurate when applied out of the sample it was trained on.

The Akaike's Information Criterion (AIC) can be used for determining the order of the ARIMA model, and it can be written as:

$$AIC = -2\log(L) + 2(p + q + k + 1) \tag{3.4}$$

where L is the likelihood of the data, k is either 0 or 1 depending on c in (3.3) being zero or not. The last term is just the total number of parameters of the model, including the variance of the residuals σ^2. Since the AIC formula does not involve the integration order d, this must be evaluated separately. In fact, it is the first parameter to be evaluated as all considerations about model training and

evaluation apply to the differenced time series.

Other commonly used information criteria are the corrected AIC (AIC_c), which is better suited for small samples, and the Bayesian Information Criterion (BIC), which penalises model complexity more heavily. They are all implemented in R and described in more detail in Hyndman and Athanasopoulos (2013).

By minimising one of these information criteria, it is possible to determine the best order and parameters for the ARIMA model.

3.3 Integrated nested Laplace approximation

INLA stands for Integrated Nested Laplace Approximation, and it is a method of performing Bayesian inference that was first proposed by Rue et al. (2009). More recently this approach has become widely used because it can be applied to an extensive set of problems. Even if it performs only an approximate inference, for these problems, it is much quicker than Markov Chain Monte Carlo methods (Robert and Casella, 1999), which have more commonly been used for simulation-based inference.

The methods we describe here are at the base of the model proposed by Bastos et al. (2017) that are discussed thoroughly in Chapter 5. Here we briefly summarise the INLA approach to approximate Bayesian inference which can be found in Rue et al. (2009) and Rue et al. (2017).

3.3.1 Latent Gaussian models (LGMs)

The present section is a summary of Sections 2.1 and 2.2 in Rue et al. (2017). Latent Gaussian models are an abstraction which allows to perform statistical inference for a huge class of statistical models. LGMs work under the assumption that observations $\boldsymbol{y}$ are conditionally independent given a latent Gaussian random field $\boldsymbol{x}$ and hyperparameters $\boldsymbol{\theta}_1$. The likelihood of the data $\boldsymbol{y}$ is then given by the product of the probabilities of all y_i as

$$\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_1 \sim \prod_{i \in \mathcal{I}} \pi(y_i \mid x_i, \boldsymbol{\theta}_1)$$

where $\boldsymbol{x}$ is a latent Gaussian random field. For what concerns our problem, a Gaussian random field is a Gaussian stochastic process specified by

$$\boldsymbol{x} \mid \boldsymbol{\theta}_2 \sim \mathcal{N}\left(\boldsymbol{\mu}(\boldsymbol{\theta}_2), \boldsymbol{Q}^{-1}(\boldsymbol{\theta}_2)\right)$$

where $\boldsymbol{\theta}_2$ are the hyperparamers for the Gaussian random field, $\boldsymbol{\mu}$ is the mean and $\boldsymbol{Q}$ is the precision matrix, i.e. the inverse of the covariance matrix. We call $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ the set of hyperparameters controlling the Gaussian latent field and the likelihood for the data, and we can write the posterior for the Gaussian random field as

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta}) \prod_{i \in \mathcal{I}} \pi(y_i \mid x_i, \boldsymbol{\theta}_1) \tag{3.5}$$

In order for the approximation we make in the following to be accurate, and to ensure computational efficiency, we need to make some assumptions. First, the number of hyperparameters does not depend on the dimension of the latent field n, is never larger than 20 and generally much smaller, between 2 and 5. Second, the latent field $\boldsymbol{x} \mid \boldsymbol{\theta}$ is normally distributed and approximately a Gaussian Markov Random Field (GMRF, see sec. 3.3.2) when its dimension n is high, i.e. in the range 10^3 - 10^5. Third, the data $\boldsymbol{y}$ are mutually conditionally independent of $\boldsymbol{x}$ and $\boldsymbol{\theta}$. This means that each observation y_i only depends on one component of the latent field, for example. x_i.

LGMs can be used to generalise a large class of *additive* or *generalised* linear models. In fact, for example, according to the assumptions made, the fact that each observation y_i only depends on one component of the latent field is equivalent to saying that it only depends on its linear predictor x_i, meaning that it can be used to generalise a linear model. In turn, x_i can be interpreted as the linear predictor η_i which is additive too with respects to other effects,

$$\eta_i = \mu + \sum_j \beta_j z_{ij} + \sum_k f_{k,j_k(i)}, \tag{3.6}$$

where μ is the intercept, $\boldsymbol{z}$ are fixed covariates with linear coefficients $\{\beta_j\}$, and $\{f_k\}$ are additional terms representing specific Gaussian processes where element j contributes to linear predictor i. These further f_k terms are *model components* that could be used, for example, to model auto-regressive time-series models. The fixed effects $(\mu, \boldsymbol{\beta})$ are assumed to have a joint Gaussian prior and to be independent of the model components, which are assumed to be independent among each other as

well. Since all of the pieces of model (3.6) are Gaussian, we can consider the set of stochastic variables

$$\boldsymbol{x} = (\boldsymbol{\eta}, \mu, \boldsymbol{\beta}, \boldsymbol{f}_1, \boldsymbol{f}_2, ...) \tag{3.7}$$

which has a Gaussian joint distribution, and could be interpreted as the latent Gaussian random field of an LGM as described previously in this section. The hyperparameters $\boldsymbol{\theta}$ include the parameters of the likelihood and of the model components.

3.3.2 Gaussian Markov random fields (GMRFs)

The present section is a summary of Sections 2.3 and 2.4 in Rue et al. (2017). A GMRF $\boldsymbol{x}$ is a Gaussian stochastic process where the elements x_i and x_j are conditionally independent given the remaining elements $\boldsymbol{x}_{-ij}$ for quite a few $\{i, j\}$'s. For a more detailed introduction to GMRF see Rue and Held (2005). An advantageous consequence of this property is that it results in zeros for pairs of conditionally independent values in the precision matrix (the inverse of the covariance matrix). Having sparse matrices provides a huge computational benefit compared to calculations involving dense matrices.

Additive models including GMRFs have the property that the precision matrix of the joint distribution for $\boldsymbol{x}$ such as (3.7) consists of sums of the precision matrices of the fixed effects and the other model components. This is one of the key reasons why the INLA approach is very efficient. In fact, in the INLA algorithm the joint distribution of the latent field is formed many times, and being able to work with sparse matrix avoids many computationally costly matrix operations.

3.3.3 Laplace approximations

The present section is a summary of Section 2.5 in Rue et al. (2017). The Laplace approximation is a technique for approximating integrals of the type:

$$I_n = \int_x \exp(nf(x))\, dx$$

as $n \to \infty$. If x_0 is the point in which $f(x)$ has its minimum, we can approximate $f(x)$ with a Taylor expansion around its minimum up to the second order and we

can write I_n as

$$\begin{aligned} I_n &\approx \int_x \exp\left(n\left(f(x_0) + \frac{1}{2}(x-x_0)^2 f''(x_0)\right)\right) dx \\ &= \exp(n(f(x_0)) \int_x \exp\left(\frac{n}{2}(x-x_0)^2 f''(x_0)\right) dx \\ &= \exp(n(f(x_0))\sqrt{\frac{2\pi}{-nf''(x_0)}} = \widetilde{I_n} \end{aligned} \tag{3.8}$$

where we integrated the Gaussian integral. Whilst doing so, the error turns out to be *relative* and of order $\mathcal{O}(n^{-1})$.

Let us assume that we would like to calculate a marginal distribution $\pi(\gamma_1)$ from a joint distribution $\pi(\boldsymbol{\gamma})$.

$$\begin{aligned} \pi(\gamma_1) &= \frac{\pi(\boldsymbol{\gamma})}{\pi(\boldsymbol{\gamma}_{-1} \mid \gamma_1)} \\ &\approx \left.\frac{\pi(\boldsymbol{\gamma})}{\pi_G(\boldsymbol{\gamma}_{-1}; \boldsymbol{\mu}(\gamma_1), \boldsymbol{Q}(\gamma_1))}\right|_{\boldsymbol{\gamma}_{-1}=\boldsymbol{\mu}(\gamma_1)} \end{aligned} \tag{3.9}$$

Here $\pi(\boldsymbol{\gamma}_{-1} \mid \gamma_1)$ has been approximated by a Gaussian. Tierney and Kadane (1986) show that given n replicated data from the same parameters $\boldsymbol{\gamma}$, it is possible to compute posterior marginals with a *relative* error of $\mathcal{O}(n^{-3/2})$ assuming the numerical error to be negligible.

There are two issues with the underlying assumptions, though, when we want to apply it to our problem. First, we usually do not have replicated observations from the same model. More commonly we only have one observation from any single model, but can have multiple observations from similar models. Second, having only one realisation for each observation in the random effect(s) in the model means that the size of $\boldsymbol{\gamma}$ grows with n.

If instead of saying we have *replicated observations from the same model* we say *several observations from similar models*, where "similar" is used in a loose sense, it is possible to overcome these issues in a certain sense. Of course, the closer a posterior is to a Gaussian, the more accurate will the results be. In this context, it is necessary for the posterior to be uni-modal, and certainly, it helps if it is symmetric.

3.3.4 INLA

The present section is a summary of Section 3 in Rue et al. (2017). When performing Bayesian inference, the most complicated thing one has to do is to find the posterior marginals. Here we want to demonstrate how to reformulate a problem for LGMs as a set of subproblems that can be solved with Laplace approximation.

As an example, consider a model

$$\eta_i = g(\beta) u_{j(i)} \tag{3.10}$$

where $y_i \mid \eta_i \sim \text{Poisson}(\exp(\eta_i))$, $i = 1, ..., n$, $\beta \sim \mathcal{N}(0, 1)$, $g(\cdot)$ is a well-behaved monotone function and $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$. The dimension of $\boldsymbol{u}$ does not depend on that of $\boldsymbol{\eta}$, and all u_js are observed roughly the same number of times. In our example, the data are distributed as a Poisson, so they are rather far from a Gaussian. To calculate the posterior marginals, then, we have a product of a Poisson and a Gaussian, which is not Gaussian. Unfortunately, we can only use the Laplace approximation if the density is almost Gaussian.

As we said before, we want to consider a set of subproblems where, instead, we can use the Laplace approximation.

The posterior for β can be written as

$$\pi(\beta \mid \boldsymbol{y}) \propto \pi(\beta) \int \prod_{i=1}^{n} \pi\left(y_i \mid \lambda_i = \exp\left(g(\beta) u_{j(i)}\right)\right) \times \pi(\boldsymbol{u})\, d\boldsymbol{u} \tag{3.11}$$

The integrand is a Poisson-count correction of a Gaussian prior, so we can assume that it is close to a Gaussian. This means it can be calculated using Laplace approximation.

The posterior for $\boldsymbol{u}$, and in particular for all u_j and for all values of β, can instead be written as

$$\pi(u_j \mid \boldsymbol{y}) = \int \pi(u_j \mid \beta, \boldsymbol{y}) \times \pi(\beta \mid \boldsymbol{y})\, d\beta \tag{3.12}$$

which can be calculated directly since β is one-dimensional.

The only missing bit, then, is the posterior for u_j given β, which we need to calculate

the posterior for u_j using (3.12). Similarly to (3.11)

$$\pi(\boldsymbol{u} \mid \beta, \boldsymbol{y}) \propto \prod_{i=1}^{n} \pi\left(y_i \mid \lambda_i = \exp\left(g(\beta) u_{j(i)}\right)\right) \times \pi(\boldsymbol{u}) \tag{3.13}$$

which should be close to Gaussian and thus can be calculated using Laplace approximation.

We have shown, then, how to break the problem of finding the posterior marginals by considering subproblems for which we could use Laplace approximation. Using this approach means accepting to deal with more complexity, but the good thing is that more complicated calculations are avoided through conditioning and numerical integration. This approximation is fast to compute, with little loss of accuracy. This strategy can be applied to LGMs when we replace β with $\boldsymbol{\theta}$ and $\boldsymbol{u}$ with $\boldsymbol{x}$. For a more detailed procedure of how to address these steps see Rue et al. (2017).

3.4 Nowcasting models evaluation

Evaluating the performance of nowcasting models is not a very straightforward process, especially when such models are to be used by policymakers, who might be trained in another discipline, or when they are consumed by the general public. Different metrics could in principle be used, and each of them could give useful insights on how the model is performing. The problem is that, without some expert knowledge, interpreting the results cannot be done only relying on one of those metrics because very often they could lead to contradictory conclusions. There has been some recent and less recent literature discussing the advantages and disadvantages of these metrics. Makridakis (1993) discusses how the appropriateness of a any metrics we use must be evaluated based on how effectively it provides information on post-sample accuracy, and that there is not a best method, but the best method needs to be related to the purpose of forecasting. In particular, he highlights that it is important to distinguish between academic research, which is more interested in large scale accuracy studies and is in general more concerned with averages, and reporting the performance of forecasting methods in business, government or military applications, where stakeholders are more interested in knowing what happens in specific cases and particular periods of time. Hyndman and Koehler (2006) also show that commonly used model evaluation metrics degenerate in particular situations that commonly occur. They instead propose a new metric, the mean absolute

scaled error (MASE) for comparing forecast accuracy across multiple time series. Cleger-Tamayo et al. (2012) also explore the use of different metrics in the particular case of recommendation systems, showing again that depending on the purpose of the model, the evaluation of the best model needs a different approach. Finally, Tofallis (2015) and Reich et al. (2016b) introduce some metrics that are particularly suited for our problem. In particular, they introduce the logarithmic error and relative metrics that we discuss in this section.

Here we describe the most commonly used metrics to evaluate forecasting models.

MAE. The Mean Absolute Error is probably the most common metric used for the evaluation of nowcasting and forecasting models. The reason is that it has a very intuitive meaning and can be easily computed without particular constraints. It is defined as

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{3.14}$$

where $\boldsymbol{y}$ are the true values and $\hat{\boldsymbol{y}}$ are the model's estimates.

RMSE. The Root Mean Squared Error is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{3.15}$$

where $\boldsymbol{y}$ are the true values and $\hat{\boldsymbol{y}}$ are the model's estimates. This is very similar to the MAE, but it gives more importance to points further from the expected value.

MAPE. The Mean Absolute Percentage Error is one of the most commonly used metrics for evaluating the relative error of nowcasting and forecasting models. This metric is also very intuitive. The size of the error relative to the size of the estimated values can be understood immediatly. It is defined as

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{3.16}$$

where $\boldsymbol{y}$ are the true values and $\hat{\boldsymbol{y}}$ are the model's estimates. There are two main issues with this metric. The first is that this metric is not defined any more if just one of the actual values is 0. This situation could happen fre-

quently when considering count data, which is the case of the current study, especially if there are periods of low or no incidence which is very often the case off-season in small cities. The second issue is that this function is not mathematically symmetric. For example, suppose we have two estimates for the same time i. One of them is half the true value and the other one is twice the true value. While using the MAPE, their relative error is different. When the estimate is only half of the true value, the absolute percentage error is $|y_i/2 - y_i|/y_i = 0.5$ while when the estimate is twice the true value, it is $|2y_i - y_i|/y_i = 1$. Furthermore, for a related reason, it is bounded below but not above. In fact, for estimates that are smaller than the true value, the absolute percentage error can go from 0 when the estimate is equal to the true value, to 1 when the estimate is 0. For count data, we cannot have estimates smaller than 0. On the other hand, for estimates that are higher than the true value, the absolute percentage error can go from 0 when the estimate is equal to the true value, to ∞ because there is no limit to how much higher than the true value estimates can be. Because of this asymmetry, using this metric for model selection, in general, favours models that underestimate the expected value compared to models that overestimate it. Thus it is necessary to keep this in mind when using the MAPE to compare or select models.

Recently, Tofallis (2015) proposed another metric which provides a more robust way of comparing relative errors; a method that avoids some of the problems that MAPE has such as asymmetry and lower bound.

LOG(Q). This metric is defined as

$$\begin{aligned} \text{LOG(Q)} &= \frac{1}{N}\sum_{i=1}^{N}\left(\log(\hat{y}_i) - \log(y_i)\right) \\ &= \frac{1}{N}\log\left(\prod_{i=1}^{N}\frac{\hat{y}_i}{y_i}\right) \end{aligned} \tag{3.17}$$

where $\boldsymbol{y}$ are the true values and $\hat{\boldsymbol{y}}$ are the model's estimates. This metric solves one of the issues that affected MAPE. In fact it is symmetric apart from the sign. This means that it is not biased towards over- or under-prediction in the same sense MAPE is. This makes it very useful to compare how the errors are distributed around zero, which corresponds to a perfect fit. Reich et al. (2016b) proposed a generalisation of this, in particular as a generalisation of

the MAE with a general scaling function. Following the concept behind MAE, he proposed the following

$$\begin{aligned} \mathrm{LOG(Q)} &= \frac{1}{N}\sum_{i=1}^{N} |\log(\hat{y}_i + 1) - \log(y_i + 1)| \\ &= \frac{1}{N}\sum_{i=1}^{N} \left|\log\left(\frac{\hat{y}_i + 1}{y_i + 1}\right)\right| \end{aligned} \tag{3.18}$$

This is more suited to represent the average distance from the true value, and also solves the problem of being undefined when one of the values is zero.

Finally, Reich et al. (2016b) proposed the idea of a relative metric in order to compare two different models.

relMAE. The relative MAE between model A and model B considered in the same time window is defined as

$$\mathrm{relMAE}_{A,B} = \frac{\mathrm{MAE}_A}{\mathrm{MAE}_B} \tag{3.19}$$

This is a very useful way to directly compare the average errors of two models and easily quantify how much one model is better than another.

In the same way we can extend this definition also to the other metrics defined before by (3.15), (3.16) and (3.18). We then have:

relRMSE. The relative RMSE between model A and model B considered in the same time window is defined as

$$\mathrm{relRMSE}_{A,B} = \frac{\mathrm{RMSE}_A}{\mathrm{RMSE}_B} \tag{3.20}$$

relMAPE. The relative MAPE between model A and model B considered in the same time window is defined as

$$\mathrm{relMAPE}_{A,B} = \frac{\mathrm{MAPE}_A}{\mathrm{MAPE}_B} \tag{3.21}$$

rel LOG(Q). The relative LOG(Q) between model A and model B considered in

the same time window is defined as

$$\text{rel}\,\text{LOG(Q)}_{A,B} = \frac{\text{LOG(Q)}_A}{\text{LOG(Q)}_B} \tag{3.22}$$

When trying to improve a nowcasting or forecasting model, the error on the point estimate is not the only thing we aim to reduce. Another critical factor that is often overlooked but that is kept in high consideration by practitioners is the width of prediction intervals. Having smaller prediction intervals means that the model is more precise, and the more precise it is, the more useful is the information about the point estimate. In analogy with the MAE, the following can be defined:

MPI. The Mean Prediction Interval is defined as the average width of the 95% confidence intervals across the considered time period.

$$\text{MPI} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}^i_{97.5\%} - \hat{y}^i_{2.5\%}\right) \tag{3.23}$$

where $\hat{y}^i_q$ is the q-quantile of the predicted value $\hat{\boldsymbol{y}}$ at time step i.

relMPI. Analogously to the relative MAE, the relative MPI between model A and model B considered in the same time window is defined as

$$\text{relMPI}_{A,B} = \frac{\text{MPI}_A}{\text{MPI}_B} \tag{3.24}$$

The MPI is very useful because it can give us an idea of what the best and worst case scenarios are in a given week in terms of the expected number of dengue cases. To better understand how reliable is the MPI at describing the variability around the predictions, i.e. how well the extrema of the 95% prediction intervals in each week represent the best and worst case scenarios, we also calculate the percentage of weeks where the observed dengue case count falls within the 95% predcition interval around the estimated value.

When this number is close to 95%, it means that the extrema of the 95% prediction interval reasonably represent the best and worst case scenarios in terms of expected number of dengue cases. When this number is instead smaller than 95%, it means that more often than 5% of the times our estimate of the worse or best case scenarios are incorrect, and we observe a dengue case count higher than the worse case scenario or smaller than the best case scenario. Having unreliable prediction intervals means

that the policymaker is likely to allocate an insufficient amount of resources to address an outbreak or, on the contrary, to waste resources because they allocated too many.

The time series of the dengue case count is characterised by a sequence of peaks and troughs. The error metrics we outline here are affected by the model's performance during both peaks and troughs. However, accurate, precise information may be of most use to policymakers during epidemics when case counts are high. We, therefore, carry out sub-analyses in which we focus specifically on model accuracy and precision during periods of epidemics. To identify periods of epidemics, we apply the *Moving Epidemic Method* (MEM) (Vega et al., 2013), to historical data for Rio de Janeiro. This is a method which can be used to determine the minimum number of dengue cases per week that would be expected during epidemics. By applying this methodology to the official dengue case count data, we obtain an epidemic threshold of 550 dengue cases per week for the city of Rio de Janeiro. This value has been obtained from Fiocruz, and it has been calculated on hystorical data of the number of dengue cases in Rio de Janeiro.

In conclusion, then, the way to proceed is not to find the best metrics to evaluate models automatically and make decision-based upon them but to have possibly more metrics that can provide a clear idea of what the differences between models are. Given that the typical end user of the product of this work could be someone without technical expertise, clarity and interpretability should be favoured. For this reason, it is often useful to use relative metrics, so that it is clear how much better or worse one model is compared to another.

CHAPTER 4

The adaptive nowcasting model

The problem we seek to solve in this thesis is that of estimating the number of new dengue cases in Rio de Janeiro starting from severely delayed official data, as previously described in section 3.1.1. The first approach we consider to do so in this chapter is that of auto-regressive models. Auto-regressive models have been around for decades, and they are among the most commonly used methods to predict quantities for which historical data are available. The more one looks into the future, the higher the uncertainty on the prediction.

Nowcasting is an approach that allows to gain valuable information about the present which would not be otherwise available because data are available only up to a point in time in the past. Having this information allows to make more informed decisions based on better knolwedge of the current situation. In the present chapter we focus on nowcasting the number of dengue cases in Rio de Janeiro for the week that just ended.

Since we want to focus on producing a model that is operationally realistic, we consider a particular type of auto-regressive model. These models are known as adaptive nowcasting models and have been successfully used in the past to nowcast diseases (Preis and Moat, 2014; Yang et al., 2017). We evaluate the possibility of using such methods in the context of dengue in Rio de Janeiro, given the characteristics of the data we describe in Chapter 3, and we also assess the impact of using online data from Google and Twitter on their performance.

We first provide a simple explanation of how adaptive nowcasting models work, and we then present the results for different models using different data sources.

4.1 Methods

In this section, we detail the models analysed in the present chapter. The models we consider all seek to deliver weekly estimates of dengue case counts in Rio de Janeiro. We carry out our analysis using epidemiological weeks, which are defined as starting on the Sunday. When a week spans two different calendar years, it belongs to the year in which more days of the week fall. As such, if the calendar year begins on a Monday, a Tuesday or a Wednesday, the epidemiological year is considered to have started on the final Sunday of the previous calendar year. Otherwise, the epidemiological year starts on the first Sunday of the calendar year. Each epidemiological year therefore has either 52 or 53 epidemiological weeks.

We investigate whether rapidly available data on Google searches and tweets relating to dengue in Rio de Janeiro can enhance weekly estimates of the number of dengue cases in Rio de Janeiro reported to doctors in the previous week. We therefore compare the following models:

Baseline. We first consider the adaptive nowcasting model which was applied by Preis and Moat (2014) to nowcast influenza outbreaks in the US. The reason it is called *adaptive* is because it is trained every week when new data become available, with a moving window that always includes the last n weeks before the current week. This means that the parameters of the model and the order of the model itself (p, d, q) are recalculated at each time step, and there is no memory of what the model and the data were at a time previous to the last n weeks. This time window needs to be sufficiently long to be able to train the ARIMA model, but sufficiently small so that recent changes in the relationship between the predictor variables and the outcome variable can be captured. For this reason, we do not consider training windows larger than a few months, and thus we do not include any seasonality in our models.

The adaptive nowcasting model is based on a crucial assumption, which is that the data relating to the previous n weeks $y_{t-n}, ..., y_{t-1}$ are fully known at the end of week t. In other words, when making a nowcasting at week t, all data relating to the weeks up to $t-1$ are available. This assumption does not hold in the case of dengue in Rio de Janeiro, and later in this chapter we discuss the steps we have to take to account for this.

The algorithm of the adaptive nowcasting model is straightforward and can

be summarised as follows:

1. Consider week t;
2. Train an ARIMA model with the previous n weeks, using the data $y_{t-n}, ..., y_{t-1}$, with automatic selection of the order p, d, q based on the AIC metric. We do this using the R function `auto.arima()` in the *forecast* package (Hyndman et al., 2018). We also train our model on the logged variable $y_t = \log n_t$ where n_t is the actual number of dengue cases at time t. As dicussed in Section 3.2.1 we do this for stability and to guarantee that our estimates remain positive. The model we obtain is described as

$$y'_t = c + \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + ... + \phi_p y'_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q} \tag{4.1}$$

   where y'_t is the d_{th}-order differenced time series of the number of dengue cases, p is the order of the auto-regressive component, and q is the order of the moving average component.
3. Make an out-of-sample forecast of the value of y_t using the ARIMA model that was just trained. The actual value of the number of dengue cases can be calculated as $n_t = \exp y_t$.
4. Move to the following week $t \to t+1$ and start again from the beginning.

We use the first twenty weeks of data in 2012 for training only, and begin generating estimates on epidemiological week 21 of 2012, which started on Sunday 20th May 2012. The width n of the training window is an hyper-parameter of the model and needs to be optimised separately.

Because of how it is built, the adaptive nowcasting model can be enhanced with external regressors in the same way an ARIMA model can. To make a prediction in week t, though, it is essential that the values of the external regressors are also known at week t.

Google (Dengue). This model is the same as the baseline model, with data on Google searches related to the topic of *dengue* added as an external regressor. The model we obtain can be described as follows:

$$y'_t = c + \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + ... + \phi_p y'_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q} + \gamma G'_t \tag{4.2}$$

where G'_t is the d_{th}-order differenced time series of the logged Google search volumes.

Twitter. This model is the same as the baseline model, with data on the volume of tweets that express personal experience of dengue added as an external regressor. The model we obtain can be described as follows:

$$y'_t = c + \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + ... + \phi_p y'_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + ... + \theta_q e_{t-q} + \alpha T'_t \tag{4.3}$$

where T'_t is the d_{th}-order differenced time series of the logged number of Twitter posts.

Google (Dengue) + Twitter. This model is the same as the baseline model, with data on Google searches related to the topic of *dengue* and the volume of tweets that express personal experience of dengue added as external regressors. The model we obtain can be described as follows:

$$y'_t = c+\phi_1 y'_{t-1}+\phi_2 y'_{t-2}+...+\phi_p y'_{t-p}+e_t+\theta_1 e_{t-1}+\theta_2 e_{t-2}+...+\theta_q e_{t-q}+\gamma G'_t+\alpha T'_t \tag{4.4}$$

where G'_t is the d_{th}-order differenced time series of the logged Google search volumes and T'_t is the d_{th}-order differenced time series of the logged number of Twitter posts.

Naive. Following Yang et al. (2017), this model uses the number of dengue cases cases relating to week $t-1$ which are known at the end of week t as the estimate of the number of dengue cases in week t.

4.1.1 Delay correction

As previously described in Section 3.1.3, the main obstacle to being able to produce accurate estimates of the current number of dengue cases is the fact that official data are severely delayed.

With reference to Figure 4.1, let us call week t the last full week from Sunday to Saturday. This means that if today is Sunday, week t ended yesterday, week $t-1$ ended the day before last Sunday and we are currently in week $t+1$. The original assumption of the adaptive nowcasting model is that, while nowcasting week t, official data are completely available up to the end of week $t-1$.

Here we present an operationally realistic baseline model that only uses data that are available at the time estimates are made. The data we use to train the model, in this case, are not all the data relating to the period up to the end of week $t-1$.

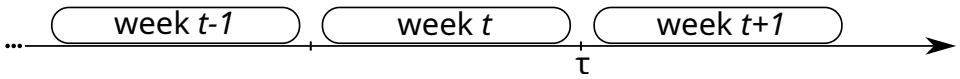

Figure 4.1: Timeline of an operationally realistic analysis. Because of how data are collected, time τ is when we obtain official data and when we perform our analysis. We define as week t the last full week which precedes time τ, starting on a Sunday and ending on the following Saturday. Week t is the week for which we want to produce an estimate of the number of dengue cases. Online data are also available for week t at time τ, as discussed in Section 3.1.2.

Instead, they are all the data relating to the period up to the end of week t which are available at the end of week t. This is precisely the situation we deal with in the real world, and it is much more complicated. Instead of having complete and reliable data up to a certain week in the past, we have incomplete data up to the last full week, with varying degrees of incompleteness. Only a small fraction of the data relative to week t are known, and more data are available for previous weeks the more we look back to the past. This is well depicted by the delay curve reported in Figure 3.2.

If all the data available at the end of week t are aggregated by week, the time series we obtain is a reasonable approximation of the complete notified dengue case count only at the beginning of the training window, but it is very different from the complete notified dengue case count at the end of the training window. A baseline model trained with this time series is not able to provide a reasonable estimate of the number of dengue infections in week t. Before we can train the baseline model, in an operationally realistic setting, it is necessary to apply a correction to the data. Furthermore, when using this operationally realistic baseline model, data relating to week t are not discarded. In the model described in Section 4.1 there were not any data available relating to week t. In our scenario, instead, there are some available data relating to week t and all available cases are used for training.

The approach followed by Codeço et al. (2016) for estimating the number of dengue infections in week t, briefly detailed in Section 2.4, consists of estimating a correcting function based on a fixed training set and then using this function to correct the data every week. Here, we follow a different approach. The official data relating to cw weeks preceding week t are used to calculate a delay curve which is then used to correct the data. This delay curve is similar to the one in the bottom panel of

Figure 3.2, but it is only relating to the considered period. In particular if

$$(d_i)_{\ 0\leq i\leq n} \tag{4.5}$$

is the empirical cumulative distribution function of the known fraction of dengue cases as a function of the delay in weeks, i.e. the delay curve, then the correction is calculated as

$$\left(y'_{t-i} = \frac{y_{t-i}}{d_i}\right)_{0\leq i\leq n} \tag{4.6}$$

where y'_t are the corrected values while y_t are the original values. It is important to note that cw does not necessarily need to be equal to the number of weeks in the training set dw of the adaptive nowcasting model or the order p on which the ARIMA model is trained. The reason for this is that in periods with low dengue case count there might not be enough cases to calculate this curve appropriately. For this reason, it might be more convenient to fix the number of cases instead. In this way, cw will be variable, and during peaks it will be small, thus representing only a small number of weeks, while outside the epidemic season it will be larger, possibly representing a number of weeks higher than the training window.

This is the starting point for using the adaptive nowcasting algorithm described previously in this section. After correcting the available case count for all the weeks in the training set, it is then possible to train a baseline model or one enhanced with online data.

4.2 Results

To evaluate the capabilities of the adaptive nowcasting model for our problem, here we follow a progressive approach starting from the simplest case and proceeding to the more complicated ones. First, we reproduce the model used by Preis and Moat (2014), and then we consider an operational situation where not all the information is available at the time we make our estimate.

Figure 4.1 illustrates the timeline of our model's operation. This structure applies to all the analyses performed in the present chapter. In general, we are at time τ every time we make an estimate of the number of dengue cases in week t, and online data relating to all weeks previous to week t are always assumed to be available at time τ. Instead, the availability of official data varies depending on the model.

In all the cases we consider in the present chapter, we use a naive model as a benchmark, and when we consider the more operationally realistic case, we also compare all models to the model previously used for estimating dengue case counts by the InfoDengue system in Rio de Janeiro.

4.2.1 Models with complete data

The first set of models we consider is based on the assumption that data are available relating to all weeks up to week $t-1$, and we want to estimate dengue case counts for week t. This is a good approximation in other settings, such as in the case of flu in the US (Preis and Moat, 2014), but it does not hold for dengue in Rio de Janeiro as we have shown in Section 3.1.3. Furthermore, there is, in general, a small amount of data relating to week t which is known but not taken into account in such kind of models. So, clearly, the models we describe here are not operationally realistic in this particular context, but it is good to consider them because they are less complicated than other models we explore in later chapters and because we can use them as benchmark for successive models.

Here, we first consider a baseline model which only uses official data. Because we deal with count data that may vary by orders of magnitude, the ARIMA model is trained on the logged time series. This has two main effects. The first one is to prevent the estimated values to be negative. Using the time series as it is may, and does, lead to this result in more than a few circumstances when the average count is close to 0. Even if the estimated logged values turn out to be negative, exponentiation makes them positive again. The second reason is that the variance of the error is typically the same variance of the sample since the forecast is just one step ahead. Using the original time series would make errors in the peaks too small, and errors in the descending part of the peak too big. The log transformation, instead, deals with this making the confidence intervals' boundaries both positive, and the width bigger during the peaks and smaller off the peaks.

Here we face the first choice of hyperparameters. One of the things we need to choose in training this adaptive nowcasting model is the width of the sliding window. We need to select a number of weeks which is enough to train an ARIMA model, but not too big such that recent changes in the relationship between the predictor variables and the outcome variable can be captured. We consider sliding windows with widths dw between 10 and 20 weeks.

Then, following the approach of Preis and Moat (2014), we add online data as external regressors. In particular, we use two data sets. The first one is the Google search volume relating to the topic *dengue* and the other one is the number of Twitter posts as described in Section 3.1. Both data sources are available up to week t. These allow us to build three different models that we call *Google (Dengue)*, where we only use Google data as external regressor; *Twitter*, where we only use Twitter data; and finally *Google (Dengue) + Twitter*, where we use both the data sets in tandem.

We compare the performance of all these models to that of a naive model where the dengue case count in week t is the value of the last known data point, i.e. the number of dengue cases relating to week $t-1$ that are available at the end of week t.

Table 4.1: Accuracy of all dengue nowcasting models compared to the naive model for the case of complete data with training window of $dw = 20$ weeks. To compare the models, we use the relative Mean Absolute Error (relMAE), relative Root Mean Squared Error (relRMSE), relative Mean Absolute Percentage Error (relMAPE) and the relative logarithmic error (relLOG(Q)), as defined in Section 3.4. All values are relative to the naive model, and therefore the values of all metrics are 1 for the naive model. We also report the actual values of the metrics for the naive model in parentheses. For example, under relMAE, we give the true MAE for the naive model in parentheses. We see that the relMAE and relRMSE are lower than 1 for the *Google (Dengue)* and *Google (Dengue) + Twitter*, but if we consider the relMAPE or relLOG(Q) we find that the naive model is the best performing model.

Model	relMAE	relRMSE	relMAPE	relLOG(Q)
Naive	1 (193.6)	1 (462.2)	**1** (0.290)	**1** (0.263)
Baseline	1.096	1.103	1.046	1.121
Google (Dengue)	**0.827**	**0.682**	1.115	1.111
Twitter	1.235	1.308	1.276	1.260
Google (Dengue) + Twitter	0.919	0.820	1.142	1.144

Table 4.1 shows a comparison of different error metrics between all models for the case of complete data with a training window of 20 weeks, which is the value of dw for which we observe the highest accuracy. When we say complete data, we mean that the models are run under the assumption that all data for weeks previous to the one we are predicting are available, and online data are available also for week t, which is the one that we are trying to predict.

