Instituto Tecnologico y de Estudios Superiores de Monterrey

Monterrey Campus

School of Engineering and Sciences

Predicting Influenza in Latin America: Using Voting Ensembles to Combine Google Search Activity and Geo-spatial Synchronicities from Historical Flu Activity

A thesis presented by

César Leonardo Clemente López

Submitted to the
School of Engineering and Sciences
in partial fulfillment of the requirements for the degree of

Master of Science

in

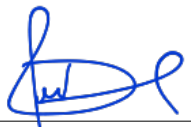
Computer Science

Monterrey, Nuevo León, December, 2019

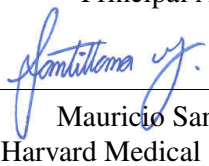
Instituto Tecnologico y de Estudios Superiores de Monterrey

Campus Monterrey

The committee members, hereby, certify that have read the thesis presented by César Leonardo Clemente López and that it is fully adequate in scope and quality as a partial requirement for the degree of Master of Science in Computer Sciences.



Jose Carlos Ortiz Bayliss
Tecnológico de Monterrey
Principal Advisor



Mauricio Santillana
Harvard Medical School
Co-Advisor

Santiago Enrique Conant Pablos
Tecnológico de Monterrey
Committee Member

Iván Mauricio Amaya Contreras
Tecnológico de Monterrey
Committee Member

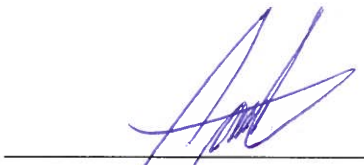
Rubén Morales Menéndez
Associate Dean of Graduate Studies
School of Engineering and Sciences

Monterrey, Nuevo León, December, 2019

Declaration of Authorship

I, César Leonardo Clemente López, declare that this thesis titled, Predicting Influenza in Latin America: Using Voting Ensembles to Combine Google Search Activity and Geospatial Synchronicities from Historical Flu Activity and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.



César Leonardo Clemente López
Monterrey, Nuevo León, December, 2019

Dedication

(Only 1 page) Thanks for all your unconditional confidence, support, patience, and encouragement. You were my main motivation for pushing through this work.

Acknowledgements

During the course of my Master's degree, many things have happened, and I would not have been able to make it without the help of my family, friends, colleagues. I am grateful to Dr. Ortiz-Bayliss for taking me as his graduate student and his flexibility and support throughout this 2-year period. I am also really grateful with Dr. Mauricio Santillana for giving me the opportunity to work with him and have a first-time experience of what it is to work in a multi-cultural and high-skilled environment, always thriving for excellence in everything we do. I would also like to thank my parents, Elvia and Victor, for being there when I needed them, unconditionally, and supporting me the only way a parent knows how. I would like to thank Andrea, my girlfriend, for understanding when I had to leave out of the country due to my research projects and the times where I had to prioritize finishing some work and not spending some time together. Finally, I would also like to express gratitude to all the great people in the Intelligent systems department, for their strict feedback to polish this work into what it is now, for Tec de Monterrey's and its tuition support, and CONACyT's scholarship grant, which helped me pay for my living during this period. For everyone else those who have been with me, I am grateful.

Predicting Influenza in Latin America: Using Voting Ensembles to Combine Google Search Activity and Geo-spatial Synchronicities from Historical Flu Activity

by

César Leonardo Clemente López

Abstract

Novel influenza surveillance systems in the U.S. that combine historical influenza reports from hospitals, along with external influenza-related sources (such as web-based databases and reports from neighboring locations, among others) have shown to generate flu estimates in advance to the official health reports published by healthcare-based systems. However, in Latin America, several systems that harnessed web-based data sources in the past, such as Google Flu Trends, have shown that novel influenza surveillance are yet to deliver acceptable flu estimates. In this work, we aim to show that improved influenza estimates can be achieved for various countries in Latin America, and even in more refined geographic-scales in Mexico, by implementing methodological changes on the way that information sources such as Google Search activity and historical influenza activity reports are combined. A methodological framework which dynamically combines different influenza tracking techniques via a voting ensemble system is presented and used to generate improved influenza activity estimates in 15 countries in Latin America, 5 regions and 27 states in Mexico. Our results show that the voting ensemble outperforms the 3 different techniques implemented to harness the predictive power of local and external data sources, reaching lower error and higher correlation scores. This methodology may prove helpful to local public health officials who organize health interventions aimed at mitigating influenza outbreaks, and its adaptive power may also prove useful to extend its scope by also tracking other diseases such as Dengue and Zika.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Influenza poses significant health challenges to developing countries in Latin America, having the highest mortality of any respiratory infectious disease in the young and elderly [23]. In Mexico, more than 100 thousand patients that visited their physician exhibited influenza like illness (ILI) or severe acute respiratory infections (SARI) between January 2010 and December 2013 [4]. Out of these patients, around twenty thousand were confirmed with the influenza virus. In order to address this issue, Official Health Institutes create disease activity indicators that allow the localization of outbreaks and morbidity control of the infected with preventive measure purposes. World wide, the World Health Organization (WHO) aggregates activity reports from various national influenza centers and publishes country-level information regarding the number of reported, tested, and confirmed influenza weekly cases. Mexico's official epidemiological surveillance organization, Sistema Nacional de Vigilancia Epidemiológica (SINAVE), keeps track of several diseases such as Dengue, Zika, Tuberculosis and influenza at local, state and national level. However, while these systems provide historical ILI or influenza-confirmed case activity, reports are usually delayed by a week or more, limiting the information available to health officials.

To alleviate this timegap, multiple research teams have proposed complementary methods to estimate and forecast flu activity in real time (referred to as "nowcasting") and have overcome these delays by using various techniques [1, 24, 31]. These techniques include incorporating a variety of digital data sources such as search engine trends [6, 21, 29], Wikipedia [7, 8], social networks and crowdsourcing [14, 15, 25] and neighboring historical spatial and temporal synchronicities of influenza activity [2, 10, 27]. Although these models have been successfully implemented on data-rich countries such as the United States, there is no successful implementation on countries where the data is rather scarce. Moreover, a reliable system that successfully leverages digital data streams

to monitor influenza at finer scales is not yet available for developing nations in Latin America.

An early attempt at large-scale nowcasting started in 2008 with Google Flu Trends (GFT), an online tool using Google search activity as predictors to produce state and national level influenza activity estimates. While this approach was technologically innovative at its first deployment, its methods drew criticism from researchers when the model incorrectly predicted the 2009 A-H1N1 flu pandemic and the 2013 flu season in the United States [12, 22]. GFT also produced flu estimates in a broad range of countries, and was finally discontinued in 2015. A more recent study by Pollet et al. showed that estimates in Latin American countries have yielded poor results [19].

1.1 Problem Definition and Motivation

Influenza surveillance is the event of determining the timing, location and magnitude of outbreaks by monitoring the frequency and progression of clinical case incidence [12]. The main objective of influenza surveillance is to identify changes in incidence, either in the form of an acute outbreak or a change in long-term trends, their verification, assessment and investigation to enable public health control.

Statistical models that estimate influenza activity (such as the proportion of total visits to a clinician that relate to influenza or the weekly number of processed specimen cases tested for influenza) support health-care institutions in the development of preventive strategies with the aim to mitigate disease transmission and efficiently allocate resources to attend patients in a prompt manner. As an example in Mexico, more than 130 thousand patients that visited their physician exhibited ILI or SARI in a four-year time span, and efforts from national health officials to reduce the infection rates based on the surveillance data they collected from SINAVE [4]. However, this information is usually delayed by one or two weeks, reducing the ability of decision makers in selecting the most adequate disease prevention strategies.

Effective influenza surveillance methods that estimate disease activity in advance to the official health reports aim to overcome this time delay, and aid in development of preventive campaigns to mitigate disease transmission, reducing the incidence and avoiding expenses from basic protocol tests to emergency room stays [23].

An effective predictive methodology for influenza forecasting would reduce the economic burden that the disease generates on a yearly basis and work as a first step to extend such methodology into tracking other diseases.

Current public health surveillance data from Mexico has the potential to generate

real-time surveillance systems that incorporate external sources of information. The establishment of SINAVE in 1995, an online web-tool where official health institutes throughout the country submit information regarding different diseases with the objective of producing surveillance, and the increasing online activity of users in the Internet generates an ideal environment for digital influenza surveillance. However, previous analyses of influenza surveillance in the U.S. have shown that influenza activity may be highly heterogeneous across locations within a country and that external sources of information such as Google search activity trends that spur from, perhaps, a community's overreaction can often be misinterpreted by influenza activity trends.

1.2 Objectives

As explained by the Centers of Disease Control and Prevention (CDC), epidemic surveillance can be described by two main components: the indicator-based component, which refers to data collected through official health surveillance systems, and the event-based component, which refers to data gathered from sources of different external sources, such as social media, crowdsourced campaigns, news, among others. By considering Internet as a source of health-related information, along with the increasing advances in information technologies that allow official health-institutes to gather large amounts of data related to disease diagnosis, has led to the creation of Digital Infectious Disease Surveillance, a new form of surveillance that utilizes disease related information from various data sources. In this type of surveillance, the novelty comes from successfully combining the two previously defined components of epidemic surveillance [13]. In this work, the main objective is to combine both indicator-based and event-based components of surveillance to develop a data-driven and self-correcting method that delivers timely flu activity estimates one week ahead of time of the official health reports. To achieve this general objective, the following particular objectives are considered:

1. Identify whether there is a standard, automated way of extracting information that serves as data to perform digital ILI surveillance in Latin American countries.

2. Investigate the feasibility of combining historical flu activity and flu-related Internet activity in Latin American countries.

3. Explore the advantages of incorporating influenza activity from neighboring locations to improve nowcasting.

4. Investigate novel influenza surveillance methods shown to successfully leverage these data sources in the United States, and explore their possibility of extension

and improvement in Latin American countries, where data is more sparse.

5. Design and implement a voting ensemble method that leverages the predictive power of baseline models, with the purpose of increasing accuracy and robustness.

6. Propose a novel real-time methodology that successfully combines official health reports and Internet-based data sources, which are the only available sources of information, for monitoring ILI activity.

1.3 Hypothesis

A data-driven voting ensemble method that combines the predictive power of different models based on their historical performance can generate influenza activity estimates that are more robust to the variation in the performance of individual models in Latin American countries.

The following research questions were used to guide this work:

- Is Internet-based data a feasible source of information to conduct influenza activity surveillance in Latin American countries?
- How should web-mined data be combined to produce influenza activity surveillance estimates? How should we mine such data?
- How can influenza activity events that originate in the geographical vicinity of surveillance locations aid in nowcasting?
- What is the proper way to design a voting ensemble that extracts and combines relevant features from independent sources of information?

1.4 Goals

Influenza surveillance aids national health institutions measuring morbidity and mortality rates, and to detect changes in public health in the most possible timely manner. Moreover, advances in information technology have led to people being able to search for news, symptoms and other health-related activities through the Internet, making the web a valuable resource of information in disease surveillance. With the purpose of evaluating the possibility of influenza nowcasting in Latin America, the goals this work expects to accomplish are:

- Revising the current methods to extract Google search volumes and verify their applicability to Latin American countries. Then, generate a useful feature database for each location in which a surveillance system will be implemented.

- Mine historical influenza activity from SINAVE and WHO to produce some standardized datasets for different countries in Latin America and the states in Mexico.

- Explore the historical influenza interactions between a local surveillance point and its neighbors to analyze the possibility of combining them into a surveillance tool

- Explore and design a voting ensemble technique that successfully combines different sources of data that include official health reports and Internet-based data sources for monitoring ILI activity.

- Generate the proper set of coding tool to facilitate the research and reproducibility of the experiments conducted in this investigation.

1.5 Solution overview

In this thesis, we present the first nowcasting system that combines historical influenza activity, Internet search volumes, and geo-spatial synchronicities in Mexico and other countries in Latin America, by extending on the work of Lu et al. [10]. As a first step, a methodology named AutoRegression with General Online information (ARGO) [30], which combines historical influenza activity reports and Google search activity, is implemented. Next, a learning ensemble approach, called Net, that leverages local and neighbor influenza activity estimates along with historical official health reports (structural spatio-temporal synchronicities) is introduced. Finally, we implement a voting ensemble and show that, by combining the individual predictive power of each model, it is possible to consistently reach higher accuracy and lower error in more cases than using the models separately. We demonstrate this novel methodology's capabilities by generating retrospective estimates at state and regional geographic level for Mexico, and national level for several countries in Latin America.

1.5.1 Contributions

This work focuses on developing nowcasting methodologies that accurately predict influenza activity in Latin American countries. The main contributions from this work are:

1. Expanding Internet-based disease surveillance for various countries in Latin America through two different official health-care institution databases: Flunet and SINAVE, and two non-official sources of information: Google Correlate and Google Trends.

2. Conducting a study of the novel influenza methodology in Latin American countries and a multi-scale surveillance case for Mexico.

3. Generating a programming tool in python that helps mining the web sources that we use for disease surveillance (Google Correlate, Google Trends) automatically.

4. Providing a novel, adaptive approach that improves feature selection of data coming from Google search queries.

1.5.2 Justification

The availability for a reliable estimate of future incidences of influenza disease is a crucial factor in the improvement of decision making during potential pandemics. Important decisions such as vaccination prioritizing, resource allocation and timing of resource deployment must be justified on strong statistical evidence and robust prediction models [9]. While web-based tools such as Flunet and SINAVE routinely collect and aggregate information on weekly flu activity, these reports involve a delay of at least 1-2 weeks in Latin America, limiting the ability for a timely response to unexpected epidemic outbreaks. Multiple research teams have proposed complementary methods to estimate influenza activity in real-time for data-rich countries such as the United States. However, an accurate system using Internet search activity to monitor influenza activity in developing nations is not yet available. More recently, a self-correcting, adaptive methodology named ARGO [30], became state-of-the art in the task of real-time disease forecasting by combining several disparate sources such as historical official reports, digital-health records, and Google search activity [10, 29].

1.6 Thesis Overview

The remainder of this document is organized as follows. Chapter 2 introduces the theoretical framework and the nature of the information sources that will be used for the thesis. Chapter 3 describes *argotools*, a programming library that was developed to execute the experiments in this research to ensure reproducibility of results. Chapter 4 presents the set of preliminary experiments that assess the feasibility of applying Internet search activity trends as a proxy of influenza activity at National scale for eight countries. Chapter 5 describes a voting ensemble that, based on the historical performance of the different models

that participate in the ensemble, combines their predictive power to generate more robust results and achieving lower out-of-sample error. Chapter 6 presents the conclusions and future work based on the research conducted.

Chapter 2

Background and Related Work

In this section, we present the origin and format of the data sources, along with the mathematical formulation for ARGO, Net and AR. This chapter starts by mentioning historical influenza surveillance data and how it was collected from the web-based databases from FluNet and SINAVE. Then, the mathematical model formulation for various multivariate linear methods (AR, ARGO and Net) and the voting ensemble are presented. Finally, we formulate the definitions of several error types used in the benchmarking of point-wise estimates that are used when comparing the performance of the models.

2.1 Data Sources

In this work, two indicator-based surveillance data sources (SINAVE and FluNet) and two event-based sources (Google Correlate and Google Trends) are presented. The indicator-based data sources are used in this study as both ground-truth and features for our predictive models. Google Correlate and Google Trends provide us with features that we hypothesize as adequate proxies for ILI activity.

