Universidad Autónoma de San Luis Potosí
Facultad de Ingeniería
Centro de Investigación y Estudios de Posgrado

# Computational forecasting methodology for Acute respiratory infection using artificial neural networks and search terms

Para obtener el grado de:
Doctorado en Ciencias de la Computación

Presenta:
Daniel Alejandro Gónzalez Bandala

Asesor:
Dr. Juan Carlos Cuevas Tello
Co-Asesor:
Dr. Daniel E. Noyola Cherpitel

San Luis Potosí, S. L. P. Octubre de 2020

**Universidad Autónoma de San Luis Potosí**
**Facultad de Ingeniería**
**Centro de Investigación y Estudios de Posgrado**

# Método de predicción computacional para infecciones respiratorias agudas utilizando redes neuronales artificiales y términos de búsqueda

Para obtener el grado de:
Doctorado en Ciencias de la Computación

Presenta:
Daniel Alejandro Gónzalez Bandala

Asesor:
Dr. Juan Carlos Cuevas Tello
Co-Asesor:
Dr. Daniel E. Noyola Cherpitel

San Luis Potosí, S. L. P. Octubre de 2020

I would like to dedicate this thesis to my beloved wife Elisa, my beautiful daughter Rebeca, both of my loving parents and my amazing God.

But what saith it? “The Word is nigh thee, even in thy mouth and in thy heart,” that is, the word of faith which we preach: that if thou shalt confess with thy mouth the Lord Jesus, and shalt believe in thine heart that God hath raised Him from the dead, thou shalt be saved. For with the heart man believeth unto righteousness, and with the mouth confession is made unto salvation. For the Scripture saith, “Whosoever believeth in Him shall not be ashamed.” For there is no difference between the Jew and the Greek, for the same Lord over all is rich unto all who call upon Him. For “whosoever shall call upon the name of the Lord shall be saved.” (Romans 10:8–13)

# Abstract

The study of infectious disease behavior has been a scientific concern for many years as early identification of outbreaks provides great advantages including timely implementation of public health measures to limit the spread of an epidemic. This thesis proposes a methodology that merges the predictions of (i) a computational model with machine learning, (ii) a projection model, and (iii) a proposed smoothed endemic channel calculation. The predictions are made on weekly Acute Respiratory Infections (ARI) data obtained from epidemiological reports in Mexico, along with the usage of key terms in the Google search engine. The results obtained with this methodology were compared with state-of-the-art techniques resulting in reduced Root Mean Squared Percentage Error (RMPSE) and Maximum Absolute Percentage Error (MAPE) metrics, achieving a MAPE of 21.7%. This approach uses readily available data, such as Internet search terms and routine disease surveillance data (ARI data) and could be extended to detect and raise alerts on possible outbreaks on ARI as well as for other seasonal infectious diseases on several regions.

# Resumen

El estudio del comportamiento de enfermedades infecciosas ha sido, por muchos años, de interés científico debido a grandes ventajas provistas por una identificación temprana de brotes, como puede ser la implementación de medidas de salud pública a tiempo para limitar la expansión de una pandemia. Esta tesis propone una metodología que integra predicciones de (i) un modelo computacional basado en aprendizaje automatizado, (ii) un modelo de proyección y (iii) la propuesta del cálculo de canales endémicos suavizados. Las predicciones se hacen con datos semanales de Infecciones Respiratorias Agudas (IRA) obtenidos de reportes epidemiológicos en México, además datos sobre el uso de palabras clave en el motor de búsqueda de Google. Los resultados obtenidos son comparados con técnicas del estado del arte obteniendo en valores reducidos en métricas: la Raíz de Error Cuadrático Medio Porcentual (RMSPE, por sus siglas en ingles) y el Error Absoluto Porcentual Máximo (MAPE, por sus siglas en ingles). Específicamente, obteniendo un MAPE de 21.7%. Esta metodología utiliza datos disponibles al día, como lo son los términos de búsquedas por Internet y datos rutinarios de vigilancia (IRA) y puede ser extendida para detectar y enviar alertas en presencia de posibles brotes de IRA, así como otras enfermedades infecciosas estacionales en distintas regiones.

# Contents

# Índice general

# Introduction

Acute Respiratory Infections (ARI) are one of the main causes of morbidity and mortality in the world, particularly in children under 5 years old and adults over 65 years old. It has been estimated that 156 million acute lower respiratory infections[1] occur worldwide every year and almost 2.4 million deaths are estimated to have occurred associated with these infections in 2016 [104, 80, 9, 62, 34]. Therefore, the development of an effective monitoring and response system for infectious diseases is still a challenge [83, 114].

The most frequent pathogens that cause ARI are Respiratory Syncytial Virus (RSV), human metapneumovirus, rhinovirus/enterovirus, influenza viruses, parainfluenza 1-4, adenovirus, coronavirus, *Streptococcus pneumoniae*, and *Mycoplasma pneumoniae* [68, 103]. ARI show a seasonal pattern where RSV and influenza viruses are the major pathogens that contribute to this behavior. Changes in circulating viral strains of these viruses may result in winter ARI epidemics. In addition, introduction of novel influenza strains or other viruses into the human population can lead to the emergence of pandemics.

Despite the health and economic burden of RSV, there is currently no vaccine or effective antiviral treatment against this virus. In contrast, there are several antivirals and vaccines available for influenza. While mortality associated with influenza has been reduced since the introduction of influenza vaccine, this virus remains an important cause of ARI [93].

Health surveillance around the world has become a subject of primary concern, due to the continuous emergence of infectious disease outbreaks, including those associated with ARI such as pandemic influenza, Severe Acute Respiratory Syndrome (SARS) and, more recently, SARS-CoV-2 [118]. A detailed understanding of transmissible disease dynamics and the accurate forecast of disease outbreaks translate in reductions in the loss of human lives and money savings by avoiding desperate measures during health contingencies [98, 107, 2].

Development of an effective monitoring and response system for infectious diseases is a challenge, because it requires multiple components, periodical updating, and availability

[1] Such as pneumonia, influenza, acute bronchitis or any infection affecting lungs or airways.

to clinicians. As a result, different approaches have been developed to address this problem [83, 114].

Several systems which collect epidemiological information from informal sources like non-government reports, news reports and field agents have been proposed as potential Early Warnings Systems (EWS). ProMED [20], GOARN [56], GPHIN [61], Argus [63], BioCaster [24], EpiSPIDER [43, 50], PREDICT [60] are some examples. ProMED proved the usefulness of EWS with the early report of the SARS epidemic in 2003 in mainland China. In February 10, 2003, a ProMED report became the earliest public alert of a disease which would later be known as SARS and which would ultimately affect in excess of 8,000 individuals worldwide and kill more than 900 [112].

The quick growth in the Internet coverage during the last decade has brought the availability of an open and direct communication channel for epidemiological surveillance. The wide availability of Internet access by the general public has led to an increase in the use of online sites to obtain information concerning diseases as well as seeking medical advice. Patients may use search engines to look for their symptoms or as a self-diagnosis tool, while physicians may use web searches for available Internet resources. The use of search engines by the general public and physicians creates trends of terms, which could match the temporal occurrence of diseases and allow for potential detection of outbreaks at early stages, before traditional surveillance methods identify them [113, 55]. In the past years, the use of Internet search engines and social media platforms for surveillance and forecast of diseases has been widely studied [31, 36, 89, 96, 111, 100, 85, 116, 101, 109, 110, 94, 19, 74, 110, 31, 29, 85, 100]. Infectious diseases such as dengue [7, 37, 55, 110, 110], Influenza-Like-Illnesses (ILI) [74, 17, 113, 36, 48, 29, 86, 85, 57, 100, 109, 116, 117, 40, 111, 84, 109, 41], tuberculosis [73, 51], typhoid [115], diarrhea[69], varicella [69, 96], Lyme disease [89], dermatologic diseases[101], Ebola[107, 30], gonorrhea[47] and Zika [94] are among those that have been studied more frequently with this approach due to their seasonal and epidemic behaviors.

The most common method used by health authorities to detect outbreaks are the Endemic Channels (EC). Specifically, endemic channels represent the amount of cases within an expected normal range, and anything above this moving threshold could point to a developing outbreak [6]. The EC are calculated using the historic behavior from 5 or more years of an infectious disease (normally avoiding years affected by an outbreak), it is a simple and fast way to generate somewhat reliable thresholds allowing to detect a presence/absence of an outbreak. When using endemic channels there is a risk of not detecting new behaviors in the data, and limitations associated with abnormally high historic means and the variation in the seasonal timing that often lead to inaccurate detections[11].

The related work to this research includes Santillana et al. [85] who, in 2015, used six different data sources to predict 2013 and 2014 epidemics; the predictions are made by three different Machine Learning algorithms to perform multivariate regression, including stacked linear regression, Support Vector Machines (SVM) and AdaBoost with decision tree regression, and predict one, two and three weeks in the future. This research only reports different data sources and different regression methods, and the performance is evaluated with Correlation, RMSE, MAPE and Hit Rate [85]. Additionally, Volkova et al. [100], in 2017, employed only two data sources from 2011 to 2014. They propose a Long Short-Term Memory (LSTM), which is a Recurrent Neural Network. They trained the LSTM models on two seasons (2012–2013) and tested on the 2014 season. They also employed the same metrics used by Santillana et al. [85], except the hit rate. Their models are capable of predicting weekly ILI dynamics and forecasting up to several weeks in advance [100].

In 2017, Xu et al. [109] proposed four independent models on ILI data from Hong Kong and integrate their results. They use three datasets, mentioning that measuring the quality of the data falls outside the scope of the research. Even though this is a very interesting proposal, the data and metrics used do not make it possible to compare the proposed methodology with this research.

Therefore, previous research only report what data has been used and the best regression models for a specific period of time. Few has been said about data retrieval, data preprocessing and feature extraction (called data acquisition process). They also do not use endemic channel information.

The main contribution of this research is a methodology composed of the data acquisition and the computational model. Other contributions of this research include: i) A method to automatically select the search terms associated to the data source available; ii) A smoothed endemic channel calculation; iii) A predictive calculation made by merging the forecasting of an artificial neural network, the projection of a sum of sines model and the proposed smoothed endemic channels.

Overall, this research presents a methodology capable of making accurate predictions of ARI activity with data obtained from epidemiological reports, along with search terms usage derived from the Google search engine. The combined use of epidemiological, Machine Learning, and forecasting techniques allowed us to develop a computational model that is capable of accurately predicting ARI trends. Adaptations of this model might prove useful for timely detection of outbreaks at early stages and before they become a major health burden. This research is part of a bigger project called Mexican Infectious Disease Analysis and Surveillance mapping application (MIDASmap; http://midasmap.uaslp.mx/) which is under development and will be available online.

# Research question

How accurately can a computational model predict acute respiratory infection (ARI) outbreaks by using search terms along with Machine Learning algorithms compared with state-of-the-art outbreak forecasting models?

# Objectives

## Main objective

To develop a computational model to forecast ARI cases through Machine Learning algorithms along with statistical techniques and Internet search term trends.

## Specific objectives

- To gather and test a set of most used state-of-the-art, specific and general purpose, forecasting and statistical techniques.

- To define, monitor and analyze most correlated search trend topics with historical infectious disease reported cases.

- To propose a computational model based on Machine Learning algorithms for forecasting.

- To measure the significance of results and compare the model with current methods.

# Research methodology

An iterative methodology is proposed in order to achieve these objectives and ensure the completion of this research (see Figure 1). The initial idea of a forecasting model that uses data from Internet search engines was first focused on zoonosis, but after the first iterations of the methodology, and as the state of the art was reviewed, the definition started focusing only on ARI due to several constrains in the availability of the data; then, as the iterations continued, different ARI datasets were tested until the one defined in Chapter 2 was found to be adequate for this research. After that, iterations continued, mostly to test each stage and module of the proposed methodology, until a final structure was found and tested to establish it as the solution.

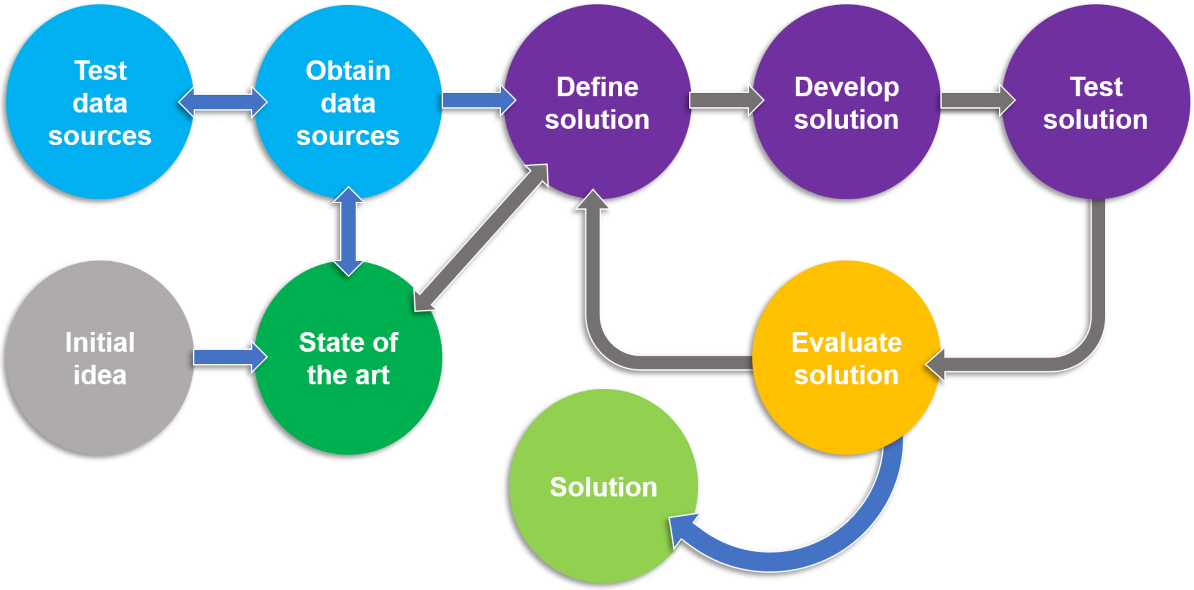


Figure 1 Research methodology used in this project.

This document presents the process of following the methodology described above, describing the findings, the difficulties found and the proposals made during the completion of this research.

## Thesis outline

The first chapter focuses in the state of the art in Epidemiological surveillance systems and the forecasting methods used for Acute respiratory infections, the second chapter presents the two main datasets used and their analysis to continue with the chapter 3 where the architecture of the proposed methodology is introduced, describing all the involved methods and calculations made to fulfill the specified objectives. The fourth chapter shows the results obtained and a comparison with the most similar investigations found, finally the fifth chapter contains the conclusions for this research and the planned future work. It is worth noting the inclusion of two Appendices, Appendix A focused in concepts and topics related to forecasting infectious diseases, and Appendix B showing all considered methods and techniques that, for some reason, were discarded.

# Chapter 1

# Epidemiological Surveillance and Acute Respiratory Infections Forecasting

This chapter focuses on two state-of-the-art branches: epidemiological surveillance and acute respiratory infections forecasting. The first explores successful epidemiological information systems, and the latter analyzes forecasting systems used on ARI, along with their approaches and data sources. The greatest difference with the Epidemiological surveillance systems and the Acute respiratory infections forecasting is that the former works as an automated collector of, mostly online, data, and with this information reports and alerts are created to provide an overview of what is happening in specific regions, some of them are capable of processing worldwide information and not only focused on ARI, but in a whole set of diseases. the latter, on the other hand, are researches focused on enhancing the ARI forecasts by exploring different approaches, many of them with no immediate intention to become an automated system, but to add knowledge to the solution of the ARI forecasting problem.

## 1.1 Epidemiological surveillance

The gathering, study and analysis of patterns, causes, and effects of health and disease conditions in populations, is known as Epidemiology, and it shapes the decision making and evidence-based practice of health authorities. Epidemiology is the cornerstone of Health Surveillance Systems (HSS) and outbreak forecasting systems, as it is needed to gather, analyze and process historical data in order to predict outbreaks with the best accuracy possible. Table 1.1 shows a quick overview of certain characteristics of the most relevant Early Warnings Systems (EWS) that have been developed to provide novel approaches to traditional HSS; a deeper description is given below, with some other related works that may not be as significant, but are still similar to what this thesis tries to achieve.

Table 1.1 Comparative table of EWS proposed for health surveillance

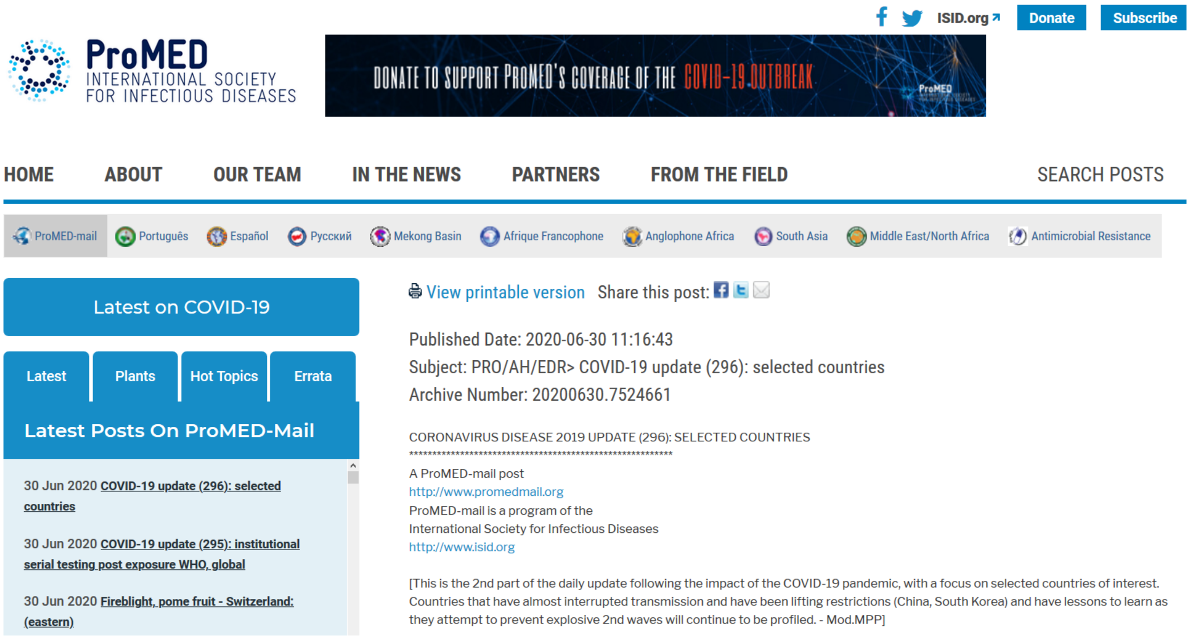
| | | | | **Sources** | | | | |
|---|---|---|---|---|---|---|---|---|
| **Name** | **Year** | **Public** | **Origin** | **RSS** | **Web Pages** | **Official Reports** | **Informal Reports** | **Written Media** |
| ProMED | 1993 | Y | US | N | Y | Y | Y | Y |
| GPHIN | 1994 | N | CAN | Y | Y | Y | N | N |
| EIN | 1997 | Y | US | Y | Y | Y | Y | N |
| GOARN | 2000 | Y | WHO | N | N | Y | N | Y |
| BioCaster | 2006 | Y | US | Y | Y | N | N | N |
| EpiSpider | 2006 | Y | US | Y | Y | Y | N | N |
| HealthMap | 2006 | Y | US | Y | Y | Y | Y | Y |
| PREDICT | 2009 | Y | US | N | N | Y | N | Y |

### 1.1.1 Program for Monitoring Emerging Diseases (ProMED)

Created in 1993 by an international group of 35 senior scientist and health experts, the goal of ProMED is to provide an effective global system of infectious disease surveillance that could give early warning about outbreaks of new diseases, as well as known ones [20](Figure 1.1 shows the ProMED website screenshot). Until 1996, its most notorious characteristic was its email network with about 4,500 subscribers from over 100 countries.

Today, ProMED is an Internet-based reporting system focused on rapid diffusion of information on outbreaks of infectious diseases and acute exposure to toxic substances that affect human health, including those in animals and plants grown for human food or animals feed, providing reliable and constantly updated information about the threats to human, animal, plant health and food health worldwide. Sources of information include reports in the media, official reports, online summaries, local observers, among others, with no political restrictions. A group of experienced experts in the field of human, plant and animal diseases monitors, reviews and examines reports before they are sent to the network. Reports are distributed by e-mail to subscribers and published directly on the ProMED website. Currently, ProMED has more than

70,000 subscribers in at least 185 countries. ProMED has opened way to many other similar health surveillance systems, and even provides data for some of them.



30 Jun 2020 COVID-19 update (296): selected countries

30 Jun 2020 COVID-19 update (295): institutional serial testing post exposure WHO, global

30 Jun 2020 Fireblight, pome fruit - Switzerland: (eastern)

Published Date: 2020-06-30 11:16:43
Subject: PRO/AH/EDR> COVID-19 update (296): selected countries
Archive Number: 20200630.7524661

CORONAVIRUS DISEASE 2019 UPDATE (296): SELECTED COUNTRIES
********************************************************

A ProMED-mail post
http://www.promedmail.org
ProMED-mail is a program of the
International Society for Infectious Diseases
http://www.isid.org

[This is the 2nd part of the daily update following the impact of the COVID-19 pandemic, with a focus on selected countries of interest. Countries that have almost interrupted transmission and have been lifting restrictions (China, South Korea) and have lessons to learn as they attempt to prevent explosive 2nd waves will continue to be profiled. - Mod.MPP]

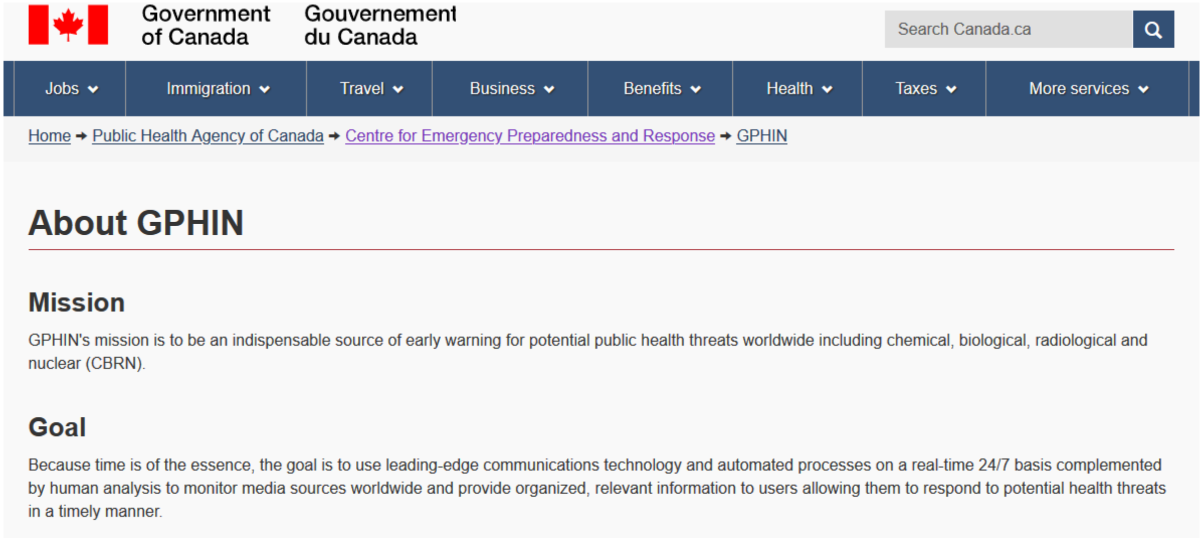
Figure 1.1 ProMED Home page, June 2020 (https://ProMEDmail.org/) [20]

### 1.1.2 The Global Public Health Intelligence Network (GPHIN)

GPHIN began in Canada in 1994 as an Internet system for outbreak alerts with restricted access, which refers to information regarding public health alerts that might be of potential international importance. Instead of relying on the contribution of subscribers, GPHIN collects information during disease outbreaks and other public health surveillance resources throughout the world. The two main sources for its news information are Factiva and Al Bawaba (in Arab language). GPHIN subscriptions are restricted to organizations with health mandates. It has a restricted website[1] and alerts through e-mail (Figure 1.2). The GPHIN system keeps track of six key areas: epidemics, biological agents, chemistry, the environment, radioactivity and natural disasters.

GPHIN has shown its significance by making reports that have mobilized other networks within GOARN in many occasions [61]. It scans items published in the official languages of the World Health Organization (WHO) and uses automated translation software to translate non-English articles to English, and English articles into Russian, French, Spanish, Arabic,

[1] https://gphin.canada.ca/cepr/aboutgphin-rmispenbref.jsp



About GPHIN

Mission

GPHIN's mission is to be an indispensable source of early warning for potential public health threats worldwide including chemical, biological, radiological and nuclear (CBRN).

Goal

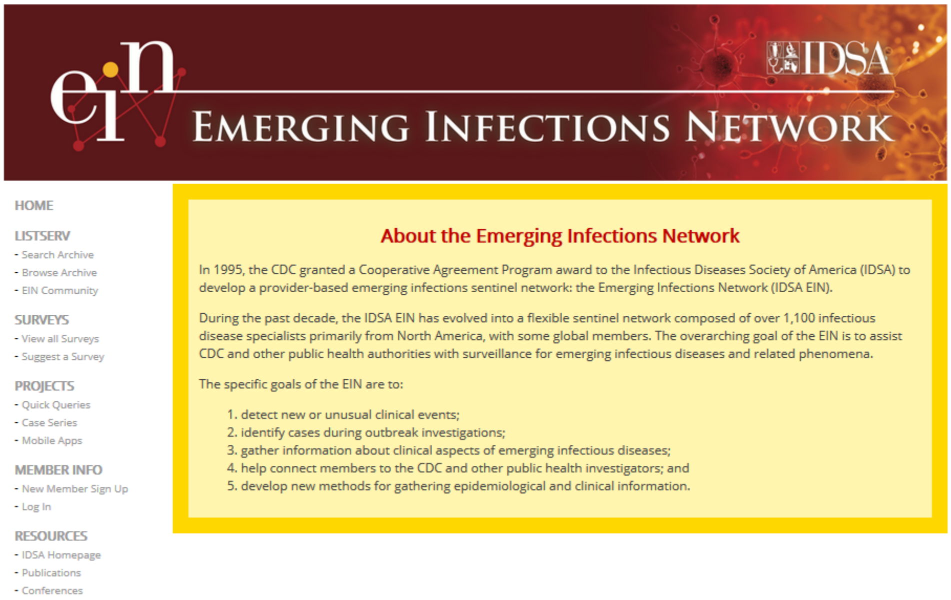
Because time is of the essence, the goal is to use leading-edge communications technology and automated processes on a real-time 24/7 basis complemented by human analysis to monitor media sources worldwide and provide organized, relevant information to users allowing them to respond to potential health threats in a timely manner.