When we compare the models using the relative MAE or relative RMSE, we observe that the baseline model has a higher error than the naive model, with an MAE of 212.2 and an RMSE of 509.8 compared to an MAE of 193.6 and an RMSE of 462.2 for the naive model. The *Google (Dengue)* and *Google (Dengue) + Twitter* exhibit a smaller error instead, with an MAE of 160.1 and an RMSE of 315.2 for the *Google (Dengue)* model and an MAE of 177.9 and an RMSE of 379.0 for the *Google (Dengue) + Twitter* model. This is not the case, though, when we consider the relative MAPE or relative LOG(Q) metrics, which indicate that no model performs better than the naive model, with an MAPE of 0.290 and a LOG(Q) of 0.263.

As we highlighted in Section 3.4, different metrics show different results and lead to different conclusions. For example, all metrics in 4.1 suggest that the baseline and *Twitter* models have higher errors than the naive model. The MAE for the *Twitter* model is 23.5% higher than the naive model, at 239.1 cases. Instead, if we look at the *Google (Dengue)* or *Google (Dengue) + Twitter* models, we can see that different metrics show these errors go in different directions. According to the relative MAE and RMSE these errors are smaller than for the naive model. However, the *Google (Dengue)* model exhibits an MAPE of 0.323 while and a LOG(Q) of 0.292 while the *Google (Dengue) + Twitter* model has an MAPE of 0.331 and a LOG(Q) of 0.301. In both cases, the errors are higher than the naive model.

This discrepancy might be explained considering that the first two metrics, i.e. MAE and RMSE, have a scale, while MAPE and LOG(Q) are scale-free because they are defined as ratios of quantities of the same scale. In metrics with a scale, since smaller values are summed together with higher values, the higher the value, the higher the variability, the higher the error, and the more it counts within these metrics. On the other hand, in scale-free metrics points in the peaks count as much as points in the troughs.

In the top panel of Figure 4.2 we can see a direct comparison of the time series of the different models. We see that the only two lines that are easily distinguishable from the group are those relative to the baseline and *Twitter* models, which are also the ones that appear to be worse than the naive model when considering relative MAE and relative RMSE as shown in Table 4.1. In the bottom panel of the same figure, we compare the time series of the absolute errors for the naive model (y-axis pointing down) and the *Google (Dengue)* model (y-axis pointing up), which is the best performing model in this case.

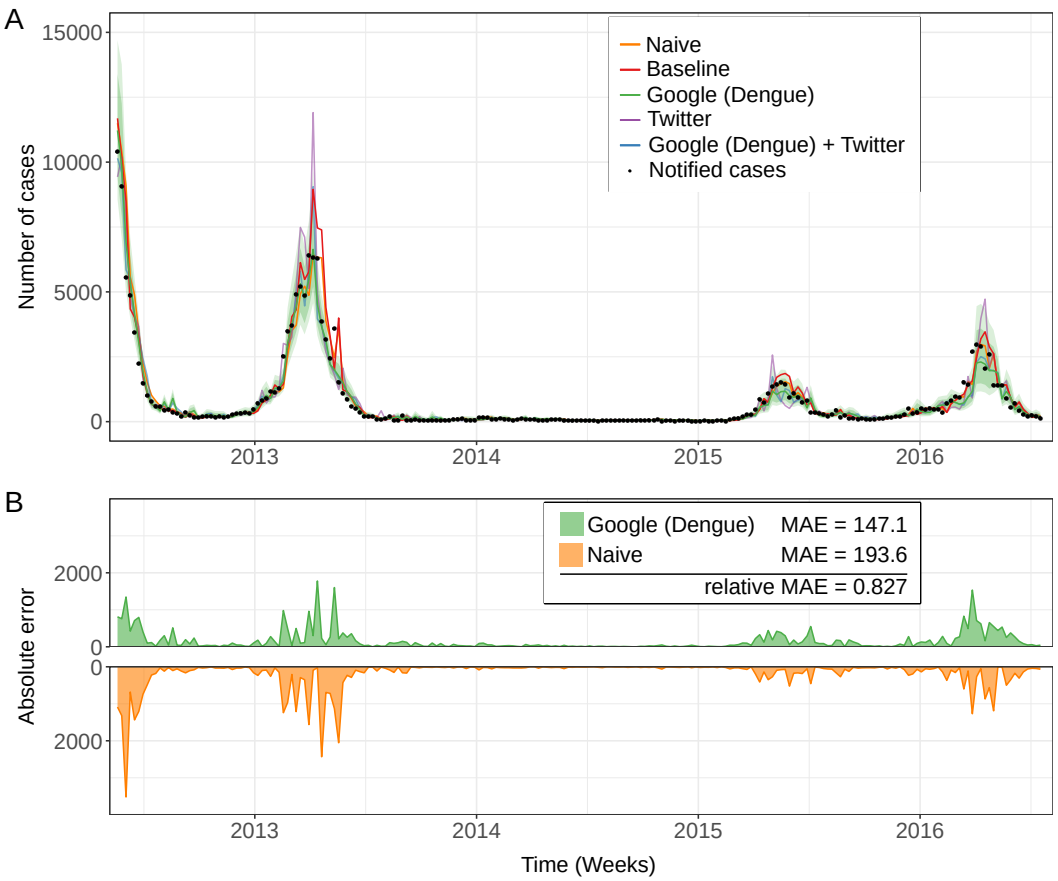

Figure 4.2: Comparison of the different models considered in the case of complete data with training window of $dw = 20$ weeks. (A) We compare the performance of the naive model with that of the baseline model and that of models drawing on data from Google and Twitter. In black, we depict official data on the total number of dengue cases, while the green shaded areas represent the 80% (dark green) and 95% (light green) prediction intervals for the *Google (Dengue)* model. (B) We compare the weekly absolute error for the naive model and the *Google (Dengue)* model. We can see that the absolute error is generally smaller during the epidemic seasons in 2012 and 2013 for the *Google(Dengue)* with respect to the Naive model. When looking at the 2015 epidemic season, the difference is not easily visible, and in 2016 it seems that the naive model outperforms the *Google(Dengue)*.

Based on these considerations, the results reported in Table 4.1 and Figure 4.2 seem to suggest the following: errors are generally smaller during epidemic seasons for the *Google (Dengue)* and *Google (Dengue) + Twitter* models compared to the naive and baseline models, and they are generally higher during the rest of the year. Let us take Figure 4.2 as an example. We see that during the 2012 and 2013 epidemic seasons, which are much more intense than the other epidemic seasons, the absolute errors are much smaller for the *Google (Dengue)* model compared to the naive model. In the same epidemic seasons we expect also the absolute percentage errors to be smaller for the *Google (Dengue)* model compared to the naive model. When we calculate the MAE, these epidemic seasons are those who weigh more, and the smaller absolute errors in these epidemic seasons result in a smaller MAE. Nevertheless, we observe a higher MAPE. When calculating the MAPE, all points weigh exactly the same. This implies that the absolute percentage errors outside of the epidemic seasons we considered must be on average higher for the *Google (Dengue)* model compared to the naive model. From this, it follows that the same is true for absolute errors.

There is something more that we could do to better analyse these data. Specifically, it is much more important that we have better performance during the epidemic season than that we avoid a slightly worse performance outside the epidemic season. We could therefore look at only considering weeks with high dengue case counts in our analysis, or in other words, weeks in which it would be considered that an epidemic was taking place.

Using the Moving Epidemic Method (Vega et al., 2013), we determine the epidemic threshold for Rio de Janeiro to be 550 dengue cases per week. We then calculate the MAE considering only weeks with a number of dengue cases higher than the epidemic threshold.

These results are presented in Table 4.2. We observe that when considering only weeks a number of dengue cases above the epidemic threshold the *Google (Dengue*) model provides the highest accuracy according to all the metrics we use. Specifically, the *Google (Dengue*) model exhibits an MAE of 411.0, 25.6% smaller than the naive model at 552.4. Similarly, the MAPE for the *Google (Dengue*) is 0.216, 20% smaller than for the naive model at 0.271. During epidemic periods we also observe higher accuracy than the naive model for the *Google (Dengue* + Twitter) model, with an MAE of 471.4 cases and an MAPE of 0.249, repsectively 14.6% and 8% smaller than the baseline model.

Table 4.2: Accuracy of all dengue nowcasting models compared to the naive model during epidemic seasons for the case of complete data with training window of $dw = 20$ weeks. To compare the models, we use the relative Mean Absolute Error (relMAE), relative Root Mean Squared Error (relRMSE), relative Mean Absolute Percentage Error (relMAPE) and the relative logarithmic error (relLOG(Q)), as defined in section 3.4. Moreover, we only consider weeks with number of cases greater than the dengue epidemic threshold for Rio de Janeiro $N_{mem} = 550$. All values are relative to the naive model, and therefore the values of all metrics are 1 for the naive model. We also report the actual non-relative values of the metrics for the naive model in parentheses. For example, under relMAE, we give the true MAE for the naive model in parentheses. We see that during epidemics seasons the *Google (Dengue)* model performs better than all other models according to all metrics.

Model	relMAE	relRMSE	relMAPE	relLOG(Q)
Naive	1 (552.4)	1 (462.2)	1 (0.271)	1 (0.257)
Baseline	1.102	1.105	1.054	1.025
Google (Dengue)	**0.744**	**0.665**	**0.799**	**0.915**
Twitter	1.242	1.309	1.183	1.248
Google (Dengue) + Twitter	0.854	0.810	0.920	1.030

These results suggest that, as in the case of nowcasting flu-like illness in the US (Preis and Moat, 2014), when using complete data an adaptive nowcasting model using information from Google Trends produces better estimates with respect to a naive model, but also with respect to the baseline model which in this case seems to be outperformed by the naive model. In this particular case, we also see that there is no advantage in adding Twitter data to our model as the *Google (Dengue)* model produces better estimates than the *Google (Dengue) + Twitter* model.

After having observed that Google data helps reduce the prediction error when all official data are available, we want to investigate the more complicated and operationally realistic case of this problem, which is when we do not have complete data.

4.2.2 Model with incomplete data

Here we consider a more operationally realistic set of models which only use official data that are available when we perform the estimation. In this case, we use a different baseline model where we correct the available dengue case count as described in Section 4.1.1 before training the ARIMA models. Again, we are making out-of-

sample estimates. In fact, even though we are aware of a small percentage of dengue cases relating to week t when we estimate the dengue case count in week t, we do not use the true case number.

As we have done in the case of complete data, here we consider models with online data as well.

Table 4.3: Accuracy of all dengue nowcasting models compared to the naive model for the case of incomplete data with training window of $dw = 16$ weeks. To compare the models, we use the relative Mean Absolute Error (relMAE), relative Root Mean Squared Error (relRMSE), relative Mean Absolute Percentage Error (relMAPE) and the relative logarithmic error (relLOG(Q)), as defined in section 3.4. All values are relative to the naive model, and therefore the values of all metrics are 1 for the naive model. We also report the actual non-relative values of the metrics for the naive model in parentheses. For example, under relMAE, we give the true MAE for the naive model in parentheses. Here we also include for comparison the performance of the model previously used for estimating dengue case counts by the InfoDengue system. We find that the naive model is vastly outperformed by all other models. All models other than the naive and InfoDengue models build on the baseline model.

Model	relMAE	relRMSE	relMAPE	relLOG(Q)
Naive	1 (425.0)	1 (959.3)	1 (0.502)	1 (0.733)
InfoDengue	0.689	0.693	0.899	0.481
Baseline	0.553	0.553	0.670	0.420
Google (Dengue)	0.536	0.537	0.690	0.435
Twitter	0.541	0.538	0.719	0.440
Google (Dengue) + Twitter	**0.511**	**0.495**	**0.651**	**0.412**

Table 4.3 shows a comparison of the accuracy of all the models with the naive model in the case of incomplete data. We also report the results of the InfoDengue model, which was previously used for estimating dengue case counts by the InfoDengue system, to make a comparison with our current models. For the incomplete data analysis, we consider several sizes of training windows as well, but we only report the outcome in the case of $dw = 16$ because it shows the highest improvement in accuracy of the *Google (Dengue) + Twitter* model compared to the baseline model. For the same reason, we choose a variable number of weeks cw to calculate the delay correcting function, as defined in (4.6) such that they include 40,000 dengue cases. When we change the training window's size, we see slight variations of the values in Table 4.3 but our findings are qualitatively similar. Specifically, we still generally find that the model using both Google and Twitter data is the best performing

model, with values of the relative errors slightly smaller than those of the other models.

First, we observe that the naive model in the case of incomplete data is much worse than in the case of complete data. While in the case of complete data the MAE for the naive model is 193 cases, in the case of incomplete data it is 425 cases, more than double. This is a consequence of the fact that we are now using only data relating week $t-1$ that are available at the end of week t to estimate dengue case counts for week t. As seen previously in Section 3.1.1, this corresponds to about 50% of data on average, and thus we find bad values for the prediction error of the naive model. Furthermore, Table 4.3 shows that the naive model is vastly outperformed by all other models, regardless of the error metric.

The baseline model exhibits an MAE of 235.0 cases, about 44.7% smaller than the naive model, and all models using online data show higher accuracy than the baseline model, independent of the metric considered. The *Google (Dengue) + Twitter* model, which shows the lowest MAE, is 48.9% more accurate than the naive model, with an MAE of 217.2 cases, and 7.6% more accurate than the baseline model. For this reason, we do not consider the naive model any further in the rest of this analysis. Instead, we calculate the relative metrics relative to the baseline model. For example, this means that when calculating the relative MAE of the *Google (Dengue) + Twitter* model, we divide its MAE by the MAE of the baseline model instead of dividing it by the MAE of the naive model.

The top panel of Figure 4.3 shows a comparison of the predictions of all the different models and the actual data. We can see that the number of cases known at the moment of nowcast relative to the week being nowcasted is tiny compared to the total number of cases in that week that will be known many weeks later. Despite this, even the baseline model that only uses official data can grasp the main features of the time series, while models using online data seem to just marginally improve that fit.

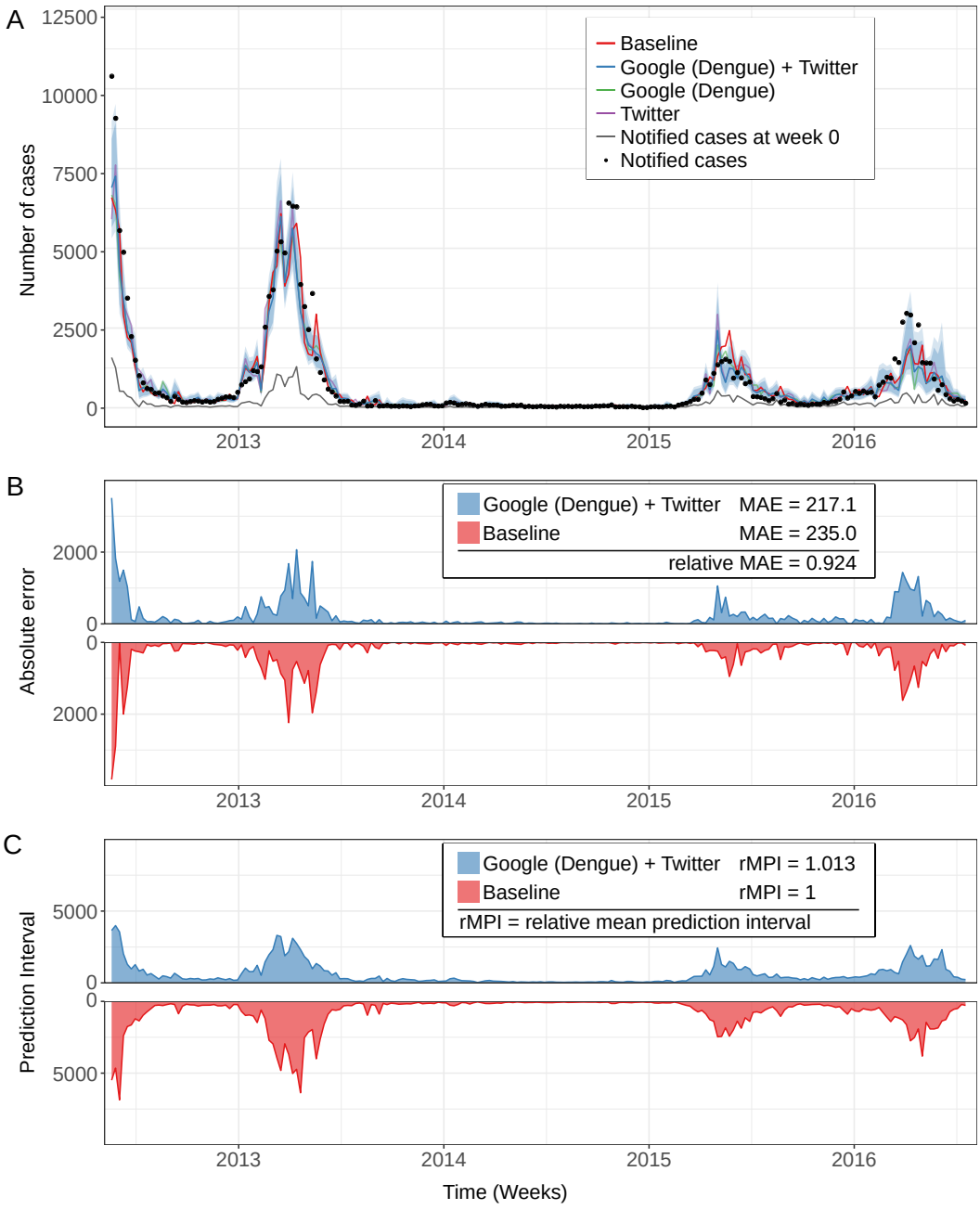

Figure 4.3: Comparison of the different models considered in the case of complete data with training window of $dw = 16$ weeks. (A) We compare the performance of the baseline model with that of models drawing on data from Google and Twitter. In black, we depict official data on the total number of dengue cases. In orange, we depict the total number of dengue cases known to the authorities by the end of each week, which constitute a tiny fraction of the total cases. The blue shaded areas represent the 80% (dark blue) and 95% (light blue) prediction intervals for the *Google (Dengue) + Twitter* model. *(continues on the following page)*

Figure 4.3: *(continues from previous page)* (B) We compare the weekly absolute errors for the baseline and *Google (Dengue) + Twitter* models. (C) We compare the weekly prediction interval width for the baseline and *Google (Dengue) + Twitter* models. We can see that both the absolute error and the prediction interval are generally smaller during the epidemic periods for the *Google(Dengue)+Twitter* with respect to the Baseline model.

Table 4.4: Accuracy of all dengue nowcasting models compared to the baseline model for the case of incomplete data with training window of $dw = 16$ weeks. We redefine all the metrics relative to the baseline model, and therefore now the values of all relative metrics are 1 for the baseline model. We also report the actual values of the metrics for the baseline model in parentheses. For example, under relMAE, we give the true MAE for the baseline model in parentheses. We see that in this case as well the inclusion of online data provides a slight advantage, with the *Google (Dengue) + Twitter* model providing a reduction of the prediction error of up to 10% or less depending on the metric considered. In addition, all the adaptive nowcasting models we consider outperform the model that was previously used for estimating dengue case counts by the InfoDengue system.

Model	relMAE	relRMSE	relMAPE	relLOG(Q)
Baseline	1 (235.0)	1 (530.7)	1 (0.336)	1 (0.307)
InfoDengue	1.245	1.253	1.341	1.147
Google (Dengue)	0.969	0.971	1.030	1.037
Twitter	0.978	0.973	1.072	1.050
Google (Dengue) + Twitter	**0.924**	**0.895**	**0.971**	**0.982**

In Table 4.4 we recalculate the results presented in Table 4.3 relative to the baseline model. We can see that, according to relMAE and relRMSE, the results are slightly more accurate than the baseline model for all the adaptive nowcasting models including online data. In particular, the *Google (Dengue) + Twitter* shows a reduction in MAE of 7.6% and a reduction in MAPE of 2.9% compared to the Baseline model. More generally, the *Google (Dengue) + Twitter* model produces more accurate estimates than the baseline model according to all the metrics we consider, with also a reduction in RMSE of 10.5% and a reduction in LOG(Q) of 1.8%.

Depending on the metrics we choose to consider, for the *Google (Dengue)* and *Twitter* models the prediction errors can either slightly increase or decrease. For example, for the *Google (Dengue)* we observe a reduction in MAE of 3.1% and an increase in MAPE of 3% compared to the baseline model. Also, the changes we observe in Table 4.4 are very small, representing differences of only a few percentage points.

We can also see that all the adaptive nowcasting models that we have analysed perform better than the model previously used for estimating dengue case counts by the InfoDengue system, which shows an MAE of 292.6 cases and an MAPE of 0.45.

In Table 4.5 we again further develop our analysis and consider only data relating to weeks where the number of weekly counts is above the epidemic threshold in Rio de Janeiro. In this case, there seem to be slightly higher reductions across all metrics. For the *Google (Dengue) + Twitter* model we observe a reduction in MAE of 9.7% and a reduction in MAPE of 11.3% compared to the baseline model.

Table 4.5: Accuracy of all dengue nowcasting models compared to the baseline model during epidemic seasons for the case of incomplete data with training window of $dw = 16$ weeks. Here we only consider weeks with number of cases greater than the dengue epidemic threshold for Rio de Janeiro. We redefine all the metrics relative to the baseline model, and therefore now the values of all metrics are 1 for the baseline model. We also report the actual values of the metrics for the baseline model in parentheses. For example, under relMAE, we give the true MAE for the baseline model in parentheses. We see that in this case as well the model *Google (Dengue) + Twitter* is the best performing model across all metrics. In this case, again, all the adaptive nowcasting models we consider outperform the model that was previously used for estimating dengue case counts by the InfoDengue system.

Model	relMAE	relRMSE	relMAPE	relLOG(Q)
Baseline	1 (679.5)	1 (976.5)	1 (0.317)	1 (0.339)
InfoDengue	1.209	1.251	1.344	1.051
Google (Dengue)	0.939	0.969	0.915	0.983
Twitter	0.961	0.973	1.028	1.011
Google (Dengue) + Twitter	**0.907**	**0.896**	**0.887**	**0.928**

Speaking, then, of the advantage of using online data such as Google searches and Twitter posts, we find that using either one of the two data sources alone might not be sufficient to provide a notable improvement. Instead, it is only when we use them together in the same model that we can observe the highest reduction across all metrics we consider.

Until this point, we have only analysed what happens to the prediction error, and we have discovered that we obtain the best results in terms of error reduction when we consider a model that uses together both Google and Twitter data. As we discussed before in Section 3.4, reducing the prediction error is not the only thing that we are

concerned about. It is also vital that we look at the width of prediction intervals. In terms of strategy and allocation of resources, it is essential to understand what the worst and best case scenarios are. We want to be able to assess what the minimum amount of resources is that will certainly be needed, and what the maximum amount of resources is that might be required in a more severe case. Here we focus on the 95% prediction intervals, and in particular we want to check two things: whether or not they change when we use a model with online data with respect to a model without online data; and whether or not their meaning is preserved, or in other words, whether the 95% prediction intervals actually contain 95% of the true points.

Table 4.6: Precision of dengue nowcasting models using *Google* and *Twitter* data compared to the baseline model for the case of incomplete data with training window of $dw = 16$ weeks. We define the mean prediction interval (MPI) as the mean width of the 95% prediction interval for all estimates generated. The MPI for the baseline model is given in parentheses. We define the relative mean prediction interval (relMPI) as the MPI for the model divided by the MPI for the baseline model. The relMPI for the baseline model is therefore 1. We see that in this case the *Google (Dengue)* and *Google (Dengue) + Twitter* have $\simeq 25\%$ smaller relMPIs with respect to the baseline model, while the InfoDengue model's relMPI is nearly double than that of the baseline. However, importantly, we can see that for all the models we introduced in the present chapter the prediction intervals do not seem to be reliable, since they all contain less than 95% of the true points.

Model	relMPI	Percentage points within 95% prediction interval		
		all	> 550	< 550
Baseline	1 (817.1)	84.7	76.2	88.4
InfoDengue	1.891	93.6	95.2	92.9
Google (Dengue)	0.760	80.7	71.4	84.5
Twitter	0.925	85.3	79.4	87.7
Google (Dengue) + Twitter	**0.750**	82.11	73.0	85.8

In Table 4.6 we use the mean prediction interval (MPI), as defined in Section 3.4, and we calculate its value for all models relative to the baseline model (relMPI). We also report the percentage of true data points that fall in the 95% prediction interval while considering the whole time series, or only weeks during epidemic seasons. Table 4.6 highlights a few important facts that were not evident before when we only looked at accuracy. First, when we consider models with online data the prediction intervals shrink. For the baseline model, we observe an MPI of 817.1. At 755.8, the MPI of the *Twitter* model is 7.5% smaller than the baseline model, while at 621.0 it is 24% smaller for the *Google (Dengue)* model compared to the baseline.

We observe the highest reduction for the *Google (Dengue) + Twitter* which has an MPI of 612.8, 25% smaller than the baseline model. Second, and most importantly, the prediction intervals of our models are inadequate to represent the variability of the true data. We observe that the Baseline model is only able to capture 84.7% of the true points within its 95% prediction intervals when considering all weeks in the period of analysis, and only 76% when we consider only weeks with a weekly case count greater than 550, i.e. during epidemics in Rio de Janeiro. Analogously, the 95% prediction interval of the *Google (Dengue) + Twitter* model only contains 82.1% of the true data points when we consider all weeks in the period of analysis, and only 73% of the true data points when we consider only weeks with a weekly case count greater than 550. This is because the correction function we use to make a first estimate of the dengue case counts does not have an uncertainty, and we cannot propagate such uncertainty in the nowcasting model. On the other hand, the InfoDengue model, which was developed by our collaborators at Fiocruz to predict the number of dengue cases based on official data only, seems to have a much bigger mean prediction interval at 1545.1, almost two times wider than the MPI of the baseline model. However, it captures the appropriate number of true points. We observe that the 95% prediction intervals of the InfoDengue model contain 93.6% of the true data points when we consider the entire period of analysis, and 95.2% of the true data points when we only include weeks in epidemic periods. For these reasons, we can conclude that adaptive nowcasting models are not appropriate to estimate dengue cases in Rio de Janeiro, because even if we can obtain reasonable point estimates, the prediction intervals are not reliable for being used in an operational setting.

4.3 Discussion

In this chapter, we have explored how we could use adaptive nowcasting models to estimate the weekly number of dengue cases in Rio de Janeiro. We chose to start our analysis with this approach because auto-regressive models are among the most commonly used methods in the literature for disease surveillance and nowcasting problems similar to ours.

We have seen that in an optimal situation, i.e. when all data relative to previous weeks are available, an adaptive nowcasting model that only uses official data is not more accurate than a naive model, but when Google searches data are used,

the accuracy of the adaptive nowcasting model is higher than that of the naive and baseline models. We reproduced the algorithm used by Preis and Moat (2014) and in particular explored the advantage of using online data such as Google search volume and the number of Twitter posts relative to dengue in Rio de Janeiro. We concluded that in the case of complete data the baseline model is outperformed by the naive model, but the naive model is outperformed by the model also using information from online data, with the best performance obtained using Google search data.

The situation in an operational and real-world case is much more complicated because of the considerable delay in the availability of dengue data in Brazil in Rio de Janeiro. Since delays can be of up to 6 months and the amount of information that one has at the end of any week for which one wishes to predict the number of dengue cases is typically of about 25%, we had to explore different solutions. First, we considered a method to correct the official data, and then we added online data to the model as external regressors. In the latter case, we found that even though we could achieve a $\simeq 10\%$ prediction error reduction when using Google and Twitter data together, the prediction intervals were not reliable and not suitable for the type of analysis we were performing. In fact, for the baseline model and all models using online data, the 95% prediction intervals only contain about 80-85% of the points, and while considering weeks with dengue case count over the epidemic threshold, they contain as few as 70% of the points. This is a problem, mainly because we want the prediction intervals to be as reliable as possible since they are an important piece of information needed by policymakers to make decisions on possible actions.

In conclusion, adaptive nowcasting models do not appear to be suitable for our particular problem in our particular setting. There are two main reasons for that. The first one is that they cannot automatically account for the severe delays in the official data, or in other words, they cannot automatically account for the high variability in the notification rate. This means that when we try to estimate the number of dengue cases using the correction function we described, we cannot calculate the uncertainty, and since we do not have any uncertainty we cannot propagate it into the ARIMA model. For this reason, the prediction intervals that we obtain are too small and are not reliable. Being able to produce such uncertainty and propagate it into the ARIMA model might help solve this problem.

The second reason is that many hyperparameters are not automatically calculated but need to be set manually. These could be fine-tuned to get the best possible numbers for accuracy, but still, they would be very particular to the case of Rio de

Janeiro. Furthermore, a considerable fine-tuning to fit the data from the past would produce models that may not generalise to the future. In principle, we could also decide to empirically expand the prediction intervals by a certain amount to account for the extra points that they cannot capture. Doing so, though, would mean fixing yet another hyperparameter, and making the model even more particular to Rio de Janeiro and to the observed data. In practice, we would be overfitting the model, and we would make that based on information that we obtained after testing the model. This means that it is not an approach we want to follow if we aim to provide an operationally realistic model.

Finally, the third reason is that Rio de Janeiro is a particular case, a large city where there is a large number of dengue cases during epidemics and where the dengue case counts during the rest of the year is always high enough that we almost never see a week with no dengue cases. As we see in Chapter 8, when we consider smaller cities the dengue case count could easily be zero for many weeks in a row, or we could have small numbers in general. Time series containing fewer case counts, where we can easily observe no dengue case in multiple weeks, are more difficult for ARIMA models to handle and this means that most of the time the algorithm ends up using an AR(1) model.

For all these reasons, in the next chapters we consider other approaches. Adaptive nowcasting models are suitable for many similar problems, and could help to give us us a quick first approximation of the quantity that we seek to estimate. Nevertheless, we are looking for more reliable estimates, and for this reason, we move our attention to a different class of models that we describe in the next chapter.

CHAPTER 5

A Bayesian nowcasting model

In the present chapter, we introduce a model to estimate the weekly number of dengue cases in Rio de Janeiro which has some important differences to the adaptive nowcasting model we presented in the previous chapter.

As summarised in the previous chapter, even though the estimates produced with adaptive nowcasting models can be considered accurate, they cannot be considered reliable. The prediction intervals produced by such models are too small and, as a consequence, they do not contain the observed notified dengue cases the appropriate number weeks. In particular, the 95% prediction intervals contain the observed notified dengue cases around 75-80% of the weeks, and this means that they do not represent the likely variation of the notified dengue fever case counts appropriately. Furthermore, quite a large number of hyperparameters need to be chosen. This means that with the adaptive nowcasting model there is also an increased risk of overfitting. In conclusion, the adaptive nowcasting model, in the particular context of nowcasting dengue in Rio de Janeiro, is not general enough to be versatile and easily transferable to other cities.

For such reasons we here consider a different model with the aim of producing an operationally realistic model which is able to automatically deal with the severe delays in our data set. We also seek to have a small number of hyperparameters, and to produce reliable estimates and prediction intervals.

The model we consider here is based on the INLA algorithm, which is described in Section 3.3. The INLA algorithm follows a Bayesian approach, which makes the model easily adaptable to different cities and different patterns. Furthermore, the model we consider automatically takes into account delays in the official data, and

it can produce estimates of the current weekly number of dengue cases based solely on the delayed data.

As in Chapter 4, after describing the model, we consider the advantages of including information coming from online data sources such as Google Trends and Twitter. We find that this produces improvements in terms of accuracy and precision, but most importantly in terms of reliability. Crucially, the model we present here is operationally realistic and can actually be used in practice, and hence can easily be implemented in the InfoDengue system.

5.1 Methods

In this section, we detail the models analysed in the present chapter. The models that we consider all seek to deliver weekly estimates of dengue case counts in Rio de Janeiro. We carry out our analysis using epidemiological weeks, which are defined as starting on a Sunday. Where weeks span two different calendar years, the week belongs to the year in which more days of the week fall. As such, if the calendar year begins on a Monday, a Tuesday or a Wednesday, the epidemiological year is considered to have started on the final Sunday of the previous calendar year. Otherwise, the epidemiological year starts on the first Sunday of the calendar year. Each epidemiological year therefore has either 52 or 53 epidemiological weeks.

We investigate whether rapidly available data on Google searches and tweets relating to dengue or other arboviruses present in Rio de Janeiro can enhance weekly estimates of the number of cases of dengue in Rio de Janeiro reported to doctors in the current week. The timeline of operation of all the models described here is the same as that described in the previous chapter and illustrated in Figure 4.1. Importantly, we carry out these investigations while taking into account the true nature of the delays in dengue case count data described in Section 3.1.3. We therefore compare the following seven models:

Baseline. We first consider a model developed by Bastos et al. (2017) that aims to infer the number of cases of dengue in the current and previous weeks using the delayed dengue case count alone. In simple terms, the model aims to estimate the number of cases of dengue that will be reported for each week with a given number of weeks delay. In other words, the approach explicitly models the

Table 5.1: Here we show a visualisation of the notification data as a matrix where each element is the number of cases that occurred in week t and were notified with a delay of τ weeks. There is a running unknown triangle that must be estimated in order to infer the number of dengue cases for the weeks for which we do not yet have complete data (Bastos et al., 2017).

Week (t)	**Delay (τ)**						**Total**
	0	**1**	**2**	**3**	**4**	**5**	
1	15	12	6	3	1	2	39
2	5	13	3	4	4	1	30
3	14	12	5	3	3	2	39
4	18	11	6	6	3	1	43
5	11	4	5	4	4	1	39
6	11	9	5	5	2	2	34
7	14	13	2	4	1	?	?
8	7	7	2	1	?	?	?
9	16	10	5	?	?	?	?
10	11	9	?	?	?	?	?
11	7	?	?	?	?	?	?

gradual delivery of information relating to dengue cases in a given week over the following weeks.

As described in Section 3.1.1, official data are available to us in the form of a list of dengue cases with a date of notification and a date of system entry. From this list, it is possible to build a table similar to Table 5.1. In this model, each unknown cell is estimated using information from previous lines (previous weeks) and columns (delay structure).

Formally, let $n_{t,\tau}$ be the number of cases that occurred in week t and were reported in week $t+\tau$, thus with delay τ. We assume that $n_{t,\tau}$ follows a negative binomial distribution

$$n_{t,\tau} \sim \mathcal{NB}(\lambda_{t,\tau}, \phi) \tag{5.1}$$

which has the following form

$$P(n_{t,\tau} = k) = \binom{\lambda_{t,\tau} + k - 1}{k} (1-\phi)^{\lambda_{t,\tau}} \phi^k \tag{5.2}$$

where the mean $\lambda_{t,\tau}$ is given by

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau \tag{5.3}$$

μ is a constant and α_t and β_τ are random effects with an auto-regressive structure

$$\begin{aligned} \alpha_t &\sim \alpha_{t-1} + \mathcal{N}(0, \eta_\alpha) \\ \beta_\tau &\sim \beta_{\tau-1} + \mathcal{N}(0, \eta_\beta) \end{aligned} \tag{5.4}$$

Parameters are fit using the INLA framework described in Section 3.3, and values of $n_{t,\tau}$ are estimated using sampling. The total number of cases at week t is then given by

$$n_t = \sum_\tau n_{t,\tau} \tag{5.5}$$

We use the first twenty weeks of data in 2012 for training only, and begin generating estimates in epidemiological week 21 in 2012, which began on Sunday 20th May 2012. The model is fit to the data again every week, using all data available from the start of 2012 until week t. The same approach is used for all of the following models, apart from the naive model.