2.1.1 FluNet

FluNet is a web-based tool created by the World Health Organization in 1997 used to perform international influenza surveillance and to allow public access for data regarding confirmed and suspected influenza activity within National Influenza centers (NICs). The WHO routinely collects and aggregates data from several NICs and makes it publicly available for everyone via visualizations, including tables, maps and graphs. FluNet has the purpose of helping track influenza viruses and interpreting epidemiological observed events.

FluNet's official health reports are of interest for digital disease surveillance, since they provide information about the Number of Received Specimens (NRS) for inspection, Number of Processed Specimen (NPS) and influenza-confirmed cases, among others, which can be used directly (or as the ratio of the two) as gold standard in influenza forecasting. For example, Table 2.1 presents an example of the data that can be extracted from FluNet.

Table 2.1: Sub-sample of the data that FluNet provides when downloading weekly influenza health-reports from their website.

Year	**Week**	*Start Date*	*End Date*	**NRS**	**NPS**	$Influenza\,confirmed$
2010	1	1/4/10	1/10/10	25	445	4
2010	2	1/11/10	1/17/10	55	434	21
2010	3	1/18/10	1/24/10	21	427	6
2010	4	1/25/10	1/31/10	24	298	1
2010	5	2/1/10	2/7/10	28	415	4
2010	6	2/8/10	2/14/10	11	410	0
2010	7	2/15/10	2/21/10	17	420	3
2010	8	2/22/10	2/28/10	8	430	1
2010	9	3/1/10	3/7/10	16	485	0
2010	10	3/8/10	3/14/10	9	553	3

2.1.2 SINAVE

The Sistema Nacional de Vigilancia Epidemiologica (SINAVE) is the official national epidemiological surveillance platform from Mexico, where all official reports from national health institutions such as Instituto Mexicano del Seguro Social (IMSS) and Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado (ISSSTE) are collected. SINAVE keeps records of emerging activity for different diseases. SINAVE records information of the number of patients either infected, non-infected, in danger and deceased, along with virus sub-type information, which can be used for disease surveillance methodologies if combined correctly.

SINAVE collects information from approximately 20, 000 health units in Mexico, and has proved to be a successful epidemiological surveillance tool, particularly for generating preventive health campaigns in the advent of a disease outbreak such as the 2009 influenza pandemic.

2.1.3 Collecting historical influenza reports from FluNet and SINAVE

FluNet collects and aggregates multiple indicators of flu activity at country level. For this study, we selected the number of processed specimens (NPSs) as the ground truth. As these specimens were taken from patients with flu-like symptoms and then sent to a laboratory for testing, we interpreted them as an indicator of suspected flu activity in the population. Weekly aggregated NPS reports were collected from January 5, 2009 to December 25, 2017 for Argentina, Brazil, Chile, Colombia, Costa Rica, Ecuador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Peru, Paraguay, Salvador and Uruguay.

Similarly, SINAVE provides health clinicians with historical influenza activity reports that are collected from the majority of official government managed health institutions. As with FluNet, we interpreted SINAVE's influenza activity reports (an aggregated time-series from all flu-related activity time-series available from SINAVE) as an indicator of suspected flu activity, and collected weekly aggregated influenza activity reports from January 05, 2010 through December 25, 2017 for all the available states in Mexico. To corroborate that the data collected from SINAVE was similar to the flu activity reported for Mexico in FluNet, data from all states in Mexico obtained through SINAVE was aggregated to generate a national-level influenza trend. Our preliminary test confirmed that both influenza activity time-series followed an almost identical trend every week during our study period.

To generate regional-level information, historical flu reports from different states were aggregated a follows:

1. North-West (NW): Baja California Norte, Baja California Sur, Chihuahua, Sinaloa and Sonora.
2. North-East (NE): Coahuila, Durango, Nuevo Leon, San Luis Potosi and Tamaulipas.
3. South-Central (SC): Ciudad de Mexico, Estado de Mexico, Guerrero, Hidalgo, Morelos, Puebla and Tlaxcala.
4. South-West (SW): Aguascalientes, Colima, Guanajuato, Jalisco, Michoacan, Nayarit, Queretaro and Zacatecas.
5. South-East (SE) : Campeche, Chiapas, Oaxaca, Quintana Roo, Tabasco, Veracruz, Yucatan.

2.1.4 Google Correlate and Google Trends

Online web search query data has proved useful in providing models of real world problems. However, many of these results relied solely on prior knowledge about the queries

that were related to the phenomenon. Google presented an online, automated method for query selection called "Google Correlate" that requires no such prior knowledge and instead, it uses a temporal or spatial pattern to determine which queries are the most similar. These search queries can then serve to build an estimate of the true value of the phenomenon. Google Correlate implements an approximate nearest neighbor algorithm over millions of candidates to produce results and picks the most similar queries based on the R^2 correlation. A table of influenza-related queries generated by Google Correlate is presented in Table 2.2.

After its success in the United States, Google Correlate has been implemented in many countries around the globe, including Argentina, Brazil, Chile and Mexico. Even though Google Correlate does not explain why only these countries are available, it is suspected that only countries that have "enough" data can be searched through with this web-tool, therefore giving us a hint of which countries in Latin America may benefit from using online search queries as an explanatory variable of influenza activity.

Google Trends works in the opposite way. Instead of looking up correlated terms based on a search term or a search term activity, Google Trends returns the activity from the search term of interest. Google Trends adjusts search data to make comparisons between terms easier. Each data point is divided by the total searches of the geography and time range it represents, to compare relative popularity.

Table 2.2: This table shows a sub-sample of how the data generated by Google Correlate looks after being normalized and zero-patched by Google.

Date	**"virus influenza"**	**"prevencion de la gripa"**	**"influenza"**	**"influenza virus"**	**"gripe"**
1/4/04	0	0	0	0	0
1/11/04	0	0	573.0499	0	0
1/18/04	0	0	0	0	0
1/25/04	0	0	0	0	1117.8653
2/1/04	0	0	0	0	0
2/8/04	0	0	0	0	508.5831
2/15/04	0	0	0	0	0
2/22/04	484.2659	0	0	484.2659	483.0183
2/29/04	0	0	0	0	776.0539

2.1.5 Collecting web-based activity from Google Correlate and Google Trends

Given their near real-time availability via the online tool, Google Trends, we selected influenza-related Internet search activity to be used in our models as proxies or predictors for flu activity. Where available and based on country-specific historical flu indicators (during the training time period of our models), we used the online tool, Google Correlate, to identify flu-related search term trends, leading to a total of 285 Spanish terms and 96 Portuguese terms for the National level and 287 terms for the state and regional levels of Mexico. A detailed list of the terms used in this work can be accessed on `http://bit.ly/2F5qpYA`). We opted to use the state level spanish search terms as features for the regional study.

After generating a search term list through Google Correlate, Google search term time-series were downloaded using the Google Trends' (GT) API.

The following strategies were used to produce a list of terms using Google Correlate:

1. We submitted state-level and national-level influenza activity reports to Google Correlate and extracted, for each state and country, a table with the most correlated terms.

2. submitted a query to collect the most correlated search terms with the word "influenza" for every location at National level (this process can also be done in Google Trends for state-level although to a more limited extent).

3. A final list of terms was created from all the tables collected for each separate Google Correlate consult. Data was filtered by hand to omit terms that, even though they exhibited a correlation with the influenza activity reports, were not semantically related to ILI symptoms.

It is relevant to add that the dates of the influenza activity reports submitted to Google Correlate ranged from January 5, 2009 through December 25, 2011 for the preliminary work and January 1, 2010 through the December 29, 2013 for our main work, with the objective ensuring that our predictions for the period of 2014-2017 were strictly out-of-sample.

2.2 Methods

This section introduces three different multivariate linear models and a voting ensemble which will be the focus of the main work developed.

The endogenous and exogenous variable coefficients in all these models are fitted by using the Least Absolute Shrinkage and Selection Operator (LASSO) [26], which is a multivariate linear regression with $L1$ regularization. $L1$ regularization is used as a way to impose coefficient sparsity.

For each model, re-training occurs on a weekly basis, updating the input features with the newest available historical flu reports and Internet activity from the most recent two years, or 104 most recent data samples in our dataset. This approach allows for re-calibration of regression coefficients in a way that adjusts the variables based on its prediction ability over the training set. The decision of using only the two most recent years of data comes from empirical observation on how these novel methodologies fit the data better to make predictions [30].

2.2.1 Autoregressive Model

An autoregressive model (AR) is used to describe a random process that changes over time. AR models are constructed on the assumption that the target value depends linearly on its own previous values. In this work, we represent influenza activity target as dependent to the last 52 weeks (referred as autoregressive lags) of its own activity.

The AR nowcast for influenza activity at week t is:

$$y_t = u_y + \sum_{j \in J} \alpha_j y_{t-j} + \epsilon_t, \epsilon_t \sim N(0, \sigma^2) \tag{2.1}$$

where:

- y_t stands for the number of processed cases at time t.
- J corresponds to the set of autoregressive lags.
- ϵ is an error term or white noise.
- u_y is an intercept term
- α_j is the j-th linear coefficient for the historical flu activity ocurring at a time y_{t-j}
- $N(0, \sigma^2)$ represents a Normal distribution of mean 0 and variance σ^2

2.2.2 Autoregressive model with General Online Information

The Autoregressive Model with General Online information (ARGO) is a dynamic multivariate linear regression model introduced by Yang et al. in 2015 [29], used to predict flu

incidence in the United States. The motivation behind ARGO comes from the assumption that unobserved flu activity at time t depends on previously observed activity, and that flu-related Google search terms at time t depend on unobserved flu activity at time t [29], as formalized by the hidden Markov model from Eq. 2.2.

$$\begin{array}{ccccccc} y_{i,1:N} & \rightarrow & y_{i,2:N+1} & \rightarrow & \cdots & \rightarrow & y_{i,(T-N+1):T} \\ \downarrow & & \downarrow & & & & \downarrow \\ X_{i,N} & \rightarrow & X_{i,N+1} & \rightarrow & \cdots & \rightarrow & X_{i,T} \end{array} \tag{2.2}$$

In the ARGO model, higher search frequencies for disease-related Google queries will be observed when the disease also presents increments in the activity, such as when people are infected or experience symptoms.

The ARGO nowcast for disease activity at week t is:

$$y_t = u_y + \sum_{j \in J} \alpha_j y_{t-j} + \sum_{k \in K} \beta_k X_{k,t} + \epsilon_t, \epsilon_t \sim N(0, \sigma^2) \tag{2.3}$$

where:

- y_t is the number of processed cases at time t.
- J corresponds to the set of auto regressive lags.
- K is the set of Google query terms.
- $X_{k,t}$ is the Google search frequency of term k at time t.
- ϵ_t is assumed to be a Gaussian white noise process with zero mean and constant variance.
- u_y is an intercept term.
- α_j is the j-th linear coefficient for the historical flu activity ocurring at a time y_{t-j}.
- β_k is the k-th linear coefficient for the k-th Google search frequency.

2.2.3 Net

Net is a multivariate linear model that works by hypothesizing that a disease's activity at location i (the location of interest where influenza activity is estimated) can be improved by looking at the neighboring locations s that exhibit similar disease behavior. These similarities between the influenza activity within location i and the neighboring locations

are referred to as spatio-temporal synchronicities, and the locations that constantly exhibit similar influenza activity to the one of location i within a specified time interval (in the practice, a two year time window is used) are used as explanatory variables along with the historical autoregressive terms of the local influenza activity. Net incorporates multiple weeks of historical influenza activity from all the available neighboring locations in a multivariable regression, giving the model flexibility to assess [10].

The Net nowcast for disease activity in week t is:

$$y_{i,t} = u_i + \sum_{j \in J} \alpha_j y_{i,t-j} + \sum_{s \neq i : s \in S} \sum_{k=0}^{3} \beta_{s,k} y_{s,t-k} + \epsilon_t, \epsilon_t \sim N(0, \sigma^2) \tag{2.4}$$

where:

- $y_{i,t}$ is the number of processed cases at time t at location i.
- α_j are the fitting coefficients of the historical influenza activity at location i.
- $\beta_{s,k}$ are the fitting coefficients of historical influenza activity of city s at time $t - k$.
- J corresponds to the set of autoregressive lags.
- S is the set of neighboring locations.
- u_i is the target's intercept term.
- ϵ_t is assumed to be a Gaussian white noise process with zero mean and constant variance.

Net detects geo-spatial synchronicities from neighbor influenza activity at time t ($k = 0$) when predicting for $y_{i,t}$. Since information from the neighboring locations at time t is not yet available to predict for $y_{i,t}$, the $y_{s,t}$ are substituted by the most recent influenza activity estimation for each of the locations s.

2.2.4 Google Flu Trends

Google Flu Trends (GFT) was a web-tool for influenza surveillance launched in November of 2008. GTF was the first approach that attempted to harness the power of online web activity as a predictor for disease activity in the United States. Google generated an univariate linear model that sought to estimate the probability that a random physician visit in a particular region is related to ILI (in other words, trying to predict ILI-related physician

visits). GFT was fit to estimate the log-odds of an ILI physician visits by using the log-odds of an ILI-related search query. The model equation is as follows:

$$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \epsilon \tag{2.5}$$

where:

- P is the percentage of ILI related visits.
- Q is the ILI-related query fraction.
- ϵ is an error term.
- β are the fitted coefficients of the linear regression.

Q is calculated based on the sum of the top 45 search terms that are most related with the CDC (Center of Disease Control and Prevention in Atlanta, Georgia) ILI data [22].

2.2.5 Voting Ensemble

If the explanatory variables used to generate Influenza activity estimation models are not perfect predictors of influenza, their performance tends to fluctuate within time. If several models are fit using different explanatory variables to predict the same target, then each of these models have abstracted different knowledge and, performance will vary from one weekly prediction to another. For example, ARGO and Net, where one bases its predictions on Google search activity and the other on Geo-spatial synchronicites, lead to different estimations for the same week based on the different patterns learnt from flu-related Internet search activity and neighboring influenza activity.

The process of consulting several "experts" in the matter of interest before making a final decision is a natural thought in the human mind, and it has been recently shown that such a way of thinking can be also used to improve machine-learning algorithms. In computer science, this process of consultation is referred to as "ensemble", where a mixture of expert models provide their answer and a higher level abstraction layer (the ensemble) learns to manipulate the input data, producing a more robust prediction. An ensemble consists of a set of independently trained prediction models that are combined with the objective of increasing robustness. The combination method relies on a rule or algorithm, which can be a step or series of steps applied onto a set of previously selected models and can range from simple heuristics to more complex algorithms that involve random or learning processes. Ensemble development is considered a meta-heuristic area that has been successfully applied to pattern recognition, machine-learning, and statistics [17].

In this work, a majority voting ensemble is implemented to combine the predictive power of AR, ARGO and Net. As shown in Figure 2.1, the ensemble approach adopts a "Winner takes all" philosophy and works by comparing each of the models' weekly performance in the short-time historical record of the k most recent weeks. The historical performance is measured based on the error (distance between the prediction and its target) and draw conflicts are resolved through coin flipping.