Figure 1.2 GPHIN Homepage, as it has restricted access it is not possible to go any further without credentials, June 2020 (https://gphin.canada.ca/cepr/aboutgphin-rmispenbref.jsp).

and Chinese. Informal assessments suggest that, on average, GPHIN processes anywhere from 2,000 to 3,000 updates per day, of which about one-quarter to one-third are discarded as irrelevant or duplicate. The early warning alerts given by GPHIN help with the challenge of effectiveness and credibility of global health media on infectious disease epidemics.

### 1.1.3 Emerging Infections Network (EIN)

Launched by the Infectious Diseases Society of America (IDSA), in cooperation with the Centers of Disease Control and Prevention. Conceived as a sentinel system to monitor infectious diseases and complement other public health surveillance efforts [95]. The following five objectives were defined: (1) identification of new or unusual clinical events, (2) identification of cases during outbreak investigations; (3) characterization of emerging infections (e.g., diagnostic methods, treatment methods, and preliminary assessment of morbidity and mortality); (4) research collaboration; and (5) communication and education. EIN relies on its members to actively identify new cases during outbreak investigations, and provides a platform for the exchange of information on emerging infectious diseases among its members. The success of EIN depends directly on the number of volunteer hospitals and the ongoing commitment of



About the Emerging Infections Network

In 1995, the CDC granted a Cooperative Agreement Program award to the Infectious Diseases Society of America (IDSA) to develop a provider-based emerging infections sentinel network: the Emerging Infections Network (IDSA EIN).

During the past decade, the IDSA EIN has evolved into a flexible sentinel network composed of over 1,100 infectious disease specialists primarily from North America, with some global members. The overarching goal of the EIN is to assist CDC and other public health authorities with surveillance for emerging infectious diseases and related phenomena.

The specific goals of the EIN are to:

1. detect new or unusual clinical events;
2. identify cases during outbreak investigations;
3. gather information about clinical aspects of emerging infectious diseases;
4. help connect members to the CDC and other public health investigators; and
5. develop new methods for gathering epidemiological and clinical information.

Figure 1.3 Emerging Infections Network (EIN) Homepage, June 2020 (http://ein.idsociety.org/).

infectious diseases physicians to identify and report potential emerging infectious diseases. EIN has an informational web page; to access the information it is required to create a free account[2].

### 1.1.4 Global Outbreak Alert & Response Network (GOARN)

The decision to establish GOARN was agreed in a WHO meeting in Geneva in April 2000, as a need to achieve coordinated approaches to address the challenges of disease outbreaks in the twenty-first century. Defined as a technology network that can contribute to the international response to infectious diseases outbreak, thus providing technical and multidisciplinary skills for major projects responsiveness to epidemics. GOARN has grown in technological partners form multiple countries and regions: Regional Office of the Americas, African Regional Office, European Regional Office, Eastern Mediterranean Regional Office, South-East Asian Regional Office, Western Pacific Regional Office. By 2012, GOARN partners had participated in 137 missions, including 1,471 deployments which totalled 31,629 person days. These missions were carried out in 79 different countries, regions or localities. Rapid support to contain the spread of viruses is the goal. GOARN has a SharePoint Website[3] that displays a warning message,

[2] http://ein.idsociety.org/

[3] https://extranet.who.int/goarn/

support request, operational updates, technical support, tips, and deployment details (Figure 1.4 shows GOARN website).

GLOBAL OUTBREAK ALERT AND RESPONSE NETWORK

Connecting health resources for better health outcomes

GOARN turns 20!

Cooperation and mutual support are key components of a successful...

Share your GOARN Story

Figure 1.4 GOARN Homepage, September 2020 [56] (https://extranet.who.int/goarn/).

### 1.1.5 BioCaster

BioCaster [25, 24, 49] is a text-based ontological analysis system to identify and monitor the spread of infectious diseases through language signals in the Web. In operation since 2006, it ceased operations on 2012, BioCaster was developed to achieve an early detection of infectious disease outbreaks (Figure 1.5). It is composed by two main components: a web/database server and a backend computer cluster equipped with text mining technology that continuously scanned hundreds of RSS newsfeeds from local and national news providers, it specialized in providing advanced research and analysis, web analytics and scientific literature for employees of public health agencies and infectious disease investigators. The system captured over 1700 RSS feeds, classified by their relevance, and placed the information on a Google map using geocoded information. Some of the sources are EurekAlert!, European Media Monitor Alerts (EMMA), Google, the Morbidity and Mortality Weekly Report (MMWR) from the Centers for Disease Control and Prevention (CDC), MeltWater, OIE, ProMED, Reuters, WHO (World Health Organization) and Vetsweb. The BioCaster ontology initially included information in eight languages focused on the epidemiological role of pathogens and their geographical locations. This platform was later upgraded to identify information contained in articles written in Arabic, Chinese, English, French, Japanese, Korean, Portuguese, Russian, Spanish, Thai and Vietnamese.

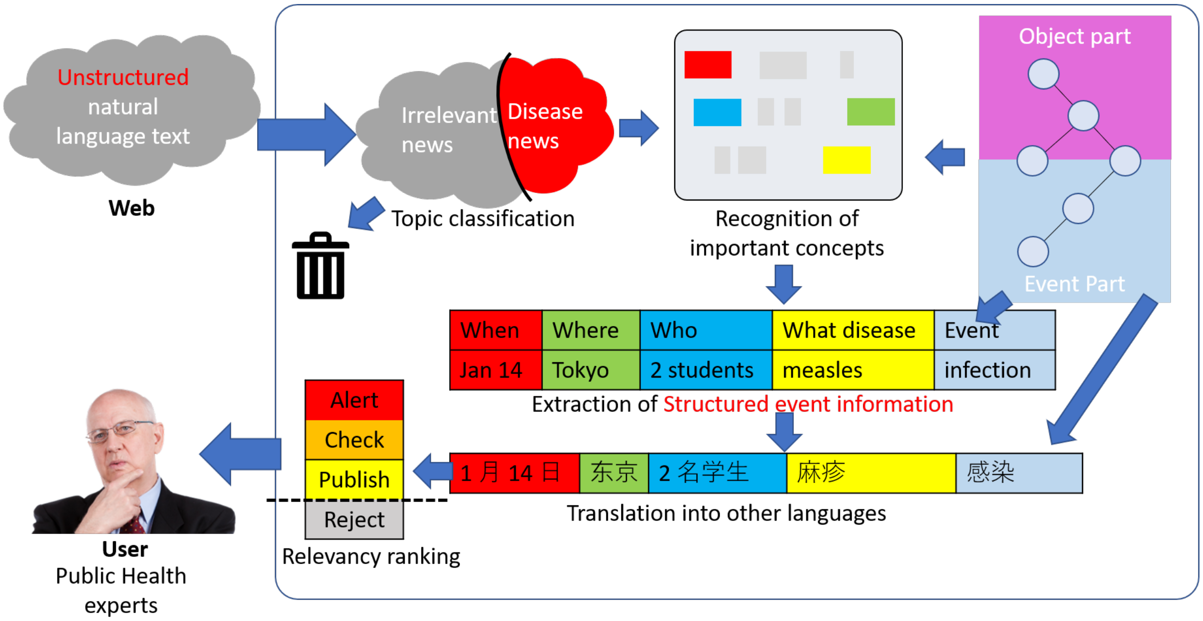


Figure 1.5 BioCaster architecture. Adapted from [49].

### 1.1.6 EpiSPIDER

EpiSPIDER [43, 50], created on 2006, first conceived as a ProMED-mail visual supplement report. It now collects information from many sources, such as Daylife, on Google, humanitarian news, and ProMED, Twitter and WHO. It uses natural language processing on structured information and stores it in a relational database, generating RSS feeds and data views in an interactive map (Figure 1.6). EpiSPIDER gets location names from reports and uses georeferencing services to plot data at a country level in the United States and Canada, and in some other selected countries. EpiSPIDER shows how the data is distributed, based on the integration of unstructured sources of media events to complement the awareness of the disease surveillance situation.

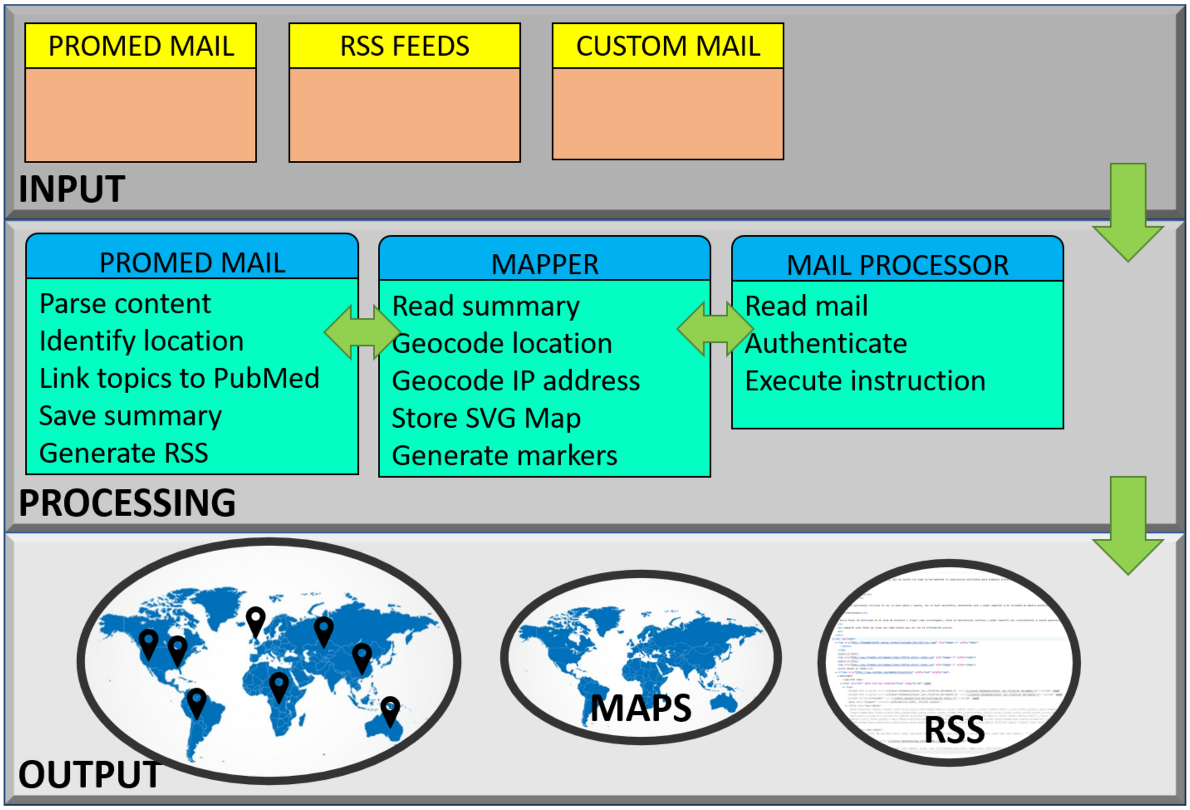


Figure 1.6 EpiSPIDER architecture. Adapted from [43].

### 1.1.7 HealthMAP

HealthMap [13, 12] is an automated system for requesting, filtering, integrating and displaying unstructured reports of illness outbreaks. Designed to collect and display new outbreaks with geographic location, time, and infectious agent. The system is a starting point for real-time information service on a wide range of emerging infectious diseases and is designed for public health officials and international travelers. The system integrates outbreak data from a variety of electronic sources: Baidu, Eurosurveillance, Google, HealthMap Community News reports, EIA (Environmental Impact Assessment), ProMED, Soso, Ocular Eye Users account, WDIN and WHO, and scans for articles in Arabic, Chinese, English, French, Portuguese, Russian and Spanish (HealthMap architecture is shown in Figure 1.7).

HealthMap uses geo-referral data to show reports on an interactive map[4] available for free since 2006. As part of the system evaluation, the system processed 778 reports during one month,

[4] http://www.healthmap.org/es/

covering 87 categories of diseases and 89 countries. HealthMAP has a significant utility in managing large amounts of information, the automated file classifier had an 84% accuracy (Correctly classifying 655 out of 778 reports) compared with ProMED alerts at 91% (Correctly classifying 188 out of 207) and Google News reports for which this was 81% (Correctly classifying 443 out of 547), as ProMED messages follow a more regular structure [33].

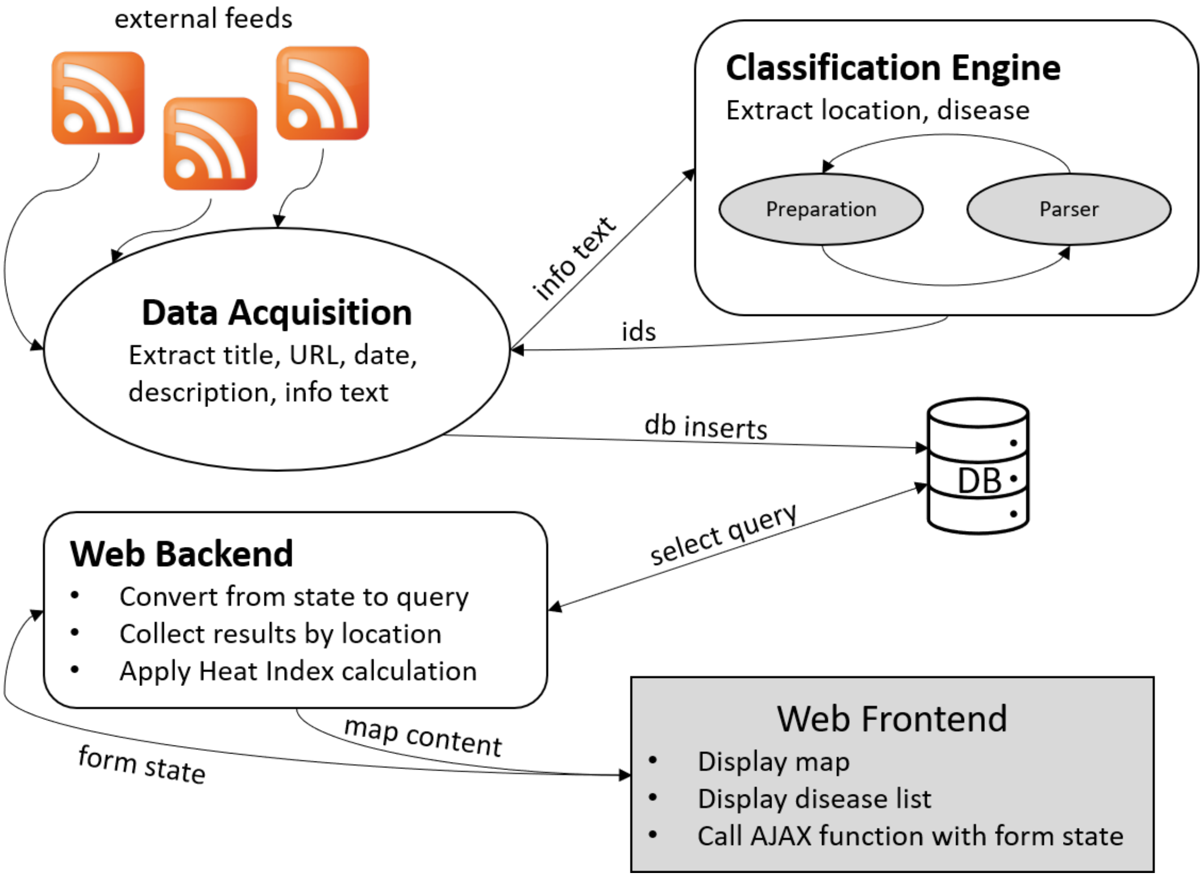
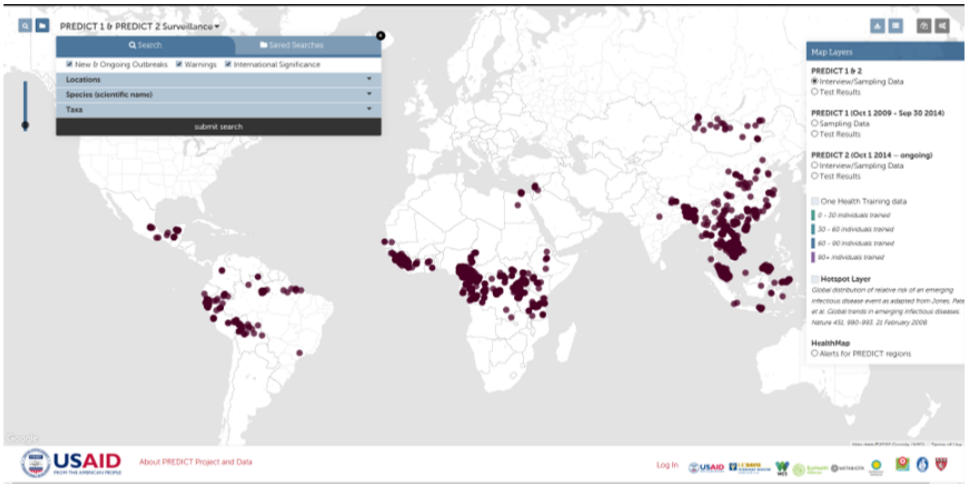

external feeds
Classification Engine
Extract location, disease
Preparation
Parser
info text
ids
Data Acquisition
Extract title, URL, date, description, info text
db inserts
DB
select query
Web Backend
Convert from state to query
Collect results by location
Apply Heat Index calculation
map content
form state
Web Frontend
Display map
Display disease list
Call AJAX function with form state

Figure 1.7 HealthMap architecture. Adapted from [33].

### 1.1.8 PREDICT

This system started in 2009 as a project of the Emerging Pandemic Threats program (EPT) of the U.S. Agency for International Development (USAID) [60]. PREDICT was meant to create the global capacity for surveillance and risk assessment for emergent and zoonotic infections with pandemical potential. It uses the One Health approach to encompass and integrate health surveillance among species, with collaborations with governments and agencies such as CDC, FAO, WHO and OIE. Along with other complementary projects: RESPOND, IDENTIFY and PREVENT, PREDICT focuses on the human-animal interface to understand the pathogen background agent in other species that come into contact with people and risk factors for the emergence of new improved zoonoses. Currently, activities are carried out in at least 20 developed countries, work includes field search and collecting samples, as well as searching for viruses, other pathogens and microbiological test results when available. By the end of 2011, this project had found almost 100 viruses (in plants, rats and non-humans primates) showing a span of viral families. A data system is used to keep and correlate data coming from a variety of sources. Its fieldwork ended in 2019 due to lack of funding, and the program was closed in March 2020 by the Trump administration.



☑ New & Ongoing Outbreaks ☑ Warnings ☑ International Significance

PREDICT 1 & 2
● Interview/Sampling Data
○ Test Results

PREDICT 1 (Oct 1 2009 - Sep 30 2014)
○ Sampling Data
○ Test Results

PREDICT 2 (Oct 1 2014 – ongoing)
○ Interview/Sampling Data
○ Test Results

□ One Health Training data
🟩 *0 - 30 individuals trained*
🟦 *30 - 60 individuals trained*
🟦 *60 - 90 individuals trained*
🟪 *90+ individuals trained*

□ **Hotspot Layer**
*Global distribution of relative risk of an emerging infectious disease event as adapted from Jones, Pate et al. Global trends in emerging infectious diseases. Nature 451, 990-993. 21 February 2008*

**HealthMap**
○ Alerts for PREDICT regions

Figure 1.8 PREDICT site information, September 2020 [60] (http://data.predict.global/).

All of these Health surveillance systems have different approaches and methodologies to try to keep track of infectious diseases and some of them even to predict outbreaks as early as possible. The most common source for information they use is ProMED, along with RSS feeds and medical reports; in no case they use Internet Search terms as their main source of information. About the most recent event, the COVID-19 outbreak there is no official report of any of these EWS being the first to detect the outbreak[108, 32, 54, 4], the most plausible reason could be that the first country affected was China (country known for its isolationism), according to what is known, the first report was an epidemiologic alert by the local health authority on December 31, 2019 [108].

## 1.2 Disease dynamics forecasting

This section presents the review of the disease dynamics research focusing on approaches and datasets used. Figure 1.9 shows a visual representation of the state-of-the-art analysis grouped in four main subsets: forecasting, Internet search engines, Machine Learning and acute respiratory infections.

### 1.2.1 Forecasting techniques

There is a broad set of techniques used in the literature for predicting outbreaks. Because not all diseases behave similarly, a single forecasting approach might not be applicable in all circumstances; nevertheless, knowledge of the most common and successful forecasting techniques is required to propose a new model. An initial subset of relevant studies are selected exclusively because of their use of forecasting techniques to predict some disease; their analysis is contained in this subsection, and represented in Figure 1.9 as the grey area of forecasting and its intersections with the other topics.

#### ARIMA models

Some of the most common techniques which have been used for outbreak prediction are the Auto Regressive Integrated Moving Average (ARIMA) [72, 73, 109, 58] and variations: ARIMA additive and ARIMA multiplicative [3], seasonal ARIMA (SARIMA) [71, 109] and ARIMA with intervention [3]. From these variations, SARIMA has consistently shown better results and is one of the most used when the only input available is a single time series.

#### Regression models

Regression models are a frequent approach in biological and statistical data analysis [7, 30, 109, 14, 107, 51], where the authors focus mainly in calculating and describing the transmission rates, correlations between relevant variables, statistical properties, and projections of the studied diseases. A wide variety of models, including linear regression [76, 71, 74], adaptative regression [14], multivariate regression [110, 113, 107, 86], multiplicative and additive regression [3], Classification And Regression Trees (CART) [55], generalized linear [109] and linear mixed model [58], Bayesian space-time regression [58], and auto logistic regression model [52] have been used.

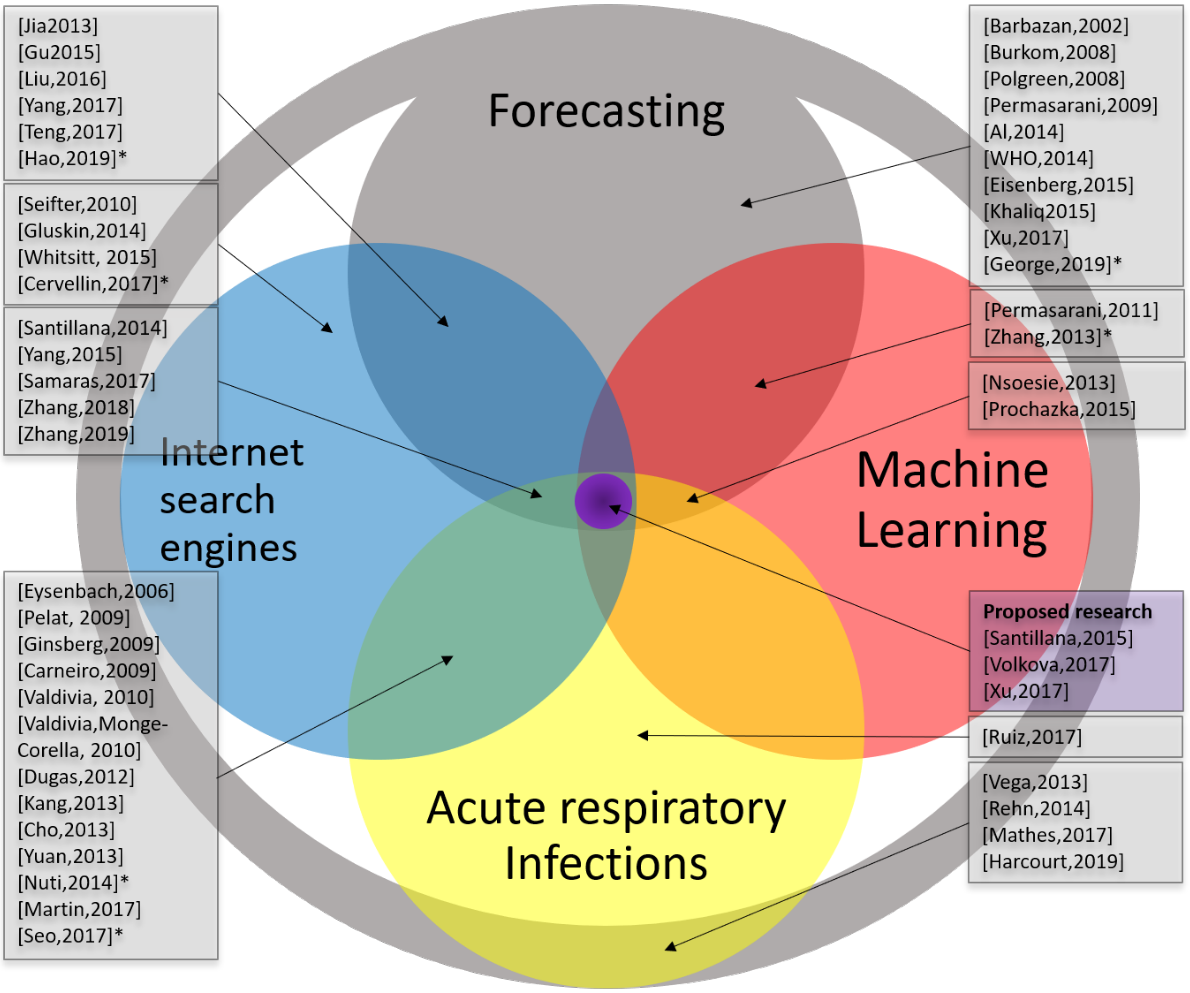


Figure 1.9 State-of-the-art summary, grouped in four main topics: Forecasting, Internet search engines, Machine Learning and Acute respiratory infections. As it can be seen, the intersection of these subsets is the topic of this thesis, where only few papers are within this category. (*) Denote surveys.