Google (Dengue). This model is the same as the baseline model, with data on Google searches related to the topic of *dengue* added as an external regressor. The mean $\lambda_{t,\tau}$ is now calculated as

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau + \log(G_t^d) \tag{5.6}$$

where G_t^d is the volume of Google searches related to *dengue* in week t.

Twitter. This model is the same as the baseline model, with data on the volume of tweets that express personal experience of dengue added as an external regressor. The mean $\lambda_{t,\tau}$ is now calculated as

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau + \log(T_t) \tag{5.7}$$

where T_t is the volume of Twitter posts in week t.

Google (Dengue) + Twitter. This model is the same as the baseline model, with data on Google searches related to the topic of *dengue* and the volume of tweets that express personal experience of dengue added as external regressors. The mean $\lambda_{t,\tau}$ is now calculated as

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau + \log(G_t^d) + \log(T_t) \tag{5.8}$$

where G_t^d is the volume of Google searches related to *dengue* and T_t is the volume of Twitter posts in week t.

Google (all diseases). This model is the same as the baseline model, with data on Google searches related to the topics of *dengue*, *Zika* and *chikungunya* added as external regressors. The mean $\lambda_{t,\tau}$ is now calculated as

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau + \log(G_t^d) + \log(G_t^z) + \log(G_t^c) \tag{5.9}$$

where G_t^d, G_t^z and G_t^c are the volumes of Google searches in week t related to *dengue*, *Zika* and *chikungunya*.

Google (all diseases) + Twitter. This model is the same as the baseline model, with data on Google searches related to the topics of *dengue*, *Zika* and *chikungunya* and the volume of tweets that express personal experience of dengue added as external regressors. The mean $\lambda_{t,\tau}$ is now calculated as

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau + \log(G_t^d) + \log(G_t^z) + \log(G_t^c) + \log(T_t) \tag{5.10}$$

where G_t^d, G_t^z and G_t^c are the volumes of Google searches in week t related to *dengue*, *Zika* and *chikungunya*, while T_t is the volume of Twitter posts in week t.

Naive. Following Yang et al. (2017), this model uses the number of dengue cases cases relating to week $t-1$ which are known at the end of week t as the estimate of the number of dengue cases in week t.

5.2 Results

Following Yang et al. (2017), we begin by comparing the accuracy of all models proposed to the accuracy of the naive model. Again, the naive model uses the known case count for the previous week as the estimate for the case count in the current week. To evaluate model accuracy, we calculate the mean absolute error (MAE) for each model. To facilitate comparison of the models, we also calculate the relative MAE (relMAE) for each model. We define the relative MAE as the MAE of a given model divided by the MAE of the naive model. The relative MAE of the naive model is therefore 1. More information on the model evaluation metrics can be found in Section 3.4.

Table 5.2 shows that the naive model is vastly outperformed by all other models. The MAE for all other models is at least 37% smaller than the MAE of the naive model. The best performing model is the *Google (Dengue) + Twitter* model, for which the relative MAE is 0.502. As the performance of the naive model is considerably worse than all other models, we disregard it for further analyses.

Table 5.2: Accuracy of all dengue nowcasting models compared to a naive model. Following Yang et al. (2017), we compare the accuracy of the naive model to all other models. We define the relative mean absolute error (relative MAE) as the MAE of a given model divided by the MAE of the naive model. The relative MAE of the naive model is therefore 1. We find that the naive model is vastly outperformed by all other models. Note that the baseline model is a more advanced model than the naive model, and is explicitly designed to account for the structure of the delays in the dengue case count data (Bastos et al., 2017). All models other than the naive model build on the baseline model. The best performing model is the *Google (Dengue) + Twitter* model (bold), which exhibits an MAE 49.8% smaller than that of the naive model.

Model	MAE	relative MAE
Baseline	267.2	0.629
Google (Dengue)	215.4	0.507
Twitter	223.3	0.525
Google (Dengue) + Twitter	**213.3**	**0.502**
Google (all diseases)	218.8	0.515
Google (all diseases) + Twitter	213.7	0.503
Naive	425.0	1

For the remainder of our analyses, we focus on comparing the models that use Google and Twitter data to the baseline model. We redefine the relative MAE as the MAE of a given model divided by the MAE of the baseline model. The relative MAE of the baseline model is therefore 1.

Table 5.3 shows that all the models enhanced with online data from either Google or Twitter outperform the baseline model. Across the full time period analysed, the baseline model exhibits an MAE of 267.2 cases. The model enhanced with data on tweets relating to dengue exhibits an MAE 16.4% smaller than the baseline model, at 223.3 cases. The model enhanced with data on Google searches relating to dengue exhibits an MAE 19.4% smaller than the baseline model, at 215.4 cases. As was already seen in Table 5.2 however, the best performing model is the *Google (Dengue) + Twitter* model, which draws on data on both Google searches and tweets relating to dengue in tandem. This model exhibits an MAE of 213.3 cases, 20.2% smaller

than that of the baseline model (Figure 5.1B).

Table 5.3: Accuracy of dengue nowcasting models using Google and *Twitter* data compared to the baseline model. We redefine the relative mean absolute error (relative MAE) as the MAE of a given model divided by the MAE of the baseline model. The relative MAE of the baseline model is therefore 1. We find that all the models using online data outperform the baseline model. The best performing model is the *Google (Dengue) + Twitter* model (bold), which exhibits an MAE 20.2% smaller than that of the baseline model.

Model	MAE	relative MAE
Baseline	267.2	1
Google (Dengue)	215.4	0.806
Twitter	223.3	0.836
Google (Dengue) + Twitter	**213.3**	**0.798**
Google (all diseases)	218.8	0.819
Google (all diseases) + Twitter	213.7	0.800

The accuracy of estimates generated by the models which additionally draw on data on Google searches relating to Zika and chikungunya is similar, with the *Google (all diseases) + Twitter* model exhibiting an MAE of 213.7 cases, 20.0% smaller than that of the baseline model. Overall, therefore, it does not appear that integrating this extra Google data relating to other arboviruses present in Rio de Janeiro improves the accuracy of estimates of dengue incidence.

The performance of the models during epidemics is of particular importance. We therefore examine whether the estimates generated by the *Google (Dengue) + Twitter* model are more accurate when considering periods of epidemics alone. Using the Moving Epidemic Method (Vega et al., 2013), we determine the epidemic threshold for Rio de Janeiro to be 550 dengue cases per week. For each week in which the final number of notified dengue cases was above the epidemic threshold, we calculate the absolute error of the estimates generated by the baseline model and the *Google (Dengue) + Twitter* model. We find that during epidemics, the baseline model exhibits an MAE of 774.8 cases. In contrast, the *Google (Dengue) + Twitter* model exhibits an MAE of 596.0 cases, 23.1% lower than the baseline model (Figure 5.2A).

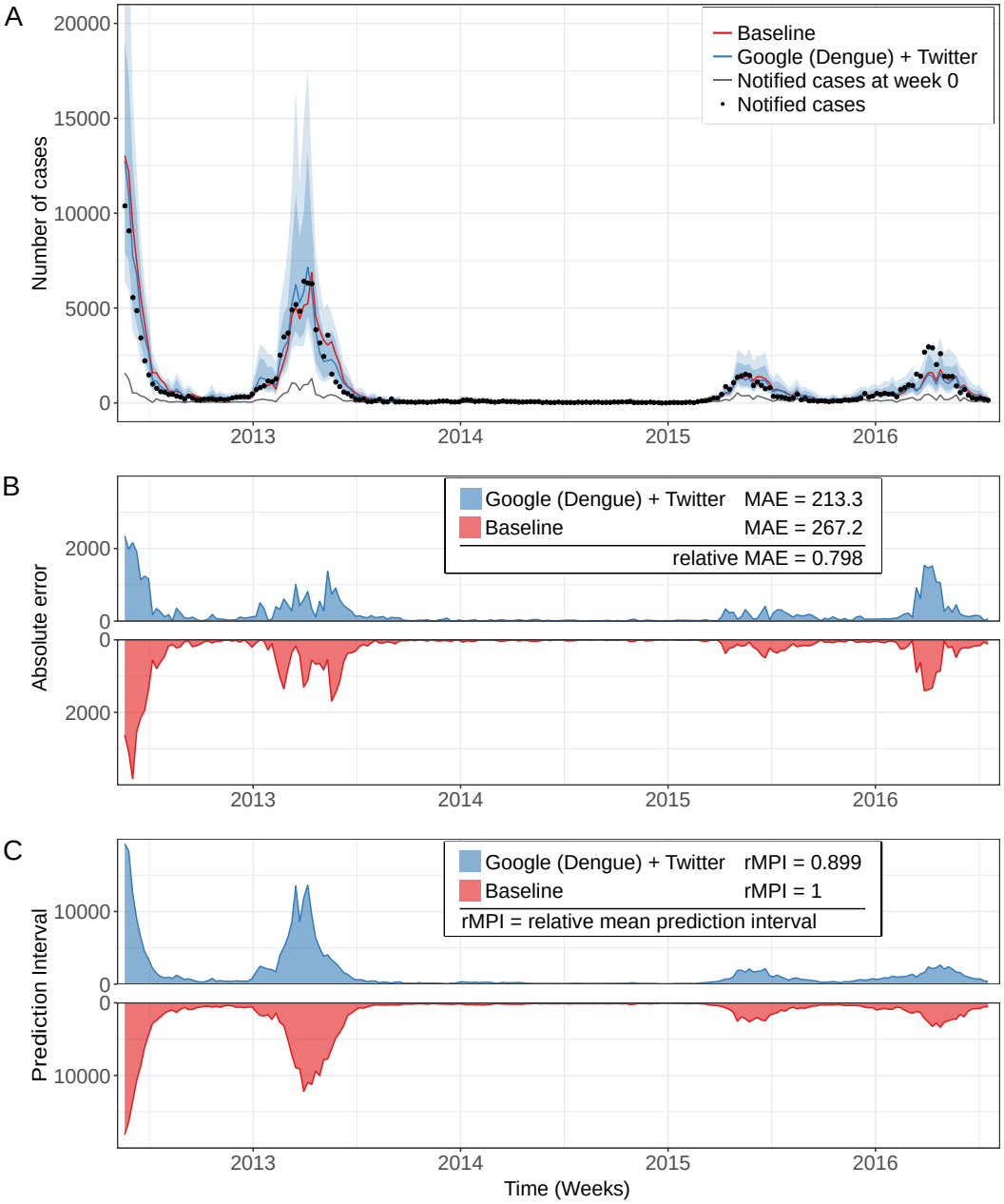

Figure 5.1: Improving accuracy and reducing uncertainty for dengue case count estimates with Google and *Twitter*. (A) We compare the performance of the baseline model with a model drawing on data from Google and Twitter. In black, we depict official data on the total number of dengue cases recorded for each week in Rio de Janeiro, from January 2012 until July 2016. In green, we depict the total number of dengue cases known to the authorities by the end of each week, which constitute a tiny fraction of the total cases. In red, we depict estimates of the number of dengue cases generated by the baseline model for each week at the end of the corresponding week. *(continues on the following page)*

Figure 5.1: *(continues from previous page)* The baseline model uses the official dengue case count data only and was designed to explicitly take into account the nature of the delays in the dengue data (Bastos et al., 2017), going beyond standard auto-regressive approaches. It is clear that this model generally succeeds in capturing the timing and magnitude of the peaks. In blue, we depict estimates of the number of dengue cases generated by the *Google (Dengue) + Twitter* model. It can be seen that the estimates enriched with Google and Twitter are often even closer to the final weekly dengue case count, in particular during the large peaks in case counts in 2012 and 2013. The blue shaded areas represent the 80% (dark blue) and 95% (light blue) prediction intervals for the *Google (Dengue) + Twitter* model. (B) We compare the weekly absolute error for the baseline model and the *Google (Dengue) + Twitter* model. While the mean absolute error (MAE) for the baseline model is 267.2 dengue cases per week, the MAE for the *Google (Dengue) + Twitter* model is lower, at 213.3 dengue cases per week. The *Google (Dengue) + Twitter* model is therefore more accurate. (C) An ideal model for estimating dengue case counts would produce accurate estimates with low uncertainty. To evaluate the level of uncertainty in the estimates produced by each model, we examine the relative mean prediction interval (relMPI) for each model. We define the mean prediction interval (MPI) as the mean width of the 95% prediction interval for the full period for which estimates are generated. We define the relMPI as the MPI for the model divided by the MPI for the baseline model. The relMPI for the baseline model is therefore 1, whereas the relMPI for the *Google (Dengue) + Twitter* model is lower at 0.899. The *Google (Dengue) + Twitter* model therefore also generates more precise estimates.

The inclusion of extra parameters in a model, such as data on Google searches or tweets, increases the likelihood of overfitting. While the analyses detailed so far have considered estimates generated out-of-sample, thereby guarding against this danger, we also calculate the Watanabe-Akaike Information Criterion (WAIC) model quality metric for each of our six models. The WAIC rewards goodness of fit while penalising models for the inclusion of extra parameters. As the model is fit each week when new data arrives, we calculate a WAIC value for each of the six models for every week.

Figure 5.2B depicts the weekly WAIC values for all six models, relative to the baseline model. A lower WAIC value indicates a higher quality model. We find that models enhanced by online data generally exhibit lower WAIC values than the baseline model. In most weeks, the lowest WAIC is again obtained by the *Google (Dengue) + Twitter* model, which draws on data on both Google searches and tweets relating to dengue in tandem.

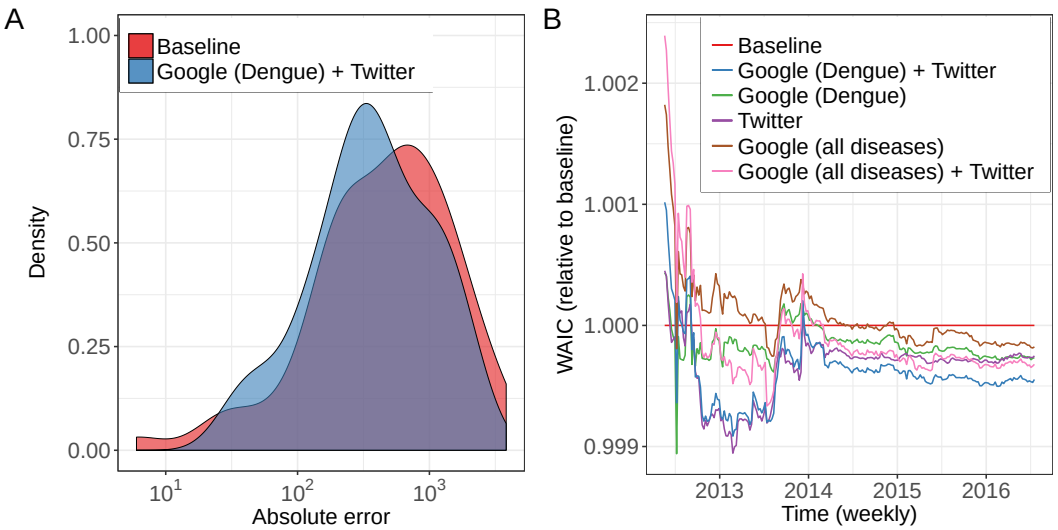

Figure 5.2: Further analyses of the quality of dengue nowcasting models including Google and *Twitter* data. (A) The performance of the models during epidemics is of particular importance. Using the Moving Epidemic Method (Vega et al., 2013), we determine the epidemic threshold for Rio de Janeiro to be 550 dengue cases per week. For each week in which the final number of notified dengue cases was above the epidemic threshold, we determine the absolute error of the estimates generated by the baseline model and the *Google (Dengue) + Twitter* model and plot the distribution using a kernel density estimate. We find that the *mean absolute error* (MAE) for the *Google (Dengue) + Twitter* model (596.0 dengue cases per week; blue) is again considerably lower than the MAE for the baseline model (774.8 dengue cases per week; red). (B) In addition to evaluating the accuracy and precision of out-of-sample estimates generated by the models, here we examine a further metric of model quality, the Watanabe-Akaike information criterion (WAIC). The WAIC rewards goodness of fit but explicitly penalises models for the presence of additional parameters, such as data on Google searches or tweets. We evaluate the quality of all six models explored in our main analysis: the baseline model (red), the *Google (Dengue)* model (green), the *Twitter* model (purple), the *Google (Dengue) + Twitter* model (blue), the *Google (all diseases)* model (orange) and the *Google (all diseases) + Twitter* model (pink). As the model is fit each week when new data arrives, we calculate a WAIC value for each of the six models for every week. To facilitate comparison of these weekly WAIC values, for each week we normalise the six WAIC values by the WAIC for the baseline model. The resulting value for the baseline model is therefore always 1 (red line). A lower WAIC indicates a higher quality model. It can be observed that the models enhanced by online data generally exhibit lower WAIC values than the baseline model. We note that, again, the *Google (Dengue) + Twitter* model (blue) performs particularly well.

An ideal model for estimating current dengue case counts would not only produce accurate estimates but would also produce precise estimates, where uncertainty

about the true value is low. We therefore examine whether dengue nowcasting models enhanced by online data generate estimates that are more precise, as well as more accurate. To evaluate the precision of estimates produced by each model, we calculate the mean prediction interval (MPI), the mean width of the 95% prediction interval for all estimates generated. To facilitate comparison to the performance of the baseline model, we also calculate the relative MPI (relMPI), which we define as the MPI for a given model divided by the MPI for the baseline model. The relative MPI for the baseline model is therefore 1.

Table 5.4 shows that the relMPI for all models enhanced by online data is lower than 1. This indicates that the estimates generated by the models enhanced by online data are more precise than those generated by the baseline model. The *Twitter* model is the most precise model, exhibiting an MPI which is 11.1% lower than the MPI of the baseline model. The *Google (Dengue)* model, drawing on data on Google searches relating to dengue, achieves a smaller but still notable improvement of 8.8%. The combined *Google (Dengue) + Twitter* model, which produced the most accurate estimates, generates the second most precise estimates, with an MPI 10.1% lower than the MPI of the baseline model (Figure 5.1C).

The precision of estimates generated by models which additionally draw on data on Google searches relating to Zika and chikungunya is again similar, with the *Google (all diseases) + Twitter* model exhibiting an MPI 9.9% lower than the MPI of the baseline model. It therefore does not appear that integrating this extra Google data relating to other arboviruses present in Rio de Janeiro improves the precision of estimates of dengue incidence.

We verify whether the 95% prediction intervals continue to reliably represent the range within which 95% of true data points fall. Table 5.4 demonstrates that whether considering all weeks, weeks with more than 550 cases (i.e., during epidemics) or weeks with fewer than 550 cases (i.e., outside epidemics), the 95% prediction intervals appear to behave as desired. In other words, this 10% improvement in the precision of estimates does not come at the cost of the reliability of the prediction intervals.

The characteristics of the dengue season in Rio de Janeiro vary from year to year. In some years, over 5 000 cases a week are reported at the height of the season, whereas in other years, the case count is much lower (Figure 3.1). Previous research has aslo highlighted that the relationship between online data and case counts may

Table 5.4: Precision of dengue nowcasting models using Google and *Twitter* data compared to the baseline model. We define the mean prediction interval (MPI) as the mean width of the 95% prediction interval for all estimates generated. The MPI for the baseline model is given in parentheses. We define the relative mean prediction interval (relMPI) as the MPI for the model divided by the MPI for the baseline model. The relMPI for the baseline model is therefore 1. We find that models using online data generate more precise estimates, reflected by lower relMPIs. The most precise model is the *Twitter* model (bold), followed by the *Google (Dengue) + Twitter* model. We also verify that the 95% prediction intervals reliably represent the range within which 95% of true data points fall. We find that whether considering all weeks, weeks with more than 550 cases (i.e., during epidemics) or weeks with fewer than 550 cases (i.e., outside epidemics), the 95% prediction intervals appear to behave as desired.

Model	relative Mean Prediction Interval	Percentage points within 95% prediction interval		
		all	> 550	< 550
Baseline	1 (1554.6)	95.0	93.7	95.5
Google (Dengue)	0.912	94.5	93.7	94.8
Twitter	**0.889**	95.4	96.9	94.8
Google (Dengue) + Twitter	0.899	94.5	95.3	94.2
Google (all diseases)	0.938	95.4	96.9	94.8
Google (all diseases) + Twitter	0.901	95.4	95.3	95.5

vary across time (Preis and Moat, 2014). We therefore investigate whether the use of online data helps deliver more accurate estimates of dengue incidence in Rio de Janeiro in each of the years covered in our analysis.

In Table 5.5, we report the relative MAE for each model for each year of analysis. We note that statistics for 2012 and 2016 are based on incomplete years, as the analyses begin in Week 21 of 2012 and end in Week 29 of 2016. We find that in 2012, 2013, 2014 and 2015, the accuracy of all models using online data is higher than the accuracy of the baseline model. Using the *Google (Dengue) + Twitter* model, the MAE is reduced by between 11% and 32%.

In 2016 however, we find that the baseline model delivers the most accurate estimates and that the MAE of estimates generated by the *Google (Dengue) + Twitter* model is 8% higher. At the same time, however, we note that the MAE for the baseline model in 2016 (369.1 cases per week) is relatively high given the size of the peak. For example, the MAE for the baseline model in 2013 was similar at 354.3 cases per week, but the peak number of dengue cases per week in 2013 was 6430 in comparison

Table 5.5: Evaluating the accuracy of dengue nowcasting models using Google and *Twitter* across different years. For each year, we define the relative mean absolute error (relative MAE) as the MAE of a given model divided by the MAE of the baseline model. The MAE is given in parentheses. In bold, we highlight the lowest relative MAE for each year. We find that in 2012, 2013, 2014 and 2015, the accuracy of all models using online data is greater than the accuracy of the baseline model. In 2016, we find that the baseline model delivers the most accurate estimates. However, Figure 5.1A shows that in 2016, the performance of the baseline model itself is notably worse than in previous years. We discuss the particular circumstances of 2016 in more detail in the text.

	Relative mean absolute error				
Model	2012	2013	2014	2015	2016
Baseline	1 (678.4)	1 (354.3)	1 (19.0)	1 (123.0)	**1 (369.1)**
Google (Dengue)	0.69	0.76	0.98	0.93	1.03
Twitter	0.74	0.78	0.91	0.96	1.03
Google (Dengue) + Twitter	0.68	**0.74**	**0.86**	**0.89**	1.08
Google (all diseases)	0.73	0.77	0.96	0.90	1.02
Google (all diseases) + Twitter	**0.65**	0.80	0.87	0.92	1.02

to a peak of 2973 cases per week in 2016. This diminished performance in 2016 can also be seen in Figure 5.1A.

Why might we observe differing results for 2016 in comparison to earlier years? A potential answer to this question can be found by examining the nature of the delays in the entry of dengue cases into the surveillance system around this period. Figure 5.3 illustrates that from January 2012 to May 2015, there was a mean delay of 4.9 weeks until 80% of dengue cases for a given week were entered into the surveillance system, with a standard deviation of 1.5 weeks. From June 2015 to December 2015 however, delays were notably reduced such that there was a mean delay of 2 weeks until 80% of dengue cases for a given week were entered into the surveillance system. From January 2016 to the end of the dataset in July 2016, the delays increased again to a mean of 4.6 weeks until 80% of dengue cases for a given week were entered into the surveillance system. This abnormally large variation in delays may have made it particularly difficult for the baseline model to correctly model the delay structure, leading to a higher baseline MAE for 2016.

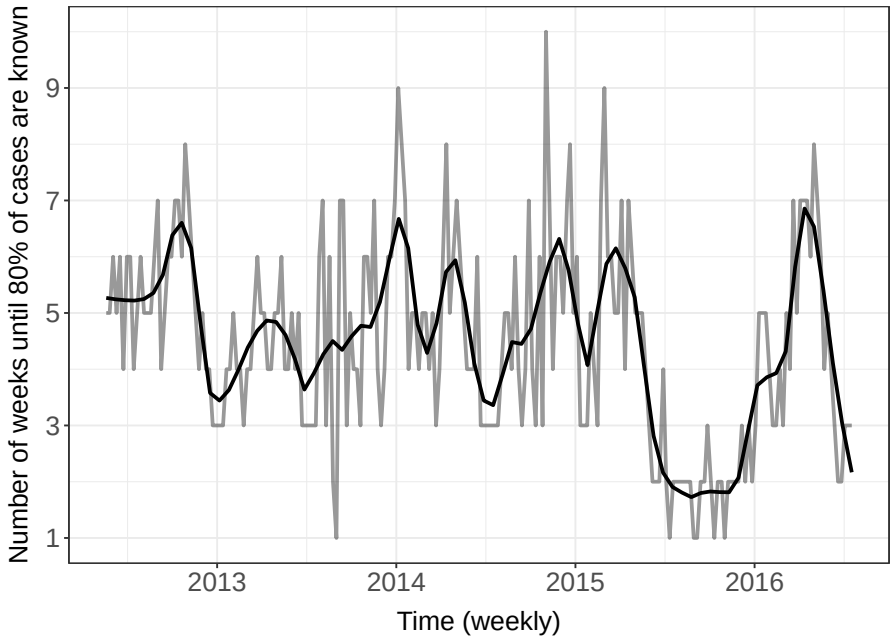

Figure 5.3: Abnormal variation in delays in the recording of dengue cases before the 2016 season. Here we depict in grey the number of weeks until 80% of cases are known, while the black line is a moving average of the grey line. In 2016, we observe diminished performance of the baseline model in comparison to earlier years (Figure 5.1A). To try to explain this finding, we investigate whether there were any changes in the nature of the delays in the entry of dengue cases into the surveillance system around this period. Here we see that from January 2012 to May 2015, there was a mean delay of 4.9 weeks until 80% of dengue cases for a given week were entered into the surveillance system, with a standard deviation of 1.7 weeks. From June 2015 to December 2015 however, delays were notably reduced, such that there was a mean delay of 2 weeks until 80% of dengue cases for a given week were entered into the surveillance system. From January 2016 to the end of the dataset in July 2016, the delays increased again to a mean of 4.6 weeks until 80% of dengue cases for a given week were entered into the surveillance system. This abnormally large variation in delays may have made it particularly difficult for the baseline model to correctly model the delay structure, leading to worse performance in 2016.

It is also worth noting that there was a Zika outbreak in Brazil during the 2016 dengue season. Zika is not only spread by the same mosquito as dengue but also shares some symptoms. Difficulty in discerning the symptoms of dengue from the symptoms of Zika before a laboratory analysis has taken place will have led to some cases of dengue being recorded as suspected cases of Zika, and vice versa. The Zika outbreak was also covered widely in the media, and it is possible that people with dengue may have searched for information relating to Zika instead.

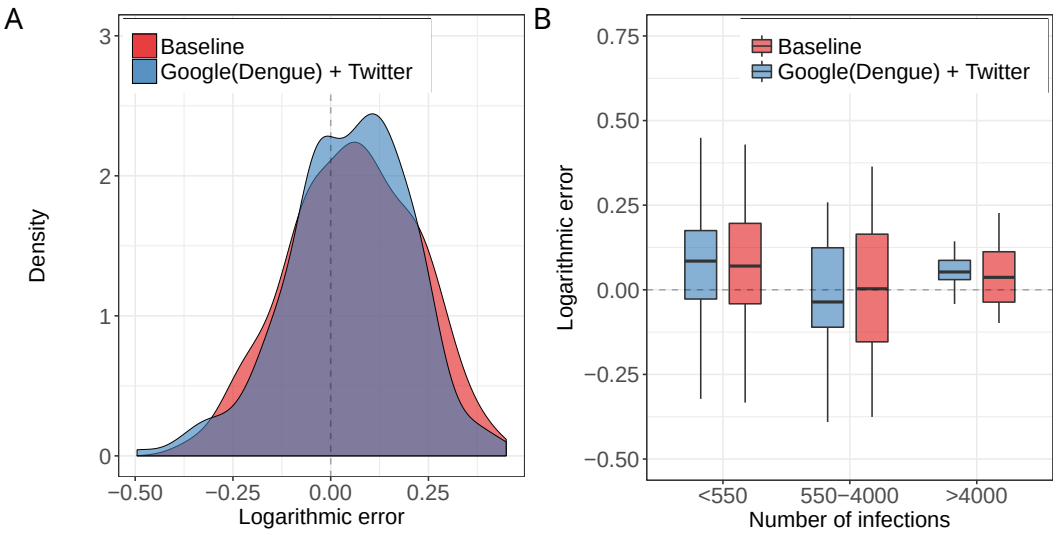

Figure 5.4: Accuracy of the baseline and *Google (Dengue) + Twitter* models according to the logarithmic error metric. (A) In evaluating the accuracy of our models, we also consider an alternative error metric, the *logarithmic error*. Unlike the mean absolute error (MAE) metric that we use in our main analysis, the logarithmic error metric takes into account whether an error of a given number of cases occurred when the true number of cases was very high or very low. The logarithmic error is defined as $\log_{10} Q$ where $Q = \hat{y}/y$, $\hat{y}$ is the predicted value and y the true value. Both the baseline and the *Google (Dengue) + Twitter* models show a clear tendency to overestimate rather than underestimate the case counts. Errors generated by the *Google (Dengue) + Twitter* model are slightly lower overall, and therefore more concentrated around 0. The error distribution is plotted using a kernel density estimate. (B) We evaluate how the error distribution differs in periods of epidemics and outside such periods. Using the Moving Epidemic Method (Vega et al., 2013), we determine the epidemic threshold for Rio de Janeiro to be 550 dengue cases per week. We also investigate how the error distribution changes when dengue case counts are particularly high. Here, we use a threshold of 4 000 dengue cases per week. We find that below the epidemic threshold, both models generally overestimate the number of dengue cases. Above the epidemic threshold, other than in periods when dengue case counts are particularly high, we find that both models tend to slightly underestimate the number of dengue cases. When case counts are higher than 4 000 a week, the models tend to slightly overestimate the number of dengue cases again, but error rates are relatively low in the context of the true dengue case counts. In all three scenarios, errors generated by the *Google (Dengue) + Twitter* model tend to be lower than errors produced by the baseline model.

Figure 3.1D shows that there was a surge in searches relating to Zika in 2016, and Figure 3.1E shows that a similar surge occurred for searches relating to a further arbovirus present in Rio de Janeiro, chikungunya. Indeed, Table 5.5 shows that for

2016, the best performing models using online data are the *Google (all diseases)* model and the *Google (all diseases) + Twitter* model, both of which additionally draw on data on Google searches relating to Zika and chikungunya. However, both models still generate estimates with errors which were 2% greater than the errors generated by the baseline model. We return to this point in the discussion.

The time series of dengue case count are characterised by a sequence of peaks and troughs. The vast differences in case counts at different points in the time series can pose challenges for the evaluation of models that seek to estimate these case counts (Hyndman and Koehler, 2006; Reich et al., 2016b).

In our main analysis, we use the MAE metric to evaluate the performance of our model. This error metric is easy to interpret, as it is measured in numbers of dengue cases. For example, across the full time period analysed in this paper, the baseline model exhibits a mean absolute error of 267.2 dengue cases per week. However, in evaluating the performance of a model, it might be desirable to consider whether an error of a given number of cases occurred when the true number of cases was very high or very low. The mean absolute error does not behave like this and allocates an error of a given number of cases the same weight at a peak and at a trough. For this reason, we also consider an alternative error metric, the logarithmic error (LOG(Q)) that was presented in 3.4. Unlike the mean absolute error, the logarithmic error is not scale-dependent: that is, it is not defined in the units of the underlying time series, and the metric takes into account the size of the corresponding true value.

The logarithmic error is defined as $\log Q$ where $Q = \hat{y}/y$, $\hat{y}$ is the predicted value and y the true value (Tofallis, 2015). Both the baseline and the *Google (Dengue) + Twitter* models show a clear tendency to overestimate rather than underestimate the case counts (Figure 5.4A). Errors generated by the *Google (Dengue) + Twitter* model are slightly lower overall, and therefore more concentrated around 0.

We further evaluate how the error distribution differs in periods of epidemics and outside such periods. Figure 5.4B shows how the error distributions vary for the two models in these three periods: periods outside epidemics, when weekly case counts are below the epidemic threshold; and two classes of periods during epidemics, firstly when weekly case counts are below 4 000, and secondly when weekly case counts are particularly high and above 4 000. We find that below the epidemic threshold, both models generally overestimate the number of dengue cases.

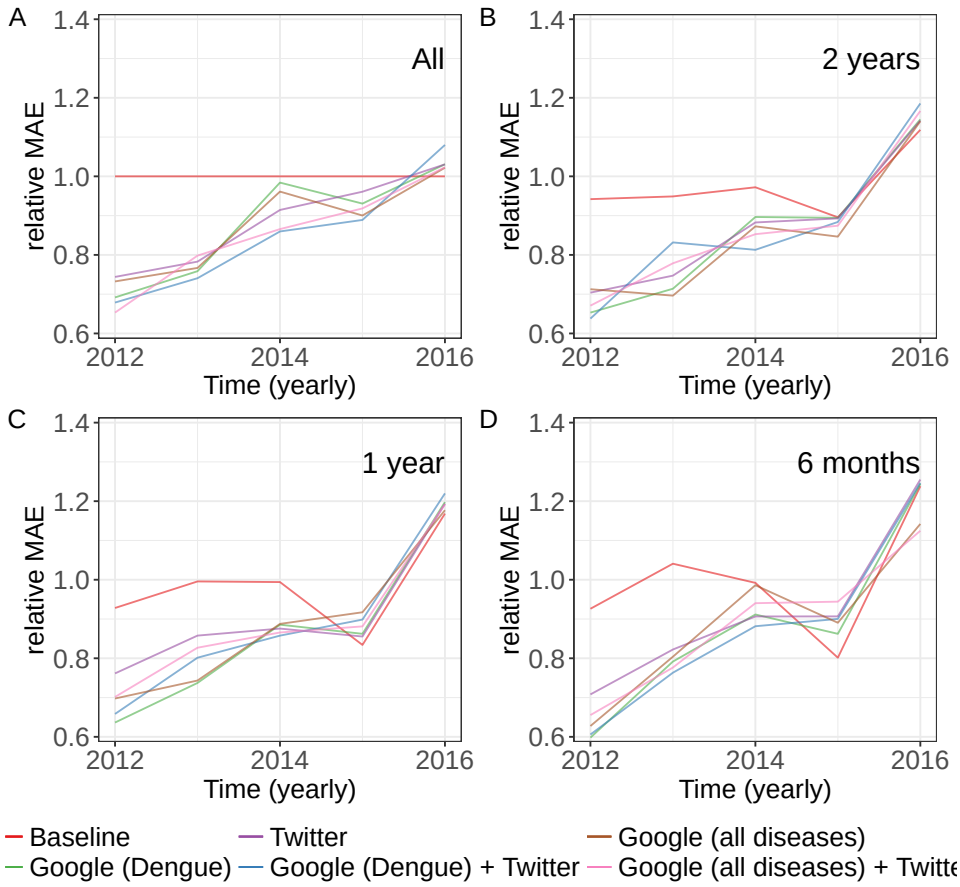

Figure 5.5: Exploring the benefits of using a sliding training window. We consider the performance of all six models explored in our main analysis: the baseline model (red), the *Google (Dengue)* model (green), the *Twitter* model (purple), the *Google (Dengue) + Twitter* model (blue), the *Google (all diseases)* model (orange) and the *Google (all diseases) + Twitter* model (pink). To evaluate performance, we calculate the relative mean absolute error (relative MAE), which we define as the mean absolute error (MAE) of a given model divided by the MAE of the baseline model when trained on all data available from the beginning of the time series to the current week. We assess the relative MAE of each model for each year from 2012 to 2016 (where data for 2016 is partial, ending in mid-July). (A) Yearly performance of all six models when trained on all data available from the beginning of the time series to the current week. As the baseline model is our reference model, the relative MAE for the baseline model is 1 for each year. (B) Yearly performance of all six models when using a sliding training window of two years. In other words, in each week, the model is trained on the most recent two years of data. (C) Yearly performance of all six models when using a sliding training window of one year. (D) Yearly performance of all six models when using a sliding training window of six months. Overall, we find little evidence of any consistent benefit from using only recent data to train the model.