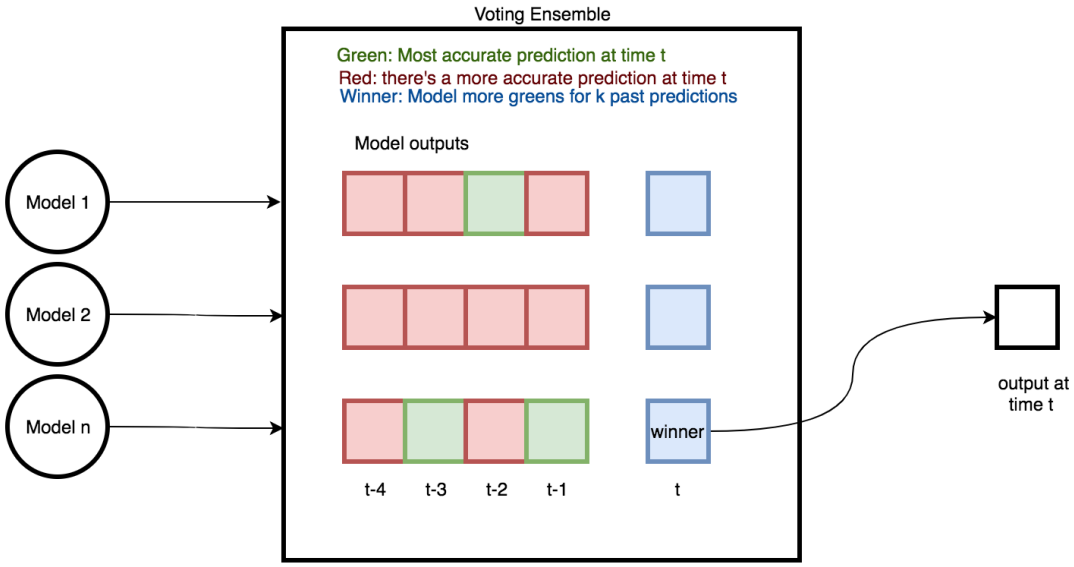


Figure 2.1: Graphical diagram that shows how a voting ensemble makes a prediction at time t from several baseline models, based on the historical predictions they have produced over time.

2.3 Benchmarking Metrics

Objective comparison of prediction models is based on standard metrics that produce correlation and error estimates. The following metrics are used in this work.

Root mean squared error. (RMSE) It is described as the square root of the mean of the squared difference between the mean and the actual observations. RMSE (Eq. 2.6) uses squaring of errors, which tend to implicitly highlight large errors from single predictions compared MAE. RMSE is expected to increase as the variance of the frequency distribution of error magnitudes increases.

$$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2]^{\frac{1}{2}} \tag{2.6}$$

Mean absolute error. (MAE) The mean absolute error (Eq. 2.7) measures the average magnitude of absolute errors between a set of predictions and its corresponding observations. It is similar to RMSE in the sense they both calculate mean model prediction error, they are both non-negative and ignore vector orientation.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y - \hat{y}| \tag{2.7}$$

Mean absolute percent error. (MAPE) This metric is similar to MAE with the exception that each of the differences are divided by the target value. MAPE (Eq. 2.8) is useful when comparing models that harness different magnitude orders.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y-\hat{y}|}{y_t} \tag{2.8}$$

As an additional step, confidence intervals for error estimates can be generated by performing an statistical bootstrapping [18].

2.4 Summary

In this chapter we presented the data and methods that are used to produce influenza activity forecasts. Historical influenza reports from SINAVE and FluNet, along with the Google search time series from Google Correlate and Google Trends were introduced. Moreover, the ARGO, Net, and voting ensemble methodologies were described. These strategies leverage the predictive power of the historical flu reports, influenza-related internet activity, and geo-spatial synchronicities between every location in the short-time period. The following chapters show the results and how these methodologies are used to successfully predict influenza for different locations in Latin America.

Chapter 3

Argotools: a Python-based Library for Digital Disease Surveillance

Internet-based digital disease surveillance has raised the interest of researchers around the globe given its utility as a tool to produce accurate and timely predictive models for disease activity. Internet-based databases have matured over the years from sporadic streams of information to near real-time updates about events in almost every health-related topic, and has reached a point where tracking diseases with useful accuracy is a possibility now. However, given the area of research is so recent, there is still an information gap between public health officials and data scientists in terms of how the data should be prepared for analysis and, therefore, several challenges that a researcher working in digital disease surveillance has to deal with. The most relevant challenges are:

1. Mining databases becomes a repetitive and, specially, a time consuming task.
2. Organizing and managing the data to map it from its original source structure to the feature-sample data-science structure.
3. The design of adequate metrics and benchmarking of models to provide an objective analysis of the different methodologies used.

In order to address this problem, we introduce Argotools, a python-based library that serves as a fully integrated framework for digital disease surveillance. Argotools was developed with the purpose of providing a set of programming tools for people aiming at reproducing the results presented in this thesis or start working in the area of digital disease surveillance.

The software libraries (shown in Figure 3.1) contained within Argotools are:

DBscrape. A library developed to mine data from various online data sources (such as Google Correlate, Google Trends, Flunet, and SINAVE) that are used for producing influenza surveillance tools in Latin America.

DataFormatter. A library designed to read data from multiple sources and merge them into an object that contains some preprocessing functions and filters that prepares the data for experimentation.

Modelling. The library used to perform data modelling and prediction that can be scaled to multiple areas.

Visualizer. A library with the purpose of benchmarking and generating visualizations of the input data (as a means for exploratory data analysis) and the results generated from the models in `Modelling`.

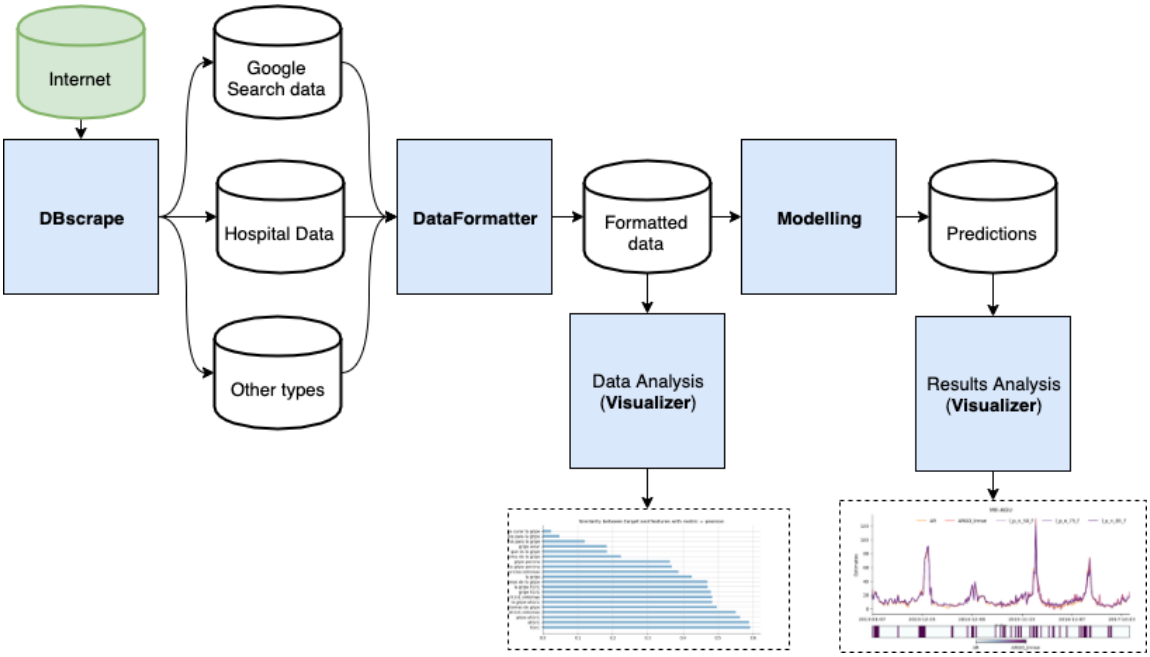


Figure 3.1: Diagram that represents the different stages within Argotools. Data is mined from the web using `DBscrape`. Independent data sources are read and given a standardized format through the `DataFormatter` library. After formatting and preprocessing the data, the modelling library is used to fit and deploy several machine learning models which can be benchmarked and analyzed via the `Visualizer` library.

All the code in Argotools was developed using Python version 3.4, and can be obtained upon request.

3.1 DBscrape

Performing digital disease surveillance using search-engine term activity involves web-based specific tasks such as accessing Google Correlate and querying for disease-related terms such as "influenza" or directly inputting time-series data from official hospital reports. In either case, the researcher has to do it manually as Google does not offer an automated tool to perform these tasks in a repetitive manner. This prior-to-research procedure may be time-consuming specially in the cases when research is conducted for many different locations (for example, looking for "influenza" time-series correlations for more than 10 countries in Google Correlate). In order to address this potential bottleneck problem, `DBscrape` was developed.

The idea behind `DBscrape` is to automate web-scraping of different publicly available data sources, such as Google Correlate, Google Trends, among others, for disease surveillance purposes (see Figure 3.2).

For this work, the following classes are currently available within `DBscrape`:

1. GC: A class created to mine Google Correlate data. It provides the user with the capacity of performing basic actions with Google Correlate, such as searching for a word and terms that are correlated and the ability to upload a personal time-series values and extract the top search terms correlated to the time-series.

2. GT: A class developed to mine Google Trends. Search terms can be downloaded and turned into a database.

3. SINAVE: A class used to automatically extract influenza-activity from SINAVE (IMSS' private health database).

`DBscrape` is built using Pandas and Selenium, and this library is open source with the purpose of encouraging other researches in expanding the number of available resources in the future.

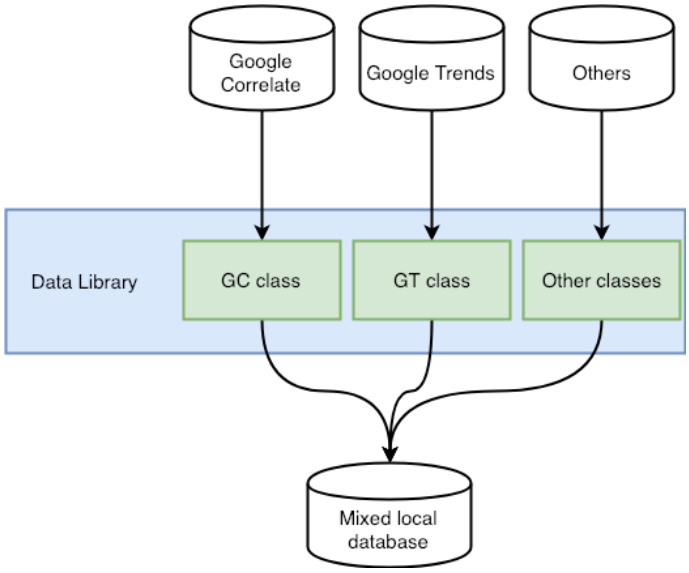


Figure 3.2: `DBscrape` is used as a tool to extract web-based data when the source lacks an option for doing this task (for example, extracting the activity of the word "influenza" for all the states within Mexico). `DBscrape` manages the online interactions and places the new files into a specified path, along with the data from other online databases.

3.2 DataFormatter

Mining web-based tools from different websites may result in a set of files that differ in their standards for presenting the data, and thus it is necessary to reorganize them into a structure that is more useful for our data science purposes, such as performing exploratory data analysis or fitting a machine learning model. The `DataFormatter` library is built to read in data inputs from web-sites such as Google Correlate, Google Trends and Flunet. The objective of this library is to be a complement from DB-Scrape in which, one mines the data, and the other reads and formats the data into a dataset ready to be used for predictive modelling,

The data library works using pandas, and stores the information specified by the user, making a distinction between target or external variables. This distinction is done to facilitate pre-processing of data and computation of autoreggressive lags (if used) in the `Modelling` library.

In this work, the data library is used to read documents from Google Correlate, Google Trends, and Flunet (see Figure 3.3).

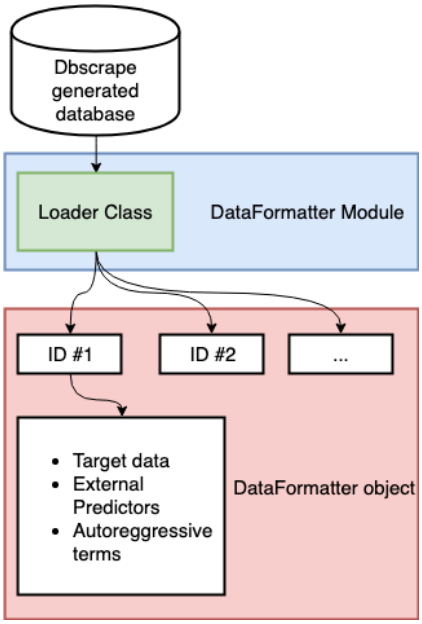


Figure 3.3: The `DataFormatter` library reads, extracts, and organizes the relevant information contained in different data files in a more effective structure contained within a Python object, ready to be used in different data science tasks.

3.3 Modelling

The `Modelling` module is a framework that facilitates predictive model generation using the ARGO methodology [29]. `Modelling` utilizes the `DataFormatter` library objects, which contains the information from various data sources, as input, and the user is able to select between different methodologies such as, ARGO, Net or a voting ensemble, among others (see Figure 3.4).

The `Modelling` library contains the following key features:

1. It follows a standard procedure which allows for multiple models to be fitted at once, for many locations, by just specifying them within a few lines of code.
2. It automatically manages autoreggressive lags within a model formulation.
3. It allows for customization of a model prior to fitting with the purpose of model comparing
4. It generates a data structure that contains all data regarding the model's estimates and model information.

5. It writes out and keeps records of the model's parameters and results.

6. It fits models that follow the same object structure than `sklearn`, a machine learning library for Python.

Throughout this project, the `Modelling` module was used to perform the fitting of AR, ARGO, Net and ensemble voting models for all the different locations of study, at different geographic resolutions.

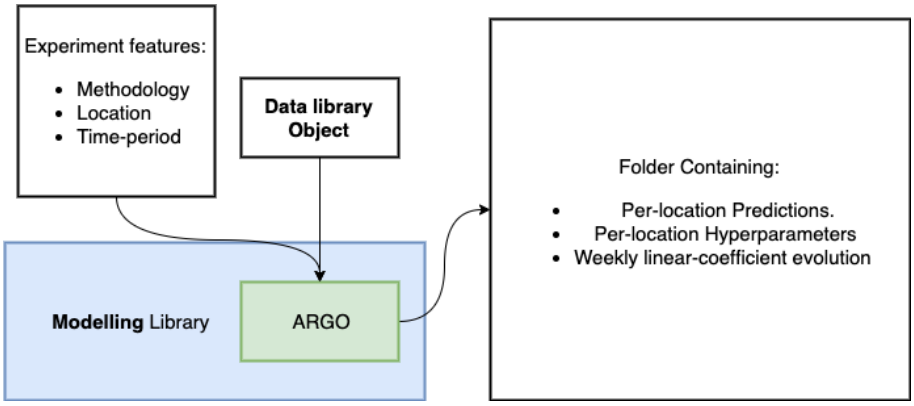


Figure 3.4: The `Modelling` library contains a set of classes that automatically deploy multivariate linear models such as ARGO for all the different locations provided within the `DataFormatter` class object. It also generates a data structure where all the results can be used for further visualization.

3.4 Visualizer

The `Visualizer` is a library that uses the data generated by both the `DataFormatter` and `Modelling` libraries to generate visualizations and benchmarking, among others. This library has the following objectives:

1. Easily perform a quick analysis of the models and their evolution in time with the ARGO methodology.

2. Allow for the straightforward comparison of models within a location and in between locations where disease surveillance models are being deployed.