### Holt-Winters Smoothing (HWS)

HWS is another commonly used technique for outbreak prediction [58] and time series smoothing[5], it also has several variations: Holt-Winters Additive [51] and Multiplicative methods [3], Normalized Holt-Winters [14], Holt-Winters Damped [58]. In most cases where SARIMA and HWS are used, the SARIMA model outperforms HWS; however, in some instances HWS does a very good job at adjusting to the time series and forecasting. In the present research HWS was tested with historical ARI data and was initially used as a projection model, but was discarded because other methods showed better results (see Appendix B for more information about this method).

### Statistical model: Directional Movement Indicator and Average Directional Index (DMI-ADX)

Proposed by Wilder [102], the Directional Movement is a system that identifies if the market is trending before providing signals for trading the trend. It is composed of three variables: the Average Directional Index (ADX), Plus Directional Indicator (+DI) and Minus Directional Indicator (-DI); these values are calculated from the time series to be analyzed. The +DI and -DI are known as the Direct Movement Indicator (DMI). This method was designed originally for commodities and daily prices; it can also be applied to stocks and to identify trends in the financial market. It helps in the decision making for financial investments minimizing risks by identifying trends.

A DMI crossover generates the bullish (tendency to increase) and bearish (tendency to decrease) signals. When the DMI+ crosses above the DMI-, a bullish signal is identified. When the DMI- crosses below the DMI+ a bearish signal is identified. The ADX does not identify the direction of a trend by itself; it only identifies the degree of strength within a trending market. The ADX measures the strength of the trend over time, regardless of the direction of the trend. It provides an additional prediction about the intensity and ranges from 0 to 100 where 0 would mean there is no trend in data, and 100 would be an extremely strong trend.

The ADX, +DI and -DI calculations involve a lot of smoothing, and the calculation starts with the sum of the first $m$ periods. Together the DMI and ADX can provide the direction an strength of a trend. Some issues when using this technique are that it is based on Moving Averages (MA), this means that it reacts slowly to low moving series (it is known as a lagging indicator), and in some instances crossover between DMI indicators can happen too often,

[5] Smoothing data removes random variation and shows trends and cyclic components, see Appendix A.

giving false alerts. It is strongly recommended to use ADX in conjunction with other reliable indicators to support the signals derived from this model (see Appendix B for more information about this method).

#### Other models

There are other models where a threshold is defined by the biological analysis made of the outbreak [7], and then used as a signal of a possible outbreak when a threshold is crossed. The decomposition model obtains forecast results by preprocessing time series obtaining some features such as: trend, seasonality, cyclicity, and irregularity [73] (see Appendix A for time series concepts). Least Absolute Shrinkage and Selection Operator (LASSO) is a penalization method used to shrink the estimation of the regression coefficients to prevent overfitting due to either collinearity of the covariates or high-dimensionality [109, 86], Moving Epidemic Method (MEM) [97, 41]. The papers using Machine Learning algorithms for forecasting [115, 74, 64, 75, 35, 85, 100, 109] are detailed in Subsection 1.2.3.

### 1.2.2 Internet search engines

Represented as a blue circle in Figure 1.9, this subset focuses on research made to analyze the Internet search engines as tools for health surveillance, real-time surveillance, and their use for forecasting focused on different diseases. The analyzed research projects focus in three search engines and one social network: Google [31, 36, 89, 96, 111, 100, 85, 116, 101, 109, 110, 94, 19, 40, 84, 18, 69, 29, 37, 53, 15, 57, 48, 21, 86, 65], Yahoo [74], Baidu [55, 31, 29, 113, 39, 47], and Twitter [85, 100]. The details of these articles is given in Subsections 1.2.1 and 1.2.3.

### 1.2.3 Machine Learning

Most of the research using Machine Learning algorithms focuses only in Artificial Neural Networks (ANN) and its variants. It is worth noting that there is an extremely wide variety of Machine Learning algorithms, yet the focus is in prediction and in ANN specifically, mostly because of the capabilities of ANN to adapt and adjust to unknown scenarios. Other Machine Learning techniques used include Kernel smoothing and Serfling's higher order models [75], a Simulation Optimization Model (SIMOP) [64] which uses the Nelder-Mead simplex method for optimization and an individual-based model for simulations. These models are not widely used or cited in the state of the art, thus they are mentioned here for informational purposes and context.

### Artificial Neural Networks (ANN)

In the articles found, the ANN are commonly used to forecast the disease cases time series. For example, in some reports Multilayer Perceptrons (MLP) are used because they are one of the most common ANN [73, 109]. A detailed description of ANN is given in Appendix A.

Feedforward Neural Network are used in the study by Xu et al. [109] to forecast influenza, with several hidden layers; they aim to make predictions for one and two weeks after the most recent data. The input is divided in three sections: meteorological variables, Google search data, and previously offline-observed ILI data. The output neurons are one-week-ahead and two-week-ahead predictions of influenza cases. A comparison of the FFNN, Radial Basis Function Neural Networks (RBFNN), and Elman Recurrent Neural Networks (ERNN) to forecast typhoid fever incidence in China [115] showed better results for the RBFNN, which can be defined as a ANN with Radial basis functions as activation functions in its nodes (neurons).

The Long Short-Term Memory Network (LSTM) is a Recurrent Neural Network (RNN) with built in memory cells to store information and exploit long range context, surrounded by gating units that can reset, read, and write information[44]. Known as a Deep Learning [6] neural network, it has internal feedback connections and is able to process sequences of data, such as time series, speech or video. It has been reportedly used for ILI forecasting [100], and is composed of a cell, an input gate, output gate and a forget gate. it was considered in the second stage of the proposed research.

Reservoir Computing is an extension of ANN in which the input signal is connected to a fixed and random dynamical system called reservoir, creating a higher dimension representation (embedding) which is then connected to a desired output via trainable units. The Echo State Network (ESN) is a RNN part of the so-called Reservoir Computing, it is composed by an input layer, a reservoir, an embedding of the input and a readout (output) layer [78]. The weights between the input and the reservoir, as well as the reservoir weights are not trainable. This model was also studied to be added to our proposed model. In both cases, the LSTM and the ESN, the conclusion was that a deeper study is needed to include them in the proposed model; several configurations were tested but none of them yielded better results than the presented for the FFNN in this thesis.

[6] Deep learning is a class of machine learning algorithms that uses multiple layers to progressively extract higher-level features from a raw input[27].

In all cases, the ANN tend to have good results and good fitting for data prediction, but the research that has been done so far does not allow to determine the best approach to use these techniques in the context of infectious disease surveillance, given that there are many variants of ANN. Nevertheless, in general, a simple ANN can be chosen, since this kind of ANN have been used and studied profoundly for a long time, and there are many enhancements made to ANN where their main differences are the modeling of the neurons and their architecture.

### 1.2.4 Acute Respiratory Infections

Denoted in Figure 1.9 as a yellow circle, papers including data from ARI are mostly on the side of analyzing whether or not Internet search engines are an accurate resource for ARI predictions, focusing on correlations and statistical analysis [31, 70, 36, 18, 96, 28, 48, 21, 113, 65, 57, 90, 19]; one paper focused on analyzing the historic behavior and trend of the ARI data [81], some other papers on predicting the ARI behavior by only using its time series [97, 76, 58, 41], and only three papers were found that intersect with the four subsets of concepts that we focus on, where they use Internet search engines and Machine Learning to forecast ARI data [85, 100, 109], which are described below.

### 1.2.5 Related work

As mentioned in the Introduction, and although all the reviewed papers provide very important information that helped shaping this thesis, there are three papers that are most related to this research, because of the approach, data and metrics used to measure their results.

The related work to this thesis includes Xu et al. [109](2017), who used four individual models: Generalized Linear Model (GLM), LASSO, ARIMA and a FFNN (which the paper claims is a deep learning algorithm, and that they are the first to use deep learning for surveillance and forecasting of infectious diseases) on ILI data from Hong Kong and used a Bayesian Model Averaging (BMA) to integrate the results. The data used is: Influenza data [7], Google search data and meteorological data (period 2011 to 2015), mentioning that measuring the quality of the data falls outside the scope of the research. The metrics used in this research are: RMSE, Correlation, Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). According to their analysis there is no clear cyclic behavior for influenza in Hong Kong (2011-2016), in years 2013 and 2014 there is no detectable influenza season and in 2012 and 2015 there are two influenza

[7] Influenza Like Illness (ILI) data (per 1000 consultations) in General Out-Patient Clinics (GOPC) in Hong Kong.

seasons. Also, they state that the Google search data is limited due to the relatively small population in Hong Kong (of 7 million). This is a very interesting proposal because they fuse the results of their methods with the BMA, but the lack of cyclicality in the ILI data, the Google search data limitations and the metrics used make it impossible to have a fair comparison of the proposed methodology with this thesis.

Another relevant paper was published in 2015 by Santillana et al. [85] who used six different data sources to predict 2013 and 2014 influenza epidemics: CDC-reported ILI[8], athenahealth[9], Google trends, influenza related Twitter microblogging, FluNearYou[10] and Google Flu Trends[11](period 2013 to 2015). The predictions are made by three different Machine Learning algorithms to perform multivariate regression, including stacked linear regression, Support Vector Machines (SVM) and AdaBoost with decision tree regression. They report the ability to predict one, two and three weeks in the future. This research only reports different data sources and different regression methods, and the performance is evaluated with Correlation, RMSE, RMSPE, Maximum Absolute Percentage Error (MAPE) and Hit Rate [85].

Additionally, Volkova et al. [100], in 2017, employed only two data sources: Defense Medical Information System (in USA) and Twitter (period 2011 to 2014). They use as baseline SVM and AdaBoost and they proposed a Long Short-Term Memory (LSTM), which is a Recurrent Neural Network. They trained the LSTM models on two seasons (2012–2013) and tested on the 2014 season. They also employed the same metrics used by Santillana et al. [85], except the hit rate. Their models are capable of predicting weekly ILI dynamics and forecasting up to several weeks in advance [100].

Their results using the metrics in common are shown in Figure 1.2, while the correlation could be used to compare different methods the information it provides is very limited, and comparisons using RMSE across different types of data would be invalid because the measure is dependent on the scale of the numbers used. Moreover, related research only report what data has been used and the best regression models for a specific period of time. Few has been said about data retrieval, data preprocessing and feature extraction (called data acquisition process). More detailed results and a comparison with the proposed model are shown in Chapter 4.

[8] Weekly number of people seeking medical attention with ILI symptoms in the United States.

[9] A medical practices management company providing nearly real-time hospital visit records.

[10] A participatory surveillance system to self-report ILI.

[11] Google Flu Trends is Unavailable from 2015.

Table 1.2 Results reported in the related work, using only the two common metrics reported in these researches. RMSE is shown not to compare but for informational purposes.

| **Method** | **Reference** | **Correlation** | **RMSE** |
|---|---|---|---|
| GLM | [109] | 0.65 | 1.97 |
| ARIMA | [109] | 0.47 | 2.14 |
| LASSO | [109] | 0.57 | 1.84 |
| FFNN | [109] | 0.63 | 1.73 |
| BMA | [109] | 0.73 | 1.53 |
| LASSO | [85] | 0.78 | 0.76 |
| SVM (RBF) | [85] | 0.89 | 0.52 |
| SVM (Linear) | [85] | 0.79 | 0.76 |
| AdaBoost | [85] | 0.92 | 0.44 |
| AdaBoost | [100] | 0.87 | 0.04 |
| SVM | [100] | 0.90 | 0.09 |
| LSTM | [100] | 0.61 | 0.01 |

# Chapter 2

# Acute Respiratory Infections Dataset and Internet Search Term Data

This chapter presents two datasets used along this research, and how they were collected, their setbacks, their analysis, and solutions found to overcome their deficiencies.

## 2.1 The acute respiratory infections data

The time series for ARI data for Mexico was generated using weekly epidemiological data reported by the Mexican Health Ministry (Secretaría de Salud), through the General Directorate of Epidemiology. This dataset is public and includes the weekly number of all medical encounters reported to the Health Ministry by healthcare facilities throughout the entire country for which the diagnosis corresponded to any of the following ICD-10[1] codes: J00-06, J20, J21, except J02 and J03 published in a weekly epidemiological report [88] (see Table 2.1). Weekly data for winter seasons included entries reported between epidemiological week 27 of 2002 and epidemiological week 26 of 2019[2].

[1] International Classification of Diseases 10th Revision.

[2] publicly available at www.genomica.uaslp.mx/Research/Bioinformatics/Bioinformatics.html

Table 2.1 ICD-10 codes for ARI and their diagnosis.

| **ICD-10 code** | **Diagnosis** |
|---|---|
| J00 | Acute nasopharyngitis [common cold] |
| J01 | Acute sinusitis |
| J04 | Acute laryngitis and tracheitis |
| J05 | Acute obstructive laryngitis [croup] and epiglottitis |
| J06 | Acute upper respiratory infections of multiple and unspecified sites |
| J20 | Acute bronchitis |
| J21 | Acute bronchiolitis |

The ARI data, as collected, are composed by year, week number and the ARI reported cases, the decision to collect the data from week 27th of 2002 to week 26th of 2019 was due to the training and testing required for the proposed methodology. An excerpt of the data is shown in Table 2.2 and some statistical properties are shown in Table 2.3.

Table 2.2 Excerpt of the ARI data as collected from the weekly epidemiological data reported by the Mexican Health Ministry (Secretaría de Salud), through the General Directorate of Epidemiology. Data is shown in two columns for formatting reasons.

| Year | Week | ARI cases | Year | Week | ARI cases |
|---|---|---|---|---|---|
| 2017 | 27 | 392,149 | 2018 | 1 | 527,679 |
| 2017 | 28 | 374,615 | 2018 | 2 | 648,268 |
| 2017 | 29 | 375,169 | 2018 | 3 | 672,728 |
| 2017 | 30 | 361,262 | 2018 | 4 | 706,863 |
| 2017 | 31 | 331,413 | 2018 | 5 | 636,584 |
| 2017 | 32 | 327,292 | 2018 | 6 | 570,324 |
| 2017 | 33 | 329,959 | 2018 | 7 | 581,180 |
| 2017 | 34 | 343,683 | 2018 | 8 | 580,628 |
| ... | ... | ... | ... | ... | ... |
| 2017 | 51 | 494,346 | 2018 | 25 | 351,992 |
| 2017 | 52 | 514,992 | 2018 | 26 | 346,265 |

Table 2.3 Statistics related to the ARI data 2002–2019.

| Minimum | Mean | Maximum | Median | Standard deviation | Variance |
|---|---|---|---|---|---|
| 190,665.00 | 452,223.67 | 1,184,372.00 | 452,185.50 | 123,443.03 | 15,238,180,975.03 |

### 2.1.1 ARI analysis

ARI are known to have seasonality, each year the same pattern is expected, but sometimes when an outbreak[3] occurs, there is an unexpected increase in the number of cases. The analysis of the ARI data can give some insight of its properties and its behavior in order to find the better approach for forecasting.

#### Fourier transform

It is important to define the seasonality of the ARI data and its length, because it will determine how the data is going to be separated and presented to the forecasting and projection techniques, the idea is to confirm that ARI data (from 2004 to 2019) has a cyclical yearly pattern the Fourier transform is used. It transforms the data from time domain to frequency domain, where the seasonality becomes evident by the accumulation of data. The frequency is defined as $1Hz$ equals 52 weeks (the expected seasonality, a year), and the amplitude measures the energy per unit of bandwidth, in this case $NumberOfCases/Hertz$ (see Figures 2.1 and 2.2).

#### Autocorrelation

This technique calculates the correlation between a time series (the ARI data) and a delayed version of itself, to check for instance independence. Its commonly used to decide if a time series can be modelled and/or used for forecast. It also can be used on the residuals from a fitted model to see if it can model all the none-random behavior of the data. This technique was used to ensure the ARI data could be modelled and to confirm the seasonal pattern in it (see Figure 2.3). Both the Fourier transform and autocorrelation analysis confirm the repetitive behavior of the data every 52 weeks (a year). The fact that the autocorrelation on the ARI data shows a sinusoidal pattern supports the idea of a cyclical behavior and a high self correlation, meaning that the time series itself contains a big amount of information that can be used to predict future

[3] According to the World Health Organization (WHO), a disease outbreak is the occurrence of disease cases in excess of normal expectancy (see Appendix A).

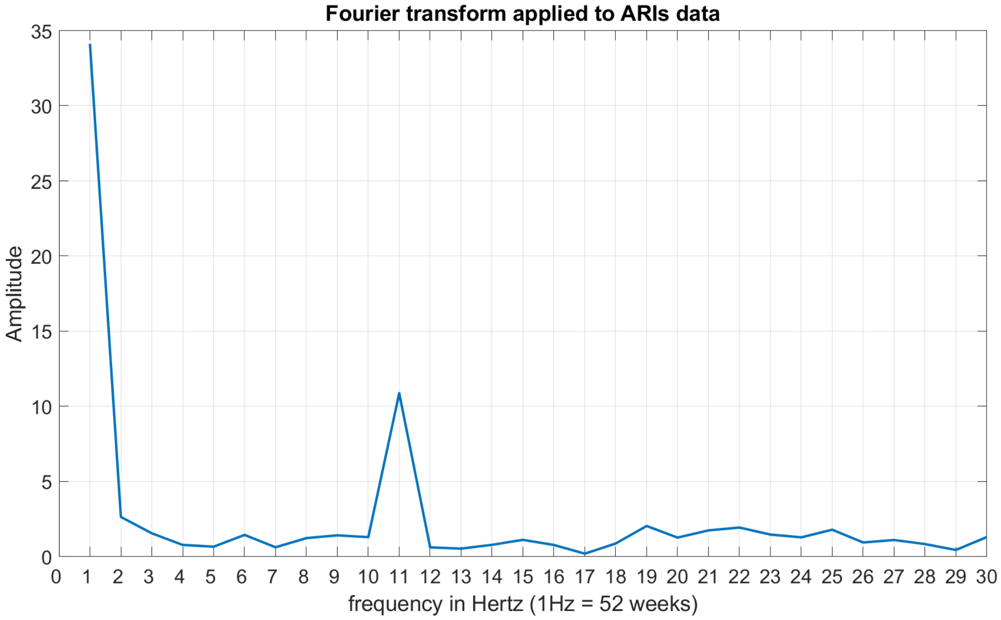


Figure 2.1 Fourier transform applied to ARI data (2004-2019) to detect seasonality. Frequency is in Hertz, where 1Hz equals 52 weeks (a year) and the amplitude measures $NumberOfCases/Hertz$.

behavior and the lag, measured in weeks, shows a cyclical pattern every 52 weeks confirming the yearly cyclical behavior of the time series.

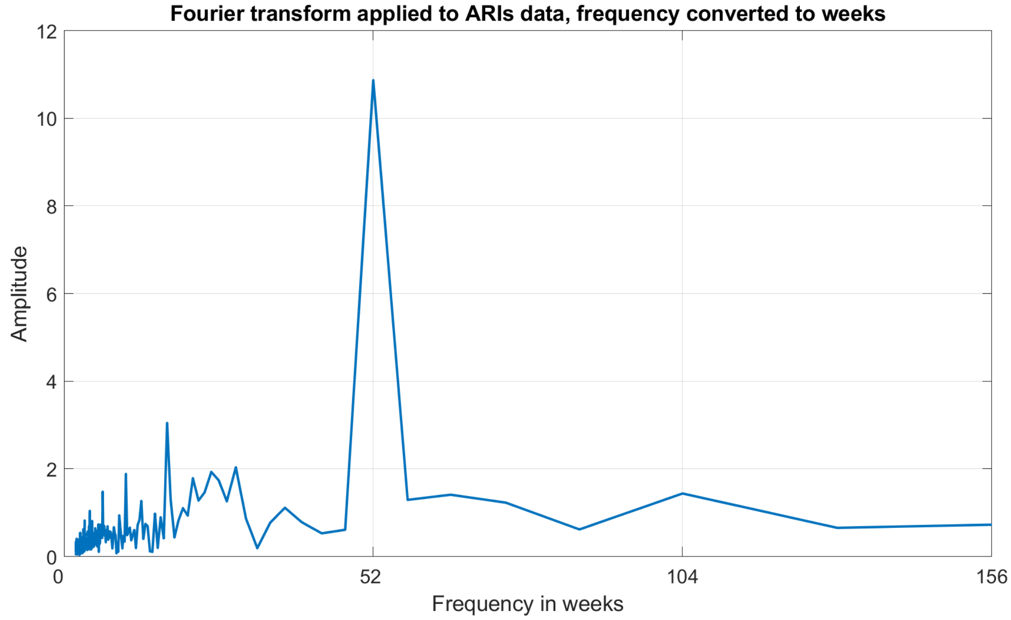


Figure 2.2 Fourier transform applied to ARI data (2004-2019), frequency is converted to weeks, showing the strength of the repeating pattern every 52 weeks. Amplitude remains $(NumberOfCases)/Hertz$.

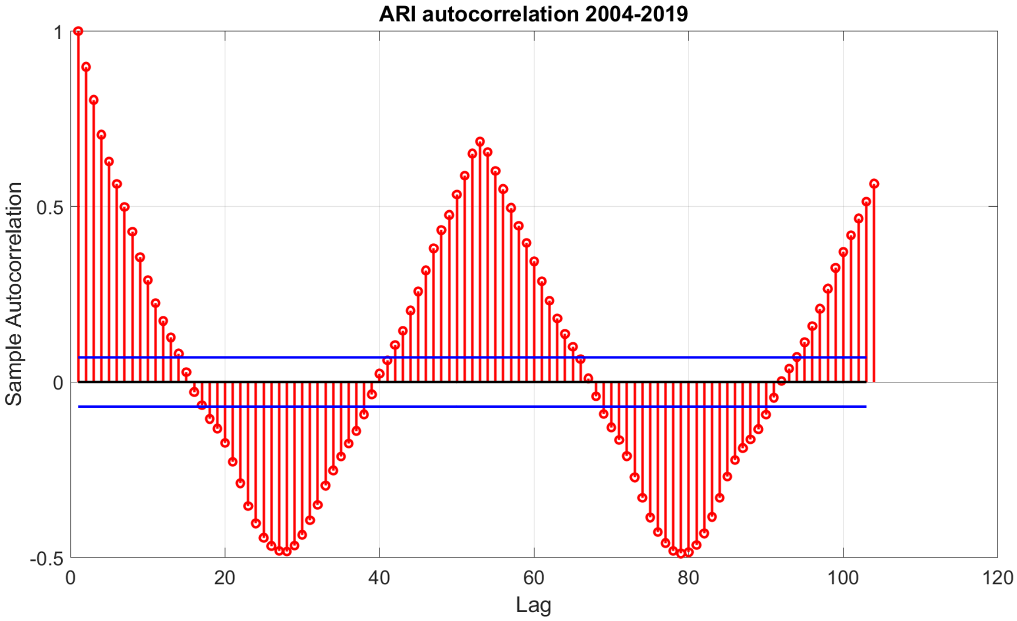


Figure 2.3 Autocorrelation analysis performed to ARI data from 2004 to 2019. The lag, measured in weeks, shows the yearly cyclical pattern repeating every 52nd week.

## 2.2 Internet search term data

The methodology proposes to use the behavior of online search engines by monitoring specific keywords, known as search terms, employed by the general public when they search for disease information to create a cross-sectional data table (see Appendix A). Google Trends[4] (GT) is used to retrieve the required data from the Google search engine. GT is a web service from Google that enables any user to download historical usage of any search term, given three main restrictions: Data goes as far as 2004, the data is shown to the user after normalizing its values between 0 and 100, and Google has privacy thresholds that hide information if the threshold is not surpassed[91]. The GT uses a webpage as interface, but there are programmatic tools that allow to automate the requests. In order to make the requests, it is needed to define the terms of interest, an initial set of terms was created by asking experts in biomedicine, epidemiology, virology and immunology of respiratory viruses, and computer science to define a set of twenty five possible words that people could use on the Internet search engines and could be correlated to the ARI cases (see Table 2.4).

Table 2.4 Initial list of Internet search terms proposed by experts in biomedicine, epidemiology, virology and immunology of respiratory viruses, and computer science.

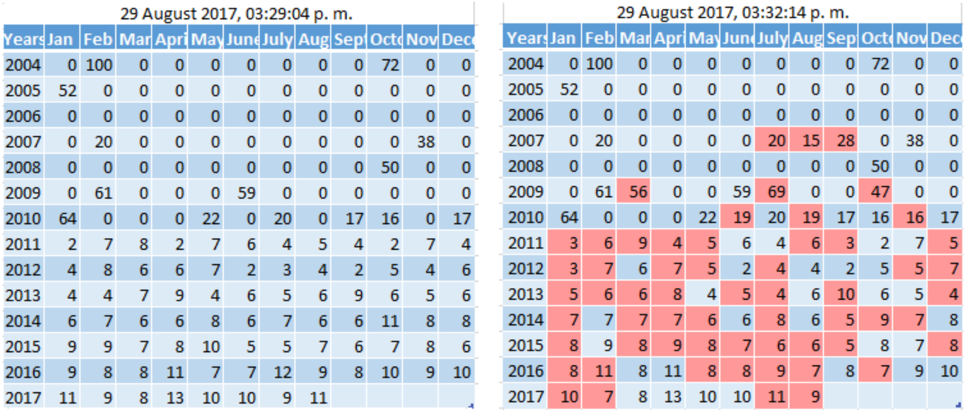
| | | | | |
|---|---|---|---|---|
| Gripa | Neumonía | Oseltamivir | Tamiflu | Vacuna |
| Gripe | Porcino | Mocos | OMS | Antibiótico |
| Flu | Porcina | Fiebre | InDRE | Antifludes |
| Influenza | Pandemia | La gloria | Alerta viajero | Gabirol |
| Catarro | H1N1 | Antigripal | Neumonía atípica | Cdc |

### 2.2.1 Internet search term analysis

The first request made to Google Trends (GT) contained the data of the terms listed in Table 2.4. To retrieve data from the GT service, it is required to define the desired range of time, location, and terms (for better results, it is better to ask for the terms one by one). This first data set helped to define some rules about the granularity of data but also showed some of the disadvantages of the GT service. A big setback is that one request to Google Trends is limited to 5 search terms. Another thing to have in mind is that Google Trends shares data from 2004 to date, and their gather-and-store information algorithms have changed through time [91];

[4] https://trends.google.com.mx/trends/

then, it would be expected that old data is less reliable than more recent data. And the most intriguing factor found yet is that the same request made at several times, may (and often does) not return the exact same data; for example, a request of the search term ‘Antibiótico’ from January 2004 to August 2017 made several times during the same day shows slight differences between requests (see Figure 2.4). This may be due to the algorithm used to fulfill the request, as GT can not go through all the raw data, it extrapolates from smaller subset to give a fairly accurate response in an acceptable time, as the problem was analyzed, some interesting facts emerged: Google Trends had updated its internal algorithms at least two times: once in 2011 and the second in 2016 [91].