Above the epidemic threshold, other than in periods when dengue case counts are particularly high, we find that both models tend to slightly underestimate the number of dengue cases. When case counts are higher than 4 000 a week, the models tend to slightly overestimate the number of dengue cases again, but error rates are relatively low in the context of the true dengue case counts. In all three scenarios, errors generated by the *Google (Dengue) + Twitter* model tend to be lower than errors produced by the baseline model.

The parameters for the models we have described so far are fit using all data available from the beginning of the time series in 2012 to the current week. This means that as each week passes, the volume of data on which the model is trained grows. In Figure 5.5, we explore whether there is any benefit to only considering recent data when fitting the model, building on a previous study into the relationship between online data and influenza incidence (Preis and Moat, 2014). In the end, we find little evidence of any consistent benefit from using only recent data to train the model.

5.3 Discussion

In this chapter, we have explored how we could build on the model introduced by Bastos et al. (2017) to estimate the weekly number of dengue cases in Rio de Janeiro, and how we could enhance its predictions using online data from Google and Twitter.

A correlation between the number of dengue cases and relevant Google searches as well as Twitter posts is not sufficient to generate more timely and more precise estimates because official data are delayed. Online data are correlated with the observed notified dengue case counts, but while we have full online data when we make a prediction, official data from previous weeks might be partially available or might become available at a later date. Recently Yang et al. (2017) have highlighted that taking delays in the official data in consideration is crucial from an operational point of view to achieve reliable predictions. They built their algorithm under the assumption that, whenever a prediction is made, official data are completely available only up to the previous data point, the previous month in their particular case. Under this assumption, their model can produce accurate predictions of the current data point, as we have also shown in Chapter 4 for adaptive nowcasting models.

Unfortunately, while this approximation might apply to many locations, it does not hold everywhere. In particular, it does not hold in Rio de Janeiro, as we have seen in Figure 3.2 and in many other Brazilian cities as we show later in Chapter 8. In fact, in Rio de Janeiro, only partial data are available relating to several weeks in the past. Nearly full data are available only for weeks more than two months in the past. Again, as we have seen in Chapter 4, it is difficult to properly take delays into account using adaptive nowcasting models, and if we do that we can make estimates that, even if accurate, are not reliable in terms of prediction intervals.

It is thus necessary to adopt a different strategy to make use of these data. The model introduced by Bastos et al. (2017) deals precisely with this aspect, and that is why we chose it as the starting point to build our model. Furthermore, Google and Twitter are very different types of data sources, and previous algorithms have not yet used them together. In line with our analyses in Chapter 4 we integrate both into our model and demonstrate that this leads to the best overall performance when compared to models that use either only one of them or none at all.

We found that data from Google Trends and Twitter, describing the volume of dengue-related searches in a given week and the number of tweets expressing personal experience of dengue, can be used to improve estimates of the current number of notified dengue infections compared to the same estimates generated using official data alone. This improvement consists of a reduction of the MAE of the prediction in models using online data compared to the baseline model by between 16% and 21% over the entire time period considered, depending on the particular type of online data source. The MAE's reduction varies however across years. In years with a very high volume of cases, we find the MAE's reduction of models using online data compared to the baseline model to be as high as 35%. However, we find that in 2016 the baseline model delivers the most accurate estimates and that the MAE of estimates generated by the *Google (Dengue) + Twitter* model is 8% higher. There are multiple potential explanations for this finding. Figure 5.3 illustrates that there is a large variation in delays of recorded cases between the secon half of 2015 and the first half of 2016. This abnormally large variation may have made particularly difficult for the baseline model to correctly model the delay structure. On the other hand, during 2016 there was a Zika outbreak in Brazil. Zika and dengue share not only the carrier, but also some symptoms, and difficulty in discerning the symptoms of Zika from the symptoms of dengue might have been the cause of this increased number of recorded dengue cases in 2016.

Having more accurate estimates is not the only way to improve predictions. Another crucial factor is uncertainty about these estimates. Reliable prediction intervals would allow policymakers to allocate the right amount of resources to prevent or deal with an outbreak. The smaller the uncertainty is, the smaller the risk of wasting resources. Here, we have shown that another advantage of the inclusion of online data in the model is the reduction in size of the 95% prediction intervals by about 10%, while the prediction intervals still contain 95% of the true data points.

Furthermore, the model that we propose could easily be transferred to other cities and used to monitor different diseases too. Finally, the fact that online data are available more often than on a weekly basis lays the foundation for potentially generating more frequent estimates or even short-term predictions. We will come back to these points in later chapters.

CHAPTER 6

Delayed delivery of official data

The Bayesian nowcasting model presented in Chapter 5 explicitly addresses the presence of severely delayed and missing data. However, it still makes some assumptions about the amount of data available. Specifically, we assume that at the end of any given week we receive a delivery of information about cases that have been entered into the system that week. However, on occasion, case counts may be not delivered for a given particular week, or at least not delivered on time to make the estimate. Instead, for example, they might become available at the end of the following week, together with the data entered in the system during that week. We address this operational issue by removing the assumption that data are delivered every week on time.

Here we investigate how our Bayesian nowcasting model behaves in cases in which the data are not delivered. We do so by assuming that this can happen with a certain probability $p > 0$. We study how the performance of our model is affected by the fact that data are not delivered on time, and we explore the effect of varying the probability that this could happen.

When we consider a baseline model using official data only in a week in which such data were not delivered, making a prediction about the current week effectively constitutes forecasting. This is because we have only past data, up to the previous week, and we do not have current data, i.e. data relating to the current week. Thus, the material presented in the present chapter lays the foundation for a baseline forecasting model. While the current chapter focuses on the challenge of nowcasting in the face of a delayed data delivery, we explore the challenge of true forecasts in more detail in Chapter 7.

6.1 Methods

For the analysis we make in this chapter, we need to conider a variation of the baseline algorithm presented in Chapter 5 where data entered into the system in the current week are on occasion not available. The timeline of operation of all the models described here is the same as that described in the previous chapters and illustrated in Figure 4.1, where week t is the week for which we want to provide an estimate. We also consider models that use Google search volumes and the number of Twitter posts as external regressors in the same way we do in Chapter 5, with appropriate modifications to the underlying baseline model to account for the possibility of a missing data delivery. Here, we describe the modified models we use in the this chapter:

Baseline. We assume that data are not delivered at the end of week t with probability p. With reference to Figure 4.1, when data are not delivered in week t, we have official data only up to week $t-1$. Instead, since online data are not prone to this unavailability problem, they are available up to week t.

When data are not delivered, we run a slightly modified algorithm on our data, one that estimates the number of dengue cases in the current week using only data entered into the system up to the previous week. In these cases, the notification data matrix shown in Table 5.1 looks like that presented in Table 6.1b, but in this case there are no known data in the last row, the one corresponding to the current week.

The baseline model therefore estimates more unknown cells for every week for which it does not have complete data. In particular, if we are considering $m = 5$ delays in our matrix, there are m less cells available to train parameters and m more cells to estimate. This difference does not impact the complexity of the model or the computational time needed to train the model and make an estimate.

The fact that we are estimating m more cells compared to the model described in Chapter 5 means that we expect the prediction intervals to be larger, but also, of course, the point estimates to be less accurate. The higher the probability that data are not being delivered, the less accurate we can expect the estimates to be, and the wider we can expect the prediction intervals to be.

The rest of the baseline model works exactly as described in Chapter 5. We repeat the definition here for ease of reference.

Table 6.1: Visualization of the notification data as a matrix where each element is the number of cases that occurred at time t and were notified with delay τ. (a) We repeat the original Table 5.1 for ease of reference. (b) In this case, at the end of week 11 data are not delivered. Thus, compared to , here there is a bigger running unknown triangle that must be estimated in order to infer the true notification curve.

(a) Baseline model

Week	**Delay**						**Tot**
	0	**1**	**2**	**3**	**4**	**5**	
1	15	12	6	3	1	2	39
2	5	13	3	4	4	1	30
3	14	12	5	3	3	2	39
4	18	11	6	6	3	1	43
5	11	4	5	4	4	1	39
6	11	9	5	5	2	2	34
7	14	13	2	4	1	?	?
8	7	7	2	1	?	?	?
9	16	10	5	?	?	?	?
10	11	9	?	?	?	?	?
11	7	?	?	?	?	?	?

(b) Modified baseline model

Week	**Delay**						**Tot**
	0	**1**	**2**	**3**	**4**	**5**	
1	15	12	6	3	1	2	39
2	5	13	3	4	4	1	30
3	14	12	5	3	3	2	39
4	18	11	6	6	3	1	43
5	11	4	5	4	4	1	39
6	11	9	5	5	2	**?**	?
7	14	13	2	4	**?**	?	?
8	7	7	2	**?**	?	?	?
9	16	10	**?**	?	?	?	?
10	11	**?**	?	?	?	?	?
11	**?**	?	?	?	?	?	?

Let $n_{t,\tau}$ be the number of cases that occurred in week t and were reported in week $t+\tau$, thus with delay τ. We assume that $n_{t,\tau}$ follows a negative binomial distribution

$$n_{t,\tau} \sim \mathcal{NB}(\lambda_{t,\tau}, \phi) \tag{5.1}$$

which has the following form

$$P(n_{t,\tau} = k) = \binom{\lambda_{t,\tau} + k - 1}{k} (1-\phi)^{\lambda_{t,\tau}} \phi^k \tag{5.2}$$

where the mean $\lambda_{t,\tau}$ is given by

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau \tag{5.3}$$

μ is a constant and α_t and β_τ are random effects with an auto-regressive structure

$$\begin{aligned} \alpha_t &\sim \alpha_{t-1} + \mathcal{N}(0, \eta_\alpha) \\ \beta_\tau &\sim \beta_{\tau-1} + \mathcal{N}(0, \eta_\beta) \end{aligned} \tag{5.4}$$

Parameters are fit using the INLA framework described in Section 3.3, and values of $n_{t,\tau}$ are estimated using sampling. The total number of cases at week t is then given by

$$n_t = \sum_{\tau} n_{t,\tau} \tag{5.5}$$

We use the first twenty weeks of data in 2012 for training only, and begin generating estimates in epidemiological week 21 in 2012, which began on Sunday 20th May 2012. The model is fit to the data again every week, using all data available from the start of 2012 until week t. The same approach is used for all of the following models.

Google (Dengue). This model is the same as the modified baseline model, with data on Google searches related to the topic of *dengue* added as an external regressor. The mean $\lambda_{t,\tau}$ is now calculated as

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau + \log(G_t^d) \tag{5.6}$$

where G_t^d is the volume of Google searches related to *dengue* in week t.

Twitter. This model is the same as the modified baseline model, with data on the volume of tweets that express personal experience of dengue added as an external regressor. The mean $\lambda_{t,\tau}$ is now calculated as

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau + \log(T_t) \tag{5.7}$$

where T_t is the volume of Twitter posts in week t.

Google (Dengue) + Twitter. This model is the same as the modified baseline model, with data on Google searches related to the topic of *dengue* and the volume of tweets that express personal experience of dengue added as external regressors. The mean $\lambda_{t,\tau}$ is now calculated as

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau + \log(G_t^d) + \log(T_t) \tag{5.8}$$

where G_t^d is the volume of Google searches related to *dengue* and T_t is the volume of Twitter posts in week t.

In this chapter we do not consider models using online data relating to Zika and chikungunya as the previous chapter showed that their performance was generally inferior to that of models using online data about dengue.

6.2 Results

As in Chapter 5, all the models are run from epidemiological week 21 of 2012 until epidemiological week 29 of 2016. The first 20 epidemiological weeks are used to fit the initial parameters of the model.

We run the model for different values of the probability p of data not being delivered on time. For every probability p we need to run the model multiple times, it is not enough to run the model only once.

Let us consider the baseline model for simplicity. Every week we generate a random number: with probability p we run the modified baseline model described in this chapter – data have not been delivered on time – while with probability $1 - p$ we run the baseline model described in Chapter 5 – data have been received correctly.

Following this algorithm every week, if we only ran the model once, we might pick an unfortunate sequence of random numbers where most of the weeks when we run the modified baseline model – the one where data have not been delivered on time – belong to epidemic seasons. This would make our predictions much worse, because prediction errors are generally higher during epidemics and prediction intervals are generally wider.

Conversely, we might pick, by chance, a lucky run where most of weeks when we run the modified baseline model are outside the epidemic seasons. This would make our predictions much better because prediction errors are generally smaller outside the epidemic seasons and prediction intervals are generally narrower.

Thus, to prevent particularly lucky or unlucky sequences of random numbers to bias our results, for every probability p we want to have a set of runs where the missed delivery occurs at different weeks. This allows us to calculate an estimate of the error on the mean absolute error (MAE) and on the mean prediction interval (MPI).

For each value of p we would like $m = 20$ time series $\mathbf{n}^i(p)$, with $i = 1, ..., m$, and correspondingly m values of MAE and MPI. A simple approach to generating these sets of runs for each value of p would be to first generate n time series under the hypothesis that data are always delivered on time ($p = 0$, i.e. there is no chance that data are not delivered on time), which is the model discussed in Chapter 5, and then we run the code n times under the hypothesis that data are never delivered on time ($p = 1$, i.e. data are always delivered with one week delay). Every week

we then have $m = 20$ estimates made with a model where data are always delayed ($p = 1$) and $m = 20$ estimates made with a model where data are never delayed ($p = 0$).

To generate the time series for all other values of p between 0 and 1 we use a bootstrap sampling procedure. This allows us to reduce the number times we need to run the model.

Given a probability $p \in (0, 1)$ that thata are not delivered on time, for every week t we generate a set $r_t^i(p)$ of random numbers between 0 and 1, with $i = 1, ..., m$. Each of these sets allows us to generate the sample for a specific week t: when $r_t^i(p) < p$ we will take $n_t^i(p) = n_t^i(p = 0)$, while when $r_t^i(p) \geq p$ we will take $n_t^i(p) = n_t^i(p = 1)$. Once we have done so for all weeks we obtain our set of $m = 20$ time series for the dengue case count $\mathbf{n}^i(p)$, with $i = 1, ..., m$, in the whole period of analysis.

In this way, if we wanted to explore z values of $p \in (0, 1)$ we would run the model for the entire period of analysis only $2m$ times instead of zm. Generating m random numbers for all weeks and all values of p and building the zm time series would require negligible computational time.

Once all the time series have been built it is possible to calculate the MAE and MPI for each value of p by taking the mean of the MAEs and MPIs calculated on each of the single m time series with a probability p that thata are not delivered on time. We also calculate the standard deviation to give an idea of the variability of the MAEs and MPIs around their respective meran. The results of this procedure are presented in Figures 6.1 and 6.2.

The variability we observe is mainly due to sampling. In fact, the variability due to parameter's variability is much smaller, and we can observe it by looking at the standard deviation of the MAE and of the MPI for the cases of $p = 0$ and $p = 1$.

All the models using online data lead to better performance than the baseline model using only official data. We also find that the model using both Google searches and Twitter posts together at the same time is more accurate than those using the two data sources separately. Figure 6.1 depicts the mean absolute errors (MAEs) for all models as a function of the probability of data not being delivered on time. Furthermore, from Figure 6.1 we observe that the rate at which the MAE grows with p is smaller for models using online data compared to the baseline model. This is true if we consider all the weeks in the period of analysis, but also if we only look

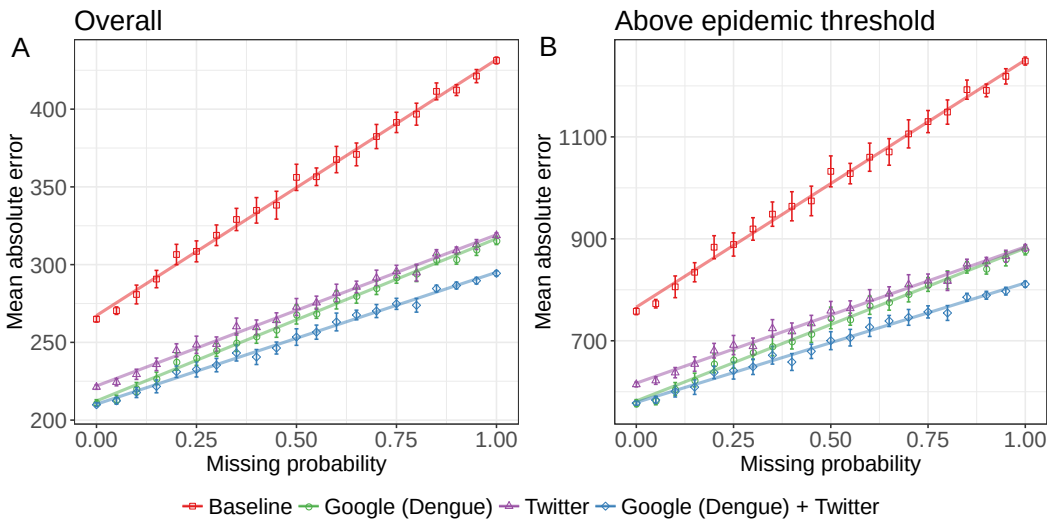

Figure 6.1: Mean absolute error (MAE) as a function of the probability of data not being delivered. For all the models considered it is possible to see a clear linear trend where the MAE increases as the probability of data not being delivered increases. All models using online data lead to lower MAEs, and hence better accuracy than the baseline model. (A) All the weeks are considered in the calculation of the MAE. (B) Only weeks where the number of cases is above the epidemic threshold are considered in the calculation of the MAE. We see that not only the accuracy of models using data from Google and Twitter is better than that of the baseline model, but also that the rate at which the MAE grows is smaller for models using online data.

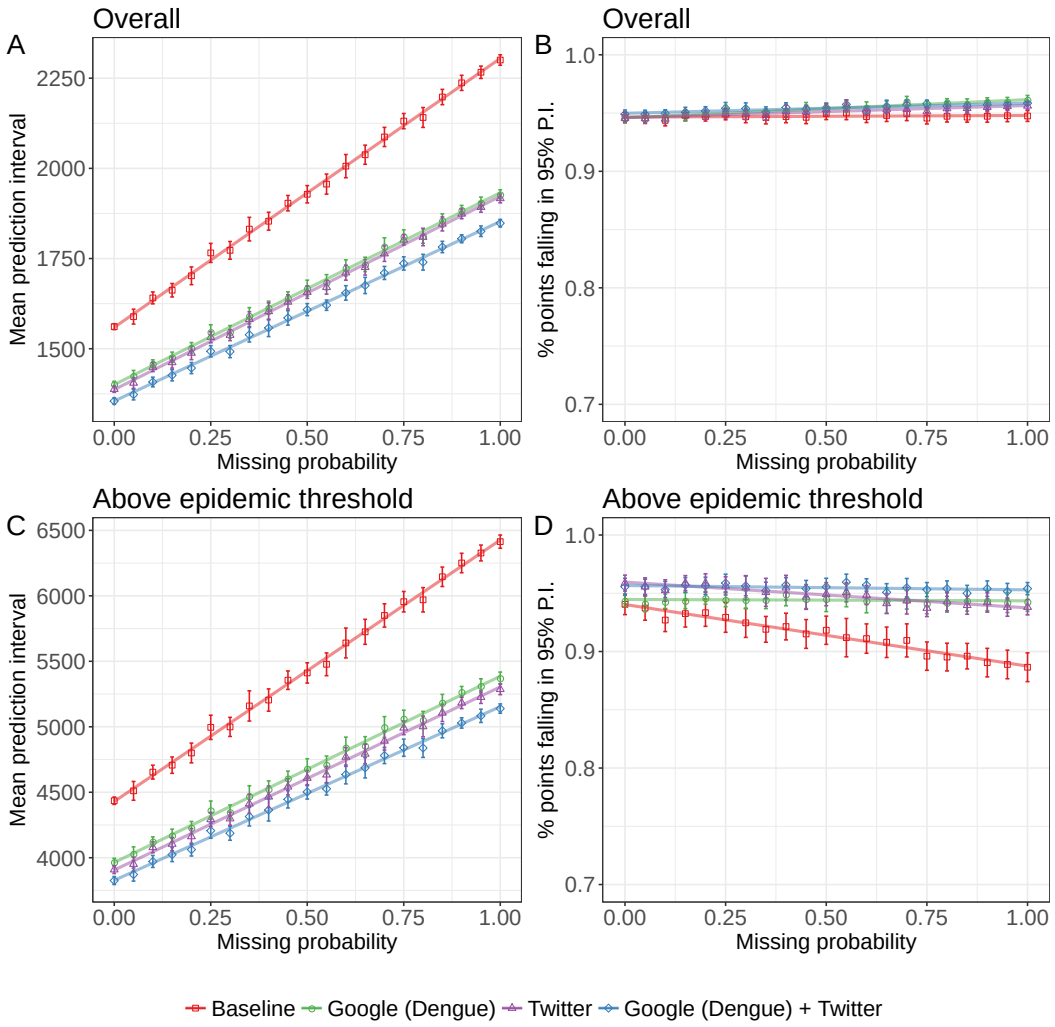

Figure 6.2: Mean prediction interval (MPI) as a function of the probability of data not being delivered. For all the models considered it is possible to see a clear linear trend where the MPI increases as the probability of data not being delivered increases itself. All models using online data show a better performance than the baseline model. (A) All weeks in the period of analysis are considered in the calculation of the MPI. (B) It is also possible to see that the model is working properly in terms of prediction intervals for all models. In fact, for all models and for all values of p the percentage of weeks where the true values fall within the prediction interval is about 95%. (C) Only weeks where the number of cases is above the epidemic threshold are considered in the calculation of the MPI. (D) In this case, only for the *Google (Dengue) + Twitter* model the 95% prediction interval reliably represents the range within which 95% of true points fall.

at weeks with a number of dengue cases above the epidemic threshold.

From Figure 6.1A we see that the MAE for the baseline model is 431.3 cases while the MAE for the *Google (Dengue) + Twitter* model is 294.4 cases when data are never delivered on time, i.e. with $p = 1$, and while considering the entire period of analysis. Thus we observe an improvement in accuracy of 32% when using Google and Twitter data together in the same model. This is larger than what we observe when data are always delivered on time, i.e. with $p = 0$. With reference to Figure 6.1A, and as reported in our analyses in Chapter 5 we see that the MAE for the baseline model is 267.2 while the MAE for the *Google (Dengue) + Twitter* model is $\simeq 213.3$. In this case, we observe a reduction in MAE of only 20%. Furthermore, we can see that the MAE of the *Google (Dengue) + Twitter* model when official data are never delivered on time is just 10% higher than the baseline model when official data are always delivered on time.

The situation is similar when we only consider weeks where the number of cases is above 550, i.e. during epidemics. With reference to Figure 6.1B, the MAE for the baseline model is 1248.7 cases while the MAE for the *Google (Dengue) + Twitter* model is 810.9 cases when data are never delivered on time, i.e. with $p = 1$. The reduction in MAE, in this case, is again of 35%. On the other hand, the MAE for the baseline model is 757.8 cases while the MAE for the *Google (Dengue) + Twitter* model is 577.7 cases when data are never delivered on time, i.e. with $p = 0$. Here we observe a reduction in MAE of $\simeq 24\%$. Again, we observe that the MAE of the *Google (Dengue) + Twitter* model when official data are never delivered on time is just 7% higher than the baseline model when official data are always delivered on time.

From Figure 6.1 we can see that if we assume that data are not delivered on time 5% of the time, using online data can give our model a mean accuracy which is better than that of the baseline model at $p = 0$, i.e. when data are always delivered on time.

The situation is very similar if we consider the precision of the models. In Figure 6.2A we depict the mean prediction intervals (MPIs) when considering the entire period of analysis. All models using online data have smaller MPIs than the baseline model for every value of the probability p that data are not delivered. The rate at which the MPI grows with p is also higher for the baseline model than for the models using online data. We see that for the *Google (Dengue) + Twitter* model, which

shows the best performance, the shrinking of the prediction intervals compared to the baseline model ranges from 13.3% when data are always delivered on time, i.e. when $p = 0$, to 19.7% when data are never delivered on time, i.e. when $p = 1$. This means that the higher the probability that data are not delivered on time, the higher the reduction of the MPI of the *Google (Dengue) + Twitter* model compared to the baseline model. Nevertheless, Figure 6.2B shows that the percentage of true points falling in the 95% prediction intervals is, in fact, 95% for all values of p.

In line with our analysis for the MAE, we also see that the prediction interval for the *Google (Dengue) + Twitter* model when data are never delivered on time, i.e. with $p = 1$, is 18.4% wider than the prediction error of the baseline model when data are always delivered on time, i.e. with $p = 0$.

Our findings are a little more complicated when we look at the prediction intervals during epidemic periods only. While the behaviour of the MPI we observe in Figure 6.2C is similar to the case when all weeks are considered, it is not the same for the percentage of true data points falling in the 95% prediction interval reported in Figure 6.2D. We observe a degradation in the reliability of the baseline prediction intervals when performing one-step-ahead forecast. This is somewhat expected. The less the information available, the larger the errors. This is because with less information it is more difficult to make a prediction, and since the prediction intervals become wider, the prediction errors will vary accordingly. For the baseline model we observe that 94% of the true data points are contained within the 95% prediction interval when official data are always delivered on time, while 88.6% of the true data points are captured by the 95% prediction interval when official data are never delivered on time. Fortunately, adding data from Google and Twitter can bring the percentage of true points included in the 95% prediction intervals to 95% independent from the delay in data delivery. Intuitively, online data act as a correction for the baseline model, providing information from the current week when official data are only available up to the previous week. In other words, they make up for the lack of official data about the current week. All these results confirm that the fewer official data are available, the more important the contribution of online data.

As in Chapter 5, we can perform the same analysis for every year separately. These results are reported in Figure 6.3. We find that results for each year are qualitatively similar to the results for all years combined, apart from the last one, 2016, where the mean absolute error for baseline model is very high. In that year, then, we find that online data models do not offer many advantages in the case of low probabilities

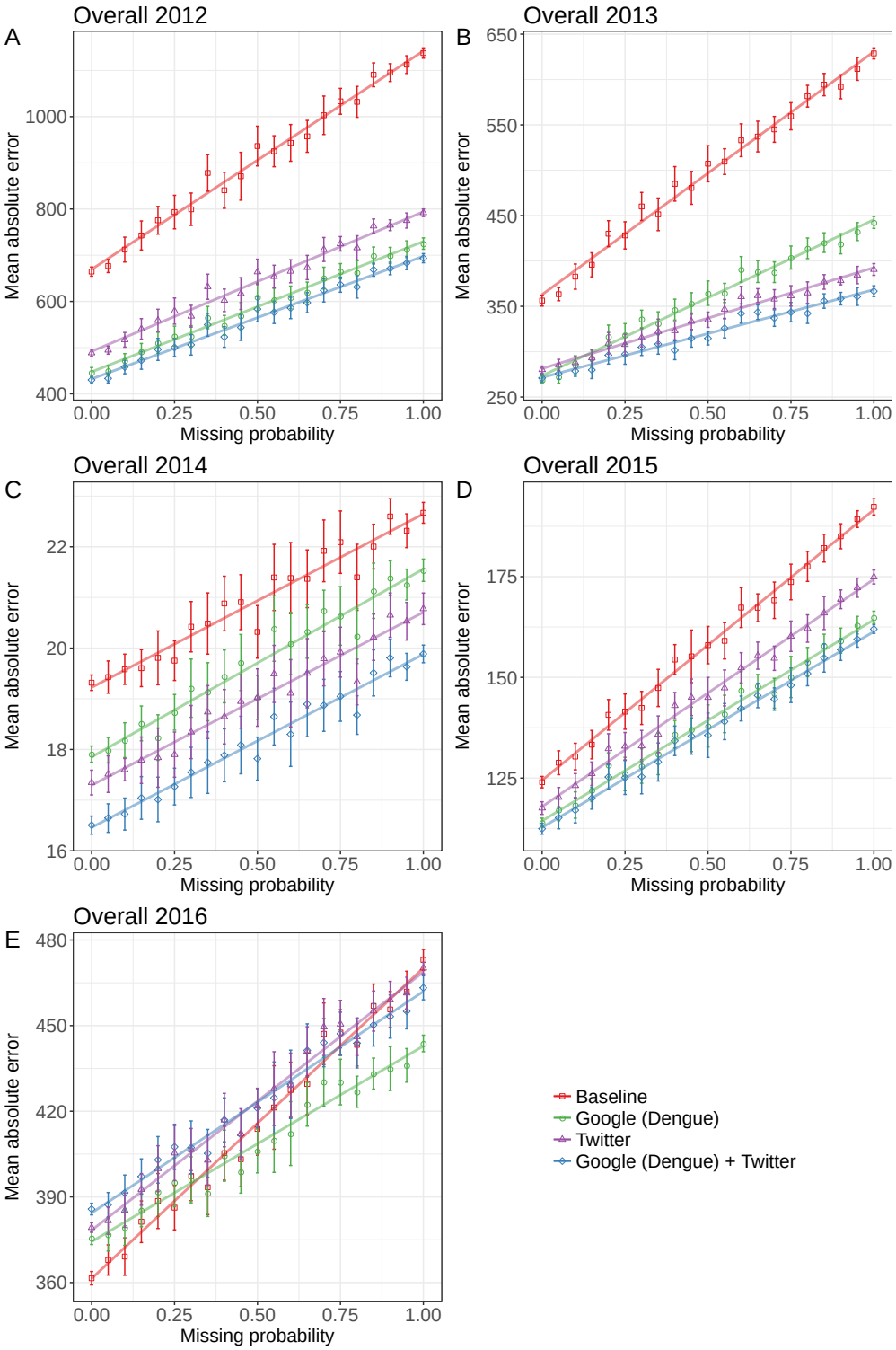

Figure 6.3: Mean Absolute Error as a function of the probability of missing data for different years. *(continues on the following page)*

Figure 6.3: *(continues from previous page)* For all the models considered it is possible to see a clear linear trend where the MAE increases as the probability of missing data increases itself. The y-axis ranges sometimes differ by an order of magnitude, and this is due to the vast differences in weekly dengue case counts. All models using online data show a better performance than the baseline model apart from the case of 2016. In this case, we do not see a noticeable improvement. This may be due to poor performance of the baseline model in that year.

of missing data, but they might become significant if data are not delivered with a very high probability, even though this is not likely to be the case in an operational situation. In quantitative terms, we can see that for years where the peak weekly dengue case counts are higher, the improvement provided by online data is more significant, while in years where the peak weekly dengue case counts are smaller, the improvement is much more marginal.

6.3 Discussion

In the real-life operation of surveillance systems like InfoDengue it might happen, for many different reasons, that official data are not delivered when they are supposed to be. Furthermore, instead of being delivered at the end of a certain week, they might be delivered at the end of the following one, together with the data entered into the system in that week. This is not a common occurrence, but it is a possibility that should be accounted for. In these situations, it is even more critical to have a reliable model because decisions need to be made even in the absence of the relevant data and we need to make the best use of the data that is available.

In this chapter, we have provided evidence that online data provide a considerable advantage over a model that is only based on official data when official data are not delivered on time. Moreover, we see that in such cases of delayed delivery of official data, online data play an even more valuable role than when official data are delivered on time. In fact, we see that the reduction of the estimation error in models using online data is higher when data are not delivered on time, with respect to the usual case when we have data delivered at the end of the current week. For example, we observe that the MAE for the *Google (Dengue) + Twitter* model is 32% smaller than the MAE of the baseline model when data are never delivered on time as compared to a 20% reduction when data are always delivered on time, as

seen in Chapter 5. Furthermore, we can see that the MAE of the *Google (Dengue) + Twitter* model when official data are never delivered is only 10% higher than the baseline model when official data are always delivered on time. If we look at the aggregate metric, the MAE, with a small probability of missing data $p \lesssim 0.1$, models using online data still display an average performance that is considerably better than that of the baseline model with complete data.

When we look at prediction intervals, we see that the model using official data becomes less reliable when data are not delivered on time, especially during periods of high dengue incidence. In contrast, the prediction intervals of the models using online data are smaller and remain reliable, always containing the observed notified case counts in 95% of the weeks.

From an operational point of view,like the model presented in Chapter 5, this modified model can be easily implemented to reduce the impact of situations where data are occasionally not delivered. Furthermore, the model presented in this chapter poses the basis for the exploration of short-term forecasting, which will be analysed in detail in the following chapter.

CHAPTER 7

Forecasting using partial online data

In all previous chapters, we have considered the problem of estimating the number of new dengue cases in the city of Rio de Janeiro for the week that has just ended. Given the severe delay in the official data, this problem is already very complicated. Here, we want to develop this analysis further and consider the problem of generating short-term forecasts of the number of dengue cases. In particular, we would like to attempt to estimate what the total weekly number of dengue cases will be at the end of the week that just started.

Figure 7.1 shows a slightly modified version of Figure 4.1 from Chapter 4 illustrating the timeline of operation of the forecasting model presented in this chapter.

In previous chapters we performed our analyses at a time τ when we obtain official data at the end of week t. What we seek to do in the current chapter is to estimate

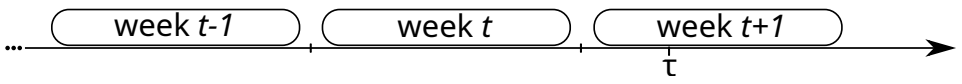

Figure 7.1: Timeline of forecasting using partial online data. Time τ is when we perform our analysis. Contrarily to the case of Figure 4.1, here time τ is not necessarily the time when we obtain official data. We define as week t the last full week which precedes time τ, starting on a Sunday and ending on the following Saturday. In this chapter we want to produce an estimate of the number of dengue cases in week $t+1$. Online data are also partially available for week $t+1$ at time τ and we want to use this partial knowledge to make more accurate predictions of the dengue case counts at the end of week $t+1$.

the number of dengue cases at the end of week $t+1$ by performing our analysis at a time τ before the end of week $t+1$ when partial online data relating to week $t+1$ are already available.