3. Provide several templates of visualizations that are specifically designed for digital disease surveillance analysis.

The `Visualizer` library is divided into two main classes: The `inputVis` class, which is designed to read all information generated within an object from a class within the `DataFormatter` module, and the `outputVis` class, which reads the information written in the folder structure generated by the `Modelling` library (see Figure 3.5).

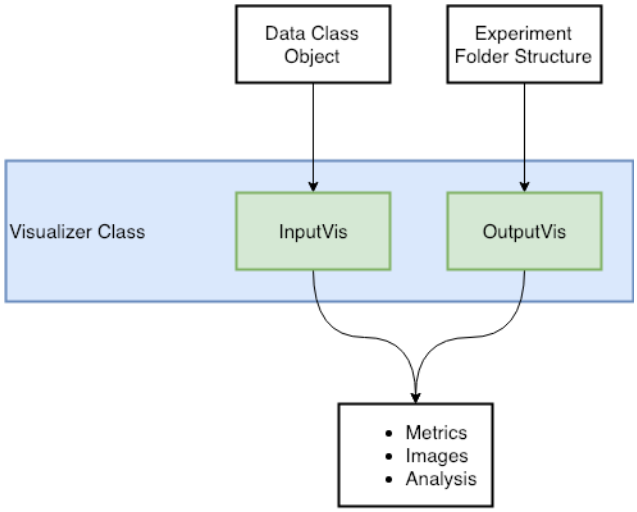


Figure 3.5: The `Visualizer` library generates metrics, graphics and visual comparison between the results of models within a location (performance) or in between (robustness and adaptability) locations. It works by calling a class and handing in the previously organized information from either the `DataFormatter` module or the `Modelling` module.

3.5 Summary

In this chapter Argotools was introduced as a python-based library for Disease forecasting models. From the initial phase of downloading external variable data using `DBscrape`, transforming it into a dataset compatible for generating predictive models using python using `DataFormatter`, to prototyping and developing with the `Modelling` library and benchmarking and quickly creating specialized graphics using the `Visualizer` class. A series of code examples for Argotools, created using Python Notebook are available at `https://github.com/LeonardoClemente/argotools-pkg/tree/master/examples`.

Chapter 4

Improved Real-time Influenza Surveillance using Internet Search Data in Latin America

A real-time methodology for monitoring flu activity in middle income countries that is simultaneously accurate and generalizable has not yet been presented. We demonstrate here that a machine learning method leveraging Internet-based search activity and past historical flu activity produces reliable flu estimates in multiple Latin American countries. The proposed model shows improvement against autoregressive models in countries where Google Correlate is available and may prove useful as a near-real-time surveillance tool.

4.1 Methods and Benchmarks

We extended the AutoRegressive model with General Online information (ARGO), a methodology originally conceived and tested to track flu activity in the United States in multiple spatial scales as a way to produce retrospective and strictly out-of-sample flu estimates individually for each country[11, 29]. This methodology is based on a multivariable regularized linear model that is dynamically recalibrated every week as new flu activity information becomes available. Besides online search information, ARGO incorporates short-term and seasonal historical flu information to improve the accuracy of predictions and mitigate the undesired effect of spikes in search activity (induced perhaps by overreaction in the population during potential health threats reported by the news). More details on this approach can be found in a study by Yang et al. [11]. Given a weekly as-yet-unseen NPS report to estimate, we used historical NPS and Google Trends information from the previous most recent two years (104 weeks) of data to calibrate ARGO and predict the

given week's NPS report.

To assess ARGO's predictive power, we built autoregressive models separately for each country that use historical flu activity from the most recent 52 weeks of activity (named AR52 throughout this paper) before predictions and generated retrospective out-of-sample estimates over the same time period. All models were built using the `glmnet` package on MATLAB version 2014a. [20, 26].

To compare the predictive ability of ARGO and AR52, we calculated Pearson correlations and the root mean square error between model predictions and the subsequently observed suspected flu cases. The added value of using Google search activity as a predictor was tested via the ratio between the mean square errors of AR52 and ARGO. For the efficiency metric, 90% confidence intervals were generated using the stationary block bootstrap method [18].

4.2 Results

Retrospective out-of-sample estimates of flu activity were produced, for each of the eight countries, from January 1, 2012 to December 25, 2016; and compared with the FluNet reported suspected cases (NPSs). Brazil's NPS data was only available until October 9, 2016. Note that because of FluNet's reporting delays, the models, which rely on past available values of FluNet and current Internet search activity, estimate current flu activity at least one week ahead of official reports.

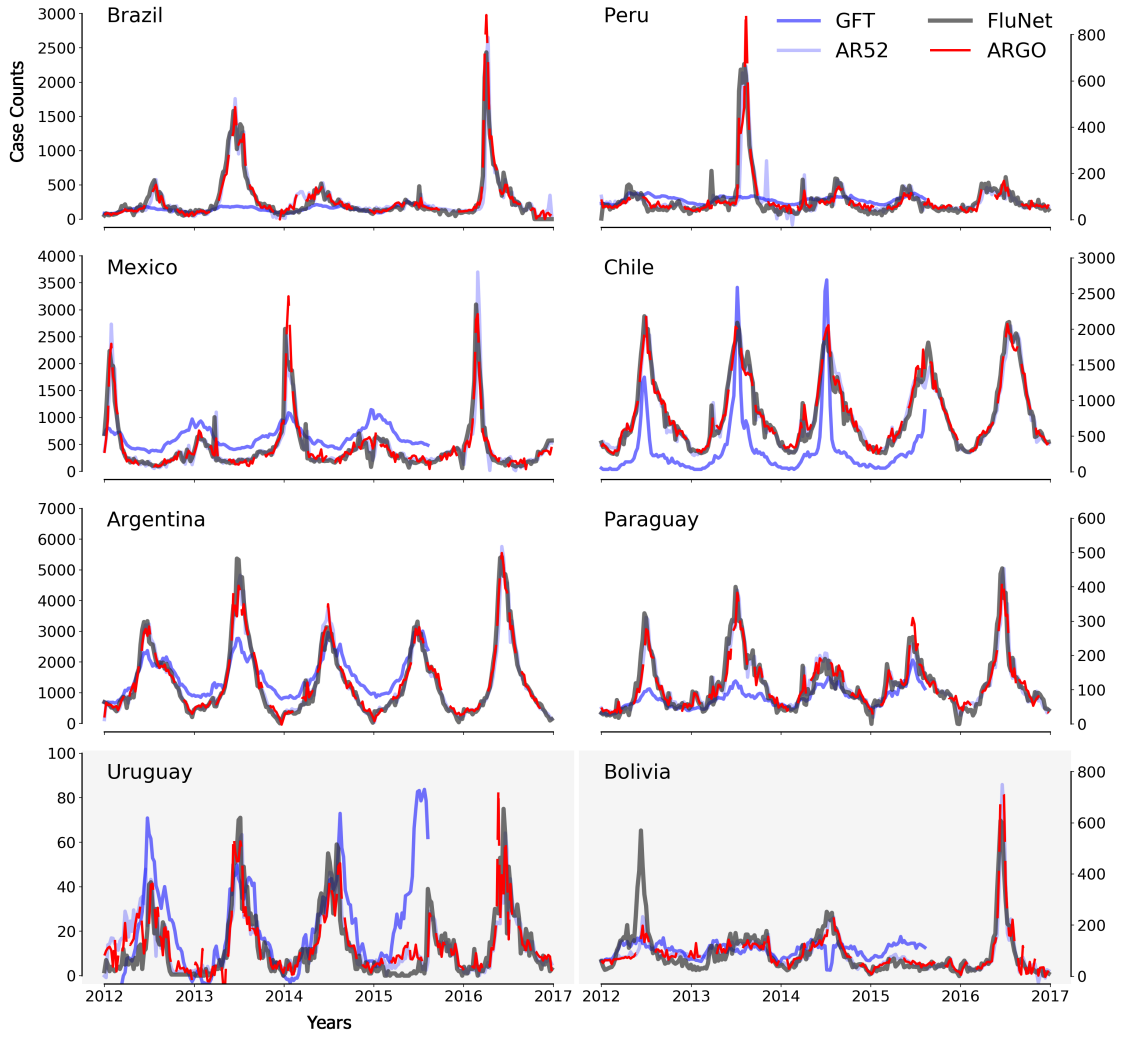


Figure 4.1: Graphical representation of the number of processed specimens (NPSs) as reported by WHO's FluNet (black), along with the NPS estimates generated by ARGO (red), AR52 (light blue), and Google Flu Trends (blue), over the whole study period of January 1, 2012 to December 25, 2016.

Figure 4.1 shows the real-time flu estimates and the subsequently observed suspected flu cases for each country. Contextually, historical GFT values (scaled to be displayed alongside with NPS values) and autoregressive estimates are also shown. The models

(ARGO and AR52) accurately predict NPS values in each country. GFT shows consistently large discrepancies when compared with the observed values, consistent with the findings reported by Pollett et al [19].

As shown in Table 4.1, ARGO shows an improvement in six countries in terms of the efficiency metric, reaching significant error reductions compared with AR52 in Brazil (155 to 104 or 33%), Mexico (243 to 184 or 24%), Peru (48 to 40 or 16%), and Chile (131 to 119 or 9%). Figure 4.3 shows a visualization of the performance of the models in the eight different models.

Table 4.1: Efficiency metric ($\frac{RMSE_{AR}}{RMSE_{ARGO}}$) for each country with 90% confidence intervals generated with the stationary block bootstrap. Scores above one indicate that ARGO incurred in less errors than AR52.

Country	Efficiency	5^{th} percentile	95^{th} percentile
Brazil	1.497	1.050	1.858
Mexico	1.320	1.023	1.778
Peru	1.197	0.976	1.322
Chile	1.104	1.031	1.205
Argentina	1.065	0.916	1.237
Paraguay	1.058	1.003	1.101
Uruguay	0.999	0.923	1.185
Bolivia	0.926	0.914	0.945

ARGO consistently outperforms GFT on Pearson correlations during the time period when GFT was active in every country and improves upon AR52 in all countries except Bolivia and Uruguay, over the whole study period, reaching significant correlation increases in Brazil (from 0.891 to 0.957), Mexico (from 0.86 to 0.92), and Peru (from 0.84 to 0.89). Refer to Tables 4.2 and 4.3 for a more in-detail explanation of the performance of the models.

4.3 Discussion

The overall improvement of ARGO over AR52 indicates that Internet search engine data, even in middle-income countries, provide increased responsiveness to changing disease trends. This improvement is clear in Brazil, Chile, Mexico, Peru, Paraguay, and Argentina, whereas in Uruguay and Bolivia, the inclusion of Google search data does not seem to improve AR52.

The availability of an online tool to select relevant flu-related terms (Google Correlate) that track historical flu activity was found to be a critical element for ARGO to improve performance over the autoregressive benchmark (Argentina, Chile, Mexico, Peru, and Brazil), suggesting that the most meaningful flu-related search queries are country-specific. In countries, such as Uruguay, where many weekly data points were missing on FluNet, ARGO's predictive ability was reduced. The best performance was seen in Brazil, Mexico, and Peru, where flu data was collected consistently every week during this study's time period (See Figure 4.2).

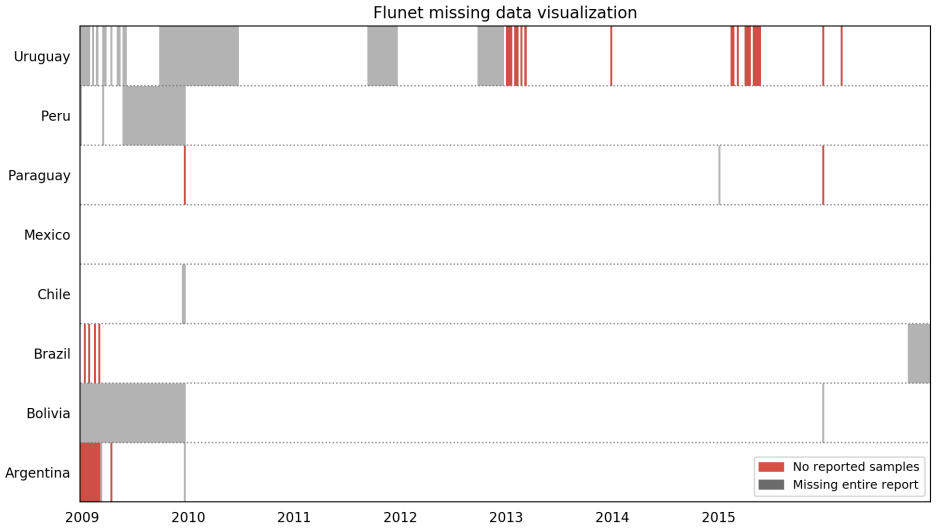


Figure 4.2: Heatmap displaying the report availability per country. For each row, red and gray vertical lines represent weeks where FluNet's NPS report was either missing or had no activity.

Based on previous research findings monitoring Dengue and Zika activity in Latin America [11, 28], we chose the number of suspected influenza cases (as captured by the

FluNet's NPSs) as the gold standard for the prediction tasks over the more standard Tested-positive proportion of confirmed influenza (computed by the ratio of the Confirmed for influenza cases and NPS). The choice was based on the intuitive fact that flu-related Google search activity is higher when more people "suspect" they may be affected by flu-like symptoms, regardless of the outcome of any lab test. As such, these models may prove useful to improve the timely allocation of resources in health care facilities in situations when increased numbers of people, with flu-like symptoms and respiratory needs, may need to be seen. It is relevant to point out that using NPS case counts as a gold standard implies that the models are not directly estimating confirmed influenza case counts but suspected Influenza-like Illness activity trends. The choice of gold standard is meaningful as it may help health care providers prepare for traffic fluctuations of patients presenting with symptoms of influenza. However, from an epidemiological perspective, more standard test-positive influenza proportions reported on previous Latin America studies [5, 19] should also be considered in future studies.

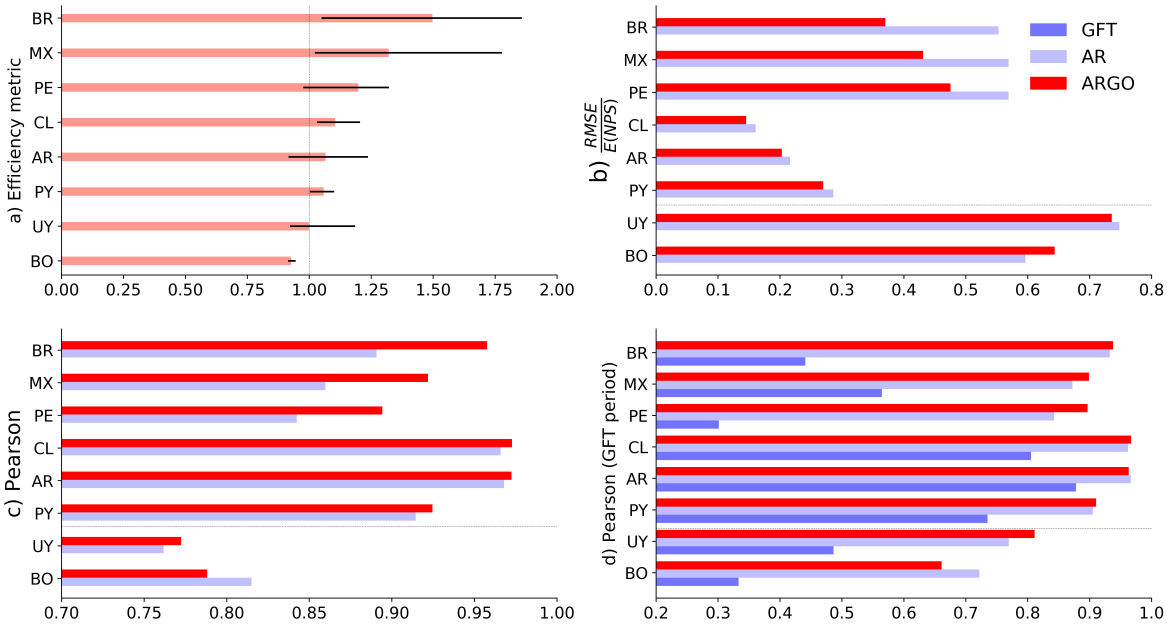


Figure 4.3: Set of bar graphs that shows the performance metrics to assess the predictive power of ARGO. (a) Efficiency metric values (salmon color) for each individual country with their respective 90 % CI (solid black line). (b) Root mean square error (RMSE) values for ARGO (red) and AR52 (light blue) during the whole study period. Each country's RMSE value is normalized by their respective average number of processed specimen (NPS) over the whole study period to avoid scale differences in visualization. c) Pearson correlation scores of ARGO and AR52 during the full period. d) Pearson correlation values for ARGO (red), AR52 (gray), and GFT (blue) during the period in the study where GFT was active.