29 August 2017, 03:29:04 p. m.

| Years | Jan | Feb | Mar | Apri | May | June | July | Aug | Sept | Octo | Nov | Dece |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2004 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 72 | 0 | 0 |
| 2005 | 52 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2006 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2007 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 38 | 0 |
| 2008 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 50 | 0 | 0 |
| 2009 | 0 | 61 | 0 | 0 | 0 | 59 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2010 | 64 | 0 | 0 | 0 | 22 | 0 | 20 | 0 | 17 | 16 | 0 | 17 |
| 2011 | 2 | 7 | 8 | 2 | 7 | 6 | 4 | 5 | 4 | 2 | 7 | 4 |
| 2012 | 4 | 8 | 6 | 6 | 7 | 2 | 3 | 4 | 2 | 5 | 4 | 6 |
| 2013 | 4 | 4 | 7 | 9 | 4 | 6 | 5 | 6 | 9 | 6 | 5 | 6 |
| 2014 | 6 | 7 | 6 | 6 | 8 | 6 | 7 | 6 | 6 | 11 | 8 | 8 |
| 2015 | 9 | 9 | 7 | 8 | 10 | 5 | 5 | 7 | 6 | 7 | 8 | 6 |
| 2016 | 9 | 8 | 8 | 11 | 7 | 7 | 12 | 9 | 8 | 10 | 9 | 10 |
| 2017 | 11 | 9 | 8 | 13 | 10 | 10 | 9 | 11 |  |  |  |  |

29 August 2017, 03:32:14 p. m.

| Years | Jan | Feb | Mar | Apri | May | June | July | Aug | Sep | Octo | Nov | Dece |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2004 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 72 | 0 | 0 |
| 2005 | 52 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2006 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2007 | 0 | 20 | 0 | 0 | 0 | 0 | 20 | 15 | 28 | 0 | 38 | 0 |
| 2008 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 50 | 0 | 0 |
| 2009 | 0 | 61 | 56 | 0 | 0 | 59 | 69 | 0 | 0 | 47 | 0 | 0 |
| 2010 | 64 | 0 | 0 | 0 | 22 | 19 | 20 | 19 | 17 | 16 | 16 | 17 |
| 2011 | 3 | 6 | 9 | 4 | 5 | 6 | 4 | 6 | 3 | 2 | 7 | 5 |
| 2012 | 3 | 7 | 6 | 7 | 5 | 2 | 4 | 4 | 2 | 5 | 5 | 7 |
| 2013 | 5 | 6 | 6 | 8 | 4 | 5 | 4 | 6 | 10 | 6 | 5 | 4 |
| 2014 | 7 | 7 | 7 | 7 | 6 | 6 | 8 | 6 | 5 | 9 | 7 | 8 |
| 2015 | 8 | 9 | 8 | 9 | 8 | 7 | 6 | 6 | 5 | 8 | 7 | 8 |
| 2016 | 8 | 11 | 8 | 11 | 8 | 8 | 9 | 7 | 8 | 7 | 9 | 10 |
| 2017 | 10 | 7 | 8 | 13 | 10 | 10 | 11 | 9 |  |  |  |  |

Figure 2.4 Same request made at different moments of a day. Each cell contains the normalized usage value returned for each month-year relation. Red marks show values that are different from the other request.

A new experiment was proposed: to retrieve the same data many times from the Google Trends service to check its congruence and reliability, separating it in three subsets: 2004-2010, 2011-2015 and 2016 to date, because of the changes made in the GT internal algorithms were made live in January 1st of 2011 and January 1st of 2016. It was found that there were many discrepancies between requests of each single term: many data values changed, which was more evident in older requests (see Figure 2.5).

At this moment the list of terms had grown to 34 terms, which were requested 21 times each; the data was gathered and analyzed, confirming that the oldest subset had greater discrepancies. Since the analysis of this data was first made by hand, watching and comparing the data in spreadsheets, the process was automated to reduce the processing time of the data. The details of how this process was automated and what tools were used is described in Subsection 3.1.1.

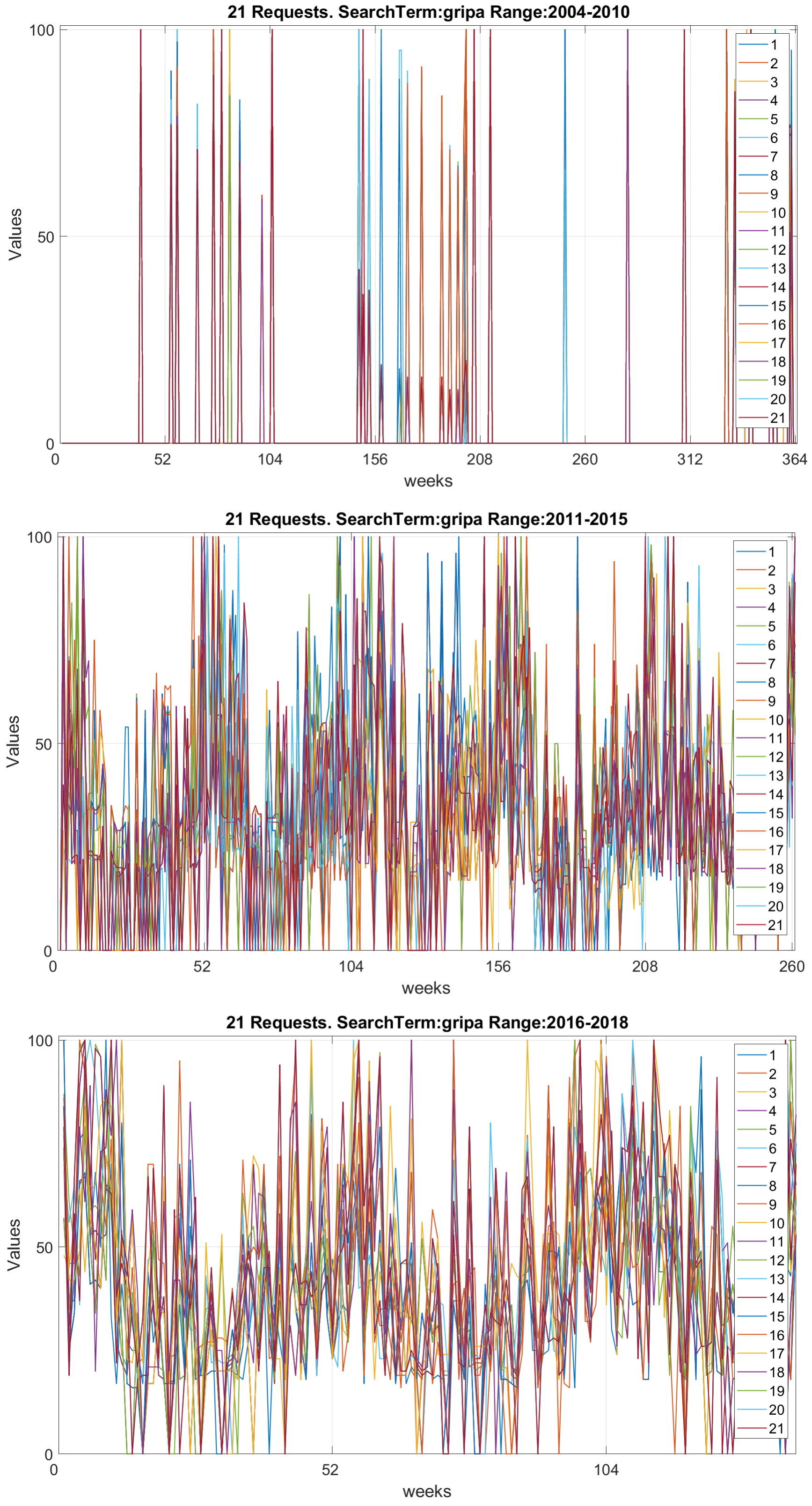


Figure 2.5 The same search term (“gripa") requested 21 times and separated in three spans of time (2004-2010, 2011-2015 and 2016-2018), vertical axis denotes the normalized values for the search term usage. The variance between requests is clearly higher in the 2004-2010 group.

The correlation between the same term requested at different times would give an idea of the similitude of these requests. The Pearson correlation was applied to the 21 requests of each term individually revealing how correlated is the term with itself on different requests, in many cases very low values were found (r>0.1) pointing to a clear discrepancy in the GT responses, thus making them unreliable. Table 2.5 shows an excerpt of the correlation matrices, showing five of the 21 requests for the term “gripa" on each of the time spans defined.

Table 2.5 An example of the self correlation tests (used term: “gripa"). Each table shows an excerpt of the correlation matrices (five of the requests) from the different span of years (2004-2010, 2011-2015 and 2016-2018), the rows and columns show the request number made for the same term. Each cell represents the correlation between the column request and the row request.

**Correlations: gripa 2004-2010**

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | **1** | 0.5470 | 0.5305 | 0.4492 | 0.4492 |
| 2 | 0.5470 | **1** | 0.9059 | 0.7094 | 0.7094 |
| 3 | 0.5305 | 0.9059 | **1** | 0.8148 | 0.8148 |
| 4 | 0.4492 | 0.7094 | 0.8148 | **1** | 1.0000 |
| 5 | 0.4492 | 0.7094 | 0.8148 | 1.0000 | **1** |

**Correlations: gripa 2011-2015**

| | | | | | |
|---|---|---|---|---|---|
| 1 | **1** | 0.4630 | 0.4737 | 0.4757 | 0.4909 |
| 2 | 0.4630 | **1** | 0.8247 | 0.7479 | 0.7299 |
| 3 | 0.4737 | 0.8247 | **1** | 0.9254 | 0.9071 |
| 4 | 0.4757 | 0.7479 | 0.9254 | **1** | 0.9747 |
| 5 | 0.4909 | 0.7299 | 0.9071 | 0.9747 | **1** |

**Correlations: gripa 2016-2018**

| | | | | | |
|---|---|---|---|---|---|
| 1 | **1** | 0.5073 | 0.4974 | 0.5348 | 0.5097 |
| 2 | 0.5073 | **1** | 0.8250 | 0.7932 | 0.7139 |
| 3 | 0.4974 | 0.8250 | **1** | 0.9167 | 0.8338 |
| 4 | 0.5348 | 0.7932 | 0.9167 | **1** | 0.9163 |
| 5 | 0.5097 | 0.7139 | 0.8338 | 0.9163 | **1** |

After this first approach to try to find the correct and unique value for each term was not successful by itself, various hypothesis were proposed:

- The terms related to ARI are the only ones with this behavior.

Table 2.6 List of terms unrelated to ARI data used to analyze Google Trends data.

| **Seasonal terms** | **Other diseases** | **Random terms** |
|---|---|---|
| “navidad" | “nauseas" | “facebook" |
| “futbol" | “diarrea" | “biblia" |
| “nfl" | “dolor" | “youtube" |
| “moda" | “cancer" | “elecciones" |
| “tendencia" | “tumor" | “becas" |
| “concierto" | “artritis" | “meme" |

- It is a problem exclusive to the studied region.
- The time of the day when the requests are made affects the outcome.
- Data from different requests have lag.
- Noise is affecting the data.

To find an answer to each of the possibilities and with the goal of being able to get the most reliable value, each of the hypothesis was addressed.

### The terms related to ARI are the only ones with this behavior

If only the terms related to ARI are affected, then, terms related to other diseases or not related to any disease would not show this behavior. A new list of terms is proposed containing six terms with expected seasonal behaviors, six terms related to other diseases, and six random terms commonly used by people (see Table 2.6). The same behavior was observed, but it was noted that terms that are greatly used show a little bit more similarities among their requests (still r>0.1) (see Figure 2.6).

### It is a problem exclusive to this region

To address the second hypothesis the requests were made also for another region. For this case the state of Tamaulipas was chosen, when this analysis was made the original idea was to focus on the State of San Luis Potosí, after the analysis the decision was to use the data from the whole country of México. The 32 terms were requested from this region but no meaningful differences were observed (r>0.2) (see Table 2.7).

Table 2.7 An example of the self correlation tests but from the Mexican state of Tamaulipas (used term: “gripa"). Each cell represents the correlation between the column and row of each pair of requests.

**Correlations: gripa 2004-2010 (Tamaulipas)**

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | **1** | 0.5470 | 0.5305 | 0.4492 | 0.4492 |
| 2 | 0.5470 | **1** | 0.9059 | 0.7094 | 0.7094 |
| 3 | 0.5305 | 0.9059 | **1** | 0.8148 | 0.8148 |
| 4 | 0.4492 | 0.7094 | 0.8148 | **1** | 1.0000 |
| 5 | 0.4492 | 0.7094 | 0.8148 | 1.0000 | **1** |
| | **Correlations: gripa 2011-2015 (Tamaulipas)** | | | | |
| 1 | **1** | 0.4630 | 0.4737 | 0.4757 | 0.4909 |
| 2 | 0.4630 | **1** | 0.8247 | 0.7479 | 0.7299 |
| 3 | 0.4737 | 0.8247 | **1** | 0.9254 | 0.9071 |
| 4 | 0.4757 | 0.7479 | 0.9254 | **1** | 0.9747 |
| 5 | 0.4909 | 0.7299 | 0.9071 | 0.9747 | **1** |
| | **Correlations: gripa 2016-2018 (Tamaulipas)** | | | | |
| 1 | **1** | 0.5073 | 0.4974 | 0.5348 | 0.5097 |
| 2 | 0.5073 | **1** | 0.8250 | 0.7932 | 0.7139 |
| 3 | 0.4974 | 0.8250 | **1** | 0.9167 | 0.8338 |
| 4 | 0.5348 | 0.7932 | 0.9167 | **1** | 0.9163 |
| 5 | 0.5097 | 0.7139 | 0.8338 | 0.9163 | **1** |

### The time of the day when the requests are made affects the outcome

Assuming the values are affected by time, an automated request was scheduled to run each hour; only two search terms were requested to avoid exceeding the number of requests allowed by Google Trends (the quota is estimated in 1,400 requests allowed every 24 hours approximately).

Time seems to affect requests in the way that requests closer in time tend to be more similar than requests made far in time (see Table 2.8). Higher correlations were obtained with this approach (r>0.4). Since previous experiments had shown some flaws in the data provided by Google Trends. It was dimmed important to check if differences in the time of the request are significant. The same request was performed daily for 5 days. The results show that from 2011 to date the differences are not significant, but it would be better to average those results for more accurate values.

Table 2.8 An example of the self correlation tests of requests made every three hours (used term: “gripa"). Each row shows the correlation of the request made at 12:00 paired with the value of the request made at the given hour and for the different span of years.

| | **gripa** | | |
|---|---|---|---|
| **Hour** | **2004-2010** | **2011-2015** | **2016-2018** |
| 12:00 | 1 | 1 | 1 |
| 13:00 | 1 | 0.8756 | 0.9093 |
| 16:00 | 0.8400 | 0.8371 | 0.9093 |
| 19:00 | 0.8455 | 0.6872 | 0.9094 |
| 22:00 | 0.8166 | 0.6757 | 0.9139 |
| 01:00 | 0.6743 | 0.6835 | 0.9139 |
| 04:00 | 0.5989 | 0.4559 | 0.5328 |
| 07:00 | 0.9775 | 0.8674 | 0.9265 |

### Data from different requests have lag.

For each term, all requests were plotted in a single graph to help recognize if there are any visible delays, lags or shifts in the data. This made evident that there is a lot of noise in the requests, in many of the terms it was impossible to detect a trend or a slope (Figure 2.7). Data does not seem to be lagged, but it looks like there is a lot of noise in some of the requests, apparently having many extreme values randomly in the series (0 and 100).

### Noise is affecting the data

To overcome the apparent noise in the requested terms, the variance is calculated for each of the requests for one term, then, only the requests with the lower variances are plotted, this showed a cleaner graph with clear identifiable trends and slopes (Figure 2.8) clearly enhancing the visual detection of the behavior and giving the chance to filter the different requests and generating a single time series for each term.

To resolve the limit of terms for each request, it was decided that each request would ask for one term at a time, also, the time range for each individual request would be a year, mainly because GT does not allow to specify the desired granularity, in this case weekly, but the service decides this depending on the time range you are requesting. Asking data one year at a time assures the granularity of the response will be weekly, as desired.

Another possible disadvantage of this service is that the search term usage is normalized between 0 and 100 for the time range you are asking, meaning that, no matter the time range you ask for, the maximum value found in this range will be 100 and the minimum 0. In this specific case this means that every year each term will have a maximum value of 100 and minimum of 0, no matter how greater or lower the data of one specific year is compared to another they will both have a maximum of 100 and minimum of 0, this represents losing valuable data of the behavior of the search term usage through the years.

To prevent this data loss, the best solution is to make one additional request for each search term, asking for the range from January 2004 to date, this request returns monthly data. With these monthly values, it is possible to apply the month value as weight for the weeks of that month to give them a proper global value, meaning that if a month has a value of 50 (out of 100) then, the values of the weeks composing that month should be multiplied by the monthly value as a percentage (0.50 in this case), as the monthly data is according to the most global time range (the normalization is applied to the whole range from January 2004 to date) the result will be a weekly value that is also according to this normalization, giving to each week the proper weight compared to the rest of the time series.

The GT data was deeply analyzed and many tests were made to the requested search terms to determinate its reliability, since several changes have been made to the internal mechanisms of the service, and Google has kept the details of these changes private, along with their security threshold[91]. It was then determined that the records from 2004 to 2010 are unreliable, mostly because their responses have very high variances and when requested several times the data is no consistent. Data from 2011 to 2015 are more reliable, data shows consistency but high variances in some terms, but the best records are found from 2016 to date, where data is more consistent and populated in comparison to the previous years. The idea of using these data to forecast the 2009 influenza outbreak is discarded to avoid, as much as possible, using search term data from before 2010. As the proposal started developing and, because it included an automated search term removing method, the list of terms escalated to 168 search terms. The decision then is to request the data, from Mexico, of the 168 search terms from the 27th week of 2008 to the 26th week of 2019 discarding the 2009 because it was affected by the outbreak.

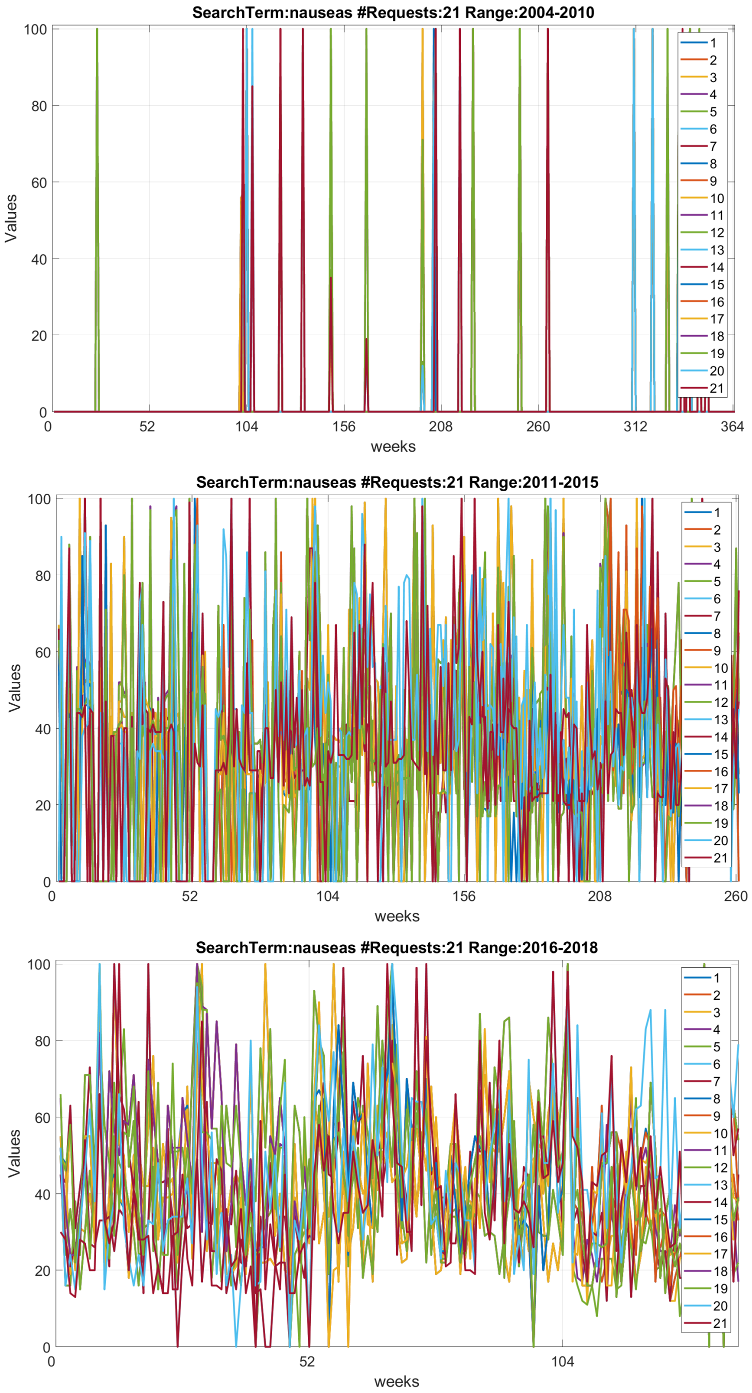


Figure 2.6 Example of ARI unrelated search term request (“náuseas") separated in three spans of time (2004-2010, 2011-2015 and 2016-2018), vertical axis denotes the normalized values for the search term usage.

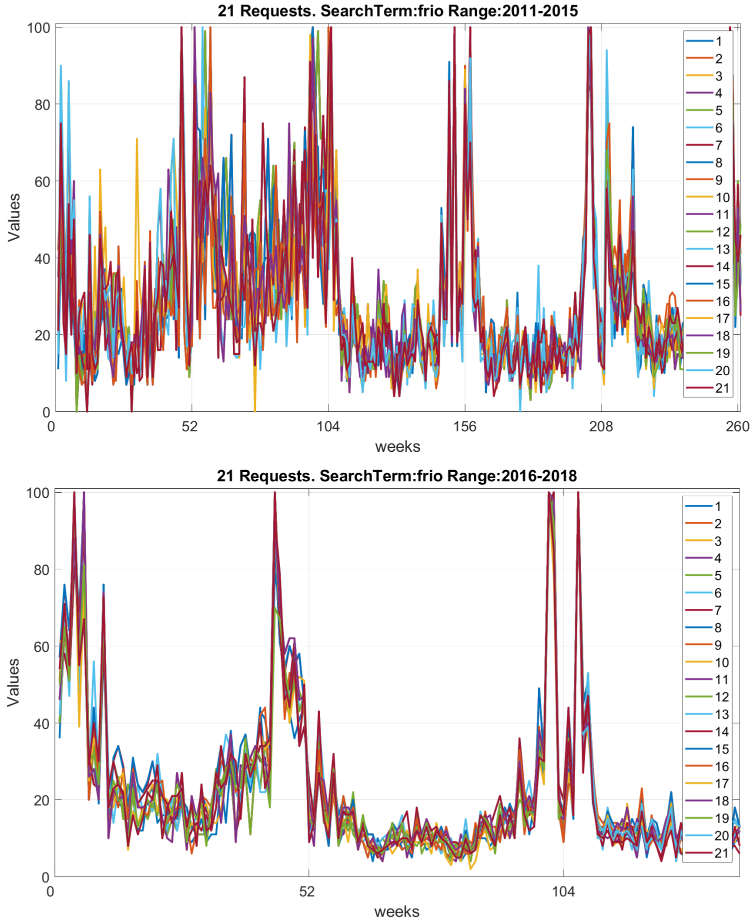


Figure 2.7 Example of data returned by Google Trends tool, no lag is detectable between requests. Term used: “frío", vertical axis denotes the normalized values for the search term usage.