Following the model development in Chapter 6, we already have the necessary tools to approach this problem if we only use official data. In fact, in Chapter 6 we have considered how we can estimate the number of new dengue cases in week t given official data only up to week $t-1$. This is what the baseline model does in that context, and it is, in fact, equivalent to the problem of forecasting the number of new dengue cases in week $t+1$ given information only up to week t, provided that official data are made available on time.

As in previous chapters, here the focus is on how online data such as Google search volumes and the number of Twitter posts can improve these predictions. In the case of forecasting, though, there is another problem that we have to take care of. Unlike the case of the model presented in Chapter 6, it is not only official data relating the next week that are not available, but Google and Twitter data too. There are multiple approaches we could follow to address this lack of information, with different degrees of complexity. We discuss each of them thoroughly in the next section.

7.1 Materials and methods

There are multiple ways we can approach the problem of using partial online data relating to week $t+1$ together with official data collected up to week t to estimate the dengue case counts in week $t+1$. In particular, this is also because for this problem we use slightly different online data compared to what we used in previous chapters. In previous chapters we used Twitter data with a weekly resolution. In this chapter, instead, we perform our analysis based on daily Twitter data.

7.1.1 Data

For the analysis we carry out in the current chapter we obtained daily Twitter data starting from week 32 of 2012 until the end of 2016, again from the *Observatorio*

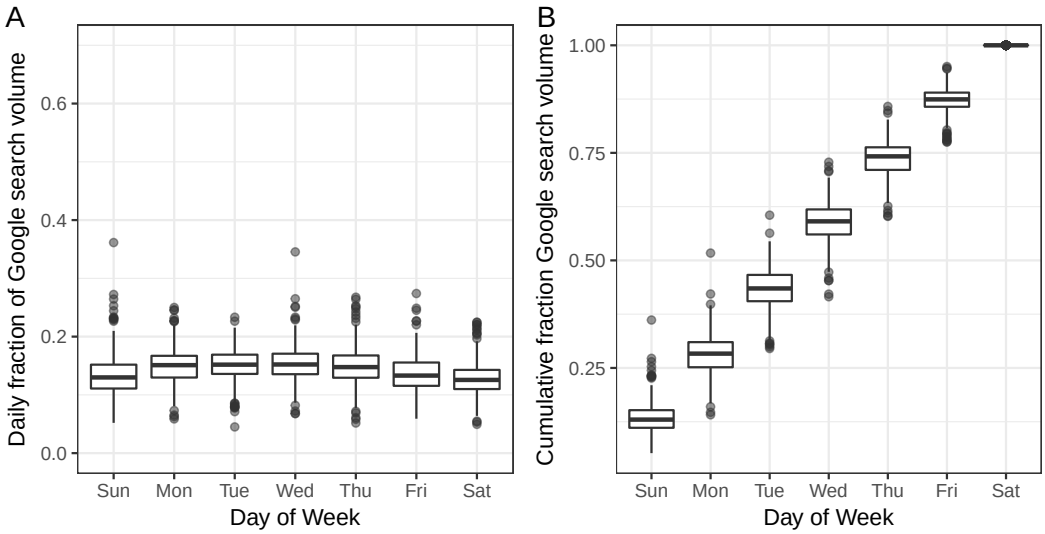

Figure 7.2: Daily distribution of Google search volume. In both figures, a box plot is built considering all the different epidemiological weeks. (A) Box plot of the day-of-week distribution of Google search volume of the topic *dengue* in the state of Rio de Janeiro. We observe that the search activity is at reasonably similar level throughout the week. (B) Box-plot of the cumulative day-of-week distribution of Google search volume of the topic *dengue* in the state of Rio de Janeiro. We observe that the cumulative search activity grows at a reasonably steady rate over the week.

da Dengue[1] via the InfoDengue[2] project (Gomide et al., 2011; Codeço et al., 2016). Apart from the finer time resolution, this data set is the same described in Section 3.1, i.e. tweets from the city of Rio de Janeiro expressing personal experience of dengue. We can also easily retrieve Google Trends data through the API for the same time interval with a daily time resolution. Thus, we have all the data necessary to carry out our analysis from week 32 of 2012 until the last epidemiological week of 2016. Since we use the first 20 weeks in the data set for training, our analysis starts at the beginning of 2013 and gets to the end of 2016.

In light of this, what we seek to do in this chapter is to exploit this better resolution of the online data to make more timely and more reliable predictions of the number of dengue cases at the end of the following week.

Online data are available on a daily basis, and we already have partial information about online activity that week on the Monday, on the Tuesday and so on. To

[1] http://www.observatorio.inweb.org.br/dengue/
[2] https://info.dengue.mat.br/

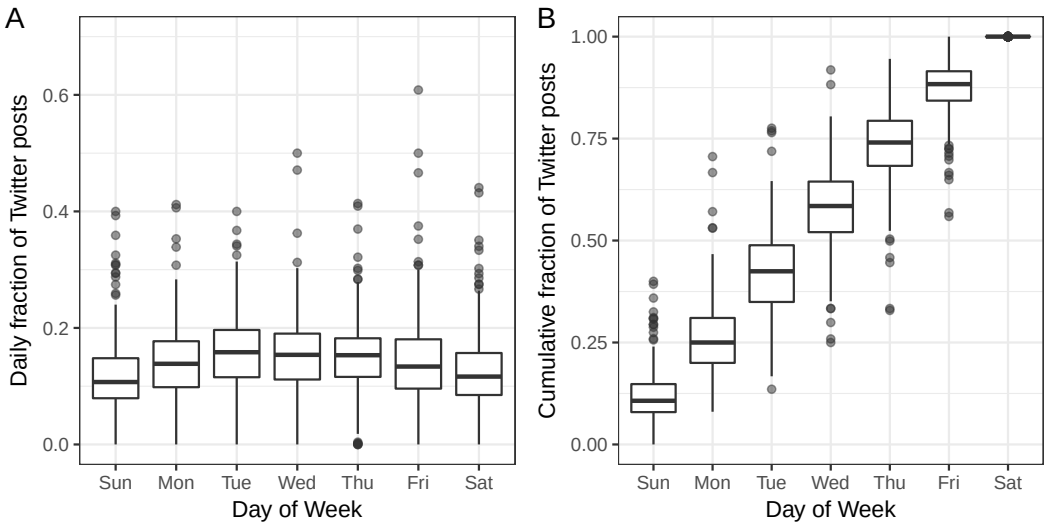

Figure 7.3: Daily distribution of the number of Twitter posts. In both figures, a box plot is built considering all the different epidemiological weeks. (A) Box plot of the day-of-week distribution of the number of Twitter posts about dengue in the city of Rio de Janeiro. We observe that on average the Twitter activity is at a reasonably similar level throughout the week. Differently from the search activity reported in Figure 7.2, here we observe a higher variability. (B) Box-plot of the cumulative day-of-week distribution of the number of Twitter posts about dengue in the city of Rio de Janeiro. We observe that the cumulative Twitter activity grows at a reasonably steady rate over the week.

understand if there is any regularity in this partial information about online activity, we look at the daily distribution of Google search volumes for the topic *dengue* in the state of Rio de Janeiro and the number of Twitter posts in the city of Rio de Janeiro, which are displayed in Figures 7.2 and 7.3.

What these figures suggest is that on average online activity is reasonably stable throughout the week. This is an average behaviour across all weeks in our period of analysis. Of course Figures 7.2 and 7.3 show that there are weeks where the online activity, specifically the daily online activity as a fraction of the weekly online activity, varies considerably over the week. This structure might be useful to predict the Google search volume or the number of Twitter posts at the end of the week by only having partial data, or just the number of dengue cases without even estimating the volume of online data at the end of the next week.

7.1.2 Forecasting methods based on online data

With reference to Figure 7.1, here we discuss three different methods to predict the number of dengue cases in Rio de Janeiro in week $t+1$. These three methods are all based on weekly official dengue case counts, Google search volumes and the number of Twitter posts relating to weeks up to week t, which are available at the end of week t. Furthermore, they also use daily Google search volumes and the number of Twitter posts relating to days in week $t+1$ previous to the time when we perform the analysis. For example, to predict the number of dengue cases in week $t+1$ at the end of the Tuesday in week $t+1$ we will only have access to data on searches up to the Tuesday, and tweets posted up to the Tuesday. We rely on such partial online data as external regressors to predict the dengue case count relating to week $t+1$.

Forecasting online data. The first method we consider is a straightforward one, where we use an auto-regressive model to make a one-week-ahead forecast of the Google search volume and Twitter post count at the end of week $t+1$ based on online data available at the end of week t. This puts us in a situation similar to that examined in Chapter 6. In fact having official data up to week t and an estimate of online data in week $t+1$ is analogous to the situation where we do not obtain official data on time at the end of week t while Google and Twitter data are available at the end of week t. We can then use the same approach discussed in chapter 6 to predict the dengue case counts at the end of week $t+1$. Estimating the parameters of an auto-regressive model is a relatively quick task compared to estimating the parameters of the INLA model. Thus, the increase in computational time of the process of estimating online data for week $t+1$ is negligible.

Estimating the total weekly online activity. The second method we consider to incorporate online data in our predictions is to estimate the number of Twitter posts and the Google search volume at the end of week $t+1$ using the daily Google search volumes and number of Twitter posts relating to days in week $t+1$ previous to the time when we make our analysis, and by exploiting the regularity of daily online activity we observed in Figures 7.2 and 7.3. Depending on what day it is when we make our prediction we need to aggregate the online data and divide by the mean fraction of available data on that weekday as shown in Figure 7.2B and Figure 7.3B. For example, if we make

our analysis at the end of the Tuesday, three full days have passed in week $t+1$. Let us look only at one of the online data sources for simplicity, for instance Google data. We find that on average, by the end of the Tuesday, the volume of Google searches since the beginning of the week corresponds to only about $\simeq 43\%$ of the weekly volume. So, if we divide the actual Google search volume generated since the beginning of the following week until the end of the Tuesday by 43% we get a reasonable estimate of the volume of Google searches we can expect at the end of the week, and we can then use it as an external regressor in the model *Google(Dengue)* presented in Chapter 6 with the probability of missing data $p = 1$. The situation is analogous if we use Twitter data or if we use both Google and Twitter data together. For exploratory purposes, and to provide further evidence on the viability of this approach, in Figure 7.4 we depict a scatter plot of the weekly Google search volumes and number of Twitter posts estimated with this approach on every weekday versus the true weekly Google search volumes and number of Twitter posts. We observe that both for Google, in Figure 7.4A, and Twitter, in Figure 7.4B, there is a strong correlation between the estimated and true volumes at the beginning of the week, and as the week progresses the correlation becomes stronger. As expected, for estimates produced at the end of the Saturday, the correlation is perfect. In this case as well, estimating the volume of online data takes a negligible amount of time compared to estimating the parameters of the INLA model. The daily distribution of online data is calculated a priori on the training period, and when estimating the volume at the end of week $t+1$ we need only to divide the current volume by average perentage of online data available at the current time of the week.

Using partial online data. The third method we consider is to use the known online data at the time when the estimate is made as external regressors. In this case, again, we use the model presented in Chapter 6 with the probability of missing data $p = 1$. In other words, if we want to make a prediction of the number of dengue cases in week $t+1$ on the Tuesday of week $t+1$, we use official data only up to the end of week t, and online data collected only up to the Tuesday of week $t+1$ for every week. With this last approach, we do not estimate the online activity at the end of the week, but we just use the available data about online activity when we perform our analysis. The only additional step required by this approach is that we need to subset and aggregate online data for all previous weeks depending on the day when we

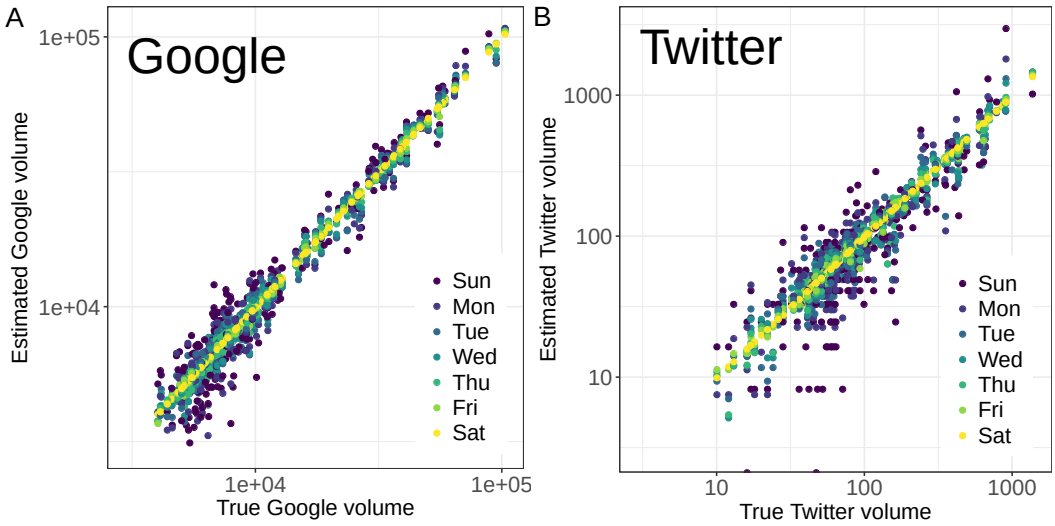

Figure 7.4: Scatter plot of online data volumes estimated with the approach of *estimating total weekly online activity* versus the true online data volumes. In both figures, a scatter plot is built considering all the different epidemiological weeks in our period of analysis, and the colour indicates the day of the week in which the estimation has been made. Both figures are in log-log scale. As expected, for estimates made at the end of Saturdays, the correlation is perfect. At the beginning of the week, on Sundays, we observe high dispersion around the line with angular coefficient 1. As the week progresses (as the colour gets lighter) the dispersion is reduced, and for estimates at the end of the week, at the end of the Saturday, we observe a perfect correlation. (A) Here we depict the scatter plot for Google search volumes for the topic *dengue* in the state of Rio de Janeiro. (B) Here we depict the scatter plot for the number of Twitter posts in the city of Rio de Janeiro. This picture provides evidence that it is possible to estimate the online activity at the end of the week by only using partial online activity early in the week. Of course, the more information is available, the more accurate is the estimate.

perform the prediction. In this case also, this further step requires a negligible amount of time compared to the estimation of parameters for the INLA model.

In terms of complexity, all three approaches are comparable, non of them requiring considerably more time to run than the standard INLA model.

The approach of *using partial online data* does not produce spurious confidence intervals or online data predictions that are not used. The only catch is that a different model needs to be used depending on the weekday we are in when we make the prediction. In fact, the whole training process of the model is based on the assumption that the prediction is always made on that particular weekday. In the next section, we explore the results obtained using all of these approaches.

7.2 Results

In this section, we present the results we obtained by following the three different approaches that have been described in Section 7.1.2.

To be able to make comparisons between models using weekly and daily online data, we need to change the operating window. Thus, as stated before, we consider weeks from epidemiological week 1 of 2013 to epidemiological week 52 of 2016.

The first approach we use to include online data is that of *forecasting online data*, i.e. predicting the Google search volume and number of Twitter posts in week $t+1$ based on known online data up to week t by using an auto-regressive model. We then use these estimates to predict the number of cases at the end of week t with the Bayesian model described in Chapter 6 corresponding to the case where official data are never delivered on time.

We consider several auto-regressive models to forecast the volume of online data, AR(p) models with several values of the parameter p. None of the models provides any improvement over a naive model. For this reason, all models using forecasted online activity to predict dengue case count in week $t+1$ use a naive model to forecast online activity in week $t+1$ based on online activity up to week t..

The results for all the models that we consider with this approach, i.e. the baseline model and the models using naively-forecasted online data, are presented in Table

7.1.

Table 7.1: Comparison of the models including online data to the baseline model when online data are forecasted. The relative MAE is the ratio between the MAE of a model and that of the baseline model. We see that using forecasted Google or Twitter data does not improve the performance of the baseline model.

Model	MAE	relative MAE
Baseline	281.0	1
Google (Dengue)	282.6	1.005
Twitter	282.2	1.004
Google (Dengue) + Twitter	284.0	1.011

Even if daily data are available, with this approach we are still bound to work at the same time scale than our official data, so we cannot take full advantage of the online data that we collect over the week. Furthermore, as we can see from Table 7.1, using forecasted online data does not grant any particular advantage in terms of accuracy with respect to the baseline model. For these reasons, we do not investigate this method further.

The second approach we consider is that of *estimating the total weekly online activity*, i.e. estimating the Google search volume and number of Twitter posts relating to week $t+1$ based on partial online activity at the time when we perform the analysis. An important thing to note is that even just after the very first day of the week we have some new online data that we could use to make more accurate predictions. We could use the distributions shown in Figure 7.2B and Figure 7.3B to estimate the volume of online data at the end of the week. This approach has the advantage of actually including further information that has been collected since the start of week $t+1$ until the end of the current day, so it can potentially help to make predictions more accurate. On the other hand, though, taking the mean value of the daily distribution means that we are ignoring all the information on the variability.

Figure 7.5 shows how the prediction error changes as a function of the weekday when we use this second approach. As in all previous chapters, the epidemiological week we use is defined as starting on the Sunday. Furthermore, estimates are made at the end of each respective day, when all the online data relative to that day have been collected. Overall we find that the prediction error tends to decrease as the week progresses and more information about online data becomes available. However, even if we only consider the first day of the week, we find that the improvement of the models using online data is already significant compared to the baseline model

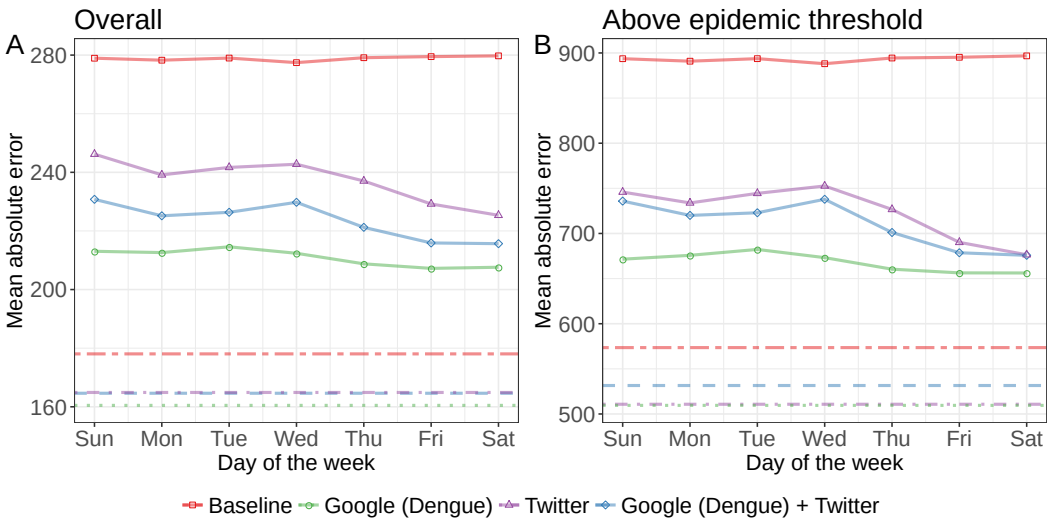

Figure 7.5: Mean Absolute Error (MAE) as a function of the weekday when online data at the end of the week are estimated from partial data. In red, we depict the mean absolute error of baseline models throughout the week. In green we depict the mean absolute error of the model using Google data, in purple that of the model using Twitter data and in blue that of the model using both Google and Twitter data together. The dashed lines spanning the entire week indicate the estimates produced once partial official data become available for week $t+1$. (A) All the weeks are considered in the calculation of the MAE. (B) We only consider weeks where the number of dengue cases is above the epidemic threshold in Rio de Janeiro. As expected, the baseline model's MAE is independent of the day of the week, since we are working on online data to get better predictions. We still observe some variability in the baseline model throughout the week because of the stochastic nature of the model we use. For all the other models considered it is possible to see a negative trend where the MAE decreases as more information becomes available as the week progresses. All models using online data exhibit a better performance than the baseline model on any weekday.

when we use this approach. From Figure 7.5A we observe a MAE of 213.1 cases for the best performing model, the one using Google data together with official data, when we perform our analysis at the end of a Sunday. The baseline model, instead, shows a MAE of 279.7 cases. Thus, we see a reduction in MAE of about 25.8% with respect to the baseline model when we use Google search volume. Even the *Twitter* model, which is the worst performing model among the models using online data, shows an MAE of 246.2 when we perform the analysis at the end of a Sunday, and a reduction in MAE of 11.8% compared to the baseline model.

The situation is similar when we consider only weeks where the dengue case count is

above the epidemic threshold in Rio de Janeiro. We can observe these results in Figure 7.5B. In this case, when we perform the analysis on a Sunday, the MAE *Google (Dengue)* model shows an MAE of 671.5 cases and still provides an improvement of 24.9% compared to the baseline model which has an MAE of 893.6 cases. During epidemics, the *Twitter* model when we perform the analysis on a Sunday shows an MAE of $\simeq 745.8$ and an improvement of 16.6% with respect to the baseline model.

We also compare the predictions of the models using online data performed on the Sunday and on the Saturday of week $t+1$ before official data about week $t+1$ become available to those of the baseline model performed at the end of the Saturday of week $t+1$ after official data about week $t+1$ become available. We observe that the mean absolute error of the *Google (Dengue)* model goes from 213.1 cases at the end of the Sunday to 207.6 cases at the end of the Saturday. The baseline model at the end of the Saturday has instead a MAE of 178.1 cases. Thus, as the week progresses, the *Google (Dengue)* model's predictions go from 19.7% to 16.6% less accurate with respect to the estimates of the baseline model calculated at the end of the Saturday. Similarly, if we only consider weeks where the case count is above the epidemic threshold in Rio de Janeiro, we observe for the *Google (Dengue)* model an MAE that goes from 671.5 cases at the end of the Sunday to 656.2 cases at the end of the Saturday, while the baseline model at the end of the Saturday shows an MAE of 873.5 cases. During epidemics, as the week progresses, the *Google (Dengue)* model's predictions go from 17.0% to 14.4% less accurate with respect to the estimates of the baseline model calculated at the end of the Saturday. This means that online data help us make estimates that are much more accurate than what we would otherwise be able to obtain just with official data but, of course, they are not as good as those we produce when we also have the official data about week $t+1$, even though they are just incomplete data as discussed in Section 3.1.3.

In line with our approach in previous chapters, we do not look at just accuracy, but we also investigate what happens to the prediction intervals. Practitioners need not only to detect an incoming outbreak, but they also need a clearer idea of the variability they can expect in the number of infections to be able to make informed decisions. Figure 7.6 shows how the prediction intervals change for all the models we consider as more online data become available during the week. When we consider predictions made on a Tuesday, we observe from Figure 7.6A that the *Google (Dengue) + Twitter* model has a mean prediction interval (MPI) of 928.2, 25% smaller than the baseline model and only 10.7% greater than the baseline model calculated at the end of the week, when partial official data about week $t+1$ are

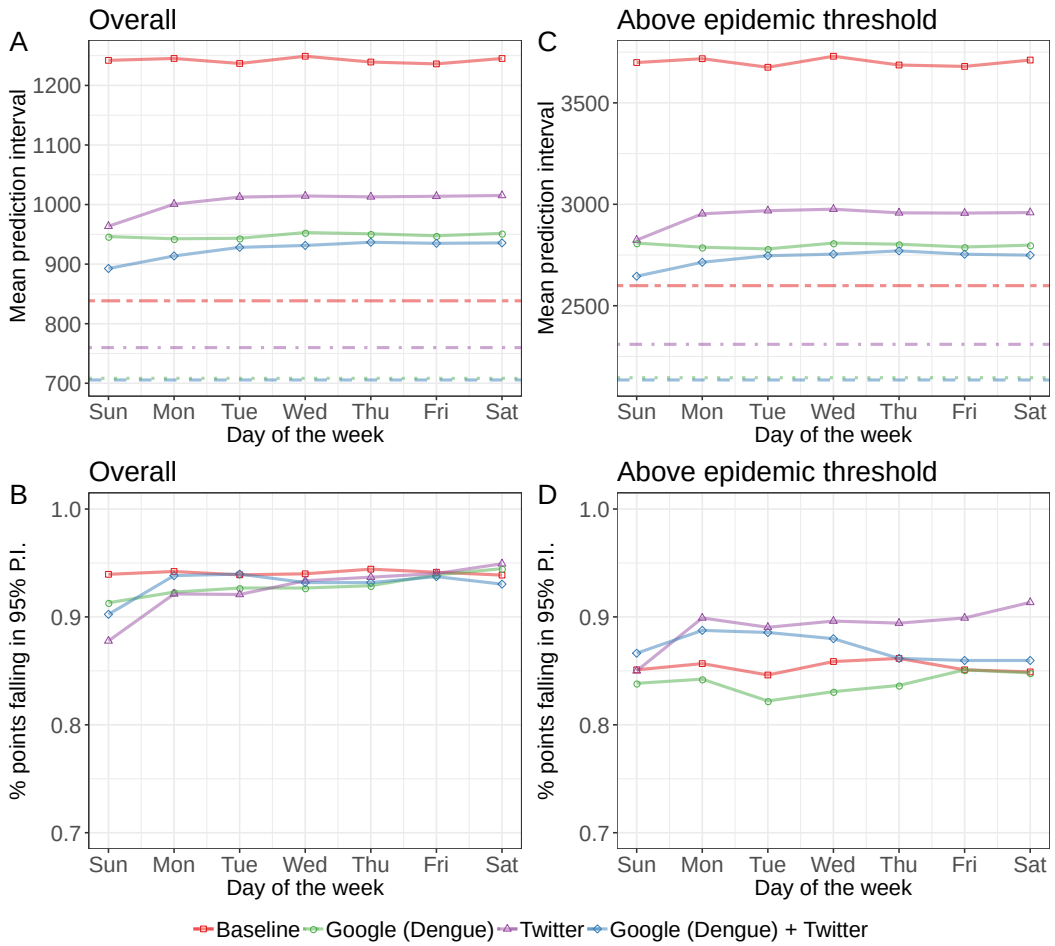

Figure 7.6: Mean prediction interval (MPI) as a function of the weekday when online data at the end of the week are estimated from partial data. In red, we depict the mean absolute error of baseline models throughout the week. In green we depict the mean absolute error of the model using Google data, in purple that of the model using Twitter data and in blue that of the model using both Google and Twitter data together. The dashed lines spanning the entire week indicate the estimations once official data become available for the current week. (A) All weeks are considered in the calculation of the MPI. As expected, the baseline model's MPI is independent of the day of the week, since we are working on online data to get better predictions. We still observe some variability in the baseline model throughout the week because of the stochastic nature of the model we use. For all models using online data, we observe a slight increase of the MPI as the week progresses. (B) Earlier in the week, for all models using online data we see that the 95% prediction intervals contain less than 95% of points. As the week progresses, the error gets smaller, and the prediction interval gets wider, thus catching approximately 95% of the data points. In particular, the MPI for the model using only Google data seems to be independent of the day of the week, while the percentage of points falling within the 95% prediction interval increases as the week progresses. *(continues on the following page)*

Figure 7.6: *(continues from previous page)* On the other hand, from Figure 7.5 we see that the MAE of the *Google* model decreases as the week progresses. From this we conclude that even though the prediction intervals remain relatively constant as the week progresses, more true points fall within them as an effect of predictions becoming more accurate. (C) Only weeks where the number of cases is above the epidemic threshold in Rio de Janeiro are considered in the calculation of the MPI. The behaviour of the MPIs as the week progresses is similar to that observed by considering all weeks in the period of analysis. (D) The situation is more complex when we consider the percentage of true points falling in the prediction intervals. In this case, the baseline model has a stable $\simeq 85\%$ percentage of points falling in the 95% prediction interval. This is explainable with the fact that, during epidemics, predictions are in general less accurate. We thus expect that if the fraction of points falling in the 95% prediction interval is smaller than expected during periods of epidemics, the fraction for weeks where the number of dengue cases is below the epidemic threshold will be higher than 95%. While for the *Google (Dengue)* and *Google (Dengue) + Twitter* models the percentage of points falling in the 95% prediction interval is below 90%, independent of the weekday, for the *Twitter* model it grows slowly as the week progresses, and it is above 90% for the whole week except for the Sunday. However, also for the *Twitter* model, the percentage of points falling in the 95% prediction interval never reaches 95%.

obtained. As we would expect, the MPI of the baseline model, i.e. the mean width of the interval around the point estimate containing 95% of the true points, remains constant as the week advances. Surprisingly, instead, while the MPI stays constant for the model just using Google data, it increases for the models using Twitter or both Twitter and Google, regardless of whether we consider all weeks or only weeks with dengue case count above the epidemic threshold in Rio de Janeiro. So, although the model becomes more accurate, the MPIs enlarge during the first part of the week.

To understand this trend we need to look at Figure 7.6C. While the fraction of true points falling in the 95% confidence interval stays consistently at about 95% for the baseline model, it shows a slightly increasing trend for all other models using online data. For example, the predictions of *Twitter* model produced at the end of the Sunday have 95% prediction intervals containing 87.8% of the true points, while the predictions of the same model produced at the end of the Saturday, before new official data are obtained, have 95% prediction intervals containing 95% of the true points.

The results in Figure 7.6D show what happens when we instead consider only weeks

with dengue case counts above the epidemic threshold. For all models the percentage of points falling in the 95% prediction interval is smaller than 95% independent of the day of the week. The *Twitter* model is the model with 95% prediction intervals containing the highest percentage of true data points, around 90% from Mondays until the end of the week when it reaches 91.3%. This is, of course, a problem because one of the characteristics that we need for our prediction intervals is for them to be reliable and, in this case, they are not.

One of the possible reasons for this behaviour is that when estimating the volume of online data at the end of the week we assume that the online activity is stable throughout the week. But looking at Figures 7.2 and 7.3 it is clear that, although on average it is true that online activity is similar in every weekday, there is considerable variability. For example, if we consider one of the outliers for the Sunday in Figure 7.2A, we can see that its value is about 0.35. The remaining 0.65 fraction of the Google search volume for the week is then split over the remaining 6 weekdays, from Monday to Saturday. This means that, on average, the fraction of Google search volume in all other weekdays needs to be just above 0.10. If we split the Google search volume evenly among all weekdays we would obtain a fraction of 0.14. Not accounting for this variability wrongly reduces the uncertainty of our models' predictions, and the result is that they are not reliable anymore.

Finally, we consider one last approach which is very similar to the one just presented but has an important difference. The difference is in the fact that instead of training the model using an estimate of the full volume of Google searches or Twitter posts for week $t + 1$, we use only a fraction of it. For example, if we wish to make short-term predictions using a model that only uses online data up to the third day of the epidemiological week, we train the model in previous weeks using online data only up to that day of the week. In other words, if we want to do a forecast at the end of the Tuesday, for example, we only use online data from Sundays, Mondays and Tuesdays both for training the model and for predicting the number of dengue cases for the week $t + 1$. The results from the approach of *using partial online data* are presented in Figures 7.7 and 7.8.

In terms of accuracy we can see that with the approach of *using partial online data* the MAEs for all models, presented in Figure 7.7, are a little bit higher than those of models using the approach of *estimating the total weekly online activity* which are shown in Figure 7.5. In both cases, the prediction error generally decreases with the day of the week as more information about online data becomes available. In

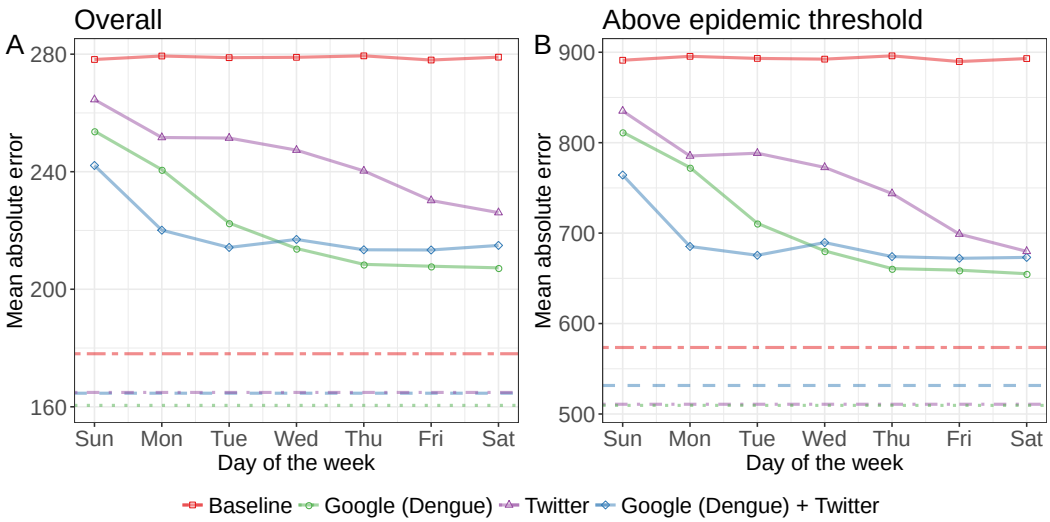

Figure 7.7: Mean Absolute Error (MAE) as a function of the weekday when using partial online data as external regressors. In red, we depict the mean absolute error of baseline models throughout the week. In green we depict the mean absolute error of the model using Google data, in purple that of the model using Twitter data and in blue that of the model using both Google and Twitter data together. The dashed lines spanning the entire week indicate the estimations once official data become available for the current week. (A) All the weeks are considered in the calculation of the MAE. (B) Only weeks where the number of cases is above the epidemic threshold in Rio de Janeiro are considered in the calculation of the MAE. As expected, the baseline model's MAE is independent of the day of the week, since we are working with online data to get better predictions. We still observe some variability in the baseline model throughout the week because of the stochastic nature of the model we use. For all the other models considered it is possible to see a negative trend where the MAE decreases as more information becomes available during the week. All models using online data exhibit a better performance than the baseline model, regardless of the weekday.

Figure 7.7, when we perform the analysis on Sundays, we find that the improvement of the models using online data compared to the baseline model is not as high as that shown in Figure 7.5. With the approach of *using partial online data*, we find that the best model is the model using both Google and Twitter data together in the first part of the week.

We see that for the *Google (Dengue) + Twitter* model the MAE results are 242.1 cases when produced at the end of the Sunday and 214.2 cases when produced at the end of the Tuesday. Thus, on the Tuesday, we observe an MAE 23.2% smaller than the baseline model. From the Wednesday to the end of the week, however,

we find that the best model is the model one using official data and Google search volume only, which is slightly more accurate than the *Google (Dengue) + Twitter* model.

The situation is similar when we consider only weeks where the dengue case count is above the epidemic threshold in Rio de Janeiro. In this case, the model using both Google and Twitter data exhibits an MAE of 764.2 cases at the end of the Sunday while it goes down to 675.5 cases at the end of the Tuesday, which is 24.3% smaller than the MAE of the baseline model. Again, from the Wednesday to the end of the week, the model using only Google search volume as an external regressor outperforms all other models.