Table 4.2: Pearson correlation for each model and country. The top performer in each time period is shown in bold. To allow comparison with GFT, correlation throughout 2015 only included the period before it was discontinued (August 9, 2015).

		Whole period	GFT period	2012	2013	2014	2015*	2016
Brazil	ARGO	**0.958**	**0.938**	**0.910**	0.953	0.705	0.441	**0.984**
	AR52	0.891	0.933	0.825	**0.957**	0.666	0.202	0.827
	GFT	–	0.441	0.625	0.766	**0.849**	**0.580**	–
Mexico	ARGO	**0.922**	**0.899**	**0.948**	0.790	**0.874**	0.890	**0.972**
	AR52	0.860	0.872	0.919	0.585	0.862	0.879	0.842
	GFT	–	0.565	0.531	**0.835**	0.617	**0.928**	–
Peru	ARGO	**0.894**	**0.897**	0.580	**0.904**	**0.694**	0.837	**0.814**
	AR52	0.842	0.843	**0.657**	0.843	0.618	**0.849**	0.804
	GFT	–	0.301	0.431	0.406	0.625	0.770	–
Chile	ARGO	**0.973**	**0.967**	**0.976**	0.963	**0.967**	**0.976**	**0.991**
	AR52	0.966	0.962	0.966	**0.965**	0.961	0.967	0.981
	GFT	–	0.805	0.888	0.850	0.809	0.940	–
Argentina	ARGO	**0.972**	0.963	0.963	0.968	0.965	0.968	**0.990**
	AR52	0.968	**0.966**	**0.969**	**0.975**	**0.966**	**0.979**	0.969
	GFT	–	0.878	0.919	0.938	0.937	0.971	–
Paraguay	ARGO	**0.925**	**0.911**	**0.928**	**0.927**	0.866	**0.914**	**0.956**
	AR52	0.914	0.905	0.918	0.920	0.822	0.874	0.931
	GFT	–	0.735	0.885	0.915	**0.879**	0.912	–
Uruguay	ARGO	**0.772**	**0.811**	0.628	**0.892**	0.857	0.050	0.681
	AR52	0.762	0.769	0.404	0.889	**0.894**	-0.072	**0.756**
	GFT	–	0.486	**0.811**	0.869	0.709	**0.183**	–
Bolivia	ARGO	0.788	0.661	0.584	0.675	0.914	0.605	**0.939**
	AR52	**0.815**	**0.722**	**0.688**	**0.789**	**0.923**	**0.630**	0.931
	GFT	–	0.333	0.641	0.607	0.001	0.610	–

Table 4.3: Root mean square error for each model and country.

		Whole period	GFT period	2012	2013	2014	2015*	2016
Brazil	ARGO	**104.38**	**94.78**	**53.78**	133.54	**92.90**	**72.52**	**155.91**
	AR52	155.87	101.03	72.23	**133.27**	101.94	76.81	316.27
Mexico	ARGO	**184.66**	**200.96**	**164.15**	**131.48**	312.47	**84.52**	**146.93**
	AR52	243.79	213.51	210.9	176.95	**290.15**	88.97	361.52
Peru	ARGO	**40.31**	**45.59**	24.83	**76.16**	**28.82**	21.89	**23.46**
	AR52	48.27	55.18	**23.62**	94.83	35.39	**20.36**	24.16
Chile	ARGO	**119.28**	**127.28**	**112.60**	141.74	**138.39**	104.47	**84.59**
	AR52	131.74	133.66	130.51	**138.86**	147.10	**103.96**	116.43
Argentina	ARGO	**274.12**	292.42	229.56	387.18	**265.30**	244.52	**217.92**
	AR52	292.01	**277.59**	**210.57**	**355.30**	289.78	**200.17**	376.24
Paraguay	ARGO	**30.81**	**31.37**	**29.19**	**36.50**	**28.83**	**29.68**	**31.75**
	AR52	32.679	32.30	29.55	36.73	31.65	29.88	38.10
Uruguay	ARGO	**10.01**	**9.27**	**11.07**	**8.91**	8.36	**8.41**	13.14
	AR52	10.17	10.32	14.10	8.92	**7.24**	8.57	**11.14**
Bolivia	ARGO	58.62	60.27	103.54	36.34	27.58	**17.88**	62.24
	AR52	**54.30**	**56.01**	**96.20**	**32.53**	**25.79**	19.91	**57.08**

Chapter 5

Implementing a Voting Ensemble for Influenza Forecasting in Latin America

Recent literature has shown the tremendous potential of digital data streams coming from the Internet as a tool for digital disease surveillance. When these streams are combined with historical influenza activity from health officials they become a powerful tool to generate accurate surveillance methodologies in data-rich countries such as the United States. In this chapter, we explore the potential that these state-of-the-art methodologies offer in more data-sparse scenarios in Latin America. We implement three state-of-the art methods based on three distinct sources of information: historical flu activity reports, Google search queries, and geo-spatial synchronicities of relevant neighboring locations, and finally combine their predictive power via a voting ensemble technique that analyzes the individual short-term historical performance with the objective of selecting the most adequate model to produce next week's prediction.

5.1 Contributions

We have generated an extensive study regarding the feasibility and predictive power of novel methodologies that incorporate external sources of information over three relevant spatial resolutions: state (27 states in Mexico), regional (five regions in Mexico), and national (15 countries in Latin America). We have also shown that the predictive accuracy of these methodologies can be further improved by the incorporation of a simple yet elegant voting ensemble technique that selects and combines the predictions of these methodologies through a short-term historical performance analysis.

Given the number of surveillance locations for this new study is almost six times in comparison to the work in section 6, different pre-processing strategies that improve the

performance of the proposed methodologies were developed. These strategies focus in increasing a model's robustness by allowing an automated local fine-tuning over time by either a) automatically adjusting the autoreggressive model features for AR, ARGO and Net to avoid predictive errors due to inconsistencies (such as lack of reports at certain point in time) in the official influenza activity reports from hospitals, b) Imposing a heuristic pre-processing method that adjusts the influence of external variables such as Google search queries and geo-spatial synchronicities prior to fitting or recalibrating a model's linear coefficients, and c) Introducing a data-driven approach that changes the strictness of such heuristic method over-time, increasing the adaptivity of our model to the predictive power of external sources of information.

5.2 Data and Methods

Three different data sources were used to work on this study: influenza activity reports from IMSS, influenza processed specimens from Flunet and Internet search frequencies from Google Trends and Google Correlate.

With the objective of separately leveraging the predictive power of internet-based activity and geo-spatial synchronicities, the following methodologies were implemented:

1. AR: An autoreggressive model consisting of only historical influenza activity. The autoreggressive model features consist in the past 52 weeks of most recent flu activity from hospital reports in the surveillance location.

2. ARGO: An ARGO model (Autoregression with General Online information) that combines flu-related internet search activity from the local population along with local official influenza activity reports.

3. Net: A model that incorporates the influence of flu activity from neighboring locations by the detection of geo-spatial synchronicities between the local flu activity and other locations at the same geographical resolution.

All models in this study were fit using the Least Absolute Shrinkage and Selection Operator (LASSO) (with a 10-fold cross validation), a multivariable linear approach that sets an $L1$ regularization constraint to the model's quadratic cost function. Our selection of LASSO as our fitting strategy is to use the $L1$ regularization as a means to mitigate the influence of weak features within our training dataset, setting their linear coefficient to 0. Moreover, each model was re-trained on a weekly basis using a constant size training dataset that contained the two most recent years of data prior to the date of prediction.

This shifting training window allows our models to re-calibrate their linear coefficients to the most recently available health reports and external data activity.

All influenza activity estimates were generated using `scikit-learn` in Python 3.4 [16]. An in-detail description of the methodology can be found in Section 4.1.

5.2.1 In-practice methodology improvements

Insight about the challenges that influenza surveillance in Latin America pose and how they originate were provided in our preliminary work in Chapter 4. These challenges frequently originate from characteristics such as:

1. The average volume of case counts within the location's official surveillance reports.
2. The degree of seasonality of the influenza activity.
3. The reporting consistency of health officials.
4. The reduction of predictive power in flu-related internet search activity when search volumes are influenced by events not entirely related to our task of predicting the number of people attending a physician exhibiting ILI symptoms (For example, an unexpected panic generated by the media).

Similarly, geo-spatial synchronicities from neighboring locations used as features for a local Net model may cause the model to behave erratically in times where the neighboring influenza activity does not keep the same synchronicity values from within the training data-set for future predictions.

The fitting and recalibration of AR, ARGO or Net consist in reorganizing the model's training dataset with the most recent 102 weeks of data. In a controlled research-scenario, data can be tidied to the point where analysis and experimentation can be optimized. In a more practical scenario, where a surveillance system is including new information from several data sources every week, the data cleaning and pre-processing has to be automated to ensure the data used to recalibrate a model is pre-processed optimally. Moreover, given the heterogeinity in the influenza activity of each surveillance location, pre-processing of data can differ for each surveillance location, leaving the data scientist with the task to find the best strategy to improve the performance of the novel methodologies for each of the locations (referred to local fine-tuning).

The following pre-processing procedures were developed and implemented in this work to implement a near real-time data cleaning and local-fine tuning of AR, ARGO and Net.

5.2.2 Managing missing Reports from Historical Flu Activity in the Autoreggressive model

The autoreggressive model used in this study consists in a 52 feature model, where each feature represents a historical flu activity report that happened in an specific week in the past. Conventionally, we name the activity reports that happened one week before 'AR1', two weeks before 'AR2', and so on. An autoreggressive model then consists in a set of autoreggressive lags AR1 through AR52 which, in other words, means that every week we look at the past 52 historical reports of influenza to make a prediction.

If the most recent historical flu report is not available in a timely manner, our most recent AR1 feature value is missing. In the dataset, the AR1 feature column contain a nan in the most recent sample (row). In this case, we fit our model for that week by removing the row in the dataset containing the NaN and using the rest of the dataset as we normally would. However, as time goes on, if the report is still missing, this NaN value will propagate towards the other autoreggressive lags, turning this strategy into a bad pre-processing choice.

In some cases, health officials may have missing reports for a continuous amount of weeks, making the Number of NaNs within the dataset bigger and continuous. If the gap NaNs is small (at most two missing reports) we perform linear interpolation to avoid removing the samples from the dataset. If the gap is too big, then we remove the samples containing the NaN values. Given the nature of the autoreggressive lags, big gaps of NaNs will reach a point where they have propagated through all the 52 features. To avoid having to remove most of the samples within our training dataset, we instead remove the features that containing the NaN gap. If the NaN gap appears in every feature, then we remove all features in exception to AR1.

Algorithm 1 Pre-processing for NaNs on 2-year window dataset

Require: $\mathbf{y}_t = \{y_t,\ y_{t-1},\ y_{t-2},\ ..., y_{t-102}\}$ (Target)
$X^{AR}_{t,102} = \{\mathbf{y}_{t-j} : j \in 1, 2...52\}$ (Autoreggressive lag features)
$X^{external}_{t,102} = \{\mathbf{x}_{e,t} : e \in E\}$ (External information features)
V : a set of non-removable columns in$X^{AR}_{t,102}$
$\mathbf{x}_{predict}$: the row vector used to predict y_{t+1}
for every column in $X^{AR}_{t,102}$ **do**
  if NaN detected within the column **and** column is not in V **then**
    Remove column
  end if
end for
for every row in $X^{AR}_{t,102}$ **do**
  if Row contains NaN **then**
    Remove row in both $X^{AR}_{t,102}$ and $X^{external}_{t,102}$
  end if
end for
for every value in $x_{predict}$ **do**
  if value is Nan **then**
    Remove feature from $x_{predict}$, $X^{AR}_{t,102}$, and $X^{external}_{t,102}$
  end if
end for

5.2.3 Adjusting the Influence of External Variables in ARGO and Net

As an empirical way to manage noisy external predictors during the recalibration process, we use the Pearson correlation coefficient between the historical flu activity and the external features as a pre-processing step to reduce the number of total variables for ARGO and Net. This approach has the advantage that it can let a variable number of terms as long as they satisfy the threshold condition. However, if the condition becomes too strict, there is a high chance that no external variables are included, reducing ARGO or Net into an AR model (which is, sometimes, a better practical choice based on the conditions and predictive power of the variables during that week). The Pearson correlation coefficient is a measure of the linear relationship between two variables and gives us insight about the potential of an external variable to be picked as a predictor for a multivariate linear model in an specific week. This pre-pocessing step is applied to flu-related Google search activity and to geo-spatial synhronicities every-time the two-year training dataset is updated.

We combined this pre-processing step with the voting ensemble scheme introduced

in Section 2.2.5 to allow both ARGO and Net to adjust the influence of the external variables every week. For a given methodology (ARGO or Net), a set of models using different thresholds is fit in real-time. The voting system then selects the model that has incurred in the least error in the past three weeks to make the prediction for the next week.

Algorithm 2 Adjusting influence of external variables

Require: $\mathbf{y}_t = \{y_t,\ y_{t-1},\ y_{t-2},\ ..., y_{t-102}\}$ (Target)
$X^{AR}_{t,102} = \{\mathbf{y}_{t-j} : j \in 1, 2...52\}$ (Autoreggressive lag features)
$X^{external}_{t,102} = \{\mathbf{x}_{e,t} : e \in E\}$ (External information features)
$M = m_i : i \in \mathbb{R}$ The set of models with different level of adjustment
$\alpha = \{\alpha_{m_i} \in [0, 1]\}$ Pearson coefficient threshold for each adjusted model
$\hat{Y} = \{\mathbf{y}_{m_i}\}$ Historical predictions for each adjusted model
Get $\mathbf{p} = \{p_k : k \in K\}$, by computing the Pearson correlation coefficient of each feature in $X^{external}_{t,102}$ and $\mathbf{y}_t$
for each m_i **do**
  Generate a new dataset $X^{adjusted}_{t,102}$ which contains only the external features such that $p_k >= \alpha_{m_i}$
  Do a pre-processing on $X^{adjusted}_{t,102}$ and $X^{AR}_{t,102}$ (see Algorithm 1)
  Merge $X^{AR}_{t,102}$ and $X^{external}_{t,102}$ into a single dataset $X_{t,102}$
Perform $LASSO(X^{AR}_{t,102}, \mathbf{y}_t)$ and predict for y_{t+1} storing the prediction in $\mathbf{y}_{m_i}$
end for
Perform a voting ensemble (see Section 2.2.5) to select which m_i prediction select for y_{t+1}

5.2.4 Voting Ensemble Approach

In order to optimally combine the predictive power of our surveillance methodologies, we implemented a voting ensemble approach based on a winner-takes-all system, as introduced in Section 2.2.5. The voting ensemble's prediction for a given week is assigned to be the prediction of the model that has incurred in the least error in the past K weeks. This parameter K has been fixed to three based on the emphasis to the most-recent short-term performance, and has also shown to be empirically successful in other works [10].