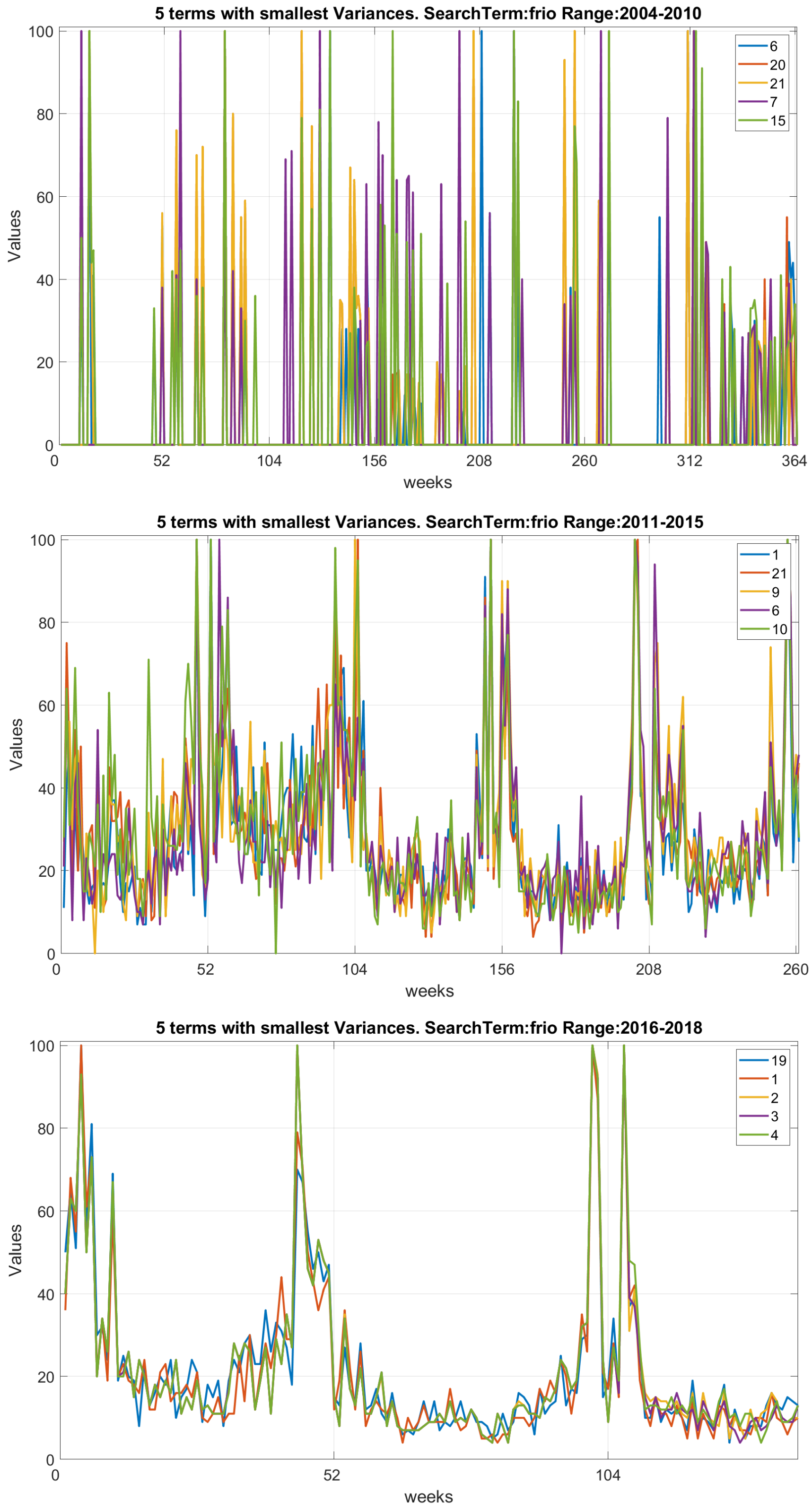


Figure 2.8 Search terms filtered requests, only 5 with the smallest variances are plotted, term used: “frío". As it can be more clearly seen in the 2016-2018 plot, the requests with the smallest variances show better quality in their data, vertical axis denotes the normalized values for the search term usage

# Chapter 3

# Forecasting Method using Artificial Neural Networks and Search Terms

After understanding and defining the two main datasets for the research, it is possible to continue with the study, definition and testing of the architecture for the proposed methodology. This chapter addresses the proposed methods and architecture of this research, defining each technique used and its role in the methodology. The methodology is divided in two stages: data acquisition and computational model, as shown in Figure 3.1. Data acquisition is divided in three modules: data retrieval, data preprocessing and feature extraction. On the other hand, the computational model has two modules: merge prediction and ARI forecast. Every module contains several tasks.

## 3.1 Data acquisition

The data acquisition stage involves three modules: data retrieval, data preprocessing and feature extraction (the data science processes) [16, 35]. A detailed diagram of this stage is shown in Figure 3.2.

### 3.1.1 Data retrieval

**ARI raw dataset**

The time series for ARI from Mexico were generated based on weekly epidemiological data reported by the Mexican Health Ministry (Secretaría de Salud), and includes the weekly number

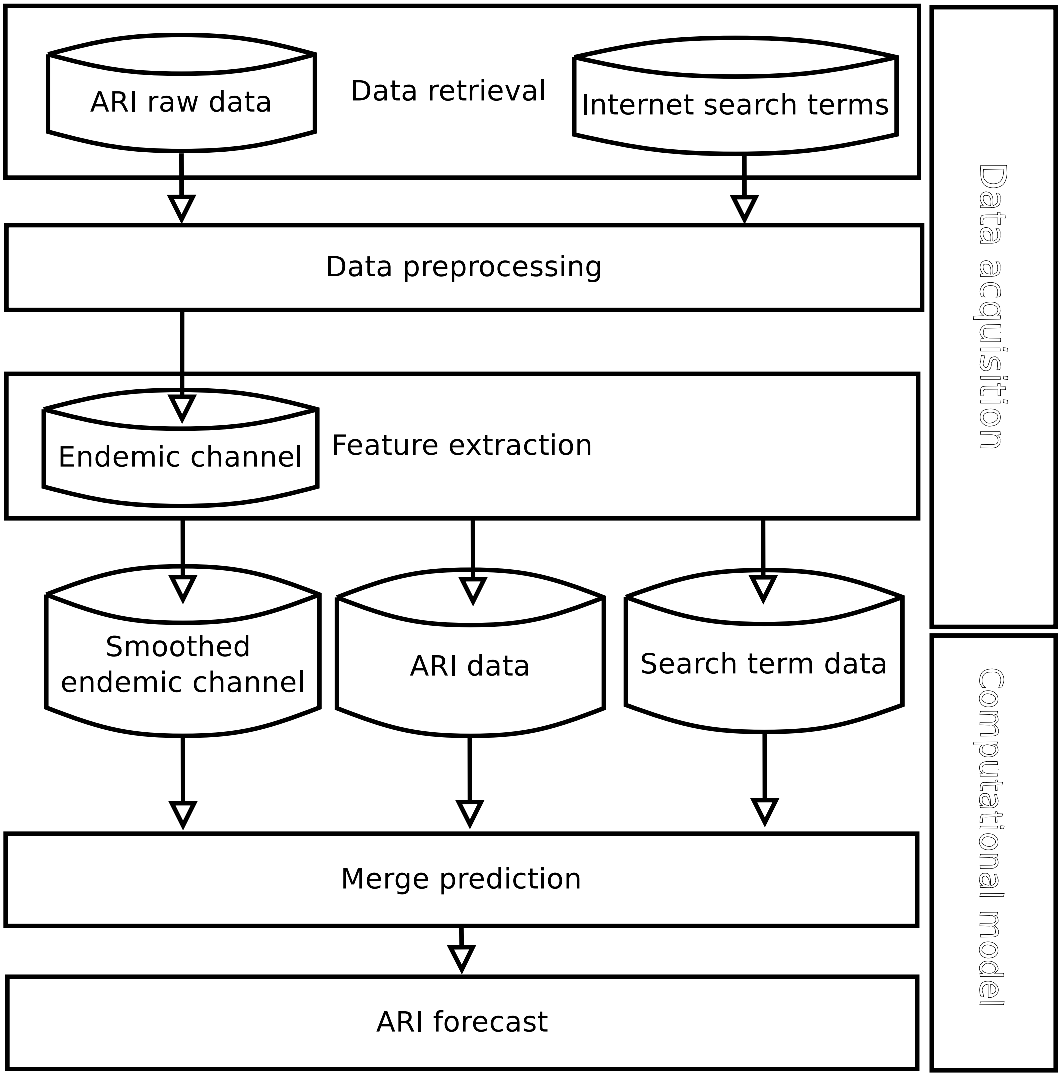


Figure 3.1 The proposed architecture is composed of two stages: the data acquisition and the computational model.

of all medical encounters reported to the Health Ministry by healthcare facilities throughout the whole country as mentioned earlier. This dataset is conformed by weekly data between epidemiological week 27 of 2008 and epidemiological week 26 of 2019 (The range between week 27 of current year to week 26 of the next year is known as winter season).

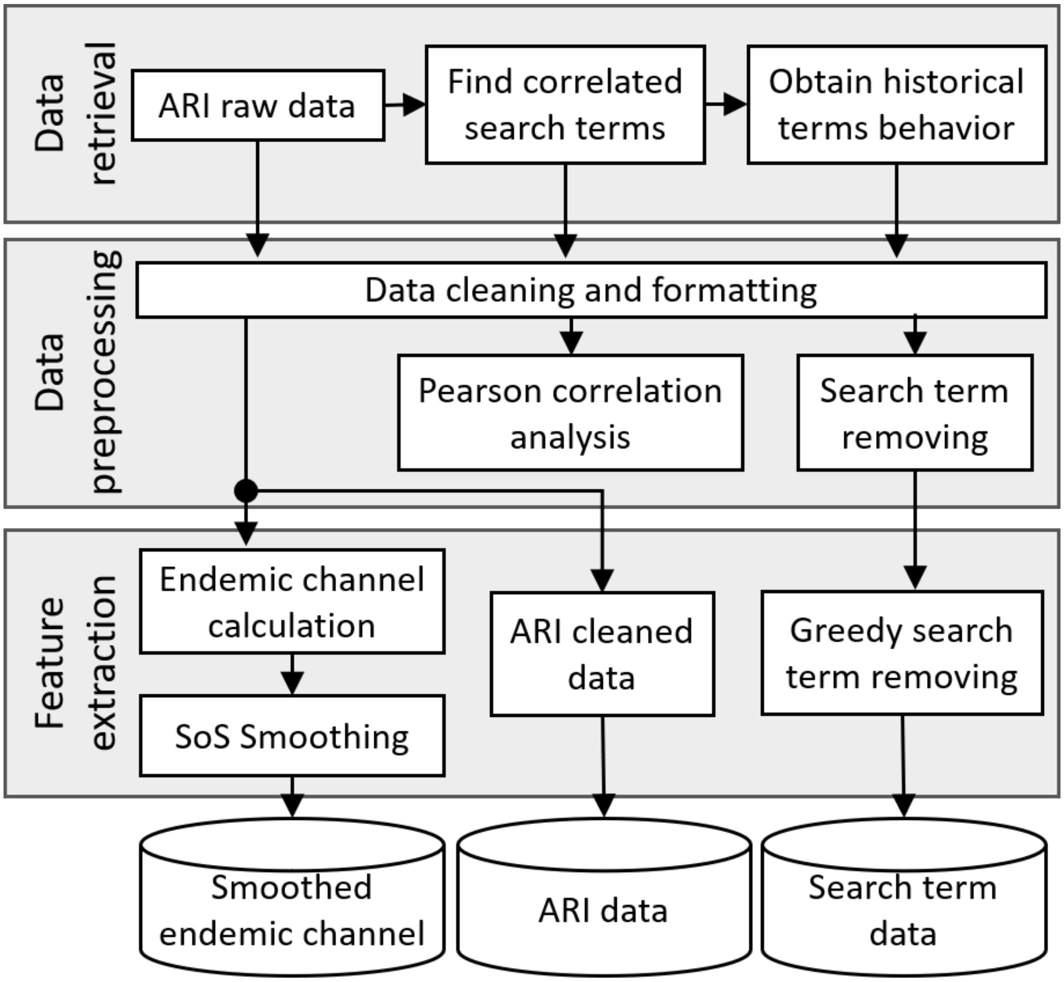

Data retrieval
ARI raw data
Find correlated search terms
Obtain historical terms behavior
Data preprocessing
Data cleaning and formatting
Pearson correlation analysis
Search term removing
Feature extraction
Endemic channel calculation
SoS Smoothing
ARI cleaned data
Greedy search term removing
Smoothed endemic channel
ARI data
Search term data

Figure 3.2 The data acquisition stage is composed by three modules: Data retrieval, data preprocessing and feature extraction.

### Find correlated search terms

The search terms needed to be a set of words that could be used on the Internet search engines and could be correlated to the ARI cases but also search terms defined by experts in epidemiology, virology and respiratory infections. As the proposed model is enabled to remove the non-significant terms, hence, a list of 168 terms is defined, composed by possible terms related to ARI data (see Subsection 2.2).

### Obtain historical terms behavior

With the list of terms defined above, the historical usage of the terms was requested to Google Trends; the data was structured weekly, and focused on Mexico. Using the same time window as ARI raw dataset, the terms were retrieved from Google Trends services using a NodeJS Software Development Kit (SDK) called “google-trends-api"[1], created by Patrick Trasborg in 2016 under the MIT License[2] (A very open and permissive software license). The script deals with the request and retrieve functionalities of the service. For each request, the service returns a JSON[3] file (A versatile structured file similar to XML files) with the response. If no data is available, the JSON structure returned is empty. At the end of this procedure there is time series for each search term requested.

## 3.1.2 Data preprocessing

### Data cleaning and formatting

In the data preprocessing stage the JSON files are parsed and used to populate tables. The tables are then cleaned from missing values, filling empty ones with zeros. The missing values appeared with the tag "<Missing>", this occurred with the 53rd week when years only had 52 weeks, or when a term was not used during a whole requested year the value for the weeks was a missing tag. For the historical usage of search terms and ARI raw data, the $53^{\text{rd}}$ week was eliminated for those years that have an additional epidemiological week (Normally when a year ends in the middle of a week, that week is assigned to the ending year, this situation generates scenarios where most years have 52 weeks but some have 53 weeks), to ensure a constant size of 52 weeks for all the years.

[1] https://www.programmableweb.com/sdk/google-trends-nodejs-sdk-patrick-trasborg

[2] https://opensource.org/licenses/MIT

[3] https://www.json.org/json-en.html

#### Pearson correlation analysis

The data is analyzed to remove the terms least correlated with the ARI dataset by using the Pearson Correlation. The Correlation between ARI data and each of the search terms is calculated and stored for future analysis, then the search terms time series is paired with the ARI dataset into one single data table.

#### Search term removing

During this process, terms with a correlation coefficient below 0.75 are removed, reducing the list of terms from 168 to 28. No negative correlations are expected, because that would mean that there could be a relation between the increase in ARI cases and the reduction of searches of some term, where the terms selected are expected to have a positive correlation with the ARI data.

### 3.1.3 Feature extraction

In this module, the ARI dataset is analyzed to extract relevant information such as the endemic channel calculation and the proposed sum-of-sines endemic channel calculation. On the other hand, the search terms are tested with a greedy algorithm to select only the most relevant search terms.

#### Endemic Channel (EC)

Many countries use a version of the endemic channels for visualising the expected case levels, based on the weekly (or monthly) average number of cases over the preceding 5 years [6]. As such, endemic channels represent the boundaries for the expected number of cases at a given time; the occurrence of cases above this moving threshold would be considered as an outbreak of a disease [45] (Detailed information about endemic channels in Appendix A).

Commonly, an endemic channel is obtained by calculating the $5^{th}$, $50^{th}$ and $85^{th}$ percentiles of a disease for each week from the 5 previous years (ignoring years affected by outbreaks); the $5^{th}$ percentile (lower endemic channel) represents the minimum number of cases expected or the area of success, the $50^{th}$ percentile is the expected behavior (the median or secure area), and the $85^{th}$ percentile (upper endemic channel) is the risk area or area of alert, surpassing these values points to possible outbreaks[10]. When using EC there is a risk of not detecting new behaviors in the data, and limitations associated with abnormally high historic means and the variation in the seasonal timing that often lead to inaccurate detections[11].

As part of these research, the use of a smoothing function applied to the endemic channels is proposed. The idea is to apply a smoothing algorithm to define the endemic channels of the ARI data and overcome some of the EC deficiencies. In order to achieve this and after testing the endemic channels need to be calculated using the lowest value for each week ($1^{st}$ percentile), the median ($50^{th}$ percentile), and the maximum value ($100^{th}$ percentile), this enables the smoothing algorithm to better detect the behavior of the ARI data and helps to reduce trend and seasonality data loss.

### Sum of sines function (SoS): Endemic Channel Smoothing

As the commonly used endemic channels have many sudden variations in their behavior, this makes them prone to false positives, i.e. if there is a sudden drop in values in the endemic channel but the current number of ARI cases has just a slight increase that makes it above the endemic channel threshold. After exploring different possible curve fitting and smoothing methods [5, 87], it was decided to propose a smoothed endemic channel by using the sum of sines function as a signal smoothing method. This method was applied to the $1^{st}$, $50^{th}$ and $100^{th}$ percentiles of the ARI time series, resulting in a set of smoothed endemic channels (Lower, median and upper endemic channels) (see Figure 3.3).

The sum of sines function can be used to model time series with historical data and project its possible future behavior. Depending on such behavior, the function may be composed by sum of several sinusoidal functions in order to fit it. A generalization is shown in the following equation:

$$\mathbf{y} = a_1 sin(b_1 x + c_1) + a_2 sin(b_2 x + c_2) + ... + a_n sin(b_n x + c_n) \tag{3.1}$$

Where $a_i$, $b_i$, $c_i$ are unknown parameters that are needed to be fitted to the data; $x$ represents the week number (time); $y$ represents the output, i.e. the smoothed endemic channel. A non linear least squares method can be used to find values for these parameters, the `fit` function in MATLAB® from Curve Fitting Toolbox™ is used for this purpose, after several tests, the best found values of the involved parameters were defined as shown in Table 3.1, and an example of the obtained results is shown in Figure 3.3.

Table 3.1 Parameters for the fit function used to smooth the endemic channels.

| **Parameters** | **Values** |
|---|---|
| Model type to fit | Sum of sines function (5 Sines) |
| Algorithm options | Non linear least squares |
| Maximum evaluations allowed | 110 |
| Maximum iterations | 400 |
| Termination tolerance on model value | $10^{-6}$ |
| Termination tolerance on coefficient values | $10^{-6}$ |
| Robust linear least-squares fitting method | Least absolute residual method |

### Greedy Search Term Removing

The goal of this task is to reduce the list of search terms, because the removal of irrelevant terms helps to minimize the input noise going into the forecasting models. Brute force algorithms test all possible combinations of search terms, but these algorithms are computationally expensive and time consuming. Instead, the selection of terms is carried out using a Greedy method that iterates choosing the best option available at the moment [26, 67] by using an Artificial Neural Network (ANN). The greedy algorithm is defined as follows:

- An initial list of inputs including only the week number
- Each term is used as input for the ANN and tested 300 times
- If there is an improvement in the results:
- The term with the lowest averaged RMSE is added to the list of inputs and removed from the list of terms
- The process is repeated until no improvement is found
- The final list of inputs contains the best inputs

The ANN used is a Feed Forward Neural Network (FFNN), The input layer starts with only the week number and one term. The best option is evaluated with the Root Mean Squared Error (RMSE) between the ARI cleaned data (hereafter ARI data, cleaned at the preprocessing module) and the output of an ANN (see Figure 3.4), after the greedy approach, the list of inputs was reduced to only 6 terms.

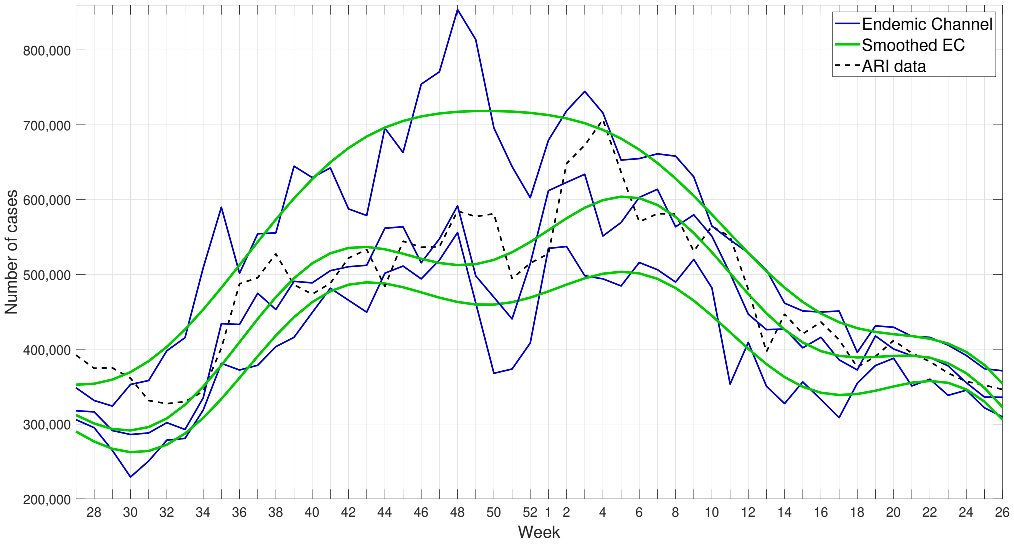


Figure 3.3 Smoothed Endemic Channels using sum of sines (SoS). At the top are the $100^{th}$ percentiles, at the middle are the $50^{th}$ percentiles and at the bottom the $1^{st}$ percentiles. The thick green lines show the smoothed channels (SoS, lower, median and upper), and the dotted line the ARI data.

### Feed Forward Neural Network

This kind of neural network is also known as backpropagation network [82, 8, 42]. They are composed of an input layer, one or more hidden layers, and an output layer. Each layer has a set of nodes, artificial neurons, and each layer is connected to the next by weights and the neurons of a layer are connected to all the neurons on the next layer. The neurons in the hidden layers have an activation function which is applied to the inputs received from the previous layer each one multiplied by its respective weight (see Appendix A for more details on the FFNN). The number of hidden layers and neurons on each of them was estimated by a grid search algorithm that tested all configurations for one and two layers with 10 to 100 neurons on each. The best configuration found (see Table 3.2), lowest RMSE in the testing stage, is a two hidden layer architecture with 60 and 20 neurons respectively (see Figure 3.5). By using the `fitnet` function in MATLAB® from the Deep Learning Toolbox™, a neural network was trained and tested 300 times using as input the week number and each of the terms and the average RMSE for every term was recorded. Subsequently the RMSE was compared between search terms, and the term with the lowest RMSE is kept. The test is repeated adding another

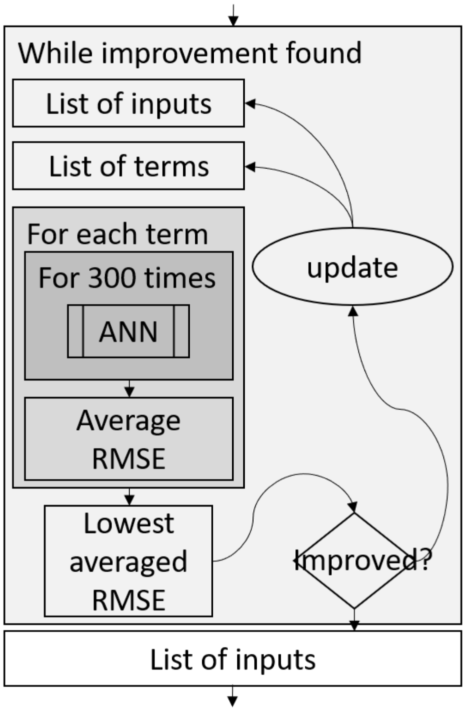


Figure 3.4 Greedy search term removing process.

term, until minimal or no improvement is found. Using this approach, the list of search terms was reduced to six keywords: asthma, bronchitis, what is flu, respiratory, cough, and respiratory tract (in Spanish correspond to ‘asma’, ‘bronquitis’, ‘que es la gripe’, ‘respiratorias’, ‘tos’ and ‘vias respiratorias’, respectively), these terms together showed the lowest RMSE with ARI data.

## 3.2 Computational model

The second stage of the proposal (see Figure 3.1), the computational model, has one module, Merge prediction, divided in three tasks: a) a forecasting model built with FFNN, b) a projection model based on SoS; and c) the merge prediction, which integrates both previous components besides the smoothed endemic channels as inputs, see Figure 3.6. The forecasting model makes its predictions based on the search terms behavior, while the projection model works in parallel with ARI data and the smoothed endemic channel, and both making a one-week projection of its expected behavior.