We then compare the predictions of these models to the estimates of the baseline model produced at the end of week $t+1$, when partial official data relating to week $t+1$ become available. We observe for the baseline model at the end of week $t+1$ an MAE of 178.1 cases. This means that, with an MAE of 214.2 cases the *Google (Dengue) + Twitter* model is 20.3% higher at the end of the Tuesday with respect to the MAE of the baseline model with partial official data on week $t+1$. Similarly, if we only consider weeks where the dengue case count is above the epidemic threshold, we see that the MAE of baseline model at the end of week $t+1$ is 573.5 cases, and thus with an MAE of 675.5 cases the *Google (Dengue) + Twitter* is 17.8% higher at the end of the Tuesday. This shows that online data from Google and Twitter can also help us make short-term forecasts that are more accurate than what we would be able to obtain using official data alone.

Looking at Figure 7.8, we observe that precision increases as the week progresses, similarly to that we observe for accuracy in Figure 7.7. When we make a comparison on the Tuesday, we observe for the *Google (Dengue) + Twitter* model an MPI 18.9% smaller than the baseline model and we see that predictions are only 20.7% less precise than those made by the baseline model at the end of week $t+1$, when official data relating to week $t+1$ become available. Although the improvement in precision we observe for the approach of *using partial online data* is smaller than the one observed for the approach of *estimating the total weekly online activity* in Figure 7.6, with the approach of *using partial online data* the MPI decreases as the week progresses, meaning that as more online data become available the precision of the models improves.

Moreover, similarly to what we observed for the approach of *estimating the total*

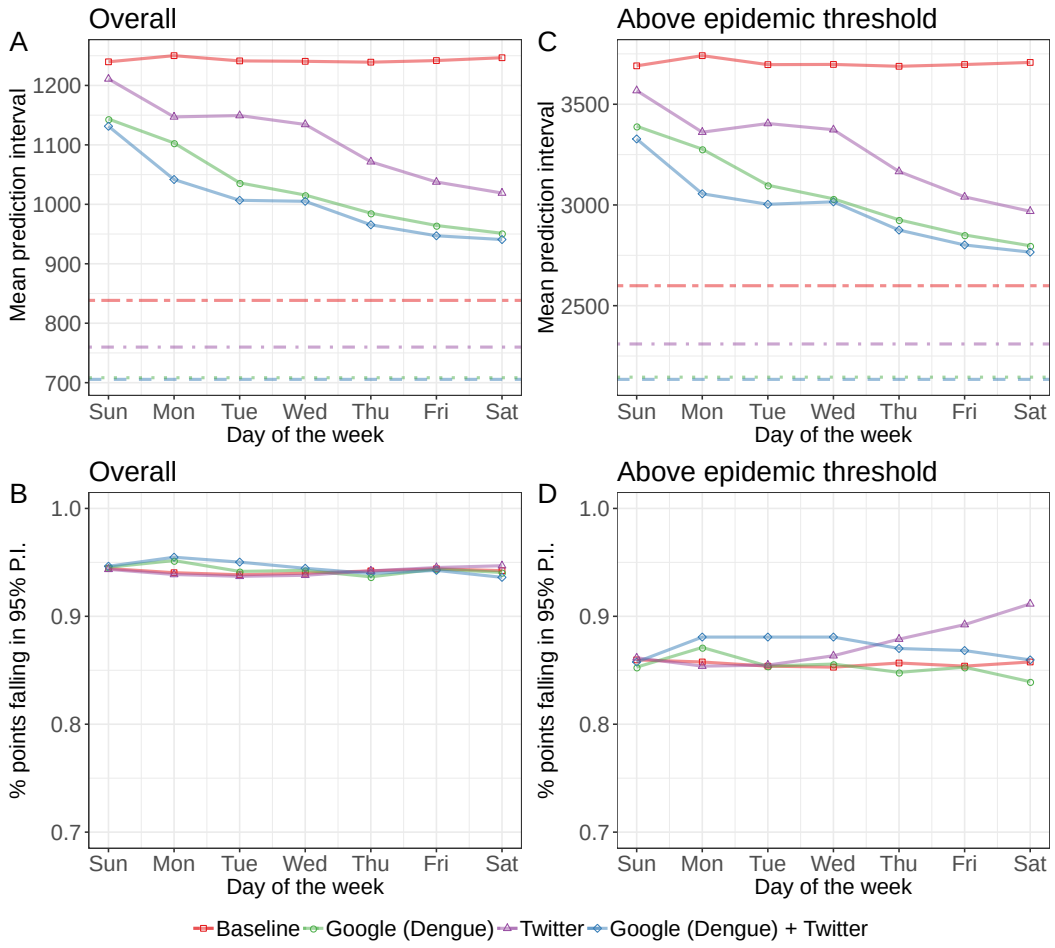

Figure 7.8: Mean prediction interval (MPI) as a function of the weekday when using partial online data as external regressors. In red, we depict the mean absolute error of baseline models throughout the week. In green we depict the mean absolute error of the model using Google data, in purple that of the model using Twitter data and in blue that of the model using both Google and Twitter data together. The dashed lines spanning the entire week indicate the estimations once official data become available for the current week. (A) All weeks are considered in the calculation of the MPI. As expected, the baseline model's MPI is independent of the day of the week, since we are working on online data to get better predictions. Furthermore, the MPI decreases with the day of the week, even though not in a linear fashion. In particular, there is a sharp decrease on the first and second day, and then the difference reduces for the rest of the week. (B) The percentage of points falling within the 95% prediction interval is approximately 95% for the whole week. (C) Only weeks where the number of cases is above the epidemic threshold are considered in the calculation of the MPI, i.e. during the epidemic periods in Rio de Janeiro. We observe a similar pattern to the case where we consider all weeks in the period of analysis. *(continues on the following page)*

Figure 7.8: *(continues from previous page)* (D) The situation is more complex when we consider the percentage of true points falling in the prediction intervals. In this case, the baseline model has a stable $\simeq 85\%$ percentage of points falling in the 95% prediction interval. This is explainable with the fact that, during epidemics, predictions are in general less accurate. We thus expect that if the fraction of points falling in the 95% prediction interval is smaller than expected during periods of epidemics, the fraction for weeks where the number of dengue cases is below the epidemic threshold will be higher than 95%. For models using online data we observe a slightly higher percentage of points falling in the 95% prediction interval. While for the *Google (Dengue)* and *Google (Dengue) + Twitter* models the percentage of points falling in the 95% prediction interval is always below 90%, for the *Twitter* model it grows slowly as the week progresses, but it never reaches 95%.

weekly online activity, in Figure 7.8D we see that the fraction of points falling in the 95% prediction interval is lower than 90% for the whole week for all models. Only for the *Twitter* model the percentage of points falling in the 95% prediction interval stays at about 90% from Monday to Saturday, when it reaches 91.1%.

By comparing Figures 7.5-7.8 we can see that with the approach of *estimating the total weekly online activity* we obtain smaller prediction errors and smaller prediction intervals earlier in the week with respect to predictions made with the approach of *using partial online data.*. For both approaches, the percentage of true data points falling in the 95% prediction intervals is approximately or slightly lower than 95% when we consider all weeks in our period of analysis, while it is between 82% and 92% when considering only weeks where the weekly dengue case count is above the epidemic threshold in Rio de Janeiro.

7.3 Discussion

In this chapter, we have considered whether it is possible to use our nowcasting model to produce a short-term forecast, i.e. to predict the number of dengue cases in Rio de Janeiro for week $t+1$, for which we do not have official data and for which we have only partial online data up to the day when we perform the analysis.

In Chapter 6 we have analysed the problem of dealing with delayed delivery of official data. We have seen that the baseline model presented in Chapter 5 can easily be extended to produce estimates of the dengue case count in week t when data are not delivered at the end of week t and we only have official data up to week

$t-1$. Producing this estimate is equivalent to generating a short-term forecast. In this chapter, we used the same approach to generate short-term forecasts of the dengue case count in week $t+1$ with official data available up to week t. Specifically, our interest here was on whether it is possible also to include online data in this analysis. In fact, contrarily to the situation described in Chapter 6 where official data are delayed but online data are available, here online data are only partially available up to the day when we produce the forecast. The challenge, in this Chapter, was to generate short-term forecasts by using only partial information on online activity.

We have explored several possibilities. The first approach we considered is *forecasting online data*. We used a simple auto-regressive model using weekly online activity up to week t to forecast online activity in week $t+1$. We have found that the best results are obtained by using a naive model to forecast online activity, and that predicting online activity does not produce any improvements in the forecast of dengue case counts compared to the baseline model which uses partial official data up to week t alone.

For the following analysis, we used daily counts of Twitter posts about dengue relating to the city of Rio de Janeiro, which were provided by Fiocruz, and daily search volumes of the topic *dengue* in the Rio the Janeiro state, which were collected through the Google Trends API. Thanks to this new dataset we were able to consider two new, different approaches.

The second approach we considered is *estimating the total weekly online activity*. We used partial data at different points of the week to estimate the total volume of both Google and Twitter data at the end of the week. In other words, we used partial online data to estimate the entire amount of online data at the end of the week both for Google and Twitter, and then we used them as external regressors in our Bayesian model. This approach produced promising results, although these need to be analysed with care. When we only consider the mean absolute error, this approach is very promising. The mean absolute error, even if we just use online data from only the first day of the week, appears to be significantly smaller than that of the baseline model for all models using online data. The model using Google data leads to a 25.8% increase in accuracy, which improves as the week progresses. If this was our only metric, we could say that our approach does an excellent job of improving the estimates.

As shown in previous chapters, however, this is not the only metric that we should consider, and there are other aspects of the estimation that we need to take into account. Our results suggest that, with this approach, the prediction intervals of models using online data tend to be smaller than what they should be. We find that the prediction intervals increase throughout the week as the prediction error lowers, but for all models using online data 95% prediction intervals contain the complete notified dengue cases slightly less than 95% of the weeks. When looking at weeks with dengue case count above the epidemic threshold, we observe a generally worse performance compared to weeks with dengue case count below the epidemic threshold. Specifically, for all models, including the baseline model, the 95% prediction intervals contain the complete notified dengue cases considerably less than 95% of the weeks.

The third approach we explored is that of *using partial online data*. Models using the approach of *estimating the total weekly online activity* have been trained with the full-week online data volume in the weeks up to week t, and then partial online data relating to week $t + 1$ have been used to estimate the full-week online data volume for week $t + 1$ to be used as an external regressor. With the approach of *using partial online data, instead*, models are trained with partial online data (for example up to the Tuesday). We then use the same proportion of online data (again up to the Tuesday) for week $t + 1$ to estimate the number of cases that will be notified during the entire week $t + 1$.

None of the three approaches is noticeably more complex than the standard INLA model, and computational time is not impacted by the further step we make to estimate online data at the end of week $t + 1$.

Similarly to the approach of *estimating the total weekly online activity*, we observe that the mean absolute errors of models using partial online data generally decrease as the week progresses. In particular, for what concerns the model using both Google and Twitter data, the prediction done at the end of the Monday is very similar to the one done at the end of the week. This provides further evidence of the advantage of using both Google and Twitter data together. In this case, either considering all weeks in the period of analysis or just weeks when the dengue case count is above the epidemic threshold, the reduction in MAE of the model using Google and Twitter compared to the baseline model is respectively of 23.2% and 24.3%, which is in line with what happened with the approach of *estimating the total weekly online activity* with partial online data. Furthermore, predictions made on the Tuesday by the

model using Google and Twitter data together with the approach of *using partial online data* are only 20.3% less accurate than those the baseline model can produce at the end of week $t + 1$ when partial official data relating to week $t + 1$ become available.

When we look at precision, if we make a comparison on the Tuesday, as we have done for the mean absolute error, we observe for the model using Google and Twitter data together with official data an MPI 18.9% smaller with respect to the baseline model and we see that its predictions are only 20.7% less precise than those made by the baseline model at the end of week $t + 1$, when official data relating to week $t+1$ become available. Contrarily to what happens with the approach of *estimating the total weekly online activity*, with the approach of *using partial online data* the prediction intervals of all models using online data progressively become smaller as the week progresses and more information is added. However, the precision intervals of model using the approach of *using partial online data* have the same problem we observed in models using the approach of *estimating the total weekly online activity* with partial online data. The percentage of points falling in the 95% prediction intervals is stable around 95% for the whole time, independently of the day when we make the estimate, when we consider all weeks in the period of analysis. If instead we only consider weeks with a weekly dengue case count above the epidemic threshold, we observe that for all models the 95% prediction intervals contain between 83% and 92% of true data points.

All models using online data, by either using the approach of *estimating the total weekly online activity* from partial online data or the approach of *using partial online data*, outperform the baseline model. The *Google (Dengue) + Twitter* model seems to offer a better compromise in terms of increase of accuracy and precision. In particular, we observe that while using the approach of *estimating the total weekly online activity*, the MAE for the *Google (Dengue) + Twitter* model is slightly higher than for the *Google (Dengue)* model, its MPI is slightly lower, and the percentage of true data points falling in the 95% prediction interval is slightly higher, although lower than 95%. Moreover, although accuracy for the *Google (Dengue) + Twitter* model using the approach of *estimating the total weekly online activity* is slightly lower than for the approach of *using partial online data*, precision is notably higher.

In conclusion, in this chapter, we have shown that it is possible to use daily online data to make predictions about the number of dengue cases notified during week $t + 1$ a few days earlier than when official data relating to week $t + 1$ are made

available. This results takes us beyond the realm of nowcasting and into the realm of forecasting. In Chapter 5 we have seen that online data from Google and Twitter together with partial official data can help us produce more accurate and more precise estimates of the number of dengue cases notified in week t. In this chapter we have seen that the baseline model presented in Chapter 5 can be extended to provide short-term forecasts of the number of dengue cases in week t. Crucially, we have shown that we can also use partial online data to make these forecasts more accurate and more precise.

The important message that we can get from this analysis is that we can make accurate and reliable estimates much earlier than at the end of the week when we use Google and Twitter data in our model. The results we observe when we use official data obtained up to week t and partial online data relating to week $t + 1$ are unsurprisingly less accurate and precise than those we obtain with the baseline model when we use official data obtained up to week $t + 1$. We observe differences smaller than 20% if we make our predictions a few days after the beginning of week $t + 1$. We again found that there is a notable advantage in using Google and Twitter data together in the same model over just using partial official data. This allows to make predictions early in the week with accuracy and precision close to those produced by the same and other models at the end of the week, providing policymakers with a crucial strategic advantage.

CHAPTER 8

Nowcasting in other cities

In previous chapters, we only considered the city of Rio de Janeiro for all our analysis. It would be useful to assess whether the algorithm we used in previous chapters works in other cities too. In order to do so, we need to carefully consider what kind of problem we might encounter when considering cities smaller than Rio de Janeiro. The main problems we can expect to encounter in smaller cities are that there might be several weeks without any notified dengue case, delays might be longer due to a less efficient system of data digitisation, or that people might not tweet or search about dengue as much as in bigger cities.

The framework we presented in previous chapters can easily be applied to other cities once this particular technical problem is taken care of. We follow the approach we introduced in Chapter 5. First, we look at the available data and perform some exploratory analysis. Then, for each city, we study how the models that use Google and Twitter compare to the baseline model that only uses official data in terms of accuracy and precision. To do so, we look at the prediction error and the width of the prediction intervals.

We perform an analysis on ten different cities from the states of Rio de Janeiro and Paraná. Our aim here is to analyse where we can potentially use our method, whether or not our results still hold when the cities and respective dengue case counts are much smaller or the delays in the data digitisation process much longer, or when the volume of Google searches and Twitter posts are much smaller.

8.1 Data

In this chapter we analyse the weekly number of dengue cases in 10 cities from the states of Rio de Janeiro and Paraná, which are presented in Table 8.1 in descending order of population.

Table 8.1: List of the municipalities considered in this analysis. The municipalities are relative to the states of Paraná and Rio de Janeiro and are ordered by population from the larger to the smaller. Data about population was retrieved from UNdata and corresponds to the 2010 census. (http://data.un.org/Data.aspx?d=POP&f=tableCode% 3A240)

City	Population	State
Rio de Janeiro	6,320,446	Rio de Janeiro
Curitiba	1,751,907	Paraná
Londrina	452,855	Paraná
Campos	356,608	Rio de Janeiro
Maringá	342,310	Paraná
São Gonçalo	337,273	Rio de Janeiro
Foz do Iguaçu	253,962	Paraná
Paranaguá	133,761	Paraná
Toledo	103,644	Paraná
Umuarama	90,105	Paraná
Resende	77,848	Rio de Janeiro

In Figure 8.1 we present a summary of this information showing the states of Rio de Janeiro (right) and Paraná (left) where the municipalities we consider are coloured according to their population. We aim to consider a broad spectrum of city sizes in terms of population while still considering cities that are sufficiently geographically separated. We only consider these two states because at the time the analysis was performed the InfoDengue system was only operating in the states of Rio de Janeiro, Paraná and Espírito Santo. However we could not find suitable cities for this analysis in the latter state. That is because we needed cities that are geographically far enough from each other but that have a large enough volume of Twitter posts to be used in our framework. Google searches are aggregated at the state level, thus would hardly constitute a limiting factor. On the other hand, Twitter posts might be not too many in a smaller city if the rate of Twitter users is very low. In these cases, we might observe very low to null count of Twitter posts relating to dengue even during outbreaks. This would make much more difficult to evaluate how Twitter posts can improve a baseline model.

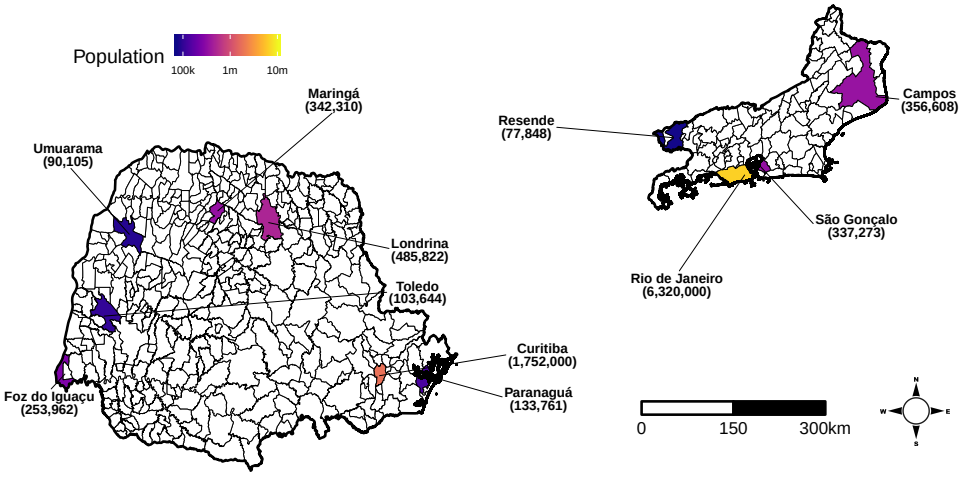

Figure 8.1: Map of the municipalities analysed in the states of Rio de Janeiro and Paraná. The map shows the states of Rio de Janeiro (right) and Paraná (left), where all the considered cities are coloured according to their population. We also include the city of Rio de Janeiro for reference. Data about population was retrieved from UNdata and corresponds to the 2010 census. (http://data.un.org/Data.aspx?d=POP&f=tableCode%3A240)

As in previous chapters, we carry out our analysis using epidemiological weeks, which are defined as starting on the Sunday. Where weeks span two different calendar years, the week belongs to the year in which more days of the week fall. As such, if the calendar year begins on a Monday, a Tuesday or a Wednesday, the epidemiological year is considered to have started on the final Sunday of the previous calendar year. Otherwise, the epidemiological year starts on the first Sunday of the calendar year. Each epidemiological year therefore has either 52 or 53 epidemiological weeks. For all these cities we have data starting from the first epidemiological week in 2012 until the last epidemiological week in 2016.

Building on the analyses produced in previous chapters, the first thing we look at is the delay distribution of the official data for all these cities, to better understand how severe the delays are, and to what extent they vary. Hence, the delays are depicted in Figure 8.2. It is possible to see that these cities display a wide range of delay patterns. When comparing the delay distributions of the new cities considered in this chapter to the delay distribution of Rio de Janeiro, also included in Figure 8.2, we can see that only a few cities display a pattern similar to that of Rio de Janeiro, i.e. with a comparably long time to reach 95% of notified cases and with a comparably small variability. In most cases, the variability is in fact much higher.

There are multiple cities in Figure 8.2 where the line indicating 95% or even 80% of the empirical distribution are still zero after several weeks. For example, for the city of Campos, the line indicating the lower end of the 80% of the empirical distribution is 0 until a delay of 8 weeks. That line marks the lower 10% of the distribution. This means that for 10% of the weeks in the period of analysis there are no data available with a delay smaller than 8 weeks in Campos. In other words, in 10% of the weeks the first notified cases will not be available earlier than two months.

From Figure 8.2 we can see that the city of Campos is the worse case in terms of delay. It is not only the city with the highest median delay, but also the city with the highest delay variability. In the data we analyse here relating to the city of Campos, for more than half of the weeks in the period of analysis, only 10% of the cases are entered in the system within a month from when they are notified, and for for more than half of the weeks no information is submitted into the system earlier than two weeks after cases are notified. For these reasons, we expect our baseline model to encounter problems in this scenario. Not only it will have to work with delayed, partial data, but in several weeks there will not be data relating to the last two or more weeks. In Chapter 6 and Chapter 7 we already analysed how the performance of our baseline algorithm is reduced when data are delayed by one week only. Here we face the possibility of delays of multiple weeks. We expect this to be the case also in other cities such as Londrina, Resende or São Gonçalo where the median delay is not so high, but their variability in delay is considerable.

The next thing we analyse is the extent to which official data are correlated with online data compared to the case of Rio de Janeiro, where we already know that our model delivers reasonable estimates and that online data provide an improvement. This analysis is reported in Figure 8.3. We should keep in mind that, as described for the case of Rio de Janeiro in Section 3.1, while official data and Twitter data are available at the city level, Google Trends data is instead available at the state level. This means that for all the cities in the same state we use the same Google Trends data.

We observe that there are no cities in which the correlation between official and Twitter data is as high as in Rio de Janeiro. However, in most of the cities, the correlation between the official data and Google data is higher than in Rio de Janeiro. This provides reason to believe that the framework we developed in Chapter 5 for the city of Rio de Janeiro may deliver useful estimates in other cities too. Specifically, we aime to analyse the performance of models using online data from Google

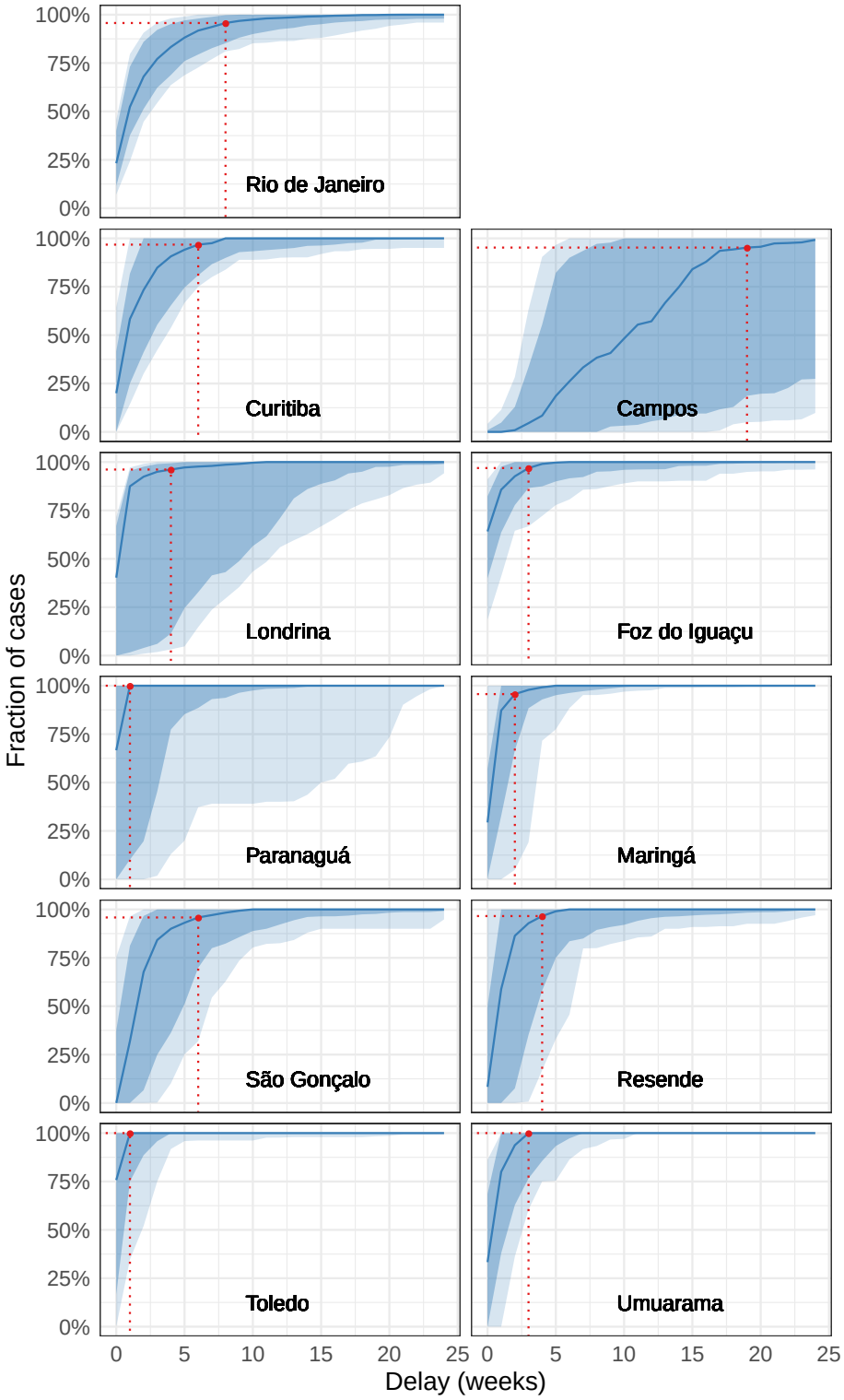

Figure 8.2: Delay curve for all cities. We examine the rate at which dengue cases for any given week are added into the system in all 10 municipalities we consider in this study.*(continues on the following page)*

Figure 8.2: *(continues from previous page)* We also include the city of Rio de Janeiro for comparison. Here we depict the empirical distributions of the delays with which dengue cases are entered into the system, over the whole time series. The blue line depicts the median fraction of cases entered into the system after a given delay. The dark shaded area indicates 80% of the empirical distribution of the fraction of cases entered into the system after the respective delay from the date of notification, and the light shaded area, instead, indicates 95% of such empirical distribution. We also highlight for every city the approximate mean number of weeks before 95% of dengue cases for a given week are entered into the system. These empirical distributions can be compared with that of Rio de Janeiro which is shown in Figure 3.2C. We can see that very few cities display delay patterns similar to those of Rio de Janeiro, i.e. with a comparably long time to reach 95% of notified cases and with a comparably small variability.

and Twitter together with delayed official data.

On the other hand, Figure 8.3 offers a partial and rather optimistic overview of the correlation between the available official and online data.

Figure 8.4 shows a visual comparison of the official data, Google data and Twitter data for all cities. All the time series have been normalised to take values between 0 and 1 to allow comparison. We can see that in smaller cities Twitter data become very noisy, and this adds more technical problems to our algorithm, especially because often the number of Twitter posts can be zero. We note that since Google data cannot be retrieved at a finer geographical resolution than state level, while official data and Twitter data are presented at city level, the Google search volumes relate to the entire state in which the city is situated. This, for some cities, causes noticeable differences between the trend of official case counts and Google search volume.

For example, São Gonçalo is geographically very close to Rio de Janeiro. For this city we observe a similar correlation between notified dengue cases and Google search volume. On the other hand, Campos and Resende are quite far away, and we in fact observe that there are peals in the Google seach volume of the state of Rio de Janeiro that do not reflect the notified dengue case counts in these cities. Despite this, for Resende we observe a correlation between the notified dengue case count and Google search volume which is higher than in Rio de Janeiro.

This means that we should take extra care when analysing results and drawing conclusions in this Chapter.

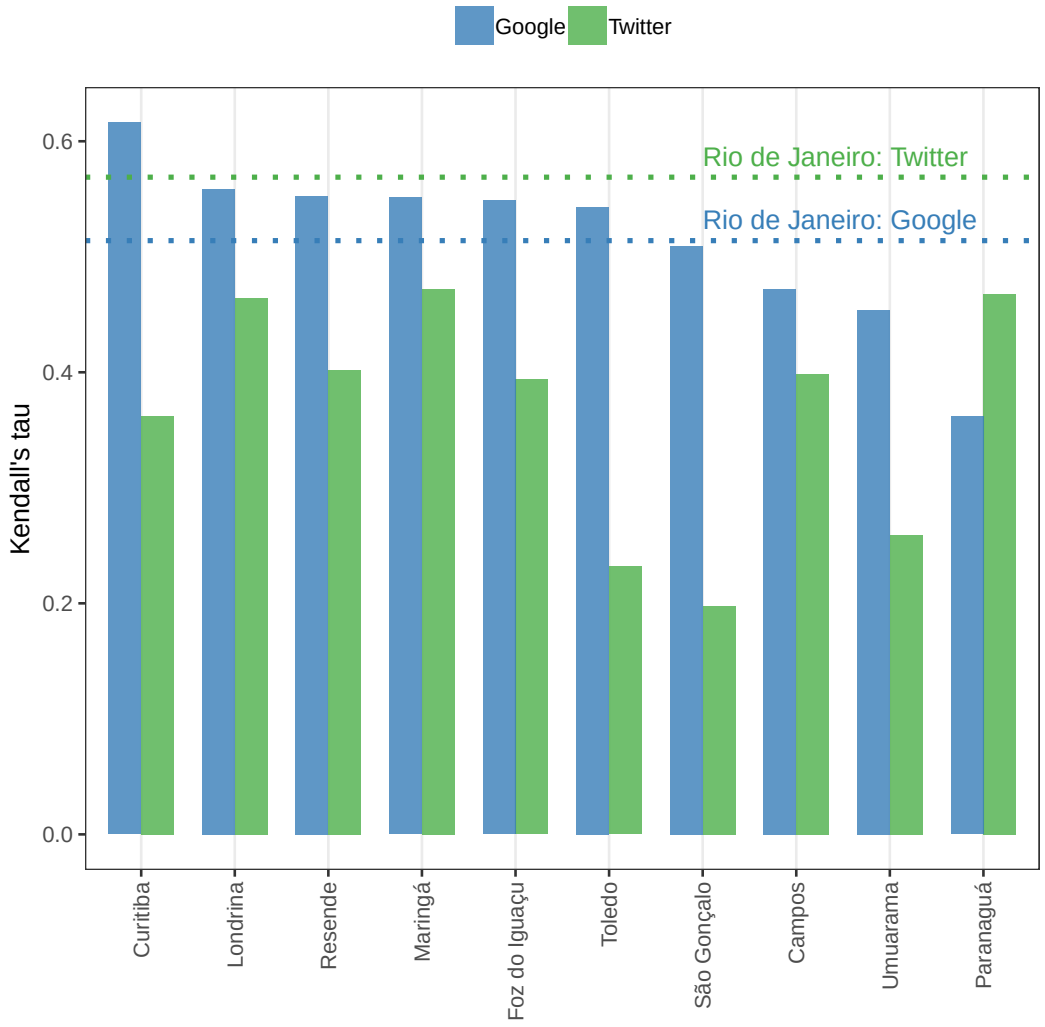

Figure 8.3: Comparison of the correlations between official data and online data for all cities. Correlations have been calculated as in Chapter 3 using Kendall's tau. The correlation τ is reported between official data and Google search volume, and between official data and the number of Twitter posts for each city. We also include the values for Rio de Janeiro for comparison, although we highlight that these values relate to a slightly shorter time window. We can see that, for most cities, the correlation with Google data is higher than that for Rio de Janeiro, while for Twitter data, no city shows a correlation as high as that for Rio de Janeiro. We can conclude that, especially in the case of Google data, these correlations are strong enough to merit further investigation of whether online data can inform estimates of current dengue case counts in these cities too.

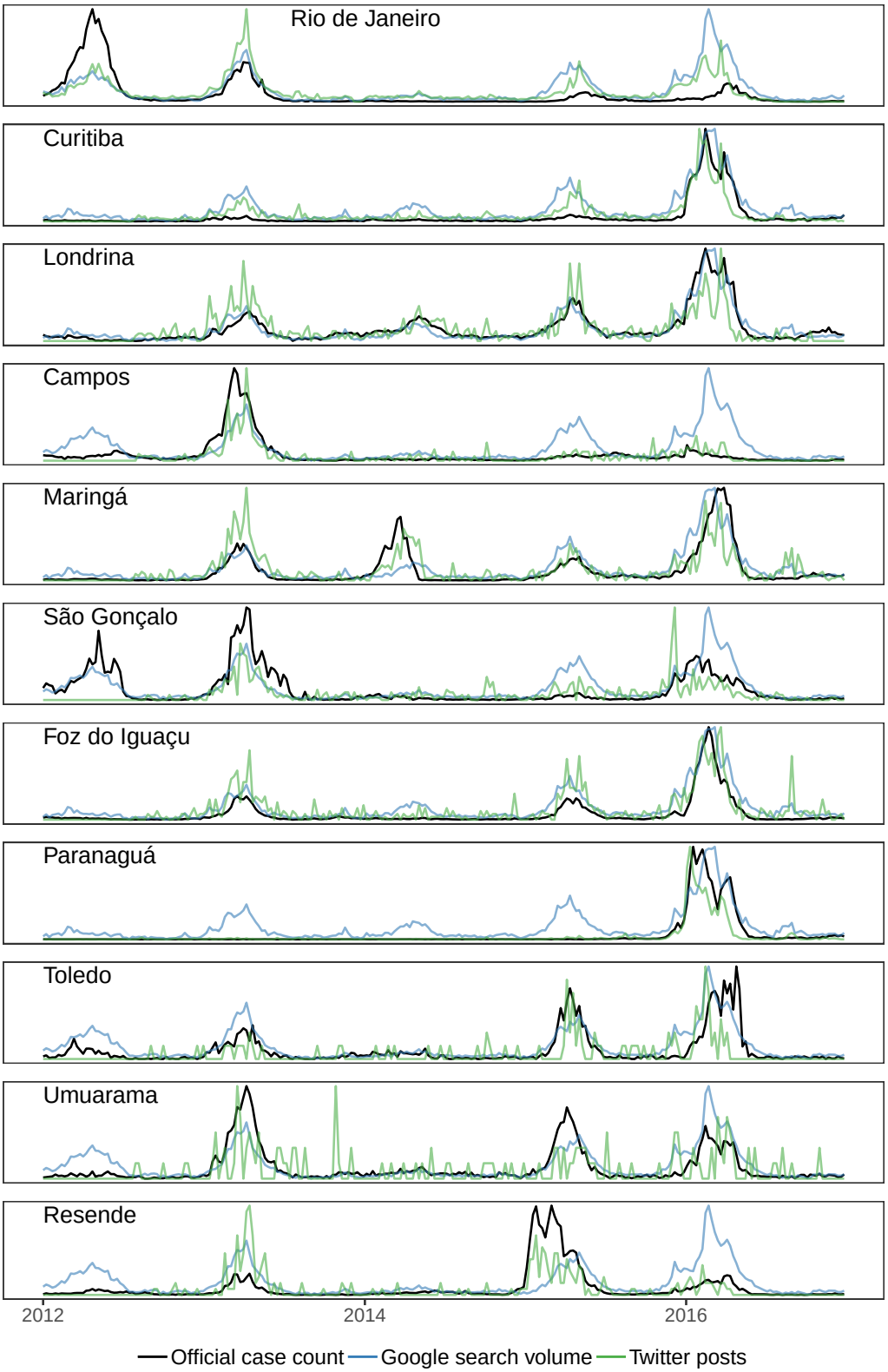

Figure 8.4: Dengue case count data compared to data from Google and Twitter. *(continues on the following page)*

Figure 8.4: *(continues from previous page)* For all cities we report the total number of dengue cases recorded in official data from January 2012 to December 2016. We also report the Google search volume of the topic *dengue* for the state including any given city, and the number of Twitter posts about dengue at the city level. All the time series have been normalised to take values between 0 and 1 to facilitate comparison. We can see that, apart from years when there are no outbreaks, peaks in official and online data in general align. We can also see that, in the smaller cities, Twitter data are very noisy. This probably explains why the Kendall correlation between official and Twitter data is much lower in those cities in comparison Rio de Janeiro.