5.3 Results

We implemented ARGO, Net and the voting ensemble, as a way to produce retrospective and strictly out-of-sample flu estimates individually for each surveillance location. The

predictive performance of the models was compared against the autoreggressive model AR. Any improvement over the AR model can be attributed to the inclusion of these external sources of information in the case of ARGO and Net, and to the capacity to identify the best performing model in time in the case of the voting ensemble. The root mean squared error (RMSE) and Pearson Correlation Coefficient were used as our benchmarking metrics. These were computed over the following time periods:

1. From the study period of January 1, 2014 through December 25, 2017; to compare AR, ARGO, Net and the ensemble model.

2. Yearly, from 2014 to 2017, with the objective of conducting a seasonal analysis to all the models.

In-detail performance scores in terms of RMSE and Pearson Correlation coefficient can be found in Tables 5.1 through 5.5. Additionally, three main Figures were produced in the results as a way to visualize the individual and overall performance of AR, ARGO, Net and the voting ensemble.

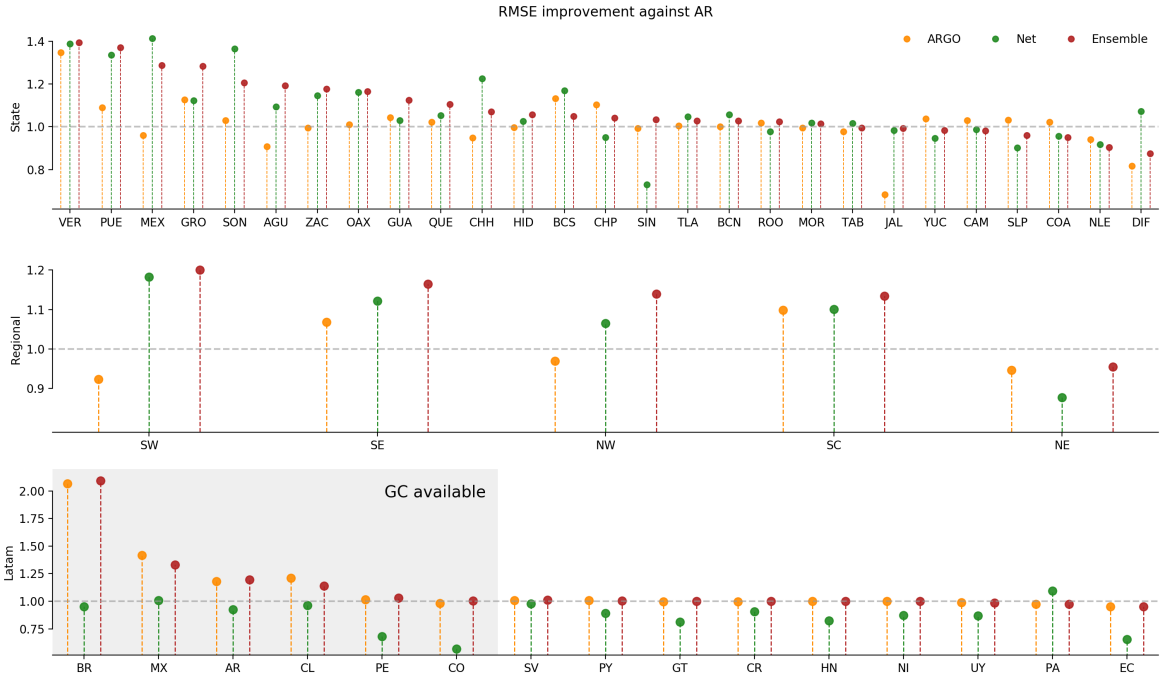


Figure 5.1: Lollipop chart that visualizes the model's performance ($\frac{RMSE_{AR}}{RMSE_{model}}$) for each country, region and state. The gray dotted line is a reference of the performance of AR in terms of RMSE. Scores above the gray dotted line represent an improvement against AR.

Figure 5.1 shows the performance of ARGO, Net and the voting ensemble over the whole time period in each of the 48 study locations, compared to the performance of AR (gray dotted line) This figure helps us to visualize the increase (or decrease) of performance of the predictive models by the incorporation of either the Internet activity or the geo-spatial synchronicities, and also the improvement that can be achieved by combining them via the voting ensemble.

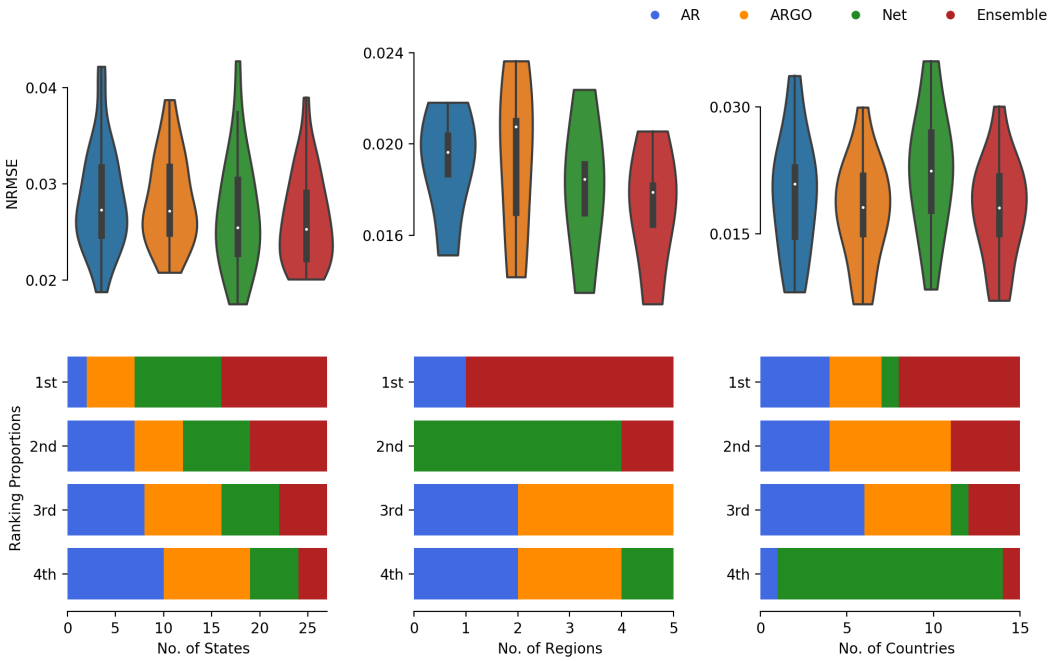


Figure 5.2: Violin plots that shows the robustness of performance of a model at different greographical scales. The inner part of the violin shows a box plot, with the median of the model's RMSE indicated by a white dot and the shape of the violin graphically shows the spread of the RMSE scores.

Figure 5.2 consists of two different visualizations that summarize the robustness of the models. The top visualization consists in violin distributions whose limits show the best and worst performance scores of the models. The violin distribution also shows the concentration of the scores of a model signaled by a bigger width and a boxplot that points the median of the RMSE scores by a white dot. The lower visualization shows the rank proportions. The first row shows, as color schemes, how many times did a model attain the first place in terms of RMSE, the second row shows how many times each model got second place, and so on.

Table 5.1: Methodology performance comparison for RMSE at state level. Lowest RMSE performance per location is signaled using bold.

Location	AR	ARGO	Net	Ensemble
SLP	11.369	**11.024**	12.612	11.839
SIN	7.385	7.435	10.129	**7.144**
ZAC	10.672	10.726	9.311	**9.07**
TAB	7.039	7.207	**6.929**	7.072
COA	8.364	**8.175**	8.75	8.796
MOR	18.132	18.216	**17.794**	17.859
AGU	9.881	10.896	9.024	**8.28**
OAX	9.207	9.107	7.921	**7.898**
QUE	10.284	10.059	9.757	**9.306**
VER	13.588	10.085	9.781	**9.746**
CHH	8.477	8.941	**6.919**	7.911
ROO	5.071	4.976	5.187	**4.95**
BCS	4.816	4.255	**4.12**	4.587
BCN	5.273	5.27	**4.99**	5.131
JAL	**15.673**	22.96	15.934	15.778
PUE	21.918	20.122	16.408	**15.979**
SON	7.214	6.998	**5.285**	5.981
GRO	6.982	6.193	6.217	**5.434**
MEX	9.491	9.894	**6.713**	7.364
YUC	2.983	**2.873**	3.153	3.034
CAM	5.45	**5.296**	5.516	5.558
CHP	3.357	**3.039**	3.532	3.22
TLA	10.398	10.352	**9.935**	10.107
HID	7.939	7.966	7.739	**7.508**
GUA	5.066	4.854	4.917	**4.501**
NLE	**10.487**	11.137	11.432	11.608
DIF	22.189	27.173	**20.705**	25.386

Table 5.2: Methodology performance comparison for Pearson correlation at state level. Highest coefficient performance per location is signaled using bold font.

Location	AR	ARGO	Net	Ensemble
SLP	**0.88**	**0.88**	0.87	0.87
SIN	0.85	0.84	0.75	**0.86**
ZAC	0.82	0.81	0.85	**0.87**
TAB	0.72	0.7	**0.73**	**0.73**
COA	0.86	**0.87**	0.84	0.84
MOR	0.85	0.84	**0.88**	**0.88**
AGU	0.83	0.79	0.86	**0.89**
OAX	0.83	0.84	**0.88**	**0.88**
QUE	0.85	0.85	**0.87**	**0.87**
VER	0.79	**0.88**	**0.88**	**0.88**
CHH	0.83	0.83	**0.89**	0.87
ROO	0.63	0.63	0.63	**0.65**
BCS	0.73	0.76	**0.79**	0.73
BCN	**0.88**	0.87	0.86	**0.88**
JAL	**0.93**	0.9	**0.93**	**0.93**
PUE	0.84	0.84	**0.87**	**0.87**
SON	0.83	0.84	**0.91**	0.88
GRO	0.71	0.76	0.78	**0.83**
MEX	0.82	0.81	**0.92**	0.9
YUC	0.57	**0.58**	0.5	0.54
CAM	0.34	**0.37**	0.35	0.31
CHP	0.66	**0.7**	0.61	0.66
TLA	0.84	0.83	**0.87**	0.86
HID	0.9	0.9	0.9	**0.91**
GUA	0.86	0.87	0.86	**0.88**
NLE	**0.91**	0.89	0.89	0.89
DIF	0.91	0.87	**0.92**	0.9

Table 5.3: Methodology performance comparison for RMSE at regional level. Lowest RMSE performance per location is signaled using bold.

Location	AR	ARGO	Net	Ensemble
SE	25.498	23.882	22.728	**21.6**
SW	44.577	48.29	37.688	**36.42**
SC	66.834	60.812	60.714	**59.177**
NE	**25.839**	27.313	29.449	26.937
NW	20.567	21.219	19.308	**17.758**

Table 5.4: Methodology performance comparison for Pearson correlation at regional level. Highest coefficient performance per location is signaled using bold font.

Location	AR	ARGO	Net	Ensemble
SE	0.88	0.89	0.9	**0.91**
SW	0.93	**0.94**	**0.94**	**0.94**
SC	0.91	**0.93**	**0.93**	**0.93**
NE	**0.93**	**0.93**	0.91	**0.93**
NW	0.9	0.9	0.92	**0.93**

Table 5.5: Methodology performance comparison for RMSE at national level. Lowest RMSE performance per location is signaled using bold.

Location	AR	ARGO	Net	Ensemble
PE	19.906	19.604	29.289	**19.34**
SV	15.332	15.212	15.681	**15.184**
PY	29.121	**28.871**	32.687	28.959
NI	**24.746**	**24.746**	28.326	**24.746**
GT	12.155	12.188	14.961	**12.133**
CR	20.028	20.127	22.079	**20.02**
CO	21.298	21.672	37.425	**21.179**
EC	**19.433**	20.427	29.708	20.426
UY	**9.512**	9.617	10.965	9.652
PA	29.204	29.948	**26.678**	30.025
AR	315.746	267.78	340.908	**264.177**
MX	239.861	**169.32**	238.266	180.457
HN	**13.281**	**13.281**	16.155	**13.281**
BR	184.109	89.075	193.656	**87.988**
CL	116.33	**95.985**	120.989	101.981

Table 5.6: Methodology performance comparison for Pearson correlation at national level. Highest coefficient performance per location is signaled using bold font.

Location	AR	ARGO	Net	Ensemble
PE	0.82	**0.83**	0.66	**0.83**
SV	**0.72**	**0.72**	0.71	**0.72**
PY	**0.92**	**0.92**	0.9	**0.92**
NI	**0.87**	**0.87**	0.82	**0.87**
GT	**0.64**	**0.64**	0.51	**0.64**
CR	**0.6**	**0.6**	0.52	**0.6**
CO	**0.94**	**0.94**	0.82	**0.94**
EC	**0.92**	0.91	0.82	0.91
UY	**0.82**	**0.82**	0.76	**0.82**
PA	0.88	**0.89**	**0.89**	**0.89**
AR	0.97	**0.98**	0.96	**0.98**
MX	0.86	**0.93**	0.87	0.92
HN	**0.68**	**0.68**	0.53	**0.68**
BR	0.77	**0.96**	0.75	**0.96**
CL	**0.98**	**0.98**	0.97	**0.98**

5.3.1 ARGO

Figure 5.2 shows ARGO being on the third position (next to the ensemble model and Net) of number of states having the least RMSE. ARGO improves over AR in 15 out of 27 states, and has significant improvements over AR on Veracruz and Baja California Sur. On a year to year basis, ARGO is the third model with the most second places (next to the ensemble model and NET). ARGO performs similarly to AR with 53 out of 108 yearly periods beating AR in terms of RMSE but falling in the last place of all models with most yearly periods on the first and second rank. ARGO's underperformance on Jalisco and Distrito Federal arise from overshooting away from the real target values on the outbreaks of December, 2016 and January, 2014; respectively. An analysis of the regional level study shows that ARGO underperformed for every region, reaching at most the third lowest RMSE every time. Yearly analysis also shows underperformance of ARGO by only being the model with the least RMSE in only two yearly periods out of 20.

At national level (see Figure 5.1), ARGO consistently improves in Argentina, Brazil, Mexico and Chile, and then follows to perform almost similarly than AR. The whole period analysis shows ARGO as being the third model with least RMSE and second at having the most periods with least and second to least RMSE.Year to year analysis displays ARGO (next to AR) as the second model with least RMSE in most yearly periods but the model with most yearly periods with least and second to least RMSE. Overall, ARGO greatly improved in countries where Google Correlate service was available, in exception to Colombia, whose data collecting process did not yield any relevant search terms correlated with its national influenza reports. For all the other countries, ARGO performed similarly to AR.