Table 3.2 Parameters used for the ANN and the *fitnet* function.

| **Parameters** | **Values** |
|---|---|
| Training function | Conjugate gradient backpropagation with Fletcher-Reeves updates |
| Hidden layers | [ 60 , 20 ] |
| Input/output processing | Normalize inputs/targets to fall in the range [-1,1] |
| Input/output processing | Remove rows with constant values |
| Train epochs | 1,000 |
| Train test data division | Divide up every sample |
| Validation and testing | [ 52 weeks, 52 weeks ] |
| Training | Remaining weeks |
| Performance function | Mean square error |

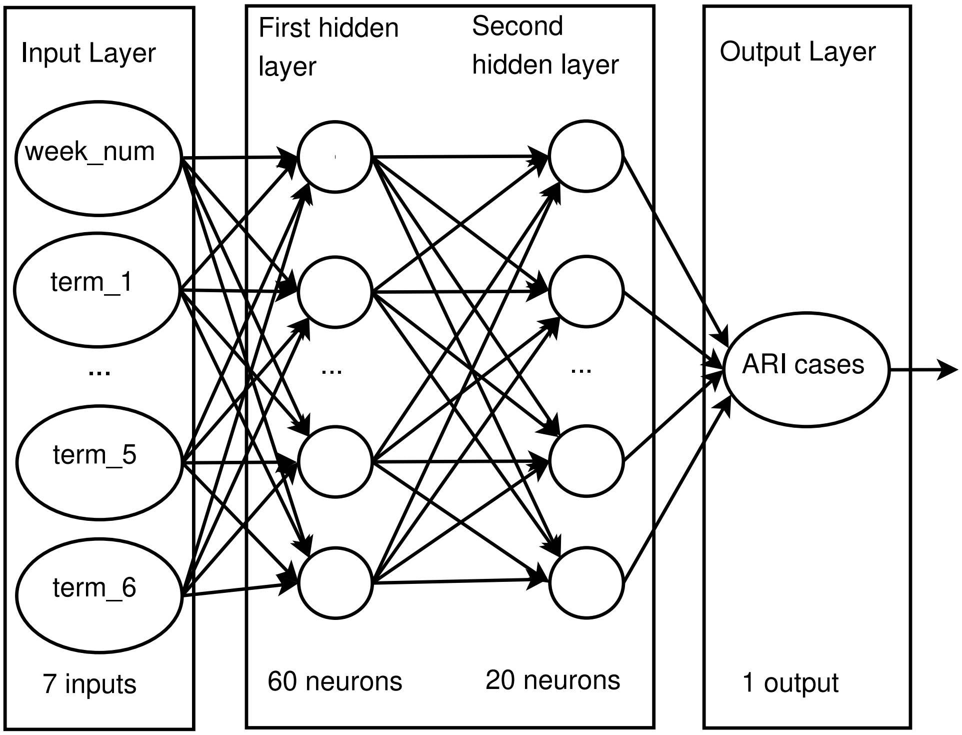


Figure 3.5 Feedforward Neural Network (FFNN) architecture

### 3.2.1 Forecasting Model: FFNN

The FFNN is used to make predictions of the expected number of ARI cases for the next weeks by receiving the recent usage of the list of search terms. This type of network is composed of an input layer, one or more hidden layers, and an output layer each composed by a set of nodes,

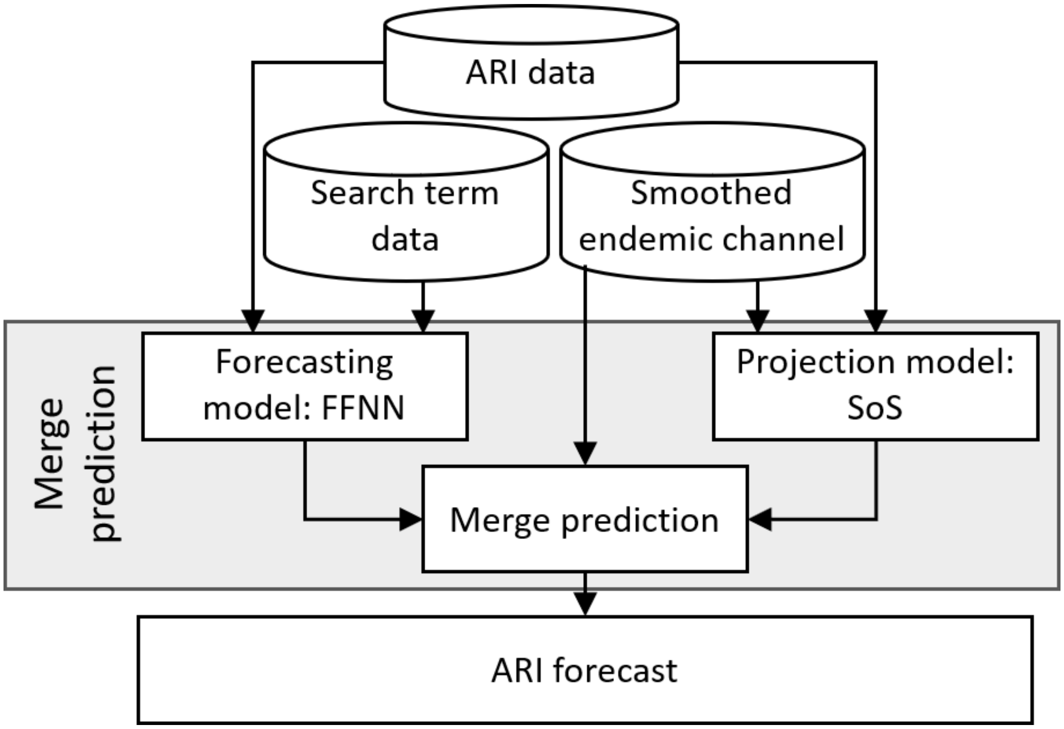


Figure 3.6 The computational model, which receives the ARI data, the search terms data and the smoothed endemic channel data as inputs, and is composed by a Forecasting model working in parallel with a Projection model to feed the third module of this stage, the Merge prediction.

called neurons. The number of hidden layers, their number of neurons and the parameters used are described in Subsubsection 3.1.3. The lowest averaged RMSE in the testing stage resulted in a two hidden layer configuration with 60 and 20 neurons respectively. The FFNN receives as input the current week number (1-52) and internet usage for each of the six search terms during that week, and it is trained to predict the ARI data for the next week. An example of the response of the FFNN for the 2017-2018 winter season is shown in Figure 3.7.

### 3.2.2 Projection Model: SoS

After studying and testing the SoS function to calculate the endemic channels, it was found that it could be fitted to project its behavior, tests showed that it could make reasonable predictions when projecting the function for several months, the only thing found that could be a restriction is that the function required 9 or more years of history to make the best predictions, for this case there is enough historical data available, a sum of sines function was used to model the behaviour of the ARI cases throughout the years.

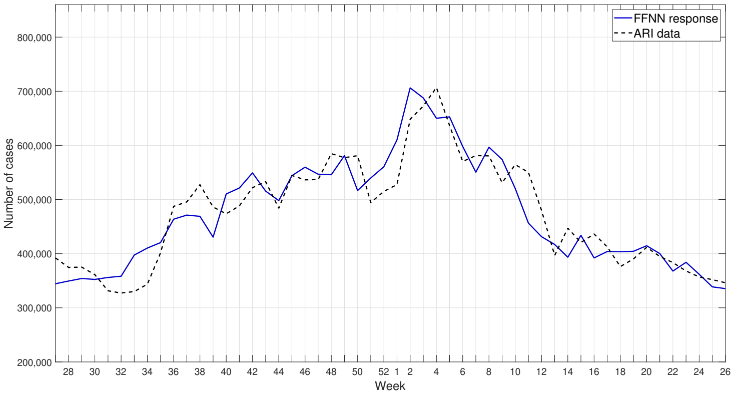


Figure 3.7 ARI data and FFNN response forecasting one week in advance for the 2017–2018 winter season.

More specifically, a sum of five sines function was used, as denoted in Equation 3.1, selected for this task because of the results obtained in section 3.2.2). The resulting parameters obtained by the `fit` function are shown in Table 3.3. The fitted model is then used to project its behavior. Several tests were performed in order to define the time span that resulted in the best prediction, and in order to simplify the comparison with the related work [85, 100], it was decided to adopt a one-week forward setup for our projections. The resulting fitted model for the ARI data is shown in Figure 3.8 and the one-week projection for the winter season 2017-2018 in Figure 3.9.

Table 3.3 Values of the adjusted parameters using the non linear least squares fitting method. Each row *i* represents the parameters affecting the *ith* sine

| *i* | *a* | *b* | *c* |
|---|---|---|---|
| 1 | 2.168e+04 | 0.002802 | 0.316 |
| 2 | 8886 | 0.004113 | 2.828 |
| 3 | 4021 | 0.1207 | 1.491 |
| 4 | 746.3 | 0.02359 | 1.264 |
| 5 | 1155 | 0.3624 | -0.6893 |

The purpose of these two models (Forecast and projection) is that the forecasting model would predict unexpected increases in ARI cases because it is fed by the search term usage, and the projection model would show the expected behavior of the ARI cases, assuming no unusual behavior. The idea is to join these two models and the boundaries given by the Smoothed endemic channels (lower and upper) to generate an accurate prediction as described below.

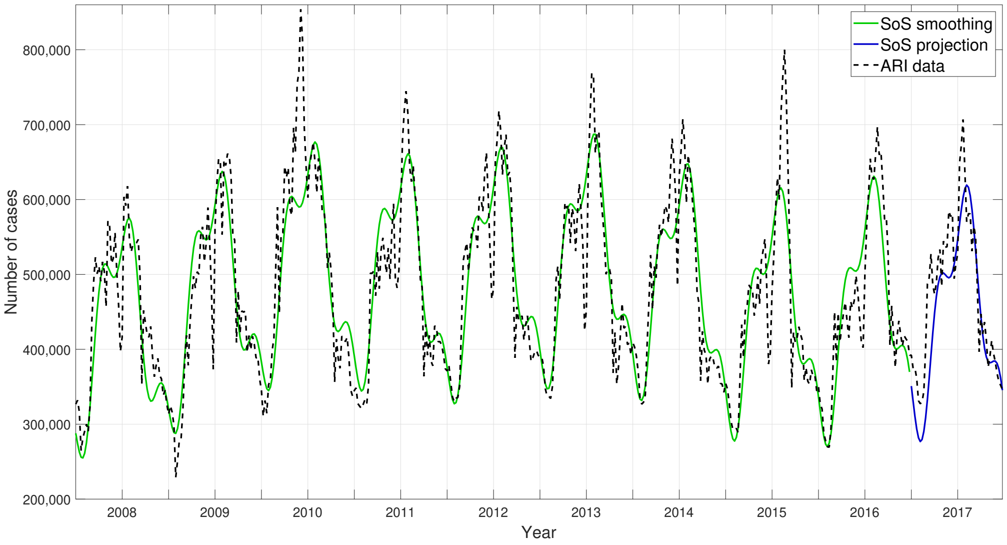


Figure 3.8 ARI data and sum of sines (SoS) fitted model.

### 3.2.3 Merge Prediction

For the merge prediction, a linear equation was used. This equation involves the results from the forecasting model and the projection model along with the lower and upper smoothed endemic channels, as a weighted sum.

$$\mathbf{\hat{y}_t} = w_1 x_{1_t} + w_2 x_{2_t} + w_3 x_{3_t} + w_4 x_{4_t} \tag{3.2}$$

where, $x_{n_t}$ is each of the responses, forecasting model, projection model, lower endemic channel value and upper endemic channel value, respectively. Parameters $w_1$ to $w_4$ are the weights given to each response and they are calculated fitting the equation to minimize its RMSE using the responses and the ARI data from the immediate previous year; the resulting values are shown in Table 3.4.

The resulting parameters in Table 3.4 show the percentage of participation of each of the responses, the reason behind these values, specifically the low value on the projection model, is that it is not desired that the output remains in the center of the endemic channels, but rather that the changes detected by the forecasting model force the output to increase, but in a controlled manner, which is the participation of the endemic channels, to control the output in these sudden

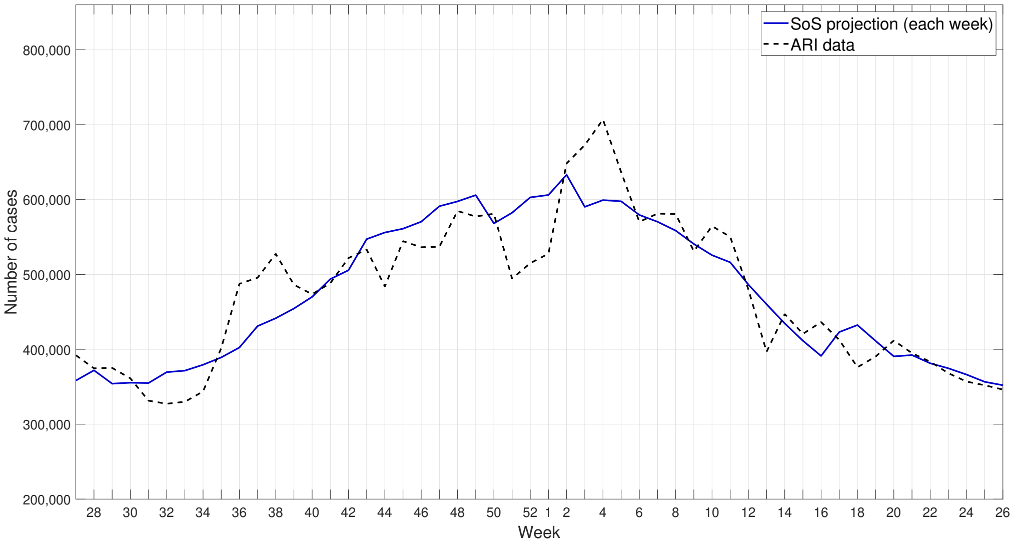


Figure 3.9 ARI data and sum of sines one-week projection for 2017–2018 winter season.

Table 3.4 Response variables and parameter values using responses from the 27$^{th}$ week of 2017 to 26$^{th}$ week of 2018

| **Variable** | **Response** | **Parameter** | **Value** |
|---|---|---|---|
| $x_1$ | Forecasting model | $w_1$ | 0.37 |
| $x_2$ | Projection model | $w_2$ | 0.045 |
| $x_3$ | lower endemic channel | $w_3$ | 0.38 |
| $x_4$ | upper endemic channel | $w_4$ | 0.205 |

changes of the forecasting model. The following chapter will discuss these and other details on the results.

# Chapter 4

# Results

This chapter presents the results from the proposed methodology fed with ARI data and Internet search terms. The recorded results are focused on the last stage of the methodology, the computational model (Section 3.2), i.e. forecasting model, projection model and the merge prediction. The metrics used are described below, followed by the definition of the experiments and results from the forecasting model and the merge prediction. In order to measure and compare the performance and accuracy of the proposed methods, four metrics have been chosen: Pearson correlation coefficient (*r*), Root Mean Square Error (RMSE), Root Mean Squared Percent Error (RMSPE), and the Maximum Absolute Percentage Error (MAPE) (the discarded metrics tested are detailed in Appendix B).

## 4.1 Accuracy metrics

Metrics that have been reported previously in similar applications were used in order to measure the results of the computational model [85, 100], including the RMSE, RMSPE, Pearson correlation and MAPE. Each of which are described below.

### 4.1.1 Pearson correlation

Also known as bi-variate correlation, it is widely used to measure the linear correlation between two variables. The result obtained in the computational model for each week was paired with the ARI data to measure their correlation as a metric to measure the accuracy of the forecasting model.

### 4.1.2 Root Mean Square Error (RMSE)

RMSE is a measure of the distribution of residuals and reflects how data concentrates around the predicted model. This metric is also used to choose the best fitting network in our proposed model. It is worth noting that the RMSE is not a normalized metric, because it depends on the units of measurement, and therefore it cannot be used to compare with models using a different dataset.

$$\textbf{RMSE} = \sqrt{\frac{1}{n}\sum (y_i - x_i)^2} \tag{4.1}$$

### 4.1.3 Root Mean Squared Percent Error (RMSPE)

RMSPE measures the difference between predicted and real values as a percentage, the advantage against the common RMSE is that the RMSPE is not bound to the units of measurement, meaning that RMSE is not recommended to compare models using different input data, while the RMSPE is always represented as a percentage, allowing comparisons.

$$\textbf{RMSPE} = \sqrt{\frac{1}{n}\sum \left(\frac{y_i - x_i}{y_i}\right)^2} \times 100 \tag{4.2}$$

### 4.1.4 Maximum Absolute Percent Error (MAPE)

MAPE is most commonly used to evaluate forecasts. It has valuable statistical properties. It makes use of all observations and has the smallest variability from sample to sample. MAPE is also often useful for purposes of reporting, because it is expressed in generic percentage terms. It has some setbacks with outliers, but it is still a common metric for forecasting models [92].

$$\textbf{MAPE} = \left(\underset{i}{max}\frac{|y_i - x_i|}{y_i}\right) \times 100 \tag{4.3}$$

Unusually large errors can affect MAPE and RMSPE. However, they share some useful characteristics, since the denominator is the real expected value, the result is not affected by the unit of measurement of the series. This makes the MAPE and RMSPE a good metric for comparing the performance of a forecasting method on different series or the performance of many methods on one series. The RMSE was used to compare the method within the methodology but, as this metric is bound to the unit of measurements, it could not be used to compare with other reported models.

## 4.2 Forecasting Model

The FFNN was tested to assess one-week-in-advance forecasting of the ARI data for the winter seasons encompassed between 2015 and 2019 in a yearly fashion. Where data from 2008 to 2018 was used in order to train the models and 2015 to 2019 to test them, i.e. data from winter seasons between 2008 and 2015 were used as training data to forecast the 2015-2016 winter season (see Tables 4.1 and 4.2).

To evaluate the accuracy of the forecasting model, several experiments (*Exp1* to *Exp6*) were performed. These included assessment of the training window size, and the evaluation of the FFNN with and without retraining of the network every 13 weeks (a quarter of a year, denoted as Q1, Q2, Q3 and Q4 in Figures). This means that for each Experiment two kinds of training will be tested: 1)Training once for the whole year (*Exp No retrain*) and 2)Once a quarter has been evaluated, retrain the network with the new information available (*Exp Retrain*).

Tables 4.1 and 4.2 show the details of the training and testing sets for all experiments to define the best window length. Since the initial weights of the FFNN (parameters) are initialized randomly, each experiment was executed one thousand times creating the same number of networks, then each FFNN was tested with data from the $1^{st}$ week of the starting year to the $26^{th}$ week of ending year. The training concluded when the best FFNN was selected (the one with the lowest RMSE) to be used in the final testing. It is worth noting that the year 2009 was omitted from these tests because of the Influenza A(H1N1) pandemic.

Table 4.1 Training sets. Each test consisted of twelve experiments, six with retraining and six without retraining the network.

| **FFNN** | **Start year (on week one)** | **End year (on week 26)** | **Window (years)** |
|---|---|---|---|
| *Exp1* | {2008,2010,2011,2012} | {2015,2016,2017,2018} | 6.5 |
| *Exp2* | {2010,2011,2012,2013} | {2015,2016,2017,2018} | 5.5 |
| *Exp3* | {2011,2012,2013,2014} | {2015,2016,2017,2018} | 4.5 |
| *Exp4* | {2012,2013,2014,2015} | {2015,2016,2017,2018} | 3.5 |
| *Exp5* | {2013,2014,2015,2016} | {2015,2016,2017,2018} | 2.5 |
| *Exp6* | {2014,2015,2016,2017} | {2015,2016,2017,2018} | 1.5 |

The final testing was divided in quarters, each of them composed of 13 weeks, but the results are grouped by year. On the retrained version of the tests, the network was retrained by adding the most recent data along with part of the previous training data (When retraining a network it is recommended to present some previous data to the network in order to avoid

Table 4.2 Testing sets. Each test consisted of twelve experiments, six with retraining and six without retraining the network.

| FFNN | Start year (on week one) | End year (on week 26) |
|---|---|---|
| *Exp1* | {2015,2016,2017,2018} | {2016,2017,2018,2019} |
| *Exp2* | {2015,2016,2017,2018} | {2016,2017,2018,2019} |
| *Exp3* | {2015,2016,2017,2018} | {2016,2017,2018,2019} |
| *Exp4* | {2015,2016,2017,2018} | {2016,2017,2018,2019} |
| *Exp5* | {2015,2016,2017,2018} | {2016,2017,2018,2019} |
| *Exp6* | {2015,2016,2017,2018} | {2016,2017,2018,2019} |

overfitting[1] with the new presented data). The version without retraining used the same network during the four quarters of each winter season test.

The metrics of the results are shown in Table 4.3, where the best three results for each of them are in bold fonts. An example of the results for the retrained FFNN are shown in Figures 4.1, and without retraining in Figure 4.2, where it is visually clear that the *Exp6 No retrain* consistently keeps close to the ARI data, compared to the other Experiments, the reduced error-based metrics (RMSE, RMSPE and MAPE) shown in Table 4.3 also confirm its accuracy. These results show that *Exp6 No retrain* have consistently more accurate results compared to the other experiments (*Exp1–5*). For this reason the *Exp6 No retrain* is selected as our forecasting model.

## 4.3 Merge Prediction

The merge prediction is composed of the forecasting model (FFNN), the projection model (SoS), and the endemic channels merged in a linear equation that reduces the error-based metrics. Results of this prediction for the 2017–2018 season are shown in Figure 4.3. The results of all seasons involved in the study are shown in Figures 4.4–4.7, were results from the FFNN model, the SoS model, and the smoothed endemic channels can be compared for each season, metrics of each season can be seen in Table 4.4. In these figures it is possible to see the SoS projection constantly following the trend of the ARI data, with no seasonality generating sudden changes; the *Exp6 No retrain*, shows a constant variation going above and below the ARI data, but being good at detecting its sudden increases; the endemic channels (lower and upper) simulate a threshold where the ARI data mostly keeps inside; finally, the merge prediction combines the

[1] See Appendix A.

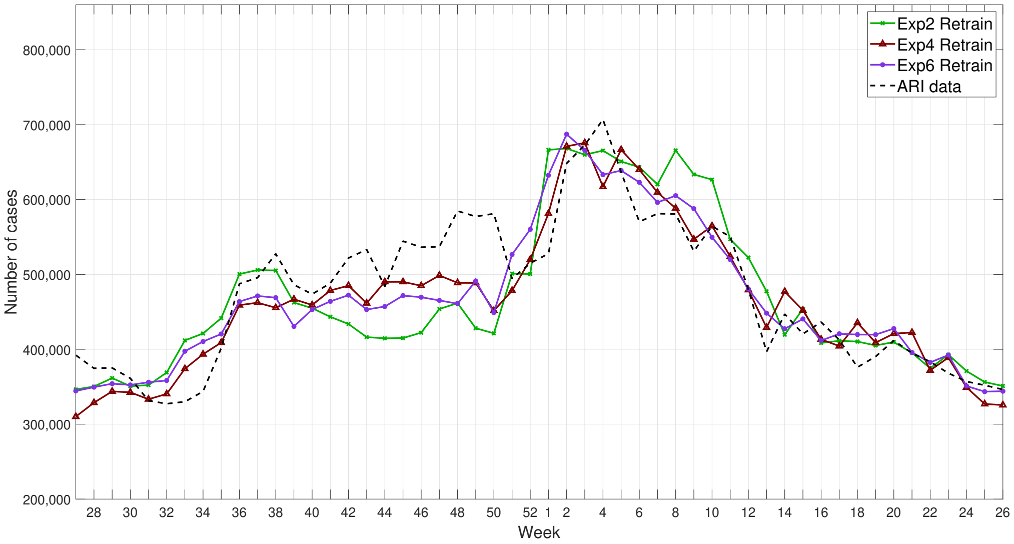


Figure 4.1 Example of results for retrained networks for the 2017-2018 winter season compared with ARI data.

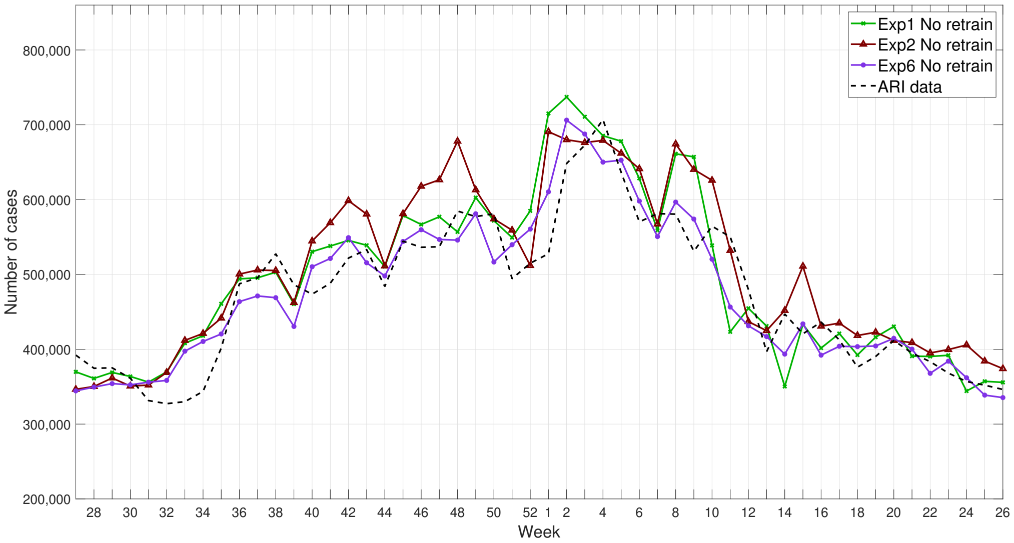


Figure 4.2 Example of results for networks without retraining for the 2017-2018 winter season compared with ARI data.

Table 4.3 Averaged metrics for all experiments with and without retraining. Best values are in bold fonts.

| FFNN | RMSE | CORR | RMSPE | MAPE |
|---|---|---|---|---|
| *Exp1 Retrain* | 86139.27 | 0.72 | 19.76% | 54.09% |
| *Exp2 Retrain* | 89181.04 | 0.72 | 19.35% | 49.98% |
| *Exp3 Retrain* | 86744.66 | 0.70 | 19.46% | 54.73% |
| *Exp4 Retrain* | 80932.07 | 0.77 | 18.60% | 51.83% |
| *Exp5 Retrain* | 80915.79 | 0.77 | 18.56% | 49.91% |
| *Exp6 Retrain* | **74482.54** | 0.79 | **17.42%** | **45.50%** |
| *Exp1 No retrain* | 86875.26 | **0.94** | 20.07% | 52.46% |
| *Exp2 No retrain* | 87912.19 | **0.93** | 19.68% | 50.91% |
| *Exp3 No retrain* | 90353.82 | 0.90 | 20.55% | 55.80% |
| *Exp4 No retrain* | 82922.59 | **0.92** | 19.20% | 47.68% |
| *Exp5 No retrain* | **80360.40** | 0.76 | **18.49%** | **47.43%** |
| *Exp6 No retrain* | **74169.01** | 0.82 | **17.35%** | **42.38%** |

qualities of all the previous signals following closely the ARI data even when it has sudden changes. The results show the capability of the model to predict the ARI data one week in advance.

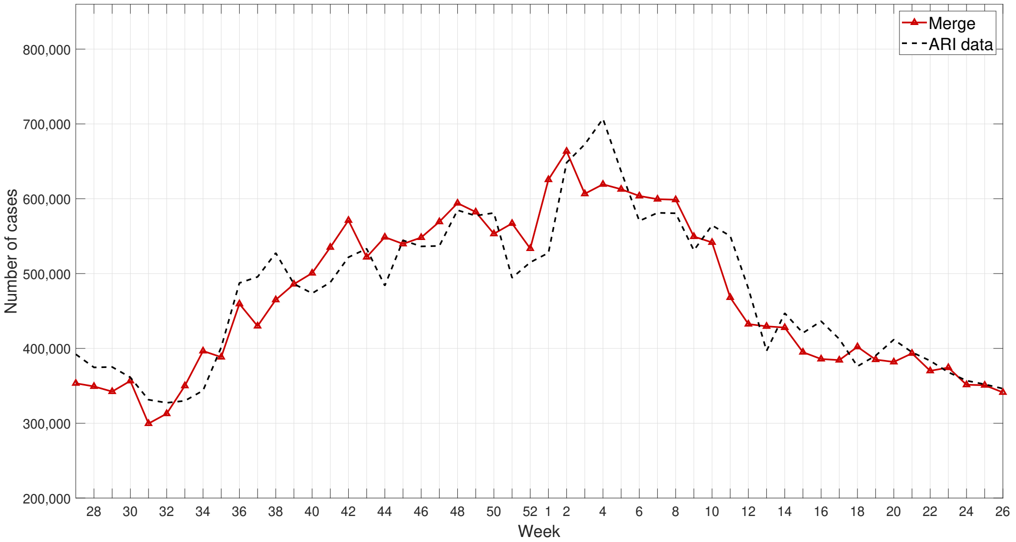


Figure 4.3 Example of results for the merge prediction for the 2017–2018 winter season compared with ARI data.