8.2 Methods

To use the baseline model we introduced in Chapter 5 to generate dengue case count estimates for the cities considered here, we need to make some small modifications as the volume of dengue case count data in the cities we consider in this chapter is much smaller than in Rio de Janeiro. As a result, it is entirely possible that in some weeks there are no notified cases of dengue, as no patients report dengue symptoms, and therefore no dengue case is entered into the system, and no other cases relating to these weeks is later confirmed to be a dengue case. In these circumstances, a simple Negative Binomial distribution is not appropriate to represent the probability distribution of $\lambda_{t,\tau}$ dengue cases occurring in week t with delay τ, as described in Chapter 5. Instead, we need a probability distribution that accounts for a relatively large number of zero counts. For this reason, we use a zero-inflated variation of the Negative Binomial distribution, which does exactly that. We detail this model in Section 8.2.1.

The tools we have to evaluate models in Chapter 5 are relative metrics that allow us to compare different models in the same time window, relative to the same complete official data. Here, we would also like to consider a comparison between cities to assess where our models are working best. For this reason, in Section 8.2.2 we introduce some metrics that allow for this kind of comparison.

8.2.1 Zero-inflated models

In this chapter we investigate whether rapidly available data on Google searches and tweets relating to dengue in all the cities we consider can enhance weekly estimates

of the number of dengue cases reported up to week t, that is the week for which we want to produce an estimate. As we anticipated before, however, we need to modify our baseline model to account for an excess of zero-count data. We therefore compare the following models:

Baseline. Zero-inflated models describe random events that contain an excess of zero-count data in the unit of time, in our case, weeks. They were first introduced by Lambert (1992) to extend Poisson models, and were later used by Greene (1994) for the case of Negative Binomials. An excess of zero-count data is often observed when cities are small and the volume of dengue cases outside the epidemic season is not very high. Having used a Negative Binomial for the city of Rio de Janeiro, here we use a zero-inflated Negative Binomial (ZINB) which is composed of two components that contain two different zero-generating processes. One component is just a binary distribution that generates zeros with a certain fixed probability π. The second process is a Negative Binomial that generates counts normally, and of course, some of these can be zeros.

As in Chapter 5, let $n_{t,\tau}$ be the number of cases that occurred in week t and were reported in week $t+\tau$, thus with delay τ. We assume that $n_{t,\tau}$ follows a zero-inflated negative binomial distribution

$$n_{t,\tau} \sim \mathcal{ZINB}(\lambda_{t,\tau}, \phi, \pi) \tag{8.1}$$

which has the following form

$$\begin{aligned} P(n_{t,\tau} = 0) &= \pi + (1-\pi)(1-\phi)^{\lambda_{t,\tau}} \\ P(n_{t,\tau} = k) &= (1-\pi)\binom{\lambda_{t,\tau}+k-1}{k}(1-\phi)^{\lambda_{t,\tau}}\phi^k, \quad k > 0 \end{aligned} \tag{8.2}$$

where the mean $\lambda_{t,\tau}$ is given by

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau \tag{8.3}$$

μ is a constant and α_t and β_τ are random effects with an auto-regressive structure

$$\begin{aligned} \alpha_t &\sim \alpha_{t-1} + \mathcal{N}(0, \eta_\alpha) \\ \beta_\tau &\sim \beta_{\tau-1} + \mathcal{N}(0, \eta_\beta) \end{aligned} \tag{8.4}$$

Parameters are fit using Bayesian methods, and values of $n_{t,\tau}$ are estimated using sampling. The total number of cases at week t is then given by

$$n_t = \sum_{\tau} n_{t,\tau} \tag{8.5}$$

We use the first twenty weeks of data in 2012 for training only, and begin generating estimates in epidemiological week 21 in 2012, which began on Sunday 20th May 2012. The model is fit to the data again every week, using all data available from the start of 2012 until week t. The same approach is used for all of the following models, which work exactly as in Chapter 5 but are now based on this baseline model.

Google (Dengue). This model is the same as the baseline model, with data on Google searches related to the topic of *dengue* added as an external regressor. The mean $\lambda_{t,\tau}$ is now calculated as

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau + \log(G_t^d) \tag{5.6}$$

where G_t^d is the volume of Google searches related to *dengue* in week t.

Twitter. This model is the same as the baseline model, with data on the volume of tweets that express personal experience of dengue added as an external regressor. The mean $\lambda_{t,\tau}$ is now calculated as

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau + \log(T_t) \tag{5.7}$$

where T_t is the volume of Twitter posts in week t.

Google (Dengue) + Twitter. This model is the same as the baseline model, with data on Google searches related to the topic of *dengue* and the volume of tweets that express personal experience of dengue added as external regressors. The mean $\lambda_{t,\tau}$ is now calculated as

$$\log(\lambda_{t,\tau}) = \mu + \alpha_t + \beta_\tau + \log(G_t^d) + \log(T_t) \tag{5.8}$$

where G_t^d is the volume of Google searches related to *dengue* and T_t is the volume of Twitter posts in week t.

8.2.2 Model evaluation

Once the time window is fixed, all the methods described in Chapter 3 are suitable to compare models that refer to the same city. Of these methods, the Mean Absolute Percentage Error (MAPE) and the Logarithmic Error (LOG(Q)) would also be suitable for considering models that refer to different cities. These metrics consider ratios between estimated and complete observed values and hence are already scale-free metrics. This is important for metrics that compare performance in different cities. Cities of different size probably have a comparably different dengue incidence, so to compare the performance of a model in cities of different size, metrics will have to be normalised. For the reasons presented in Chapter 3, we will not use the MAPE. Furthermore, since we used the Mean Absolute Error (MAE) and Mean Prediction Interval (MPI) in Chapter 5 for Rio de Janeiro, we wish to be able to compare the results for the cities we consider here using those metrics.

For this reason, here we introduce two new metrics that are simple extensions of those we have already used before, the MAE and MPI. As we have mentioned before, the MAE has a scale, and to be able to compare the MAEs of different cities, in the same time window, we need to normalise this scale.

nMAE. We define the normalised Mean Absolute Error as

$$\text{nMAE} = \frac{\sum_{i=1}^{N} |\hat{y}_i - y_i|}{\sum_{i=1}^{N} y_i} \tag{8.6}$$

where $\boldsymbol{y}$ are the true values and $\hat{\boldsymbol{y}}$ are the model's estimates. This is just the MAE divided by the mean value of y in the period of analysis. We note that this definition is equivalent to that of a weighted Mean Absolute Percentage Error, where the weights are the actual values of the time series y_i.

Similarly, it is possible to extend the definition of the MPI to create a normalised metric.

nMPI. We define the normalised Mean Prediction Interval as

$$\text{nMPI} = \frac{\sum_{i=1}^{N} \left(\hat{y}^i_{97.5\%} - \hat{y}^i_{2.5\%}\right)}{\sum_{i=1}^{N} y_i} \tag{8.7}$$

where $\boldsymbol{y}$ are the true values and $\hat{\boldsymbol{y}}$ are the model's estimates. Again, this is

just the MPI divided by the mean value of y in the period of analysis.

We use these two metrics to compare models across cities in the following analysis.

8.3 Results

To present the results for the locations we consider in this chapter, we follow the approach we used in previous chapters. We first compare the estimate errors of the different models using a normalised version of the MAE, described in Section 8.2.2. When considering any single city, the ratio between the normalised MAE of any model and that of the baseline model still retains its meaning of relative MAE as it is used in previous chapters. In making this comparison, it is useful to keep in mind the results we obtained for the city of Rio de Janeiro and use them to make a more informed comparison. For this reason, we also report metrics relating to Rio de Janeiro. When looking at these metrics, we need to keep in mind that for Rio de Janeiro they have been calculated for a slightly shorter time interval. Nevertheless, the normalised metrics allow for a meaningful comparison even in this situation.

Figure 8.5 shows a comparison of the accuracy of the various models for the various cities. We also report these results in Table 8.2.

For all cities and all models, we report the value of the normalised MAE, which allows us to compare the performance of models in the same city but also across different cities. Only two cities among those we analyse have a normalised MAE for the baseline model that is smaller than that of Rio de Janeiro. The nMAE of the baseline model is 0.104 for the city of Foz do Iguaçu and 0.238 for the city of Toledo, while the baseline model in Rio de Janeiro has an nMAE of 0.335. Comparing this information with the delay profiles in Figure 8.2, we can see that the cities of Foz do Iguaçu and Toledo show a fraction of cases already known in the earliest week which is higher than that of Rio de Janeiro. Specifically, for the city of Toledo 75% of cases are known in week 0, 68% in Foz do Iguaçu and 25% in Rio de Janeiro.

We note that in Figure 8.2, other cities such as Londrina, Paranaguá or Umuarama also exhibit a higher median fraction of cases entered into the system with zero weeks delay than Rio de Janeiro (blue line). From Table 8.2 we see that for Umuarama, the normalised MAE of the baseline model is 0.390, not much higher than for Rio de Janeiro. For the cities of Londrina we observe an nMAE of the baseline model

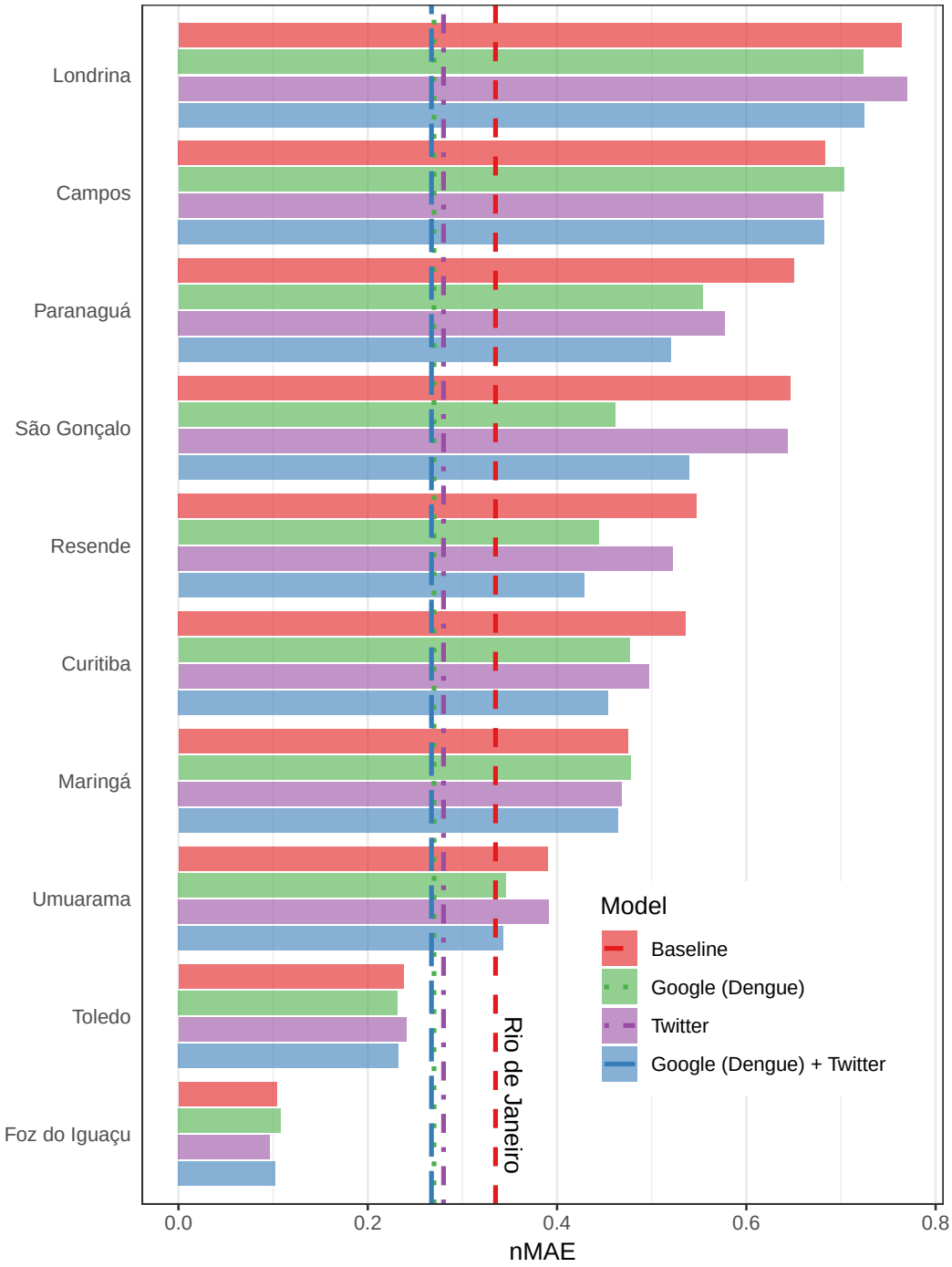

Figure 8.5: Accuracy of dengue nowcasting models using Google and Twitter data compared to the baseline model for all cities. Here we show the normalised Mean Absolute Error (nMAE) for all models and all cities. The normalised MAE is defined as the MAE of a model divided by the mean value of the time series in the same period of time, as defined in (8.6). *(continues on the following page)*

Figure 8.5: *(continues from previous page)* The dashed lines indicate the values for the city of Rio de Janeiro, included for comparison. The normalised MAE allows us to compare the errors to those in Rio de Janeiro. We can see that in most of the cities the normalised MAE is greater than the normalised MAE in Rio de Janeiro. In the cases in which it is smaller, the models using online data do not seem to be better than the baseline model. Furthermore, we can see that the accuracy of models using online data is generally higher than the respective baseline models, and when it is not higher, it is comparable to the baseline models.

Table 8.2: Accuracy of dengue nowcasting models using Google and Twitter data compared to the baseline model for all cities. Here we report the normalised Mean Absolute Error (nMAE) for all models and all cities. The normalised MAE is defined as the MAE of a model divided by the mean value of the time series in the same period of time, as defined in (8.6).

City	*Baseline*	*Google*	*Twitter*	*Google + Twitter*
Rio de Janeiro	0.335	0.270	0.280	0.267
Londrina	0.764	0.724	0.770	0.724
Campos	0.684	0.703	0.681	0.682
Paranaguá	0.651	0.554	0.577	0.520
São Gonçalo	0.647	0.461	0.644	0.539
Resende	0.547	0.444	0.522	0.429
Curitiba	0.536	0.477	0.497	0.454
Maringá	0.475	0.478	0.468	0.464
Umuarama	0.390	0.346	0.391	0.343
Toledo	0.238	0.231	0.241	0.232
Foz do Iguaçu	0.104	0.108	0.097	0.102

of 0.764 and for Paranaguá it is 0.651. These values are at least twice as great than for Rio de Janeiro. In other cities such as Maringá or Curitiba, around 25% of cases are entered into the system at week 0, which is comparable with Rio de Janeiro.

We suggest that the main reason for which their normalised MAEs are much higher than that of Rio de Janeiro might be the variability of the known fraction of cases at week 0 in these cities, rather than in the median fraction itself. The clearest example is the comparison between the normalised MAEs for the cities of Paranaguá and Toledo. For both cities, the median fraction of cases entered into the system at week 0 is around 75%. By week 1, a median of 100% of the cases have been entered into the system, suggesting that, on average, all cases are already in the system after only one week's time. Looking at the respective delay curves in Figure

8.2, however, it seems clear why the performance of the baseline model in the two cities is so different. Recall that in Figure 8.2 the dark blue area represents 80% of the weeks in our period of analysis where cases are entered into the system after the respective delay from the date of notification, while the light blue area represent 90% of such weeks. Thus, the lower edge of the dark blue area marks the lower 10% of the distribution while the lower edge of the light blue area marks the lower 2.5% of the distribution.

In the city of Toledo, we note that in 90% of the weeks the fraction of cases known at week 0 is higher than 20%, and in 97.5% it is more than zero. From Figure 8.2, looking at the city of Paranaguá, we see that only in less than 90% of weeks the fraction of cases known at week 0 is greater than zero since the lower edge of the dark blue area is at zero in week 0. This means that in more than 10% of the weeks, there is no information in week 0. Looking at week 1, instead, for the city of Toledo we see that while on average we have complete information, in 97.5% of the weeks, we have at least about 35% of the information, and in 90% of the weeks at least 75% of information. In the city of Paranaguá, the situation is much worse. In fact at the end of week 1 in 10% of the weeks we have less than about 10% of the information, and in some weeks we might still not have any information. These results show that, although two cities might exhibit a similar median delay curve, i.e. a similar fraction of cases entered into the system as time progresses, the variability of such delay curve plays a much more important role in determining the performance of a baseline model. The higher the variability, the poorer the accuracy.

Because of the high variability of delay curves, for most of the other cities we consider we observe an accuracy of the baseline model that in general is much lower than for Rio de Janeiro, Toledo or Foz do Iguaçu. Figure 8.5 shows that for some cities there is a marginal advantage in using online data, with reductions in the MAE of as much as about 5%. For Londrina the nMAE of the baseline model is 0.764 while the nMAE of the *Google (Dengue) + Twitter* model is 0.724, a reduction in MAE of just 5.3%. For Maringá the nMAE of the *Google (Dengue) + Twitter* model is 0.464, only 2.3% lower than the the nMAE of baseline model which is 0.475. For Campos, Foz do Iguaçu and Toledo the improvement in MAE while using online data is similarly small.

For other cities, instead, we can see improvements that are greater. For example, for Paranaguá the nMAE of the baseline model is 0.651 while the nMAE of the *Google (Dengue) + Twitter* model is 0.520, 20% smaller than the baseline model. For

Resende the nMAE of the baseline model is 0.547 while the nMAE of the *Google (Dengue) + Twitter* is 0.429, 21.6% smaller. In São Gonçalo the nMAE of the baseline and *Google (Dengue)* models are 0.647 and 0.461 respectively, an increase in accuracy of 28.8%.

Recall that for Rio de Janeiro the *Google (Dengue) + Twitter* is 20% more accurate than the baseline model. Here we see that, while for some cities the improvement is minimal, such as for Campos, Foz do Iguaçu, Toledo, Maringá and Londrina, for other cities, such as Paranaguá, Resende and São Gonçalo, the improvement in accuracy is comparable to or even higher than that for the city of Rio de Janeiro.

In conclusion, we can see that the variability of the delay curve indicating how quickly official data are progressively obtained over time plays a crucial role in the accuracy of estimates generated. Online data may help to improve the accuracy of the model considerably, but if the official data are too delayed, the baseline model will display errors that are too big for online data to compensate.

A similar analysis can be done for precision. Figure 8.6 shows a comparison of the precision of the various models and the various cities we consider. Results on precision are also reported in Table 8.3. For all cities and all models, we report the value of the normalised MPI, which allows us to compare the performance of models in the same city but also across different cities. As for the case of the normalised MAE, when considering any single city, the ratio between the normalised MPI of any model and that of the baseline model still reflects the relative MPI as defined in previous chapters.

Here the situation is a bit more complicated than in the case of normalised MAE. In 5 out of 10 cities (Curitiba, Londrina, Campos, São Gonçalo and Resende) , the baseline models produce prediction intervals that are not able to capture the appropriate number of points, and in general, using online data does not solve this problem. A possible explanation of this might be found in the high variability of the delay curves in these cities. In Figure 8.2 we can see that for Curitiba, Londrina, Campos, São Gonçalo and Resende the dark blue area representing 80% of the distribution around the median is very wide. Of course, this is not the only possible explanation, and likely there are multiple combined effects. Another possible concurring effect might be sought in how the delays are distributed over time, if they are close to or far from epidemic periods. Recall that, as we have noted in Figure 5.3, this is a possible explanation for poor performance of the baseline

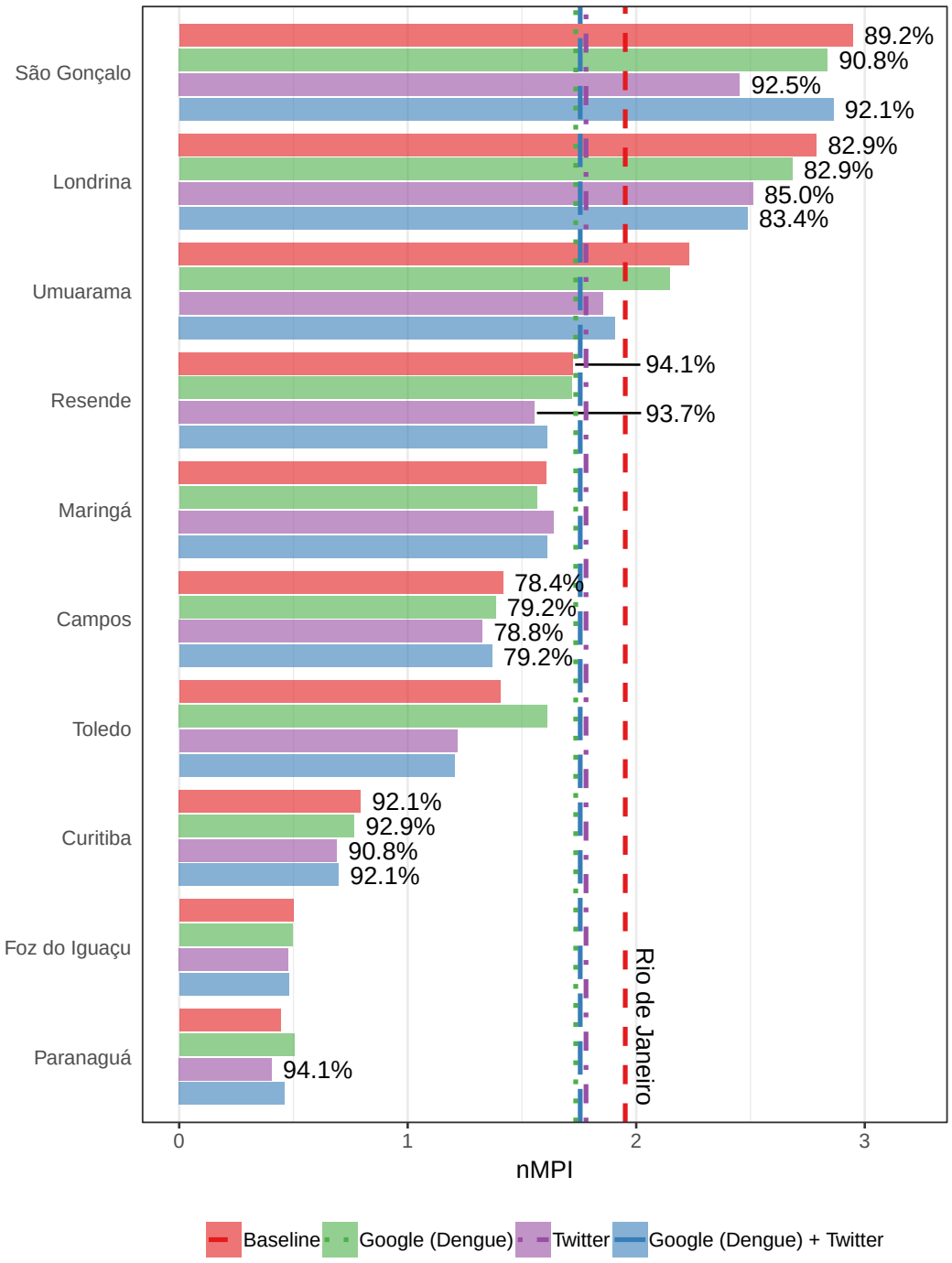

Figure 8.6: Precision of dengue nowcasting models using Google and Twitter data compared to the baseline model for all cities. Here we show the normalised Mean Prediction Interval (nMPI) for all models and all cities. The normalised MPI is defined as the MPI of a model divided by the mean value of the time series in the same period of time, as defined in (8.7). *(continues on the following page)*

Figure 8.6: *(continues from previous page)* The dashed values indicate the values for the city of Rio de Janeiro, included for comparison. The normalised MPI allows us to compare the errors to those in Rio de Janeiro. For the models that have 95% prediction intervals that include less than 95% of the points, we report this percentage on the right of the bar. We can see that for all cities the MPI of at least one of the models using online data is smaller than the baseline model. Furthermore, we see that in general if for a city the baseline model's 95% prediction intervals contain the complete notified dengue case counts in less than 95% of weeks, this is also the case for models enhanced with online data. However, we observe that, for models using online data, the 95% prediction intervals generally contain the complete notified dengue case counts in more weeks than for the baseline model.

Table 8.3: Precision of dengue nowcasting models using Google and Twitter data compared to the baseline model for all cities. Here we report the normalised Mean Prediction Interval (nMPI) for all models and all cities. The normalised MPI is defined as the MPI of a model divided by the mean value of the time series in the same period of time, as defined in (8.7).

City	*Baseline*	*Google*	*Twitter*	*Google + Twitter*
Rio de Janeiro	1.952	1.735	1.780	1.755
São Gonçalo	2.947	2.837	2.453	2.863
Londrina	2.789	2.684	2.512	2.487
Umuarama	2.232	2.147	1.853	1.908
Resende	1.723	1.719	1.556	1.610
Maringá	1.608	1.564	1.638	1.610
Campos	1.418	1.384	1.326	1.368
Toledo	1.405	1.611	1.217	1.205
Curitiba	0.795	0.766	0.690	0.698
Foz do Iguaçu	0.500	0.495	0.478	0.479
Paranaguá	0.443	0.503	0.405	0.462

model in Rio de Janeiro in 2016.

For some cities, this misbehaviour of the prediction intervals comes hand-in-hand with notably impaired accuracy of the estimates. For example, for the city of Campos (baseline nMAE = 0.648, nMPI = 1.418, 78.4% points in the 95% prediction interval), the accuracy of the baseline model is much worse than that for Rio de Janeiro (baseline nMAE = 0.335, nMPI = 1.952, 95.0% points in the 95% prediction interval), and although the normalised MPI is smaller for Campos than for Rio de Janeiro, the prediction intervals for Campos are not able to capture the appropriate number of points. We also note that the addition of online data produces

only a marginal improvement in accuracy and precision. As another example, for the city of Londrina (baseline nMAE = 0.764, nMPI = 2.789, 82.9% points in the 95% prediction interval; *Google (Dengue) + Twitter* nMAE = 0.724, nMPI = 2.487, 83.4% points in the 95% prediction interval), we can see that using online data does, in fact, provide an advantage in terms of accuracy and precision. However, the normalised MAE and MPI remain much higher than those for Rio de Janeiro, and the prediction interval does not capture 95% of the true data points.

Comparing Figure 8.5 and 8.6 we observe that for all the cities we considered in this study there is at least one model using online data that has lower mean absolute error and mean prediction interval than the respective baseline model. Furthermore, we also observe that for all cities where the percentage of true data points within the 95% prediction interval of the baseline model is lower than 95%, the percentage of true data points within the 95% prediction interval is instead higher for models using online data. This confirms once again the advantage of using online data together with official data to make estimates of the current dengue case count also in cities other than Rio de Janeiro.

In drawing these conclusions we must bear in mind the limitations of our data. Recall that Google search volumes are at the state level rather than at the city level. This means that in some years Google search volumes might not be representative of the weekly number of notified dengue case (thus not correlated with it). In such cases, when the correlation between official data and Google data is not very high, the model using Google data probably reduces to the baseline model. This might be the reason why in some cases we observe comparable performance in models using official data alone and models using official and online data.

Furthermore, online data represent only a subset of the actual population, i.e. internet user, and more specifically in our case people using Google to make searches and Twitter to interact with their friendship network. This is definitely yet another source of error in the models using online data. In Figure 8.5, for example, we observe for the cities of Campos, Maringá, Toledo and Foz do Iguacu comparable MAEs when considering the baseline model or models also using online data. In these cases, the variability in the delays we observe in Figure 8.2 is much higher for Campos than for the other cities, while correlation with the Google search volume we observe in Figure 8.3 is much lower for Campos than for the other cities. This has a clear impact on the baseline model, in fact we observe a much higher normalised MAE for Campos, but the relative improvement of using online data is comparable

in all four cities.

As specified in Section 3.1.3, these limitations often exist and cannot be removed. The only thing we can do is be aware of them and critically analyse our results correspondingly.

8.4 Discussion

In this chapter, we explored whether and how the methods that we developed in Chapter 5 for Rio de Janeiro can be applied to other Brazilian cities of different size. We observed that the population size, while it might affect the performance of our models, does not seem to be the most important factor.

We observed that in most of the cities we considered the variability of the notification delay curve is much higher compared to Rio de Janeiro. Only a few cities in Figure 8.2 have a dark blue area smaller or comparable to that of Rio de Janeiro. Recall that the dark blue area represents 80% of the weeks in our period of analysis where cases are entered into the system after the respective delay from the date of notification, while the light blue area represent 90% of such weeks.

While the number of confirmed dengue cases is generally well correlated with Google search volumes of the topic *dengue* in most of the municipalities we considered, the correlation is usually smaller than the one observed in Rio de Janeiro. This is reflected in our modeling results. We observe that in general, across all cities, using the number of Twitter posts as an external regressor does not provide any considerable improvement with respect to the baseline model, but using Google search volumes or both Google search volumes and the number of Twitter posts in tandem, instead, may provide a relatively higher improvement. In all the cities we considered, the estimates produced with models using online data are either more accurate or, in the worst cases, comparable to the baseline model.

This is due to the fact that online data are not well correlated with official data every year. This is especially true when considering Google search volumes because we obtain such data at the state level but our analysis is at the city level, as official and Twitter data are. For this reason, it might sometimes happen that the trend of the weekly number of dengue cases for a city is considerably different from that of the state the city is in. In these period the contribution of Google data in the model

becomes negligible for that city. The model using Google data essentially reduces to the baseline model in those cases.

This is why we highlight once more that results must be critically evaluated, and we must be aware that the fact that in general there might be an advantage in using online data does not mean that it would always be the case.

We also observe a shrinking of the prediction intervals in all cities for models using online data. In the few cases where we do not observe a shrinking of the prediction intervals, the prediction intervals are comparable to those of the respective baseline models. Together with a reduction of the mean prediction interval, in general, we also observe an increase in the percentage of points within the 95% prediction interval when it is lower than 95% for the baseline model, such as for Curitiba, Londrina and São Gonçalo, Campos and Resende.

From these results we can draw three main conclusions.

First, the variability of the reporting delay seems to crucially affect the accuracy of our models, but it might be more correct to say that it is the factor that mostly affect the accuracy of our models for which we have data. In fact, in terms of data that are used, our model is very simple. We only consider official data, i.e. notified dengue cases, Twitter posts and Google search volumes. A lot of other data and important information are left out of this analysis.

For example:

- To make our model more accurate we should consider the suspected cases of dengue that were later removed by this data set because laboratory analyses did not confirm them to be cases of dengue. This would surely affect our precision, but at the moment, because there is no backlogging information, we cannot possibly assess the impact of these changes.
- We mentioned in Section 3.1.3 that another problem of official data is under-reporting. This means that not all people affected by dengue do actually go to a clinic or to the hospital. This means that when we estimate the number of notified cases, we still do not have a complete picture of the situation. On the other hand, these people might instead search for information about dengue on Google or write about it on Twitter.
- Unfortunately, data is not available at a finer geographical scale. If we could

have data about where a case is notified we could have a much more accurate idea of what is happening in the city, in particular for very large cities where the situation could be different in boroughs that are far from each other.

- We are not including information about temperature, humidity, closeness to water.

All of these data would make our models probably more accurate but surely more complex. Adding further data would certainly be a valuable direction for improvement, but in terms of the model that we analyse in this study, the estimates we produce can be considered a first approximation of the notified dengue case counts.

Second, online data are quick to retrieve and can be easily implemented building on the baseline model with a negligible increase in complexity. Furthermore, online data generally provide more accurate and more precise estimates. Nevertheless, it is important to be aware of the limitations of the online data we use and to always be critical in examining and comparing estimates produced including them. At the current state, online data do not affect estimates the same way for all cities. In bigger cities, cities with a higher rate of internet users, cities where the subset of internet users is more representative of the whole population, these methods might prove more useful. For what concerns Google data more specifically, cities in which the Google search volume for *dengue* is more similar to that of the entire state might be more positively affected. This is something that we should be aware now, but that might probably become differently relevant in the future if Google decides to make data available at the city level or even at a finer scale.

Third, our sample of cities for this study has been chosen to cover the entire size range in terms of population. From our results we can observe that the size of a city is not a crucial factor in determining the performance of our nowcasting models, but there might be other characteristics of a city that might be influential. For example, the number of clinics, the number of personnel in each of those, the easiness for people of getting to these places or the level of education of patients might influence their decision either to go or not to go or when to go to a clinic when they develop the first symptoms. These and other factors such as the distance from the digitisation centres, the availability of internet connection or of computers and appropriate software might influence the time it takes for clinics to make their data available for research. External factors such as and similar to those we have listed, which could be only partially dependent from the size of a city, might help

explain what we observe in official data relating to the different cities.

The research included in this chapter could strongly benefit with the implementation of model inclusion approaches such as those discussed in Section 2.2, and in particular similarly to what Xu et al. (2017) have done in their work. Here we have shown that online data can contribute to improving estimates, but we have also shown that many factors can impact their value also depending on time and location. Future research should investigate if and how model fusion can produce better estimates.

CHAPTER 9

Conclusions

Dengue is one of the most widespread and fastest growing mosquito-borne diseases in the world. Its symptoms are usually similar to those of ordinary influenza but, unfortunately, in some cases it can be fatal (World Health Organization Regional Office for South-East Asia, 2009). More severe forms of dengue have deadly symptoms such as haemorrhagic fever, and given the high number of dengue cases, this leads to a high rate of casualties due to dengue every year. In Brazil, nearly 1.5 million cases were recorded in 2016 alone. Of these, 861 were confirmed cases of severe dengue and 8,402 were dengue cases supected to be severe. In 2016, there have been 642 confirmed deaths by dengue, nearly 7% of the suspected and confirmed severe dengue cases (Brasil. Ministério da Saúde. Secretaria de Vigilância Epidemiológica, 2017). Furthermore, since dengue is typically a seasonal disease, it is something that usually needs to be dealt with every summer season.