5.3.2 Net

Net improves over AR in 18 out of 27 countries in terms of RMSE (Figure 5.2), having great improvements over AR and ARGO in Puebla, Mexico city, Sonora and Veracruz, but also underperforming in states like Sinaloa, Coahuila, San Luis Potosi and Nuevo Leon. On a year to year analysis, Net's performance shows improvement over all models in 41 yearly periods in terms of correlation and 33 in terms of RMSE. Net is also the second model (next to the voting ensemble) to have the most scores in first and second place. For the yearly periods where Net performed poorly, Net was also the model with most scores in fourth place. Nonetheless, Net was the second model with least yearly RMSE scores, just next to the voting ensemble. However, Net did not improve in states that had either really noisy curves and uncharacteristic activity such as Sinaloa's flat outbreak spike in 2015, San Luis Potosí first weeks after 2014's outbreak.

Looking at the performance at regional level of Net over the whole period (Figure

5.1), it became the second model (just next to the voting ensemble) to perform and get the second to least RMSE scores in four out of five regions. Net successfully combines information from neighboring countries, showing that short-term synchronicities and geospatial structures still work within a regional scheme.

Net underperformed and scored as the model with highest error both in the whole period and year to year analysis. Net does not succesfully abstract any meaningful information based on the neighbor's synchronicities and past influenza activities, often overfitting and overpredicting just like in Peru and Honduras in 2017.

5.3.3 Voting Ensemble

As shown in Figure 5.2, the voting ensemble performed better than AR in 19 out of 27 states, and also turned out to be the model with least RMSE compared to the other models (11 states) and the model with most performances in first and second place. On a year to year comparison, the voting ensemble yielded as the second model with the most yearly periods having least RMSE and the top model with most yearly periods with least or second to least RMSE. The voting ensemble shows to be consistent when combining both ARGO and Net. In particular, The voting ensemble failed to reduce overshooting for Yucatan in 2014 and Mexico City in 2016, but successfully reduced overshooting in various yearly outbreaks such as Puebla, Guerrero, Estado de Mexico and Tlaxcala in 2016.

The voting ensemble becomes the model with lowest RMSE in four out of five regions by analyzing the whole period (Figure 5.1), only falling second in terms to RMSE to AR in the North-east region. The special case of NE at regional levels shows a situation where the ensemble model did not improve over the baseline AR. Nonetheless, the model is still able to reach the second lowest RMSE of all. The main factor that caused the ensemble model to have a bigger RMSE than AR is attributed to the fact that, during the epidemic season at the start of 2014, the voting ensemble selects Net over the AR model, which for the following three weeks incurred in less accurate predictions. Given the fact that 2014 is the highest epidemic year of them all, three weeks with less accurate predictions accounted for a noticeable increase in terms of RMSE (see Figure 5.3).

The decisions taken by the voting ensemble on which model will predict next week, as presented in Figure 5.4, demonstrate that the selection rule based on short-term historical performance is a simple, yet effective way to reduce prediction error. We can see that for several cases at state level, the ensemble is able to detect a reduction in performance for the previous three weeks and successfully avoids increasing error by switching to another model (for example, both ensembles at MX-BCH and MX-SIN show this behavior for the epidemic outbreak of 2016-2017).

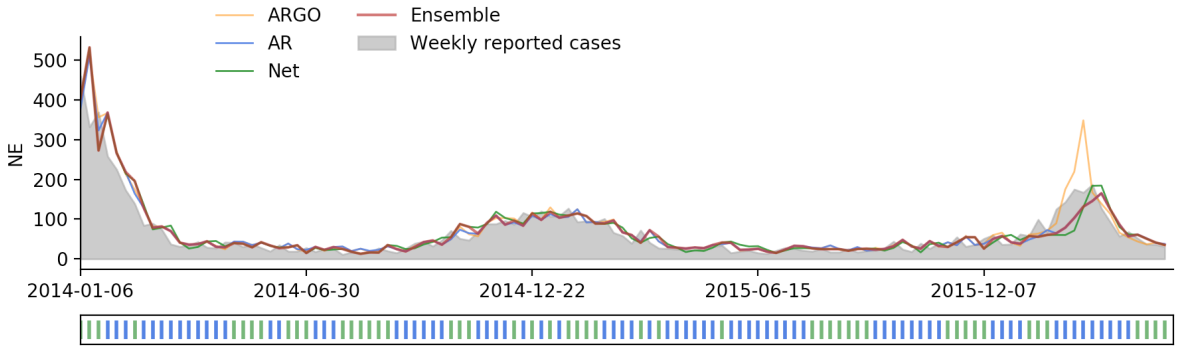


Figure 5.3: Retrospective regional-level predictions for the north-eastern region. The sequence of colored bars below the estimates show the evolution of the ensemble's decision over time.

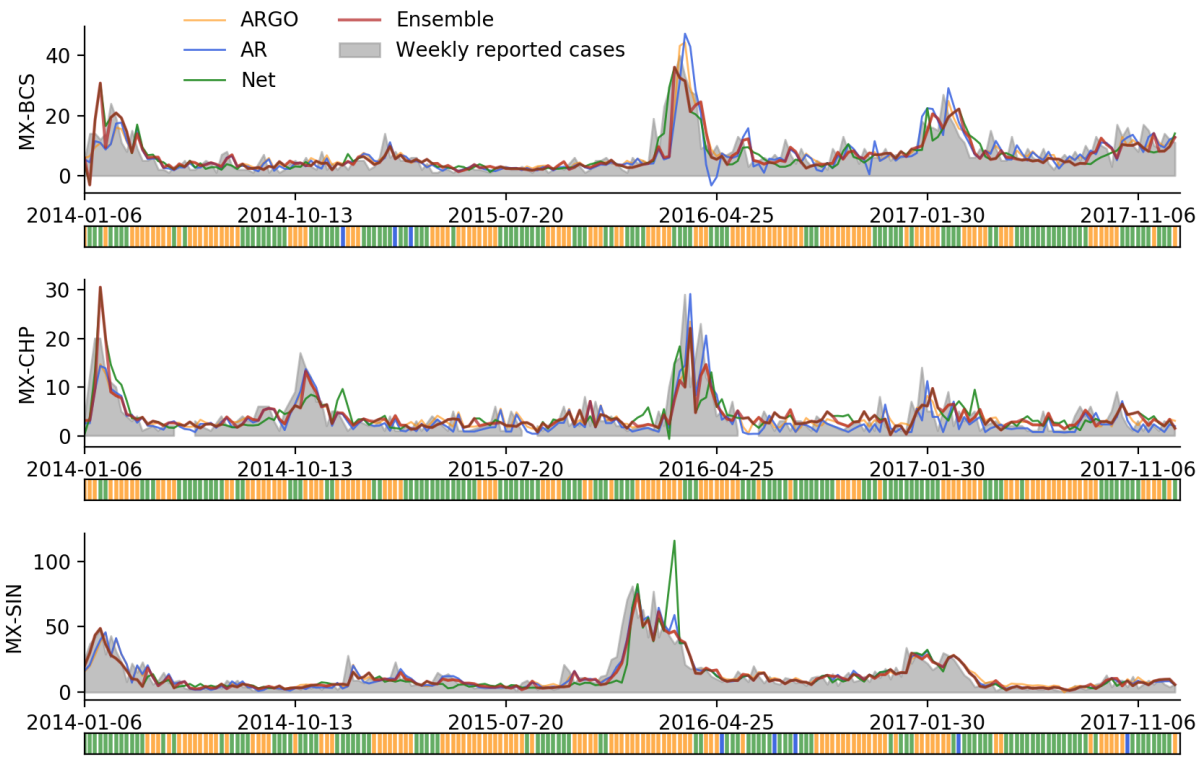


Figure 5.4: Retrospective state-level predictions for three different states in Mexico. The colorbar below the estimates show the evolution of the ensemble's decision over time.

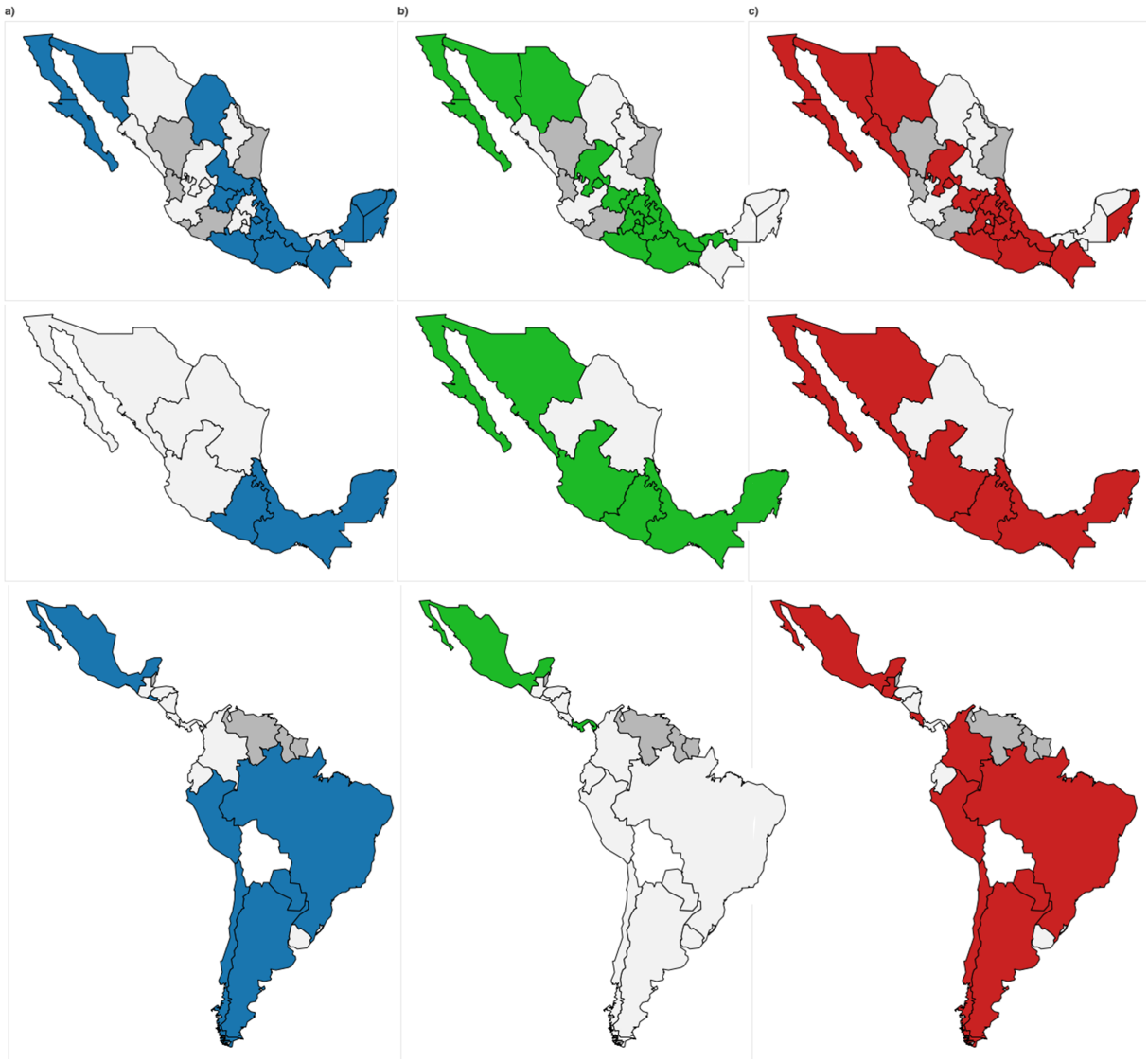


Figure 5.5: Geographical heat-maps that show the locations where each model improved upon AR for each geographical level. From top to bottom, the state, regional and country levels are shown, and from left to right, the model order is ARGO, Net and the voting ensemble.

5.4 Discussion

The results in this study show the potential that incorporating external variables such as flu-related Internet information and geo-spatial synchronicities of flu reports offer to digital influenza surveillance in Latin America. Moreover, we show that the combination of these sources of information through a voting ensemble offer a robust way to leverage the strengths and weaknesses that each method (AR, ARGO and Net) may experience based on the local characteristics of each study location over-time. We believe that this study may be useful for health officials as positive evidence for the establishment of a real-time tool for influenza surveillance that incorporates different sources of information.

5.4.1 Predictive Power of Search Term Frequencies

The success in the combination of Internet search trends and historical flu activity at national scale may be predicted in advance by measuring how well our selected ground truth correlates with Google Correlate, a fact that agrees with previous findings [3]. In this study, the data preparation process involved querying Google search terms that correlated well with our historical flu activity using only data prior to the study period. In practice, we could replace step as an iterative process that happens every new week, with the purpose of guiding our automated process as to which new Google search terms, not taken in account yet within our current pool, are becoming relevant (i.e. the change of name of a medication that is widely used to treat influenza), and temporally discarding the terms that do not show strong correlation.

The inability to access Google search term frequencies at regional level limits ARGO's predictive power, in comparison to the ARGO models generated at state and national levels. Google Correlate delivers only national level data, whereas Google Trends can deliver both national and state level information. Using national level trends is not a feasible option based on the fact that trends that do not belong to the location to predict are included within the time-series and often misguide ARGO. Using state-level trends becomes a more feasible option, but given Google does not provide information regarding their normalization and zero-patching process, an additional strategy may be necessary prior to combining the Google search trends into a regional time-series.

Figure 5.5 shows a geographical map containing information of each model performance in comparison to AR for the three different geographical-resolutions. The majority of the states that did not improve by incorporating Internet search trends were either in the south east or north-east regions and had neighboring states that were excluded from the study because of the bad quality in their historical flu activity reports.

Regional predictions can be an useful proxy for states with poor influenza surveillance. In Figure 5.5, we can see that both Net and the ensemble models are capable of producing estimates that improve over AR at the south-east and north-east regions, which account for most of the states that excluded from the state-level study.

5.4.2 Analysis of the Inclusion of Geo-spatial Synchronicites on State-level and Regional-level Estimations

Results for Net showed that spatio-temporal synchonicities did not provide predictive power to the models at national level, whereas at regional level and state level, they increased the performance. States that showed unusual influenza activity curves during the study period had the least improvement or underperformed by incorporating neighbor information. Some examples are the curves from Campeche, Yucatan and Quintana Roo. As geographic resolution increases from state-level to regional-level, the predictive power of incorporating neighboring activity becomes more evident. Net results show a big improvement in the south east, south and north east regions (see Figure 5.5). Given it is more common for people to travel within their own country than internationally and, as public transport such as airplanes or electric trains become more accessible to people, extending their potential of transmission to longer distances, we believe the inclusion of geo-spatial synchronicities that are not geographically close to the locations of prediction is a good strategy that resulted in the improvement of Net in Mexico. We let the L1 regularization and pre-proccessing prior to fitting a model decide which locations were the most useful at a particular point in time instead of setting a geographical limitation on what state's influenza activity can influence another,

5.4.3 The Voting Ensemble as a Method for Increasing Robustness and Performance

Our ensemble methodology succeeded at combining the methodologies that separately harnessed local historical flu activity, Internet search activity and geo-spatial synchronicities, as shown in the improvement at all geographical resolutions (see Figure 5.2).

Internet search behavior, the epidemic seasonality and geographical heterogeneity are different for every location, creating a robustness challenge for digital disease surveillance. Our ensemble model shows great adaptability to almost every scenario given its capacity to switch from one model to another in a relatively short-time window, and even though it is not always the model with the least RMSE, it is always in the first or second place (see Figure 5.2). At national level, the combination of Google search activity and local health influenza had better results. At regional level, given the limitations that Google

imposes upon the way data is extracted, the implementation of Google search queries as a predictor becomes a bigger challenge, leaving us with neighbor influenza activity and local influenza activity as the data sources to build our ensemble on. At state level, both Google search queries and neighbor's influenza activity are available, and therefore, combining both ARGO and Net becomes a viable approach.