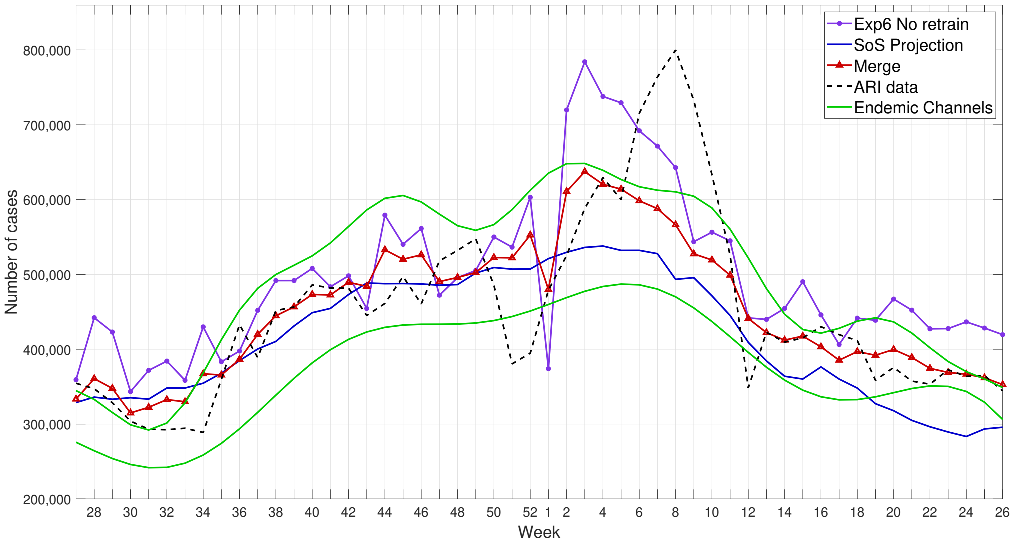


Figure 4.4 Results from the model for the year 2015–2016 compared with the real values for ARI cases.

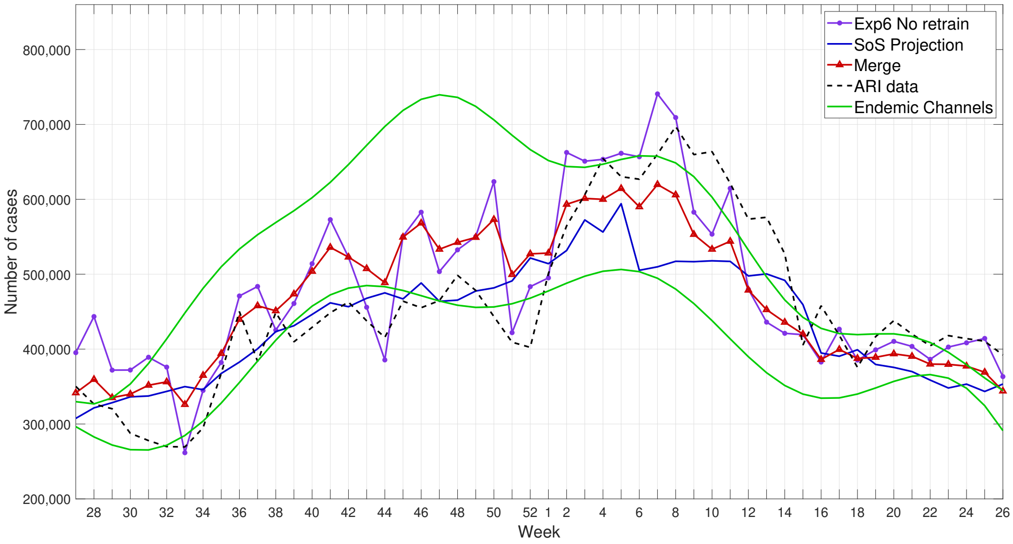


Figure 4.5 Results from the model for the year 2016–2017 compared with the real values for ARI cases.

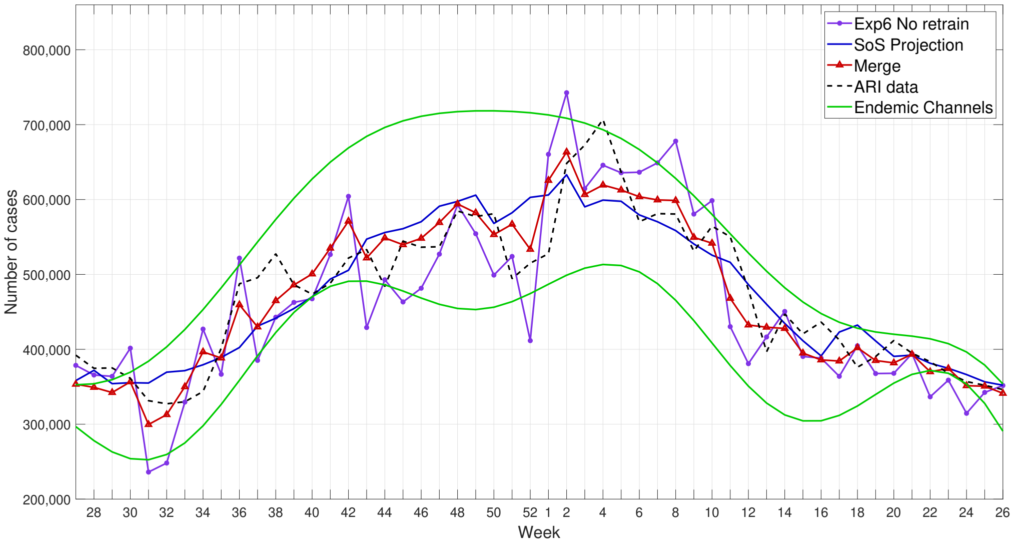


Figure 4.6 Example of all results from the model for the year 2017–2018 compared with the real values for ARI cases.

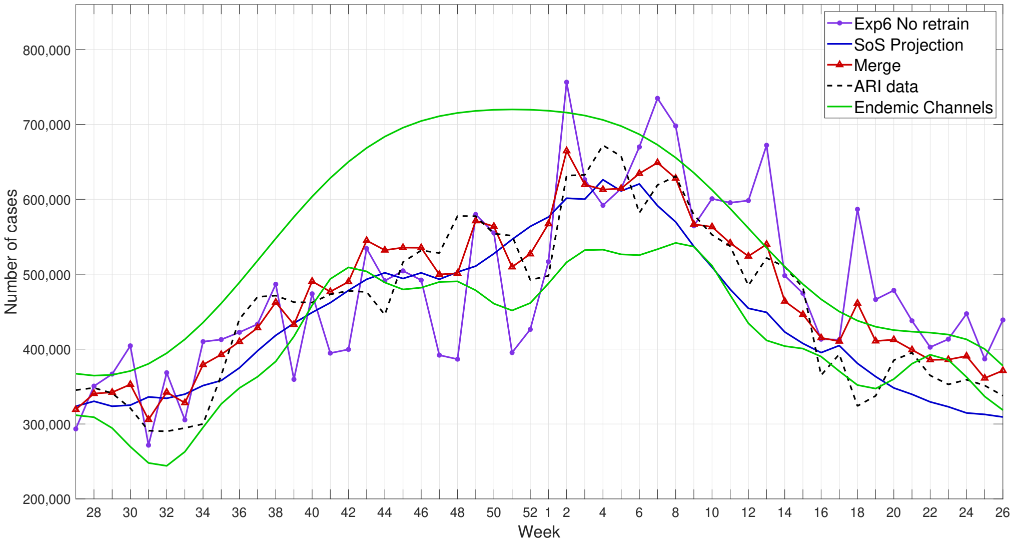


Figure 4.7 Results from the model for the year 2018–2019 compared with the real values for ARI cases.

Table 4.4 Detailed metrics for each season from 2015 to 2016.

| | **RMSE** | **CORR** | **RMSPE** | **MAPE** |
|---|---|---|---|---|
| **2015–2016** | | | | |
| *Exp6 No retrain* | 90666.4436 | 0.78 | 20.1% | 52.0% |
| SoS Projection | 87711.5075 | 0.85 | 12.5% | 31.0% |
| Merge | 70084.2756 | 0.84 | 13.5% | 35.0% |
| **2016–2017** | | | | |
| *Exp6 No retrain* | 69404.4634 | 0.82 | 16.1% | 39.0% |
| SoS Projection | 67787.1978 | 0.84 | 12.1% | 25.0% |
| Merge | 67205.4062 | 0.82 | 12.9% | 30.0% |
| **2017–2018** | | | | |
| *Exp6 No retrain* | 60084.4026 | 0.86 | 11.2% | 28.0% |
| SoS Projection | 42529.7762 | 0.91 | 8.8% | 15.0% |
| Merge | 38142.7486 | 0.93 | 7.8% | 17.0% |
| **2018–2019** | | | | |
| *Exp6 No retrain* | 83400.9898 | 0.75 | 18.4% | 81.0% |
| SoS Projection | 45056.3245 | 0.92 | 10.1% | 16.0% |
| Merge | 42022.3458 | 0.94 | 10.7% | 42.0% |
| **Averaged** | | | | |
| *Exp6 No retrain* | 75889.0749 | 0.80 | 16.4% | 50.0% |
| SoS Projection | 60771.2015 | 0.88 | 10.9% | 21.7% |
| Merge | 54363.694 | 0.88 | 11.2% | 30.9% |

This merged result has a reduced RMSE, RMSPE, and MAPE and a higher correlation coefficient (CORR) compared with FFNN and SoS individually. The average results for these metrics obtained with the three methods during the four study seasons are shown on Table 4.5. This proposal is also compared with similar models found in the state of the art, using data from Google searches [85] and Twitter [85, 100] for Influenza surveillance and ILI forecasting with Machine Learning algorithms, such as, AdaBoost, Support Vector Machines (SVM), with linear and Radial Basis Functions (RBF), and Long Short-Term Memory networks (LSTM), the best one-week forecast results were selected to compare with this methodology, as shown in Table 4.5, where the correlation (CORR) of the SoS projection and Merge prediction falls lower than the SVM, AdaBoost from [85], and the SVM reported by [100], mainly because this metric reflects only the coincidences in the direction were the original data is moving at each moment compared to the predicted data, but it does not measure the distance between the two signals nor the differences in their magnitudes, hence this metric gives limited information of what is going on with the data; RMSPE and MAPE give more information about the accuracy of the prediction. The best RMSPE value is achieved by the SVM from [85] followed by the SoS projection and the Merge prediction, something to keep in mind is that the SoS projection and Merge prediction are tested for 5 years while the SVM from [85] uses only 4 years of testing, meaning that the proposed models is tested for a bigger span of time. The most accurate MAPE metric is achieved by the SoS projection showing a 9% more accurate forecasting than the Merge prediction and 10% better than the SVM reported in [85] and after that it goes close to 20% better than the rest of the forecasting models [85, 100]. The model shows very competitive results which provide an insight of the scope of the proposed model.

Table 4.5 Comparing the merge prediction with other methods. Metrics are calculated averaging consecutive testing seasons.

| **Model** | **Reference** | **CORR** | **RMSPE(%)** | **MAPE(%)** | **Averaged Seasons** |
|---|---|---|---|---|---|
| *Exp6 No retrain* | This model | 0.80 | 16.48% | 50.10% | 2015–2019 |
| SoS Projection | This model | 0.88 | **10.92%** | **21.70%** | 2015–2019 |
| Merge | This model | 0.88 | **11.27%** | **30.90%** | 2015–2019 |
| SVM (RBF) | [85] | **0.89** | 26.90% | 137.00% | 2013–2015 |
| AdaBoost | [85] | **0.92** | 16.10% | 40.90% | 2013–2015 |
| SVM | [100] | **0.90** | **10.43%** | **31.84%** | 2011–2014 |
| AdaBoost | [100] | 0.87 | 13.35% | 43.61% | 2011–2014 |
| LSTM | [100] | 0.61 | 15.46% | 46.67% | 2011–2014 |

# Conclusions and Future Work

Over the last century the availability of vaccines and antibiotics has resulted in a significant reduction of the impact of infectious diseases in the human population worldwide. Nevertheless, infectious diseases continue to cause significant morbidity and mortality. Of special importance, the occurrence of epidemics and pandemics has resulted in major loss of life, health system saturation, and economic burdens. Early identification of outbreaks and epidemics is considered essential in order to limit the spread and effects of an infection within a community or country. Epidemiological surveillance systems are essential tools to identify the onset of outbreaks. Surveillance systems can rely on information that is obtained actively or passively. Active surveillance systems require that epidemiological information be obtained based on case finding activities which may require identifying individuals that fulfill certain case definitions or carrying out laboratory tests to detect a specific pathogen. In contrast, passive surveillance may use routinely gathered information for analysis. As a result, active surveillance systems tend to be more complete but more expensive than passive systems [99]. In both instances, analysis of the information that has been gathered is a key element in order to identify changes in disease occurrence that indicate the onset of an epidemic in comparison to fluctuations that may be considered as normal. Of interest, current computing power allows the analysis of large data sets and with the use of diverse algorithms it is possible to identify signal changes that under traditional epidemiological analyses might be difficult to observe. In addition, the expanding use of the internet has resulted in the potential use of temporal and geographic query patterns for infectious disease surveillance [1]. As a result, over the last decade there has been an increasing interest in the development of internet usage patterns to analyze infectious diseases dynamics and forecast expected behaviors and the occurrence of epidemics [22].

# Conclusions

In the present work we describe a computational model for ARI (Acute Respiratory Infections) surveillance that might allow for early detection of outbreaks. Our model is based on ARI data reported on a weekly basis to the Health Ministry in Mexico as well as the number of internet searches of a set of terms by Mexican Google users. The model has been tested with historical data, and proved to predict the behavior of ARI data for four successive winter seasons (2015–2019); the best MAPE (Maximum Average Percentage Error) results being obtained with the SoS (Sum of Sines) Projection and the merge prediction with 21.7% and 30.9%, respectively. In order to assess these results, we contrasted several metrics (Pearson correlation, RMSPE, and MAPE) with those reported with the use of other methodologies that have analyzed the behavior of respiratory infectious diseases. Unfortunately, there are no previous studies that have focused on ARI which provide similar metrics to assess the accuracy of forecasting, nor encompassing the same geographical and temporal boundaries of our study. Therefore, we included studies that have assessed specific respiratory infections (such as influenza or influenza-like illness) which have shown very good results [100, 85]. The performance of this methodology was competitive in comparison to results reported in those studies with the use of other forecasting methodologies.

The main advantage of the proposed model is the use of data that is readily available, such as Internet search terms and routine disease surveillance data (ARI data) to predict an infectious disease. Some of the previous reports that describe forecasting of respiratory infections (such as influenza infections) rely on samples obtained for virological testing rather than syndromic clinical reports [23], this means that our model does not need laboratory tests but focuses on reported symptoms. While our model could be limited when assessing the behavior of a specific microorganisms (such as influenza), it could allow for the timely identification of outbreaks when the etiological agent is unknown, such as the appearance of unusual cases of pneumonia late in 2019 in Wuhan, China, which were subsequently identified as caused by a novel viral strain (SARS-CoV-2) [118], the only restriction for this model to detect such an outbreak is the models needs to be established in the region of interest. In addition, because our system is based on routinely obtained information and does not require specific laboratory tests, it is expected to allow for surveillance of wide geographical areas, even in regions where laboratory facilities are not available. This could have immediate application in epidemiological surveillance, as a complementary methodology to already established strategies (for example, in the current SARS-CoV-2 pandemic) [105, 106]. This methodology could be adapted for use at a subnational level (such as at regional or state level). Overall, the expected usefulness of ARI analysis using this methodology includes the timely identification of an increase in the number of ill persons. The inclusion of an automatic strategy for search terms removing is also an advantage, since this allows to update the search term list and allows for inclusion of many additional terms, as the number of terms in the initial list does not matter because the model will reduce it to the minimum required for predicting results, eliminating subjectivity; nevertheless, to collect and analyze a long list of new search terms would require additional time. During the first stage of the development of our model, Google Correlate was used to obtain the initial list of potential search terms analyzing their correlation with ARI data; unfortunately, this tool was discontinued at the end of 2019. Nevertheless, potential search terms can be assessed with the use of correlation tests, such as Pearson's correlation. For the feature extraction stage, we assessed several approaches for endemic channel smoothing, including polynomial, splines, Fourier, among others [5]. In addition, several techniques to model the cyclic behavior of time series on ARI data were explored, and found that SoS is simple, and resulted in improved data fitting. Moreover, other techniques were explored to model the cyclic behavior, such as the Holt–Winters method which is used in economy; however, they are more complex and did not improve the model. This work introduces a new methodology for infectious disease forecasting using ARI data and Internet search terms. The results show that the combination of different data analysis techniques (FFNN, SoS, and smoothed endemic channels) can provide an accurate prediction for ARI data one week in advance. It is worth noting that the results obtained by this proposal are published in the International Journal of Environmental Research and Public Health (IJERPH) of the MDPI after being peer reviewed[38].

# Future work

The modular structure of the proposed model enables to change the forecasting or projection model to enhance the results. In addition, a decision-making stage is planned to be added in which the predictions are analyzed to detect and send alerts when a potential outbreak is identified. It is worth noting that the FFNN response forecasting (*Exp6 No retrain*) performed well forecasting peaks in ARI cases, this presents an opportunity of research in how to use this capability while avoiding false positives. A limitation to this observation is that our study included forecasting only for four winter seasons to be certain that this finding is reproducible in all seasons. We expect to enhance the results once there is more quality data available, this is, when the quality limitations found in the Google Trends data before 2010 pose no problem for our predictions as time passes. Appropriate interventions when outbreak signals are identified by this methodology could include targeted laboratory testing, institution of outbreak assessment and control measures, as well as mobilization of health-care supplies (such as medications or vaccines, when available). In addition, it can be adapted to be used in other countries or regions, this may be of particular use in regions where surveillance systems require strengthening. Furthermore, we plan to use this proposal for assessment of other infectious diseases that show seasonal patterns, such as gastrointestinal infections, dengue, and varicella. The final model could be used along with the endemic channels to detect possible outbreaks. The methodology can be expanded to work by regions (or states), to analyze data from other countries, or to assess the behavior of other seasonal infectious diseases. In addition, this proposal could allow the Mexican health authorities to complement traditional surveillance methods for the timely identification of ARI outbreaks. This would be a relevant application of this model, since early detection and response to an outbreak can decrease the costs and the impact associated with it.

# Bibliography

[1] Abat, C., Chaudet, H., Rolain, J.-M., Colson, P., and Raoult, D. (2016). Traditional and syndromic surveillance of infectious diseases and pathogens. *International Journal of Infectious Diseases*, 48:22–28.

[2] Abubakar, I., Tillmann, T., and Banerjee, A. (2015). Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the global burden of disease study 2013. *Lancet*, 385(9963):117–171.

[3] Al-Sakkaf, A. and Jones, G. (2014). Comparison of time series models for predicting campylobacteriosis risk in new zealand. *Zoonoses and public health*, 61(3):167–174.

[4] Ali, S. A., Baloch, M., Ahmed, N., Ali, A. A., and Iqbal, A. (2020). The outbreak of coronavirus disease 2019 (covid-19)—an emerging global health threat. *Journal of infection and public health*.

[5] Alonso, W. J. and McCormick, B. J. (2012). Epipoi: a user-friendly analytical tool for the extraction and visualization of temporal parameters from epidemiological time series. *BMC Public Health*, 12(1):982.

[6] Badurdeen, S., Valladares, D. B., Farrar, J., Gozzer, E., Kroeger, A., Kuswara, N., Ranzinger, S. R., Tinh, H. T., Leite, P., Mahendradhata, Y., et al. (2013). Sharing experiences: towards an evidence based model of dengue surveillance and outbreak response in latin america and asia. *BMC Public Health*, 13(1):607.

[7] Barbazan, P., Yoksan, S., and Gonzalez, J.-P. (2002). Dengue hemorrhagic fever epidemiology in thailand: description and forecasting of epidemics. *Microbes and infection*, 4(7):699–705.

[8] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

[9] Black, R. E., Cousens, S., Johnson, H. L., Lawn, J. E., Rudan, I., Bassani, D. G., Jha, P., Campbell, H., Walker, C. F., Cibulskis, R., et al. (2010). Global, regional, and national causes of child mortality in 2008: a systematic analysis. *The lancet*, 375(9730):1969–1987.

[10] Bortman, M. (2015). Corredores o canales endémicos y su elaboración usando planillas de cálculo.

[11] Bowman, L. R., Tejeda, G. S., Coelho, G. E., Sulaiman, L. H., Gill, B. S., McCall, P. J., Olliaro, P. L., Ranzinger, S. R., Quang, L. C., Ramm, R. S., et al. (2016). Alarm variables for dengue outbreaks: A multi-centre study in asia and latin america. *PLoS One*, 11(6):e0157971.

[12] Brownstein, J. S., Freifeld, C. C., Chan, E. H., Keller, M., Sonricker, A. L., Mekaru, S. R., and Buckeridge, D. L. (2010). Information technology and global surveillance of cases of 2009 h1n1 influenza. *New England Journal of Medicine*, 362(18):1731–1735.

[13] Brownstein, J. S., Freifeld, C. C., and Madoff, L. C. (2009). Digital disease detection—harnessing the web for public health surveillance. *New England Journal of Medicine*, 360(21):2153–2157.

[14] Burkom, H., Elbert, Y., Magruder, S., Najmi, A.-H., Peter, W., and Thompson, M. (2008). Developments in the roles, features, and evaluation of alerting algorithms for disease outbreak monitoring. *Johns Hopkins APL Technical Digest (Applied Physics Laboratory)*, 27.

[15] Butler, D. (2013). When google got flu wrong. *Nature*, 494(7436):155.

[16] Cady, F. (2017). *The Data Science Handbook*. Wiley.

[17] Carneiro, H. A. and Mylonakis, E. (2009a). Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*, 49(10):1557–1564.

[18] Carneiro, H. A. and Mylonakis, E. (2009b). Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*, 49(10):1557–1564.

[19] Cervellin, G., Comelli, I., and Lippi, G. (2017). Is google trends a reliable tool for digital epidemiology? insights from different clinical settings. *Journal of epidemiology and global health*, 7(3):185–189.

[20] Chase, V. (1996). Promed: a global early warning system for disease. *Environmental health perspectives*, 104(7):699.

[21] Cho, S., Sohn, C. H., Jo, M. W., Shin, S.-Y., Lee, J. H., Ryoo, S. M., Kim, W. Y., and Seo, D.-W. (2013). Correlation between national influenza surveillance data and google trends in south korea. *PloS one*, 8(12):e81422.

[22] Choi, J., Cho, Y., Shim, E., and Woo, H. (2016). Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC public health*, 16(1):1238.

[23] Clemente, L., Lu, F., and Santillana, M. (2019). Improved real-time influenza surveillance: Using internet search data in eight latin american countries. *JMIR public health and surveillance*, 5(2):e12214.

[24] Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., Tateno, Y., Ngo, Q.-H., Dien, D., Kawtrakul, A., Takeuchi, K., et al. (2008). Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, 24(24):2940–2941.

[25] Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R. A., Takeuchi, K., and Kawtrakul, A. (2006). A multilingual ontology for infectious disease surveillance: rationale, design and challenges. *Language resources and evaluation*, 40(3-4):405.

[26] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to algorithms*. MIT press.

[27] Deng, L. and Yu, D. (2014). Deep learning: methods and applications. *Foundations and trends in signal processing*, 7(3–4):197–387.

[28] Dugas, A. F., Hsieh, Y.-H., Levin, S. R., Pines, J. M., Mareiniss, D. P., Mohareb, A., Gaydos, C. A., Perl, T. M., and Rothman, R. E. (2012). Google flu trends: correlation with emergency department influenza rates and crowding metrics. *Clinical infectious diseases*, 54(4):463–469.

[29] Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., and Rothman, R. E. (2013). Influenza forecasting with google flu trends. *PloS one*, 8(2):e56176.

[30] Eisenberg, M. C., Eisenberg, J. N., D'Silva, J. P., Wells, E. V., Cherng, S., Kao, Y.-H., and Meza, R. (2015). Forecasting and uncertainty in modeling the 2014-2015 ebola epidemic in west africa. *arXiv preprint arXiv:1501.05555*.

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

[32] Fearnley, C. J. and Dixon, D. (2020). Early warning systems for pandemics: Lessons learned from natural hazards. *International Journal of Disaster Risk Reduction*.

[33] Freifeld, C. C., Mandl, K. D., Reis, B. Y., and Brownstein, J. S. (2008). Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2):150–157.

[34] GBD 2016 Lower Respiratory Infections Collaborators and others (2018). Estimates of the global, regional, and national morbidity, mortality, and aetiologies of lower respiratory infections in 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet. Infectious diseases*, 18(11):1191–210.

[35] George, D. B., Taylor, W., Shaman, J., Rivers, C., Paul, B., O'Toole, T., Johansson, M. A., Hirschman, L., Biggerstaff, M., Asher, J., and Reich, N. G. (2019). Technology to advance infectious disease forecasting for outbreak management. *Nature Communications*, 10(3932):2041–1723.

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

[37] Gluskin, R. T., Johansson, M. A., Santillana, M., and Brownstein, J. S. (2014). Evaluation of internet-based dengue query data: Google dengue trends. *PLoS neglected tropical diseases*, 8(2):e2713.

[38] Gónzalez-Bandala, D. A., Cuevas-Tello, J. C., Noyola, D. E., Comas-García, A., and García-Sepúlveda, C. A. (2020). Computational forecasting methodology for acute respiratory infectious disease dynamics. *International Journal of Environmental Research and Public Health*, 17(12):4540.

[39] Gu, Y., Chen, F., Liu, T., Lv, X., Shao, Z., Lin, H., Liang, C., Zeng, W., Xiao, J., Zhang, Y., et al. (2015). Early detection of an epidemic erythromelalgia outbreak using baidu search data. *Scientific reports*, 5.

[40] Hao, Z., Liu, M., and Ge, X. (2019). Evaluating the impact of health awareness events on google search frequency. *Preventive medicine reports*, 15:100887.

[41] Harcourt, S., Morbey, R., Smith, G., Loveridge, P., Green, H., Pebody, R., Rutter, J., Yeates, F., Stuttard, G., and Elliot, A. (2019). Developing influenza and respiratory syncytial virus activity thresholds for syndromic surveillance in england. *Epidemiology & Infection*, 147.

[42] Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. Prentice Hall.