A vast body of research in mathematical epidemiology exists which addresses the spreading of mosquito borne diseases such as dengue using models with roots in classic SIR models. This research also includes a component describing mosquitoes, which makes these models rather complex. On the other hand, this types of models allow researchers to investigate the underlying dynamics of dengue spreading and to evaluate the effects in the long term of possible methods to mitigate it, such as vector control and vaccines. Unfortunately, these models require complex and detailed data to be validated. Such data are usually also difficult to collect, and when they are collected they become available only long after the outbreaks they relate to.

In the present thesis, we specifically considered the situation in Brazil, and we considered the city of Rio de Janeiro in particular as a case study. In Rio de

Janeiro, at the end of any given week, only 25% of the dengue cases for that week have typically been entered into the surveillance system. In the worst cases, it can take up to about six months for all dengue cases for a given week to be entered into the system. This substantial delay in the collection of information on dengue cases reports makes the validation of mathematical models impractical while an outbreak is happening. This leaves policymakers without the information they need to make informed decisions on how to prevent and mitigate dengue outbreaks.

Because of this limit in the availability of official data, it was necessary to look at alternative data sources, including online data, to improve the estimates of the weekly number of dengue cases and make them available with a short delay. Google Trends and Twitter are two of the most popular sources of online data that have been used do nowcast disease spreading by exploiting their correlation with official data. In the case of dengue, though, a correlation between the number of dengue cases and relevant Google searches as well as Twitter posts is not sufficient to generate more timely and more precise estimates because official data are delayed. Online data are correlated with the complete observed dengue case counts, but while we have full online data when we make a prediction, official data from previous weeks might be partially available or might become available at a later date. Addressing and modeling such delays is crucial from an operational point of view. Furthermore, Google and Twitter are very different types of data sources, and previous algorithms have not attempted to use them in tandem.

For this reasons, in the present work, we looked at alternative methods that take into account delays when estimating the weekly number of dengue cases. Furthermore, we tried to create a model that is operationally realistic, and that could easily be integrated into InfoDengue[1], a nowcasting system for the surveillance of dengue fever transmission in Brazil, developed by a team of researchers at the Oswaldo Cruz Foundation (Fiocruz) in Rio de Janeiro (Codeço et al., 2016).

Throughout this thesis, our analysis has been twofold. We looked for models using online data that had higher accuracy, i.e. a smaller prediction error than the baseline model using official data only, but we also looked for models that had smaller prediction intervals than the baseline model. In fact, the usefulness of nowcasting models such as those we studied is not so much in their ability to detect an outbreak, but more in giving to practitioners a clearer idea of the variability that they can expect in the number of infections. Having smaller prediction intervals means that

[1] https://info.dengue.mat.br

the model is more precise, and the more precise it is, the more useful the information about the point estimate. We put particular attention in assessing the reliability of prediction intervals, i.e. if any reduction of the prediction intervals preserved their meaning. In other words, we want to make sure that even if they shrank, 95% prediction intervals would still contain 95% of the true data points.

Our first approach was to integrate online data with a baseline adaptive nowcasting model (Chapter 4). Because this type of model is not capable to automatically account for missing official data, we developed a methodology to correct incomplete dengue case counts before using them to train our ARIMA models. With this approach, we integrated Google and Twitter data together into the same model and demonstrated that this leads to the best overall performance compared to models that use either only one of these online data sources or none at all. However, although we found that prediction errors are reduced with this type of technique, we discovered that it does not produce reliable prediction intervals since 95% prediction intervals contain considerably less than 95% of the true data points. As this is critical information disease surveillance practitioners need to rely on, we acknowledged the potential of such kind of models but chose to opt for a different type of model that can automatically deal with missing data.

To do so, we built on a model developed by our colleagues at Fiocruz (Bastos et al., 2017) which is based on a Bayesian algorithm that can automatically take into account delayed data (Chapter 5). The model we built produces estimates of the weekly number of dengue cases in Rio de Janeiro by also drawing on online data such as Google searches and Twitter posts. The model we introduced functions on a weekly basis and at city level, and has been tested in an operationally valid setup. Furthermore, because all the assumptions this model makes about data availability are valid, it could readily be deployed.

We found that data from Google Trends and Twitter describing the volume of dengue-related searches in a given week and the number of tweets expressing personal experience of dengue can be integrated in our model to improve estimates of the current number of notified dengue infections. Again, we found that using both Google Trends and Twitter data in tandem in the same model leads to the best overall performance compared to the estimates generated using only Google Trends or Twitter data together with official data, or official data alone. This improvement is accompanied by a considerable reduction of the mean absolute error of the estimates by between 16% and 21% over the entire time period considered, depending on the

particular type of online data source. We also showed that another advantage of the inclusion of online data in the model is the reduction in size of the 95% prediction intervals by about 10%, while the prediction intervals continue to contain 95% of the true data points.

Based on this key improvement, we tried to address a range of further issues that may be encountered in dengue surveillance. The first issue we considered is a delayed delivery of official data at the end of the week. In other words, sometimes it might happen that when we produce estimates about the number of dengue cases in the week that just ended, we do not have any data about cases entered into the surveillance system in that week, whether relating to that week or to previous weeks. In Chapter 6 we investigated how this would impact our model. Again, we found that this problem can be better mitigated when using online data sources. Even though official data may be missing, online data are generally immediately available as soon as they are created, without any delay. We found that in this situation there is a notable advantage in using Google and Twitter data as well. In particular, we found that a baseline model using only official data where new official data are always delivered with a delay of one week has, on average, a prediction error about 70% higher than a baseline model using only official data where new official data are never delivered with a delay. In contrast, a model using both Google and Twitter data together where new official data are always delivered with a delay of one week has on average a prediction error only 10% higher than the baseline model where new official data are never delivered with a delay.

Secondly, in Chapter 7 we built a model to generate short-term forecasts of the weekly number of dengue cases in Rio de Janeiro. The baseline model developed in Chapter 6 naturally extends to forecasts, but to use online data in this modified baseline model to forecast dengue case counts in week $t+1$ we need to know the total online activity for week $t+1$. Of course, it is not possible to have data on the total weekly online activity before week $t+1$ ends. However, during week $t+1$ we can obtain partial online data up to the day when we perform the analysis. Specifically, we can obtain daily online data both from Twitter and Google. We found that we can produce reliable estimates of the total dengue case count for week $t+1$ days in advance. To do so, we use delayed official data obtained up to week t, complete online data from Google and Twitter relating to weeks up to week t, and partial volumes of online data from Google and Twitter relating only to the first few days of week $t+1$, before we perform the analysis. In particular, predictions made with this method on the Tuesday of week $t+1$ have on average a prediction error only

about 20% higher than that of the baseline model using delayed official data only when they are released at the end of week $t + 1$.

Finally, in Chapter 8 we tested our model in different cities across the states of Paraná and Rio de Janeiro. We again found evidence that using online data from Google and Twitter produces an increase in accuracy and precision. We found that in most of the cities we analysed, using online data from Google and Twitter reduces the prediction error, and when it does not, the accuracy is comparable. However, for every city there is at least one model using online data that outperforms the baseline model. We also found that in all the cities we analysed, using online data reduces the width of prediction intervals compared to the baseline model, and that the 95% prediction intervals of models using online data generally contain a percentage of true data points which is higher than the baseline model, or comparable to that of the baseline model. This last study, thus, confirms the advantage of using online data to complement official data and shows again that using Google data and Twitter data in tandem in the same model generally produces the highest increase in accuracy and precision compared to a model using delayed official data only.

One of the risks of building nowcasting and forecasting models is that of overfitting. It would be possible to reduce the estimate error and the prediction interval considerably by having a more complex model with lots of parameters and hyperparameters that can be adjusted. The problem would then be that such models would probably not perform as well on a different time series, such as that of a different city, or of a different time window. The baseline model we have outlined in Chapter 5 is very simple, with no hyperparameters apart from the maximum delay to consider when updating the data every week. It automatically takes into account delays, and the further inclusion of both Google and Twitter online data does not require the introduction of additional hyperparameters. Nevertheless, all the models, including the baseline, can grasp the main features of the time series and strongly outperform the naive model, i.e. a model using as an estimate of the dengue case count in week t the known number of dengue cases relating to week $t - 1$.

Models using online data alone to produce estimates of disease incidence have been strongly criticised for being prone to severely incorrect predictions (Lazer et al., 2014). In this thesis we have presented a series of models that combine official and online data, using the latter to complement the former, and not to substitute them in the nowcasting process. In this way, the risk of online data leading to vastly

inaccurate predictions is highly reduced, but performance is nevertheless improved. However, all models using online data are only extensions of the baseline model which only uses delayed official data, and the choice of the specific baseline model is thus crucial. For this reason, before using online data as external regressors, it is necessary to have a robust baseline model addressing the issue of delayed official data.

The results we presented in this thesis show that it is possible to improve the accuracy and precision of these estimates by drawing on online data such as Google searches and Twitter posts at the same time. This is critical information that is needed by public health policymakers to guide their decision making process. We provide evidence that this approach can readily be transferred to other cities in Brazil that are already monitored by the InfoDengue system. Furthermore, the model that we propose could easily be transferred to other countries that have similar surveillance systems, and potentially to monitor different diseases as well.

We must remember that, at the present moment, the available data has many other limitations. Official data are a collection of suspected dengue cases, which are also severely delayed. The delay is the factor that affects our ability to produce estimates the most, but we could still not take into account the fact that official data can be modified retroactively in the actual operation of the model. This is because, unfortunately, there are not any logged data about cases that where inserted in the list first and then removed.

Secondly, we talked in Chapter 2 about the problem of under reporting. What we have been trying to predict was not the total number of dengue cases in Rio de Janeiro, but rather the total number of notified dengue cases. This means that there is an unknown portion of dengue cases that we are not considering and that would be useful to know for policymakers. Again, unfortunately this is an intrinsic limit of the official data, and it is very difficult to infer the total number of dengue cases by knowing the total number of notified dengue cases because no clear data exist about the rate of underreporting.

Official data are not the only ones with limitations. Online data also have severe limitations that need to be taken into account when discussing the results and making decisions based on them. Google Trends data about the topic *dengue* is proprietary data that Google makes available to researchers. We do not know precisely how it is aggregated, i.e. which specific search queries are included in the topic *dengue*.

We also do not know how it is sampled, and what is the absolute volume. What we have, in fact, is a sample which is scaled according to the population of the area of interest – in this case Brazil – to give an estimate of the actual number of searches. Furthermore, the finer spatial aggregation available for Google search data is at the state level, which is too large to adequately represent single cities. This means that when considering the city Rio de Janeiro, because it might contain a large proportion of the internet users of the state, we observe that there is a good correlation between the number of dengue cases in the city of Rio de Janeiro and the Google search volume about the topic *dengue* in the state of Rio de Janeiro. As we observed in Chapter 8, the same is not always true when we consider other cities in the state. Also, we have seen this to be much less common in the state of Espírito Santo. Further research could explore the use of model fusion approaches such as those discussed in Section 2.2 in this particular case, by producing averaged estimates where all the models discussed contribute with different weights depending on their relative effectiveness at each particular time.

Google data are then more valuable when we consider the city of Rio de Janeiro. In fact, we observe a considerable increase both in accuracy and precision while also using Google data. But when in Chapter 8 we consider other cities for which this correlation is not always strong, during the periods when dengue cases are not highly correlated with Google search volumes the model enhanced with Google data basically reduces to the baseline model. In these cases, adding Google search volumes to the model does not provide any significant benefit, despite making the model slightly more complex.

On the other hand, Twitter data are available at the city level and do not have the same problem Google search volumes has. Nevertheless, Twitter's policy for making their data available is also constantly changing. This makes it much more difficult to keep the models reliable over time as assumptions behind these proprietary data are always evolving and it is necessary to constantly maintain these models for them to keep working properly.

Furthermore, concerning online data, we must keep in mind that they relate to a specific and biased subset of the population, which is *internet users that actually use Google and Twitter*. We can expect that Google search is much more widespread and used than Twitter, so there might be many more users of Google search and the user types are probably much more diversified than the Twitter's user types.

We can expect that internet users are more informed, or maybe even more well educated on average than people who do not use the internet, and for this reason they might be more likely to search for symptoms on the internet, and consequently go to a clinic when they find they might have dengue. On the other hand, there might be internet users that search for dengue symptoms or who tweet about showing dengue symptoms, without then actually notifying their condition to a clinic.

Despite this strong bias in the population producing online data, it is still very difficult to quantify their size relative to people not using internet and, more importantly, people not reporting their dengue condition to a clinic. In order to achieve that, it would be necessary to build more rich official data sets where we keep track of the people who come into clinics that searched about their symptoms on Google or posted about it on Twitter, but also of people that search and tweet but then do not go to a clinic. Of course, this is very complicated and hardly practical, and for this reason we must keep this in mind when making decisions based on results obtained with this type of data.

The removal of suspected dengue cases from the official data set after laboratory analysis excluded that they were in fact dengue cases is the one thing that makes our results slightly different from those we would obtain in the actual operation of the model for disease surveillance. Having more imprecise data as a starting point, we can expect that in the real operation of our model we could have slightly worse performances both in terms of accuracy and precision. Keeping track of which cases and when they are removed from the list of suspected dengue cases could allow us to make more accurate estimates of the performance of our model. Also, it would allow us to make analyses based on results that would be much more similar to those we would obtain in an actual operational setting.

The number of cities monitored by the InfoDengue system continues to increase. The research carried out in this thesis provides solid evidence that the inclusion of online data in the InfoDengue nowcasting model, and the implementation of the methods discussed in this thesis, could positively affect the performance of the surveillance system. Crucially, the approaches we have described are all based on realistic assumptions about data availability. As a result, these methods could be easily and rapidly deployed. Furthermore, it could easily be transferred to monitor other disease in Brazil but also in other parts of the world with similar surveillance systems.

References

Aguiar M., Ballesteros S., Kooi B.W. and Stollenwerk N. (2011). The role of seasonality and import in a minimalistic multi-strain dengue model capturing differences between primary and secondary infections: Complex dynamics and its implications for data analysis. *Journal of Theoretical Biology*, 289:181–196.

Aguiar M. and Stollenwerk N. (2017a). Dengvaxia: Age as surrogate for serostatus. *The Lancet Infectious Diseases*, 18(3):245.

Aguiar M. and Stollenwerk N. (2017b). Mathematical models of dengue fever epidemiology: multi-strain dynamics, immunological aspects associated to disease severity and vaccines. *Communication in Biomathematical Sciences*, 1(1):1–12.

Akaike H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716.

Alanyali M., Moat H.S. and Preis T. (2013). Quantifying the relationship between financial news and the stock market. *Scientific Reports*, 3:3578.

Alanyali M., Preis T. and Moat H.S. (2016). Tracking protests using geotagged Flickr photographs. *PLOS ONE*, 11(3):e0150466.

Alis C.M., Letchford A., Moat H.S. and Preis T. (2015a). Estimating Tourism Statistics with Wikipedia Page Views. In *Proceedings of the ACM Web Science Conference*, WebSci '15, pages 33:1–33:2. ACM, New York, NY, USA.

Alis C.M., Lim M.T., Moat H.S., Barchiesi D., Preis T. and Bishop S.R. (2015b). Quantifying regional differences in the length of Twitter messages. *PLOS ONE*, 10:e0122278.

Allard R. (1998). Use of time-series analysis in infectious disease surveillance. *Bulletin of the World Health Organization*, 76(4):327–333.

Althouse B.M., Ng Y.Y. and Cummings D.A.T. (2011). Prediction of dengue incidence using search query surveillance. *PLOS Neglected Tropical Diseases*, 5(8):e1258.

Asmaidi P. and Sianturi N.E.H. (2014). A SIR Mathematical Model of Dengue Transmission and Its Simulation. *IOSR Journal of Mathematics*, 10(5):56–65.

Balcan D., Colizza V., Gonclaves B., Hu H., Ramasco J.J. and Vespignani A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489.

Barchiesi D., Moat H.S., Alis C., Bishop S. and Preis T. (2015a). Quantifying international travel flows using Flickr. *PLOS ONE*, 10(7):e0128470.

Barchiesi D., Preis T., Bishop S. and Moat H.S. (2015b). Modelling human mobility patterns using photographic data shared online. *Royal Society Open Science*, 2(8):150046.

Bastos L., Economou T., Gomes M., Villela D., Bailey T. and Codeço C.T. (2017). Modelling reporting delays for disease surveillance data. arXiv:1709.09150.

Beveridge S. and Nelson C.R. (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'. *Journal of Monetary Economics*, 7:151–174.

Bhatt S., 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., William Wint G.R., 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.

Bollen J., Mao H. and Zeng X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Botta F., Moat H.S. and Preis T. (2015). Quantifying crowd size with mobile phone and Twitter data. *Royal Society Open Science*, 2(5):150162.

Brady O.J., Gething P.W., Bhatt S., Messina J.P., Brownstein J.S., Hoen A.G., Moyes C.L., Farlow A.W., Scott T.W. and Hay S.I. (2012). Refining the Global Spatial Limits of Dengue Virus Transmission by Evidence-Based Consensus. *PLOS Neglected Tropical Diseases*, 6(8):e1760.

Brasil. Ministério da Saúde. Secretaria de Vigilância Epidemiológica (2017). Monitoramento dos casos de dengue, febre de chikungunya e febre pelo vírus Zika até a Semana Epidemiológica 52, 2016. *Boletim Epidemiológico*, 48(3):1–11.

Burnham K.P. and Anderson D.R. (2004). Multimodel inference: Understanding AIC and BIC in model selection.

Campbell J.Y. and Mankiw N.G. (1987a). Are Output Fluctuations Transitory? *The Quarterly Journal of Economics*, 102:857–880. arXiv:1011.1669v3.

Campbell J.Y. and Mankiw N.G. (1987b). Permanent and Transitory Components in Macroeconomic Fluctuations. *The American Economic Review*, 77(2):111–117.

Carvalho S.A., da Silva S.O. and Charret I.d.C. (2019). Mathematical modeling of dengue epidemic: control methods and vaccination strategies. *Theory in Biosciences*.

Chan E.H., Sahai V., Conrad C. and Brownstein J.S. (2011). Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLOS Neglected Tropical Diseases*, 5(5):e1206.

Chanprasopchai P., Tang I.M. and Pongsumpun P. (2018). SIR Model for Dengue Disease with Effect of Dengue Vaccination. *Computational and Mathematical Methods in Medicine*.

Choi H. and Varian H. (2012). Predicting the Present with Google Trends. *Economic Record*, 88(S1):2–9.

Clark P.K. (1987). The Cyclical Component of U. S. Economic Activity. *The Quarterly Journal of Economics*, 102(4):797–814.

Cleger-Tamayo S., Fernández-Luna J.M. and Huete J.F. (2012). On the use of weighted mean absolute error in recommender systems. In *CEUR Workshop Proceedings*, volume 910, pages 24–26.

Codeço C., Cruz O., Riback T.I., Degener C.M., Gomes M.F., Villela D., Bastos L., Camargo S., Saraceni V., Lemos M.C.F. and Coelho F.C. (2016). Info-Dengue: a nowcasting system for the surveillance of dengue fever transmission. bioRxiv:10.1101\046193.

Conte R., Gilbert N., Bonelli G., Cioffi-Revilla C., Deffuant G., Kertesz J., Loreto V., Moat S., Nadal J.P., Sanchez A. et al. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics*, 214(1):325–346.

Copeland P., Romano R., Zhang T., Hecht G., Zigmond D. and Stefansen C. (2013). Google Disease Trends: An update. In *International Society of Neglected Tropical Diseases*, volume 2013, page 3.

Curme C., Preis T., Stanley H.E. and Moat H.S. (2014). Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences*, 111(32):11600–11605.

Curme C., Zhuo Y.D., Moat H.S. and Preis T. (2017). Quantifying the Diversity of News Around Stock Market Moves. *Journal of Network Theory in Finance*, 3:1.

Davidson M.W., Haim D.A. and Radin J.M. (2015). Using Networks to Combine "Big Data" and Traditional Surveillance to Improve Influenza Predictions. *Scientific Reports*, 5(8154):8154.

Dayama P. and Kameshwaran S. (2013). Predicting the Dengue Incidence in Singapore using Univariate Time Series Models. *AMIA Annual Symposium Proceedings*, 2013:285–92.

Deyle E.R., Maher M.C., Hernandez R.D., Basu S. and Sugihara G. (2016). Global environmental drivers of influenza. *Proceedings of the National Academy of Sciences of the United States of America*, 113(46):13081–13086.

Endy T.P. (2014). Human immune responses to dengue virus infection: Lessons learned from prospective cohort studies. *Frontiers in Immunology*, 5:183.

Engle R.F., Lilien D.M. and Robins R.P. (1987). Estimating Time Varying Risk Premia in the Term Structure: The Arch-M Model. *Econometrica*, 55(2):391–407.

Faust J., Gilchrist S., Wright J.H. and Zakrajšek E. (2013). Credit spreads as predictors of real-time economic activity: A bayesian model-averaging approach. *Review of Economics and Statistics*, 95(5):1501–1519.

Galvao P.R., Ferreira A.T., Maciel M.D., De Almeida R.P., Hinders D., Schreuder P.A. and Kerr-Pontes L.R. (2008). An evaluation of the Sinan health information system as used by the Hansen's disease control programme, Pernambuco State, Brazil . *Leprosy Review*.

Gantz J. and Reinsel D. (2011). Extracting Value from Chaos State of the Universe: An Executive Summary. Technical report.

Ginsberg J., Mohebbi M.H., Patel R.S., Brammer L., Smolinski M.S. and Brilliant L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Goel S., Hofman J.M., Lahaie S., Pennock D.M. and Watts D.J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences*, 107(41):17486–17490.

Gomide J., Veloso A., Meira W., Almeida V., Benevenuto F., Ferraz F. and Teixeira M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. *Proceedings of the ACM WebSci'11, June 14-17 2011, Koblenz, Germany.*, pages 1–8. arXiv:1010.3003v1.

Grajalez C.G., Magnello E., Woods R. and Champkin J. (2013). Great moments in statistics. *Significance*, 10(6):21–28.

Greene W.H. (1994). Accounting for excess zeros and sample selection in Poisson and Negative Binomial regression models. *NYU Working Paper No. EC-94-10*.

Hamilton J.D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57(2):357–384.

Harvey A.C. (1985). Trends and Cycles in Macroeconomic Time Series. *Journal of Business & Economic Statistics*, 3(3):216–227.

Hii Y.L., Zhu H., Ng N., Ng L.C. and Rocklöv J. (2012). Forecast of Dengue Incidence Using Temperature and Rainfall. *PLOS Neglected Tropical Diseases*, 6(11):e1908.

Hoeting J.A., Madigan D., Raftery A.E. and Volinsky C.T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401.

Hyndman R., Athanasopoulos G., Bergmeir C., Caceres G., Chhay L., O'Hara-Wild M., Petropoulos F., Razbash S., Wang E. and Yasmeen F. (2018). forecast: Forecasting functions for time series and linear models.

Hyndman R.J. and Athanasopoulos G. (2013). *Forecasting: Principles and Practice*. ISBN 978-0-98-75071-0-5.

Hyndman R.J. and Koehler A.B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Imai C., Armstrong B., Chalabi Z., Mangtani P. and Hashizume M. (2015). Time series regression model for infectious disease and weather. *Environmental Research*, 142:319–327.

Johnston J. (1963). *Econometric Methods*. McGraw-Hill, New York.

Kennedy P. (2008). *A guide to econometrics*. Wiley-Blackwell, 6th edition.

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

Kermack W.O. and McKendrick A.G. (1932). Contributions to the Mathematical Theory of Epidemics. II. The Problem of Endemicity. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 138(834):55–83.

Kermack W.O. and McKendrick A.G. (1933). Contributions to the mathematical theory of epidemics-III. Further studies of the problem of endemicity. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 141(843):94–122.

King G. (2011). Ensuring the data-rich future of the social sciences. *Science*, 331(6018):719–721.

Kooi B.W., Aguiar M. and Stollenwerk N. (2014). Analysis of an asymmetric two-strain dengue model. *Mathematical Biosciences*, 248:128–139.

Kramer A.D.I., Guillory J.E. and Hancock J.T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.

Kristoufek L., Moat H.S. and Preis T. (2016). Estimating suicide occurrence statistics using Google Trends. *EPJ Data Science*, 5:32.

Lambert D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lampos V., Miller A.C., Crossan S. and Stefansen C. (2015). Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports*, 5:12760.

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.

Lazer D., Pentland A., Adamic L., Aral S., Barabasi A.L., Brewer D., Christakis N., Contractor N., Fowler J., Gutmann M., Jebara T., King G., Macy M., Roy D. and Van Alstyne M. (2009). Computational social science. *Science*, 323(5915):721–3.

Letchford A., Preis T. and Moat H.S. (2016). Quantifying the search behaviour of different demographics using Google Correlate. *PLOS ONE*, 11:e0149025.

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 Janeiro, Brazil. *The American Journal of Tropical Medicine and Hygiene*, 79(6):933–939.

Makridakis S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527–529.

Manyika J., Chui M., Madgavkar A. and Lund S. (2017). What's now and next in analytics, AI, and automation. *McKinsey Global Institute*.

Marques-Toledo C.A., Degener C.M., Vinhal L., Coelho G., Meira W., Codeço C. and Teixeira M.M. (2017). Dengue prediction by the web: tweets are a sueful tool for estimating and forecasting dengue at country and city level. *PLOS Neglected Tropical Diseases*, page Forthcoming.

McLaren N. and Shanbhogue R. (2011). Using internet search data as economic indicators. *Quarterly Bulletin of the Bank of England*, 2011(Q2):134–140.

Meloni S., Perra N., Arenas A., Gómez S., Moreno Y. and Vespignani A. (2011). Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific Reports*, 1:62.

Mizzi G., Preis T., Bastos L., Gomes M.F.C., Codeço C.T. and Moat H.S. (in preparation). Tracking dengue in Rio de Janeiro using Google and Twitter: an operationally realistic approach.

Moat H.S., Curme C., Avakian A., Kenett D.Y., Stanley H.E. and Preis T. (2013). Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Scientific Reports*, 3:1801.

Moat H.S., Olivola C.Y., Chater N. and Preis T. (2016). Searching Choices: Quantifying Decision-Making Processes Using Search Engine Data. *Topics in Cognitive Science*, 8:685–696.

Moat H.S., Preis T., Olivola C.Y., Liu C. and Chater N. (2014). Using big data to predict collective behavior in the real world. *Behavioral and Brain Sciences*, 37(1):92–93.

Nadini M., Sun K., Ubaldi E., Starnini M., Rizzo A. and Perra N. (2018). Epidemic spreading in modular time-varying networks. *Scientific Reports*, 8:2352. 1710.01355.

Nelson C.R. and Plosser C.R. (1982). Trends and random walks in macroeconmic time series. Some evidence and implications. *Journal of Monetary Economics*, 10:139–162.

Noguchi T., Stewart N., Olivola C.Y., Moat H.S. and Preis T. (2014). Characterizing the time-perspective of nations with search engine query data. *PLOS ONE*, 9(4):e95209.

Nuraini N., Soewono E. and Sidarto K.A. (2007). Mathematical model of dengue disease transmission with severe DHF compartment. *Bulletin of the Malaysian Mathematical Sciences Society*, 30(2):143–157.

Oki M., Sunahara T., Hashizume M. and Yamamoto T. (2011). Optimal timing of insecticide fogging to minimize dengue cases: Modeling dengue transmission among various Seasonalities and transmission intensities. *PLOS Neglected Tropical Diseases*, 5(10):e1367.

Páez Chávez J., Götz T., Siegmund S. and Wijaya K.P. (2017). An SIR-Dengue transmission model with seasonal effects and impulsive control. *Mathematical Biosciences*, 289:29–39.

Pan Y., Zhang M., Chen Z., Zhou M. and Zhang Z. (2016). An ARIMA based model for forecasting the patient number of epidemic disease. *13th International Conference on Service Systems and Service Management*, pages 31–34.

Paul M.J., Dredze M. and Broniatowski D. (2014). Twitter improves influenza forecasting. *PLOS Currents*, 6:1–13.

Philip Howrey E. (1980). The Role of Time Series Analysis in Econometric Model Evaluation. In J. Kmenta and J.B. Ramsey, editors, *Evaluation of Econometric Models*, pages 275 – 307. Academic Press. ISBN 978-0-12-416550-2.

Preis T. and Moat H.S. (2014). Adaptive nowcasting of influenza outbreaks using Google searches. *Royal Society Open Science*, 1(2):140095.

Preis T. and Moat H.S. (2015). Early signs of financial market moves reflected by Google searches. In B. Gonçalves and N. Perra, editors, *Social Phenomena: From Data Analysis to Models*, pages 85–97. Springer International Publishing, Switzerland.

Preis T., Moat H.S., Bishop S.R., Treleaven P. and Stanley H.E. (2013a). Quantifying the digital traces of hurricane sandy on flickr. *Scientific Reports*, 3:3141.

Preis T., Moat H.S. and Stanley H.E. (2013b). Quantifying Trading Behavior in Financial Markets Using Google Trends. *Scientific Reports*, 3:1684.

Preis T., Moat H.S., Stanley H.E. and Bishop S.R. (2012). Quantifying the advantage of looking forward. *Scientific Reports*, 2:350.

Preis T., Reith D. and Stanley H.E. (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1933):5707–5719.

Raftery A.E., Gneiting T., Balabdaoui F. and Polakowski M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133:1155.

Ramadona A.L., Lazuardi L., Hii Y.L., Holmner Å., Kusnanto H. and Rocklöv J. (2016). Prediction of dengue outbreaks based on disease surveillance and meteorological data. *PLOS ONE*, 11(3):e152688.

Reich N.G., Lauer S.A., Sakrejda K., Iamsirithaworn S., Hinjoy S., Suangtho P., Suthachana S., Clapham H.E., Salje H., Cummings D.A.T. and Lessler J. (2016a). Challenges in Real-Time Prediction of Infectious Disease: A Case Study of Dengue in Thailand. *PLOS Neglected Tropical Diseases*, 10(6):e4761.

Reich N.G., Lessler J., Sakrejda K., Lauer S.A., Iamsirithaworn S. and Cummings D.A.T. (2016b). Case Study in Evaluating Time Series Prediction Models Using the Relative Mean Absolute Error. *The American Statistician*, 70(3):285–292.

Robert C.P. and Casella G. (1999). *Monte Carlo Statistical Methods*. ISBN 978-1-4757-3073-9.

Roussel M., Pontier D., Cohen J.M., Lina B. and Fouchet D. (2016). Quantifying the role of weather on seasonal influenza. *BMC Public Health*, 16:441.

Rue H. and Held L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. ISBN 1584884320.

Rue H., Martino S. and Chopin N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71(2):319–392.

Rue H., Riebler A., Sørbye S.H., Illian J.B., Simpson D.P. and Lindgren F.K. (2017). Bayesian Computing with INLA: A Review. *Annual Review of Statistics and Its Application*, 4(1):395–421.

Seresinhe C.I., Moat H.S. and Preis T. (2018). Quantifying scenic areas using crowdsourced data. *Environment and Planning B: Urban Analytics and City Science*, 45:567–582.

Seresinhe C.I., Preis T. and Moat H.S. (2015). Quantifying the Impact of Scenic Environments on Health. *Scientific Reports*, 5:16899.

Seresinhe C.I., Preis T. and Moat H.S. (2016). Quantifying the link between art and property prices in urban neighbourhoods. *Royal Society Open Science*, 3(4):160146.

Seresinhe C.I., Preis T. and Moat H.S. (2017). Using deep learning to quantify the beauty of outdoor places. *Royal Society Open Science*, 4(7):170170.

Sloughter J.M., Gneiting T. and Raftery A.E. (2010). Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, 105(489):25–35.

Souza R.C.S.N.P., de Brito D.E.F., Assunção R.M. and Meira W. (2015). A latent shared-component generative model for real-time disease surveillance using Twitter data. arXiv:1510.05981.

Spitzen J. and Takken W. (2018). Keeping track of mosquitoes: A review of tools to track, record and analyse mosquito flight. *Parasites and Vectors*, 11(123).

Stanaway J.D., Shepard D.S., Undurraga E.A., Halasa Y.A., Coffeng L.E., Brady O.J., Hay S.I., Bedi N., Bensenor I.M., Castaneda-Orjuela C.A., Chuang T.W., Gibney K.B., Memish Z.A., Rafay A., Ukwaja K.N., Yonemoto N. and J.L C.M. (2016). The Global Burden of Dengue: an Analysis from the Global Burden of Disease Study 2013. *The Lancet Infectious Diseases*, 16(6):712–723.

Tennakone K. and De Silva L.A. (2018). Host-vector interaction in dengue: a simple mathematical model. *Ceylon Medical Journal*, 63:58–64.

The Lancet Infectious Diseases (2018). The dengue vaccine dilemma. *The Lancet Infectious Diseases*, 18(2):123.

Tierney L. and Kadane J.B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81(393):82–86.

Tizzoni M., Bajardi P., Poletto C., Ramasco J.J., Balcan D., Gonçalves B., Perra N., Colizza V. and Vespignani A. (2012). Real-time numerical forecast of global epidemic spreading: Case study of 2009 A/H1N1pdm. *BMC Medicine*, 10(1):165.

Tofallis C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, 66(8):1352–1362.

Vega T., Lozano J.E., Meerhoff T., Snacken R., Mott J., Ortiz de Lejarazu R. and Nunes B. (2013). Influenza surveillance in Europe: Establishing epidemic thresholds by the Moving Epidemic Method. *Influenza and other Respiratory Viruses*, 7(4):546–558.

Vespignani A. (2009). Predicting the behavior of techno-social systems. *Science*, 325(5939):425–428.

Vogel G. (2018). A new dengue vaccine should only be used in people who were previously infected, WHO says.

Watson M.W. (1986). Univariate detrending methods with stochastic trends. *Journal of Monetary Economics*, 18(1):49–75.

Whitehead S.S., Blaney J.E., Durbin A.P. and Murphy B.R. (2007). Prospects for a dengue virus vaccine. *Nature Reviews Microbiology*.

Wöhling T., Schöniger A., Gayler S. and Nowak W. (2015). Bayesian model averaging to explore the worth of data for soil-plant model selection and prediction. *Water Resources Research*, (51):2825–2846.

World Health Organization (2012). Global Strategy for Dengue Prevention and Control 2012–2020. Technical report.

World Health Organization (2016). Dengue vaccine: WHO position paper. Technical Report 91.

World Health Organization (2018a). Dengue and severe dengue. Technical report.

World Health Organization (2018b). Revised SAGE recommendation on use of dengue vaccine - 19 April 2018. Technical report.

World Health Organization Regional Office for South-East Asia (2009). Dengue: guidelines for diagnosis, treatment, prevention, and control. Technical report.

Xu Q., Gel Y.R., Ramirez L.L.R., Nezafati K., Zhang Q. and Tsui K.L. (2017). Forecasting influenza in Hong Kong with Google search queries and statistical model fusion. *PLOS ONE*, 12(5):e0176690.

Yang S., Kou S.C., Lu F., Brownstein J.S., Brooke N. and Santillana M. (2017). Advances in using Internet searches to track dengue. *PLOS Computational Biology*, 13(7):e1005607.

Yang S., Santillana M. and Kou S.C. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478.