5.5 Summary

In this chapter we presented the implementation of Net, ARGO and AR as methodologies that individually harness the predictive power of external variables (flu-related Internet activity from Google Trends, geo-spatial synchronicities and official health reports from local public health organizations) in the task of influenza activity surveillance. Moreover, we presented a voting ensemble implementation that successfully combines the predictive power of these methodologies and serves as a means to overcome the heterogeneous behavior of influenza dynamics as the study locations changes geographically and over-time. We showed that each individual methodology may be adequate in specific scenarios over-time, but that cases such as the spatial-resolution highly affect their performance (ARGO in a regional scale, for example). We also demonstrated that, independently of the spatial resolution at which the voting ensemble is implemented, it has the ability to consistently reduce the prediction error.

Chapter 6

Conclusions and Future Work

This thesis describes a novel machine-learning method that accurately tracks and predicts influenza activity, as reported by SINAVE and Flu-Net, one week ahead of time. This task was done for three different geographic scales (state level and regional level for the country of Mexico, and national level for various countries in Latin America). We showed that ARGO, a model that leverages historical flu activity along with Google Internet search trends, and Net, a model that combines influenza activity and spatio-temporal trends, successfully improves over AR, a state-of-the-art autoreggressive method which relies on local historical flu activity. More importantly, we showed that, through the use of a voting ensemble that successfully selects between these three models, we can improve influenza forecasts that reach higher correlations and lower error estimates than the three of them separately, while also increasing the adaptivity.

The models described in Chapter 5 are built with various sources of information: ARGO supports its prediction based on the activity of relevant Google search terms that may represent people's concern of being infected with the disease; Net takes in account, aside of the local influenza activity, the short-term historical influenza activity of neighboring locations; and AR by itself only takes into account the local historical influenza activity, having less access to information but a better shot at predicting influenza when these external sources of information do not represent the ground's truth activity. The voting system proposed in this work is capable of improving over the three models by objectively detecting in near real-time which model has less chance of error based on their historical performance in the short-term past, generating a balance of robustness and responsiveness.

The use of Google search activity as flu activity proxies has shown to be a successful proxy of influenza activity in Latin America. However, the inability of detecting

when these predictors have a trend that originates from people's overreaction to an influenza event, and thus leading ARGO to overshooting the influenza prediction estimates. Our voting ensemble implementation is able to detect this type of problem with acceptable responsiveness and therefore avoids severe overshooting from ARGO by alternating it with a more adequate model at that specific time. Similarly, we have also shown that spatio-temporal synchronicities improved model accuracy at both state and regional scales. However, our results have also shown that the correlation between two different neighboring locations might be similar only over a certain number of seasons, thus leading to the possibility of Net incorrectly weighting neighbor activity and leading to error. In this case, the re-calibration of the weights, along with updating the training dataset with the most up-to-date available data help mitigating this problem.

In contrast to the number of confirmed influenza cases and confirmed cases proportions, as presented more commonly in epidemiological studies [23, 5], our choice of ground truth is the number of suspected cases (NPS) of influenza. The number of suspected cases is a source of information that does not only describe ILI activity, but also might be related to other diseases such as Severe Acute Respiratory Illnesses (SARI). Although this may seem like a weakness in our approach, we believe that our choice is meaningful, given that our main objective in this study is to estimate the number of people that will attend a physician presenting symptoms of influenza, a trend that is better modelled by, for example, the flu-related Internet activity collected from Google Trends and Google Correlate, based on the intuition that activity is higher when more people "suspect" they may be affected by flu-like symptoms, regardless of the outcome of any lab test.

It is relevant to add that AR, ARGO and Net rely heavily on the availability of historical flu activity reports from health officials. The inability from public health institutions in providing reports consistently poses a limitation in the predictive power of our models, and in very special cases, may turn our methodologies infeasible for specific surveillance locations. In the case when a new disease surveillance system is being introduced in official health institutions and no historical disease data is available, alternative methodologies that rely solely in external information (internet-based activity, historical data from related diseases, or geo-spatial synchronicities from surveillance locations that have information about the disease) might be an alternative. As time passes and the hospital generates data, an implementation AR, ARGO or Net could be implemented using a smaller (and increasing) training window.

In this work, we have demonstrated that a voting ensemble can increase the robustness in the task of influenza nowcasting in Latin America. However, there are still some situations that stem from the mechanics of the voting ensemble for which this methodology may not be the most adequate. This situation arises whenever a model that has dominated in accuracy during a high epidemic season for the past few weeks starts predicting

influenza activity with less accuracy than the other methodologies. Given the voting ensemble only takes in account the historical performance in the most recent short-term, the voting ensemble will keep choosing this model until the historical performance has enough data to make the switch. This is the case for the first 3 weeks in 2014 for the north-east region at regional level (see Section 5.3.3), where the voting ensemble selects Net over AR, causing the voting ensemble to incur in the same high error that Net incurred for those weeks.

6.1 Contributions

Digital disease surveillance is a recent research area, and in-practice knowledge that allows for better a understanding and fine-tuning of influenza predictive models are yet to be explored. In this work, our main contributions to the research community are:

- Generated strategies to address the challenges involved in implementing external data sources (Section 2.1.5, and Section **??**).
- Implemented novel ideas of pre-processing and hyper-parameters fine tuning that allow our models to reach higher out-of-sample results (Section 4.1, and Section 5.2.2).
- Devised a heuristic feature-selection approach which actively filters and adjusts the influence of external data-sources used in novel flu techniques (Section 5.2.3).
- Presented first multi-scale prediction study that incorporates internet-based search activity and spatio-temporal synchronicities in Mexico and the first study in Latin American countries (Chapter 5).
- Developed a python-based library that facilitates the implementation of these methods in all the steps of model research (data collection, pre-processing, EDA, model fitting and post-analysis) with both the purpose of reproducing the results presented and also aiding new researchers in the generation of results (Chapter 3).

6.2 Future Work

One of the main goals of this research was to propose and implement a methodology for ILI forecasting that could be applied to real-life scenarios. The results obtained confirm this idea on a limited set of cases. The future work of this investigation should focus on

extending this methodology to other case studies (national, state or even local) in different countries within Latin America, to confirm the robustness of the voting ensemble model.

This study presented a near-real time methodology for predicting influenza with a week in advance. However, the lack of consistency in the reporting of influenza activity within Flunet or SINAVE may result in a caveat for the implementation of a real-time disease surveillance system. As future work, we would like to extend this methodology (or propose a novel one) that is able to produce influenza estimates with two or three weeks in advance.

Our results have shown that the voting ensemble system outperformed every other model considered for this investigation in terms of RMSE and Pearson correlation. However, we think that further improvements can be achieved by finding a data-driven approach that facilitates model selection and an adaptive time window selection over which the voting ensemble can cast the votes instead of an static time-window for all the models.

Bibliography

[1] BROOKS, L. C., FARROW, D. C., HYUN, S., TIBSHIRANI, R. J., AND ROSENFELD, R. Flexible modeling of epidemics with an empirical bayes framework. *PLoS computational biology 11*, 8 (2015), 1–18.

[2] CHARU, V., ZEGER, S., GOG, J., BJØRNSTAD, O. N., KISSLER, S., SIMONSEN, L., GRENFELL, B. T., AND VIBOUD, C. Human mobility and the spatial transmission of influenza in the united states. *PLoS computational biology 13*, 2 (2017), e1005382.

[3] CLEMENTE, L., LU, F., AND SANTILLANA, M. Improved real-time influenza surveillance: Using internet search data in eight latin american countries. *JMIR public health and surveillance 5*, 2 (2019), e12214.

[4] CORTES-ALCALA, R., DOS SANTOS, G., DEANTONIO, R., DEVADIGA, R., RUIZ-MATUS, C., JIMENEZ-CORONA, M. E., DIAZ-QUINONEZ, J. A., ROMANO-MAZZOTTI, L., CERVANTES-APOLINAR, M. Y., AND KURI-MORALES, P. The burden of influenza a and b in mexico from the year 2010 to 2013: An observational, retrospective, database study, on records from the directorate general of epidemiology database. *Human vaccines & immunotherapeutics* (2018), 1–9.

[5] DURAND, L. O., CHENG, P.-Y., PALEKAR, R., CLARA, W., JARA, J., CERPA, M., EL OMEIRI, N., ROPERO-ALVAREZ, A. M., RAMIREZ, J. B., ARAYA, J. L., ET AL. Timing of influenza epidemics and vaccines in the american tropics, 2002–2008, 2011–2014. *Influenza and other respiratory viruses 10*, 3 (2016), 170–175.

[6] EYSENBACH, G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In *AMIA Annual Symposium Proceedings* (2006), vol. 2006, American Medical Informatics Association, p. 244.

[7] GENEROUS, N., FAIRCHILD, G., DESHPANDE, A., DEL VALLE, S. Y., AND PRIEDHORSKY, R. Global disease monitoring and forecasting with wikipedia. *PLoS computational biology 10*, 11 (2014), e1003892.

[8] HICKMANN, K. S., FAIRCHILD, G., PRIEDHORSKY, R., GENEROUS, N., HYMAN, J. M., DESHPANDE, A., AND DEL VALLE, S. Y. Forecasting the 2013–2014 influenza season using wikipedia. *PLoS computational biology 11*, 5 (2015), e1004239.

[9] LIPSITCH, M., FINELLI, L., HEFFERNAN, R. T., LEUNG, G. M., AND REDD; FOR THE 2009 H1N1 SURVEILLANCE GROUP, S. C. Improving the evidence base for decision making during a pandemic: the example of 2009 influenza a/h1n1. *Biosecurity and bioterrorism: biodefense strategy, practice, and science 9*, 2 (2011), 89–115.

[10] LU, F. S., HOU, S., BALTRUSAITIS, K., SHAH, M., LESKOVEC, J., SOSIC, R., HAWKINS, J., BROWNSTEIN, J., CONIDI, G., GUNN, J., ET AL. Accurate influenza monitoring and forecasting using novel internet data streams: a case study in the boston metropolis. *JMIR public health and surveillance 4*, 1 (2018).

[11] MCGOUGH, S. F., BROWNSTEIN, J. S., HAWKINS, J. B., AND SANTILLANA, M. Forecasting zika incidence in the 2016 latin america outbreak combining traditional disease surveillance with search, social media, and news report data. *PLoS neglected tropical diseases 11*, 1 (2017), e0005295.

[12] OLSON, D. R., KONTY, K. J., PALADINI, M., VIBOUD, C., AND SIMONSEN, L. Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS computational biology 9*, 10 (2013), e1003256.

[13] O'SHEA, J. Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International journal of medical informatics 101* (2017), 15–22.

[14] PAOLOTTI, D., CARNAHAN, A., COLIZZA, V., EAMES, K., EDMUNDS, J., GOMES, G., KOPPESCHAAR, C., REHN, M., SMALLENBURG, R., TURBELIN, C., ET AL. Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience. *Clinical Microbiology and Infection 20*, 1 (2014), 17–21.

[15] PAUL, M. J., AND DREDZE, M. You are what you tweet: Analyzing twitter for public health. *Icwsm 20* (2011), 265–272.

[16] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[17] POLIKAR, R. Ensemble based systems in decision making. *IEEE Circuits and systems magazine 6*, 3 (2006), 21–45.

[18] POLITIS, D. N., AND ROMANO, J. P. The stationary bootstrap. *Journal of the American Statistical association 89*, 428 (1994), 1303–1313.

[19] POLLETT, S., BOSCARDIN, W. J., AZZIZ-BAUMGARTNER, E., TINOCO, Y. O., SOTO, G., ROMERO, C., KOK, J., BIGGERSTAFF, M., VIBOUD, C., AND RUTHERFORD, G. W. Evaluating google flu trends in latin america: important lessons for the next phase of digital disease detection. *Clinical Infectious Diseases* (2016), ciw657.

[20] QIAN, J., HASTIE, T., FRIEDMAN, J., TIBSHIRANI, R., AND SIMON, N. Glmnet for matlab. *Accessed: Nov 13* (2013), 2017.

[21] SANTILLANA, M., NSOESIE, E. O., MEKARU, S. R., SCALES, D., AND BROWNSTEIN, J. S. Using clinicians' search query data to monitor influenza epidemics. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America 59*, 10 (2014), 1446.

[22] SANTILLANA, M., ZHANG, D. W., ALTHOUSE, B. M., AND AYERS, J. W. What can digital disease detection learn from (an external revision to) google flu trends? *American journal of preventive medicine 47*, 3 (2014), 341–347.

[23] SAVY, V., CIAPPONI, A., BARDACH, A., GLUJOVSKY, D., ARUJ, P., MAZZONI, A., GIBBONS, L., ORTEGA-BARRÍA, E., AND COLINDRES, R. E. Burden of influenza in latin america and the caribbean: a systematic review and meta-analysis. *Influenza and other respiratory viruses 7*, 6 (2013), 1017–1032.

[24] SHAMAN, J., AND KARSPECK, A. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences 109*, 50 (2012), 20425–20430.

[25] SMOLINSKI, M. S., CRAWLEY, A. W., BALTRUSAITIS, K., CHUNARA, R., OLSEN, J. M., WÓJCIK, O., SANTILLANA, M., NGUYEN, A., AND BROWNSTEIN, J. S. Flu near you: crowdsourced symptom reporting spanning 2 influenza seasons. *American journal of public health 105*, 10 (2015), 2124–2130.

[26] TIBSHIRANI, R. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*, 3 (2011), 273–282.

[27] VIBOUD, C., BJØRNSTAD, O. N., SMITH, D. L., SIMONSEN, L., MILLER, M. A., AND GRENFELL, B. T. Synchrony, waves, and spatial hierarchies in the spread of influenza. *science 312*, 5772 (2006), 447–451.

[28] YANG, S., KOU, S. C., LU, F., BROWNSTEIN, J. S., BROOKE, N., AND SANTILLANA, M. Advances in using internet searches to track dengue. *PLoS computational biology 13*, 7 (2017), e1005607.

[29] YANG, S., SANTILLANA, M., BROWNSTEIN, J. S., GRAY, J., RICHARDSON, S., AND KOU, S. Using electronic health records and internet search information for accurate influenza forecasting. *BMC infectious diseases 17*, 1 (2017), 332.

[30] YANG, S., SANTILLANA, M., AND KOU, S. C. Accurate estimation of influenza epidemics using google search data via argo. *Proceedings of the National Academy of Sciences 112*, 47 (2015), 14473–14478.

[31] ZHANG, Q., PERRA, N., PERROTTA, D., TIZZONI, M., PAOLOTTI, D., AND VESPIGNANI, A. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of the 26th International Conference on World Wide Web* (2017), International World Wide Web Conferences Steering Committee, pp. 311–319.

Curriculum Vitae

(Only one page) Doctoral student was born in Monterrey, México, on September 11, 1991. He earned the Physics Engineering degree from the Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey Campus in December 2008. He was accepted in the graduate programs in Computer Science in August 2017.

This document was typed in using LaTeX 2_ε[1] by César Leonardo Clemente López.

[1]The template `MCCi-DCC-Thesis.cls` used to set up this document was prepared by the Research Group with Strategic Focus in Intelligent Systems of Tecnológico de Monterrey, Monterrey Campus.