[43] Herman Tolentino, M., Raoul Kamadjeu, M., Michael Matters PhD, M., Marjorie Pollack, M., and Larry Madoff, M. (2007). Scanning the emerging infectious diseases horizon-visualizing promed emails using epispider. *Advances in disease surveillance*, 2:169.

[44] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[45] Hussain-Alkhateeb, L., Kroeger, A., Olliaro, P., Rocklöv, J., Sewe, M. O., Tejeda, G., Benitez, D., Gill, B., Hakim, S. L., Carvalho, R. G., et al. (2018). Early warning and response system (ewars) for dengue outbreaks: Recent advancements towards widespread applications in critical settings. *PloS one*, 13(5):e0196811.

[46] Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

[47] Jia-xing, B., Bcn-fu, L., Geng, P., and Na, L. (2013). Gonorrhea incidence forecasting research based on baidu search data. In *Management Science and Engineering (ICMSE), 2013 International Conference on*, pages 36–42. IEEE.

[48] Kang, M., Zhong, H., He, J., Rutherford, S., and Yang, F. (2013). Using google trends for influenza surveillance in south china. *PloS one*, 8(1):e55205.

[49] Kawazoe, A., Chanlekha, H., Shigematsu, M., and Collier, N. (2008). Structuring an event ontology for disease outbreak detection. *BMC bioinformatics*, 9 Suppl 3:S8.

[50] Keller, M., Blench, M., Tolentino, H., Freifeld, C. C., Mandl, K. D., Mawudeku, A., Eysenbach, G., and Brownstein, J. S. (2009). Use of unstructured event-based reports for global infectious disease surveillance. *Emerging infectious diseases*, 15(5):689.

[51] Khaliq, A., Batool, S. A., and Chaudhry, M. N. (2015). Seasonality and trend analysis of tuberculosis in lahore, pakistan from 2006 to 2013. *Journal of epidemiology and global health*, 5(4):397–403.

[52] Kwaku, A. B., Tan, H., Luo, K., Li, Q., Hu, S., Wu, X., and Zhou, Y. (2017). Spatio-temporal clustering analysis and its determinants of hand, foot and mouth disease in hunan, china, 2009–2015. *BMC Infectious Diseases*, 17(1):645.

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

[54] Legido-Quigley, H., Asgari, N., Teo, Y. Y., Leung, G. M., Oshitani, H., Fukuda, K., Cook, A. R., Hsu, L. Y., Shibuya, K., and Heymann, D. (2020). Are high-performing health systems resilient against the covid-19 epidemic? *The Lancet*, 395(10227):848–850.

[55] Liu, K., Wang, T., Yang, Z., Huang, X., Milinovich, G. J., Lu, Y., Jing, Q., Xia, Y., Zhao, Z., Yang, Y., et al. (2016). Using baidu search index to predict dengue outbreak in china. *Scientific reports*, 6:38040.

[56] Mackenzie, J. S., Drury, P., Arthur, R. R., Ryan, M. J., Grein, T., Slattery, R., Suri, S., Domingo, C. T., and Bejtullahu, A. (2014). The global outbreak alert and response network. *Global public health*, 9(9):1023–1039.

[57] Martin, L. (2017). A look back: investigating google flu trends during the influenza a (h1n1) pdm09 pandemic in canada, 2009–2010. *Epidemiology & Infection*, 145(3):420–423.

[58] Mathes, R. W., Lall, R., Levin-Rector, A., Sell, J., Paladini, M., Konty, K. J., Olson, D., and Weiss, D. (2017). Evaluating and implementing temporal, spatial, and spatio-temporal methods for outbreak detection in a local syndromic surveillance system. *PloS one*, 12(9):e0184419.

[59] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[60] Morse, S. S. (2012). Public health surveillance and infectious disease detection. *Biosecurity and bioterrorism: biodefense strategy, practice, and science*, 10(1):6–16.

[61] Mykhalovskiy, E. and Weir, L. (2006). The global public health intelligence network and early warning outbreak detection: a canadian contribution to global public health. *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique*, pages 42–44.

[62] Nair, H., Brooks, W. A., Katz, M., Roca, A., Berkley, J. A., Madhi, S. A., Simmerman, J. M., Gordon, A., Sato, M., Howie, S., et al. (2011). Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. *The Lancet*, 378(9807):1917–1930.

[63] Nelson, N., Brownstein, J., Hartley, D., et al. (2010). Event-based biosurveillance of respiratory disease in mexico, 2007-2009: connection to the 2009 influenza a (h1n1) pandemic. *Euro Surveill*, 15(30):19626.

[64] Nsoesie, E. O., Beckman, R. J., Shashaani, S., Nagaraj, K. S., and Marathe, M. V. (2013). A simulation optimization approach to epidemic forecasting. *PloS one*, 8(6):e67164.

[65] Nuti, S. V., Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R. P., Chen, S. I., and Murugiah, K. (2014). The use of google trends in health care research: a systematic review. *PloS one*, 9(10):e109583.

[66] O'Neil, C. and Schutt, R. (2013). *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.".

[67] Palma Méndez, J. T. and Morales, R. M. (2008). Inteligencia artificial. técnicas, métodos y aplicaciones. *Mc Graw*.

[68] Pavia, A. T. (2011). Viral infections of the lower respiratory tract: old viruses, new viruses, and the role of diagnosis. *Clinical Infectious Diseases*, 52(suppl_4):S284–S289.

[69] Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A., and Valleron, A.-J. (2009a). More diseases tracked by using google trends. *Emerging infectious diseases*, 15(8):1327.

[70] Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A., and Valleron, A.-J. (2009b). More diseases tracked by using google trends. *Emerging infectious diseases*, 15(8):1327–8.

[71] Permanasari, A. E. et al. (2009a). Prediction of zoonosis incidence in human using seasonal auto regressive integrated moving average (sarima). *arXiv preprint arXiv:0910.0820*.

[72] Permanasari, A. E., Rambli, D. R. A., and Dominic, D. D. (2009b). Prediction of zoonosis incidence in human using seasonal auto regressive integrated moving average (sarima). *arXiv preprint arXiv:0910.0820*.

[73] Permanasari, A. E., Rambli, D. R. A., and Dominic, P. D. D. (2011). Performance of univariate forecasting on seasonal diseases: the case of tuberculosis. In *Software Tools and Algorithms for Biological Systems*, pages 171–179. Springer.

[74] Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., and Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448.

[75] Procházka, B. and Kynxcl, J. (2015). Estimating the baseline and threshold for the incidence of diseases with seasonal and long-term trends. *Central European journal of public health*, 23(4):352.

[76] Rehn, M., Carnahan, A., Merk, H., Kühlmann-Berenzon, S., Galanis, I., Linde, A., and Nyrén, O. (2014). Evaluation of an internet-based monitoring system for influenza-like illness in sweden. *PLoS One*, 9(5):e96740.

[77] Rigau-Perez, J. G., Millard, P. S., Walker, D. R., Deseda, C. C., and Casta-Velez, A. (1999). A deviation bar chart for detecting dengue outbreaks in puerto rico. *American journal of public health*, 89(3):374–378.

[78] Rodan, A. and Tino, P. (2010). Minimum complexity echo state network. *IEEE transactions on neural networks*, 22(1):131–144.

[79] Rosenblatt, F., Minsky, M., and Papert, S. (1958). Perceptrons: An introduction to computational geometry. *Psychological Review*, 65:386.

[80] Rudan, I., Boschi-Pinto, C., Biloglav, Z., Mulholland, K., and Campbell, H. (2008). Epidemiology and etiology of childhood pneumonia. *Bulletin of the world health organization*, 86:408–416B.

[81] Ruiz-Matus, C., Kuri-Morales, P., and Narro Robles, J. (2017). Comportamiento de la temporadas de influenza en méxico de 2010 a 2016, análisis y prospectiva. *Gac Med Mex*, 2(153):205–13.

[82] Rumelhart, D. E., Hinton, G. E., and Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, 323(1):533–536.

[83] Saker, L., Lee, K., Cannito, B., Gilmore, A., and Campbell-Lendrum, D. (2004). *Globalization and infectious diseases, A review of the linkages*. UNDP/World Bank/WHO Special Programme on Tropical Diseases Research.

[84] Samaras, L., García-Barriocanal, E., and Sicilia, M.-A. (2017). Syndromic surveillance models using web data: the case of influenza in greece and italy using google trends. *JMIR public health and surveillance*, 3(4):e90.

[85] Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., and Brownstein, J. S. (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*, 11(10):e1004513.

[86] Santillana, M., Zhang, D. W., Althouse, B. M., and Ayers, J. W. (2014). What can digital disease detection learn from (an external revision to) google flu trends? *American journal of preventive medicine*, 47(3):341–347.

[87] Schuck-Paim, C., Viboud, C., Simonsen, L., Miller, M. A., Moura, F. E., Fernandes, R. M., Carvalho, M. L., and Alonso, W. J. (2012). Were equatorial regions less affected by the 2009 influenza pandemic? the brazilian experience. *PloS one*, 7(8).

[88] Secretaría de Salud México (2018). Boletín epidemiológico. sistema nacional de vigilancia epidemiológica. *Sistema Único de Información*, 34(1).

[89] Seifter, A., Schwarzwalder, A., Geis, K., and Aucott, J. (2010). The utility of “google trends” for epidemiological research: Lyme disease as an example. *Geospatial health*, pages 135–137.

[90] Seo, D.-W. and Shin, S.-Y. (2017). Methods using social media and search queries to predict infectious disease outbreaks. *Healthcare informatics research*, 23(4):343–348.

[91] Stephens-Davidowitz, S. and Varian, H. (2014). A hands-on guide to google data. *Tech. Rep.*

[92] Swanson, D. A., Tayman, J., and Bryan, T. (2011). Mape-r: a rescaled measure of accuracy for cross-sectional subnational population forecasts. *Journal of Population Research*, 28(2-3):225–243.

[93] Sánchez-Ramos, E. L., Monárrez-Espino, J., and Noyola, D. E. (2017). Impact of vaccination on influenza mortality in children< 5 years old in mexico. *Vaccine*, 35(9):1287–1292.

[94] Teng, Y., Bi, D., Xie, G., Jin, Y., Huang, Y., Lin, B., An, X., Feng, D., and Tong, Y. (2017). Dynamic forecasting of zika epidemics using google trends. *PloS one*, 12(1):e0165085.

[95] The Executive Committee of the Infectious Diseases Society of America Emerging Infections Network (1997). The emerging infections network: A new venture for the infectious diseases society of america. *Clinical Infectious Diseases*, pages 34–36.

[96] Valdivia, A. and Monge, S. (2010). Diseases tracked by using google trends, spain. *Emerging infectious diseases*, 16:168.

[97] Vega, T., Lozano, J. E., Meerhoff, T., Snacken, R., Mott, J., Ortiz de Lejarazu, R., and Nunes, B. (2013). Influenza surveillance in europe: establishing epidemic thresholds by the moving epidemic method. *Influenza and other respiratory viruses*, 7(4):546–558.

[98] Vizcarra-Ugalde, S. et al. (2016). Intensive care unit admission and death rates of infants admitted with respiratory syncytial virus lower respiratory tract infection in mexico. *The Pediatric infectious disease journal*, 35(11):1199–1203.

[99] Vogt, R. L., LaRue, D., Klaucke, D. N., and Jillson, D. A. (1983). Comparison of an active and passive surveillance system of primary care providers for hepatitis, measles, rubella, and salmonellosis in vermont. *American Journal of Public Health*, 73(7):795–797.

[100] Volkova, S., Ayton, E., Porterfield, K., and Corley, C. D. (2017). Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS one*, 12(12):e0188941.

[101] Whitsitt, J., Karimkhani, C., Boyers, L. N., Lott, J. P., and Dellavalle, R. P. (2015). Comparing burden of dermatologic disease to search interest on google trends. *Dermatology online journal*, 21(1).

[102] Wilder, J. W. (1978). *New concepts in technical trading systems*. Trend Research.

[103] Wong-Chew, R. M., García-León, M. L., Noyola, D. E., Gonzalez, L. F. P., Meza, J. G., Vilaseñor-Sierra, A., Martinez-Aguilar, G., Rivera-Nuñez, V. H., Newton-Sánchez, O. A., Firo-Reyes, V., et al. (2017). Respiratory viruses detected in mexican children younger than 5 years old with community-acquired pneumonia: a national multicenter study. *International Journal of Infectious Diseases*, 62:32–38.

[104] World Health Organization (2008). The global burden of disease: 2004 update. *Bulletin of the world health organization*.

[105] World Health Organization et al. (2020a). Global surveillance for covid-19 caused by human infection with covid-19 virus: interim guidance, 20 march 2020. Technical report, World Health Organization.

[106] World Health Organization et al. (2020b). Surveillance strategies for covid-19 human infection: interim guidance, 10 may 2020. Technical report, World Health Organization.

[107] World Health Organization: Ebola Response Team (2014). Ebola virus disease in west africa—the first 9 months of the epidemic and forward projections. *New England Journal of Medicine*, 371(16):1481–1495.

[108] Wu, Y.-C., Chen, C.-S., and Chan, Y.-J. (2020). The outbreak of covid-19: An overview. *Journal of the Chinese Medical Association*, 83(3):217.

[109] Xu, Q., Gel, Y. R., Ramirez, L. L. R., Nezafati, K., Zhang, Q., and Tsui, K.-L. (2017). Forecasting influenza in hong kong with google search queries and statistical model fusion. *PloS one*, 12(5):e0176690.

[110] Yang, S., Kou, S. C., Lu, F., Brownstein, J. S., Brooke, N., and Santillana, M. (2017). Advances in using internet searches to track dengue. *PLoS computational biology*, 13(7):e1005607.

[111] Yang, S., Santillana, M., and Kou, S. C. (2015). Accurate estimation of influenza epidemics using google search data via argo. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478.

[112] Yu, V. L. and Madoff, L. C. (2004). Promed-mail: an early warning system for emerging diseases. *Clinical infectious diseases*, 39(2):227–232.

[113] Yuan, Q., Nsoesie, E. O., Lv, B., Peng, G., Chunara, R., and Brownstein, J. S. (2013). Monitoring influenza epidemics in china with search query from baidu. *PloS one*, 8(5):e64323.

[114] Zeng, D., Chen, H., Tseng, C., Larson, C., Eidson, M., Gotham, I., Lynch, C., and Ascher, M. (2004). Sharing and visualizing infectious disease datasets using the wnv-bot portal system. In *Proceedings of the 2004 annual national conference on Digital government research*, page 115. Digital Government Society of North America.

[115] Zhang, X., Liu, Y., Yang, M., Zhang, T., Young, A. A., and Li, X. (2013). Comparative study of four time series methods in forecasting typhoid fever incidence in china. *PloS one*, 8(5):e63116.

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

[117] Zhang, Y., Yakob, L., Bonsall, M. B., and Hu, W. (2019). Predicting seasonal influenza epidemics using cross-hemisphere influenza surveillance data and local internet query data. *Scientific reports*, 9(1):3262.

[118] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., et al. (2020). A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine*.

# Appendix A

# Forecasting infectious diseases: Related topics and concepts

## A.1 Data analysis

Forecasting is about predicting the future as accurately as possible, given all the known and available information, including historical data and knowledge of any future events that might impact the forecasts[46]. This appendix focuses in describing topics, terms and concepts used throughout this document.

### A.1.1 Time series

Time series is a type of panel data, a multidimensional data set, whereas a time series data set is a one-dimensional panel, the panel data is the general class. As a definition a time series is a collection of observations of well-defined data items obtained through repeated measurements over time[1]. These measurements could have been gathered at a regular time intervals (such as the ARI data from this research, measured in weeks), or at irregular times. Time series can be decomposed into three components, each expressing a particular aspect of the movement of the values of the time series[2]. The components are:

[1] https://www.abs.gov.au/websitedbs/d3310114.nsf/home/time+series+analysis:+the+basics

[2] https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-32833-1_401

- Trend-cycle. A general systematic linear or (most often) nonlinear component that changes over time and does not repeat[3].
- Seasonality. A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality is always of a fixed and known frequency[46].
- Irregular variations. A non-systematic component that is nor Trend/Seasonality within the data, also known as the remainder.

### Decomposition

Refers to separating a time series into trend, seasonal effects, and remaining variability. When there is a need to extract any of these components from a time series, one possible solution is to use a method known as data smoothing. Smoothing data removes random variation and shows trends and cyclic components[4], it is even possible to extract the irregular variations when removing the smoothed signal from the original time series (see Figure A.1).

### Smoothing a time series

As seen in Figure 2.7, the trend signal would be a smoothed version of the original data. The smoothing of a signal implies the removing of the seasonal and irregular components of the data and keeping only the trend [46]. Generally smooth out the irregular roughness to see a clearer signal. For seasonal data, we might smooth out the seasonality so that we can identify the trend[5].

### Cross-sectional data

When data of one or more variables is collected at the same point in time, it is known as Cross-sectional data [6], in this research the Internet search terms data is collected as a Cross-sectional data.

[3] https://towardsdatascience.com/trend-seasonality-moving-average-auto-regressive-model-my-journey-to-time-series-data-with-edc4c0c8284b

[4] https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc42.htm

[5] https://online.stat.psu.edu/stat510/lesson/5/5.2

[6] https://www.statisticssolutions.com/time-series-analysis/

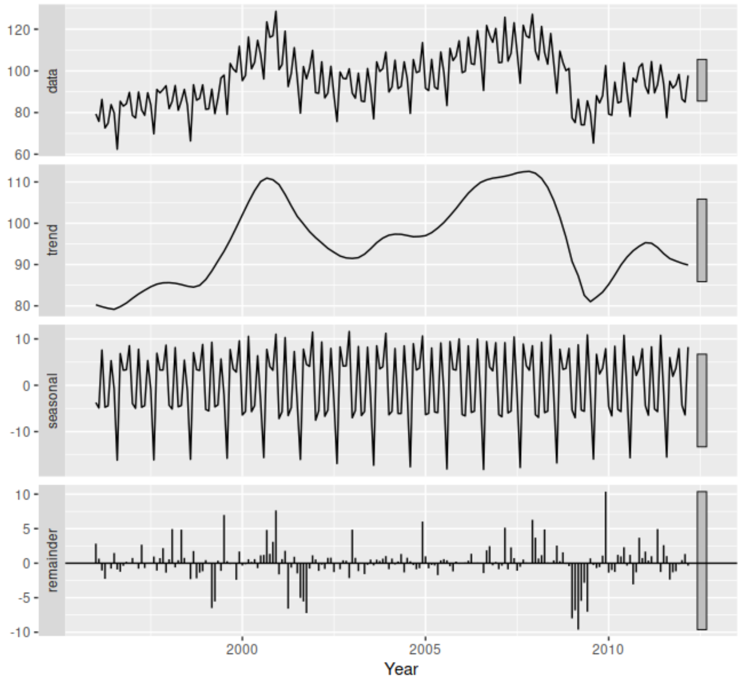


Figure A.1 An example of a decomposed time series. On top, the original data, below the extracted trend of the original data, the seasonal component, and at the bottom the remainder of the signal decomposition[46]. The addition of the three components result in the original data.

### Curve fitting in time series analysis

Curve fitting regression is used when data is in a non-linear relationship. The sum of sines function, used in this research to calculate the smoothed endemic channels and as the projection model, shows this non-linear behavior.

### Overfitting

Overfitting occurs when a statistical model or machine learning algorithm captures the noise of the data as part of the modeled behavior. In other words, overfitting occurs when the model or the algorithm fits the training data too well, so that it does not perform well on previously unpresented data.

### Data preprocessing

Data, when initially obtained, must be processed or organized for analysis. For instance, these may involve placing data into rows and columns in a table format (known as structured data) for further analysis[66].

### Data cleaning

Once organized, the data may be incomplete, contain duplicates or errors. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, identifying inaccuracy of data, overall quality of existing data, deduplication, and column segmentation[7]. There are several types of data cleaning, that are dependent upon the type of data in the set, quantitative data methods for outlier detection, can be used to get rid of data that appears to have a higher likelihood of being input incorrectly. In these research the methodology proposed dealt with missing values and empty values in the search terms data, and in formatting the ARI data.

### Disease outbreak

According to the World Health Organization[8] (WHO), a disease outbreak is the occurrence of disease cases in excess of normal expectancy. The number of cases varies according to the

[7] https://www.microsoft.com/en-us/research/project/data-cleaning/

[8] https://www.who.int/environmental$_{h}ealth_{e}mergencies/disease_{o}utbreaks/en/$

disease-causing agent, and the size and type of previous and existing exposure to the infectious agent.

### A.1.2 Endemic Channel

Endemic channels are used to identify the presence of outbreaks, by the observation of a larger number of cases in excess of what would be expected from historical time series[9]. The can be defined as a moving threshold used to survey the amount of cases within an expected normal range, where anything above this threshold could point to a developing outbreak [6]. The EC are calculated using the historic behavior from 5 or more years of an infectious disease (normally avoiding years affected by an outbreak), it is a simple and fast way to generate somewhat reliable thresholds allowing to detect a presence/absence of an outbreak. When using endemic channels there is a risk of not detecting new behaviors in the data, and limitations associated with abnormally high historic means and the variation in the seasonal timing that often lead to inaccurate detections[11].

Many countries use a version of the endemic channels for visualising the expected case levels, based on the weekly (or monthly) average number of cases over the preceding 5 years [6]. The endemic channels are used to identify the presence of outbreaks, by the observation of a larger number of cases in excess of what would be expected from historical time series[9]. As such, endemic channels represent the boundaries for the expected number of cases at a given time; the occurrence of cases above this moving threshold would be considered as an outbreak of a disease [45].

Commonly, an endemic channel is obtained by calculating the $5^{th}$, $50^{th}$ and $85^{th}$ percentiles of a disease for each week from the 5 previous years (ignoring years affected by outbreaks); the $5^{th}$ percentile (lower endemic channel) represents the minimum number of cases expected or the area of success, the $50^{th}$ percentile is the expected behavior (the median or secure area), and the $85^{th}$ percentile (upper endemic channel) is the risk area or area of alert, surpassing these values points to possible outbreaks[10].

There are a set of steps to calculate the $k^{th}$ percentile values (where $0 < k < 100$) for a time series:

1. For each week of every year we are interested in.
2. Create an array with the historic values for that week in the past five years ($a = (a_1, a_2, ..., a_n)$ where $n = 5$), ignoring values from years affected by an outbreak.

[9] http://www.searo.who.int/topics/disease_outbreaks/en/

3. Order the values in the array from smallest to largest.

4. To get the index of the $k^{th}$ percentile ($i_k$), multiply $k$ by $n$, the total of numbers in the array, $i_k = (kn)$ and if:

    - The number is not an integer, it needs to be rounded to the closest integer, $i_k = round(kn)$, then the $k^{th}$ percentile $= a_{i_k}$ for that week.
    - The number is an integer, then the $k^{th}$ percentile is the average of the elements $i_k$ and $i_k + 1$ from the array. $k^{th}$ percentile $= \overline{a_{i_k} a_{i_k+1}}$ for that week.

Some reported disadvantages on using endemic channels are: a) Interpreting the crossing of area of alert line as a "warning sign" of an outbreak (rather than as an indicator that an outbreak is effectively already underway) and initiate a delayed emergency response; b) According to Rigau-Perez et al. [77], the risk area has limited sensitivity (only 40% of such events when case numbers crossed this threshold were followed by a "massive" increase of cases in their research; using a similar predictor, Barbazan et al. [7] found a sensitivity of 66%) c) Outbreaks in previous years can result in thresholds that are too high; d) The seasonal increase in cases may come earlier than in the 5 preceding years providing the impression of an outbreak.

## A.2 Artificial neural networks

Since the first artificial neural network (ANN) models emerged [59], they have been the target of rigorous scientific scrutiny [79], due to their properties, constraints and their need to be trained with some data set, where, under some circumstances, considerable information or training time is required. Nevertheless, ANN have shown their high capabilities in a great number of everyday problems. ANN has been widely used for pattern recognition, function approximation, prediction/forecast, optimization among other applications [42, 8].

### A.2.1 Feedforward neural network

A feedforward neural network is a biologically inspired classification algorithm. It consists of a number of simple neuron-like processing units, organized in layers[10]. Every unit in a layer is connected with all the units in the previous layer. These connections are not all equal: each connection may have a different strength or weight and a bias. The weights on these connections

[10] https://www.fon.hum.uva.nl/praat/manual/Feedforward_neural_networks_1__What_is_a_feedforward_ne.html

encode the knowledge of a network, often the units in a neural network are also called nodes or neurons. The first layer has a connection from the network input. Each subsequent layer has a connection from the previous layer. The final layer produces the output. Feedforward networks can be used for any kind of input to output mapping. A feedforward network with one hidden layer and enough neurons in the hidden layers, can fit any finite input-output mapping problem. The neurons are arranged in layers, with the first layer taking in inputs and the last layer producing outputs. The middle layers have no connection with the external world, and hence are called hidden layers. Each neuron in one layer is connected to every neuron on the next layer. Hence information is constantly "fed forward" from one layer to the next, this explains why these networks are called feedforward networks. There is no connection among neurons in the same layer.

### Training algorithm

In order to work properly, the weights and biases of each connection between two neurons need to be adjusted in FFNN, this could be a massive task given that any slight change in one of these weights modifies the behavior of the connected neurons and, as a consequence, of the whole network and its outputs. The process of adjusting the weights of each connection within a network is known as the learning phase (or training algorithm). Depending on the problem to resolve the network could be used for classification or regression, in this research we are using the network to predict the behavior of a time series (ARI data) this means we are using it to resolve a regression problem. The FFNNs use a supervised learning algorithm, meaning that, to be trained, the network needs the inputs and the corresponding correct (or expected) values for the outputs. There are several training algorithms, but for this case the Backpropagation training algorithm is used to effectively train a neural network through a method called chain rule. In simple terms, after each forward pass through a network, backpropagation performs a backward pass while adjusting the model's parameters (weights and biases).

An example of a FFNN with four inputs, two hidden layers, and one output is shown in Figure A.2. The inputs can be seen as simple scalars or complex structures as vectors or matrices, the equation A.1 shows the behavior of the input neurons, where the set of activations ($a$) is equal to the input values.

$$\mathbf{x_i} = a_i^{(1)}, i \in 1,2,3,4 \tag{A.1}$$

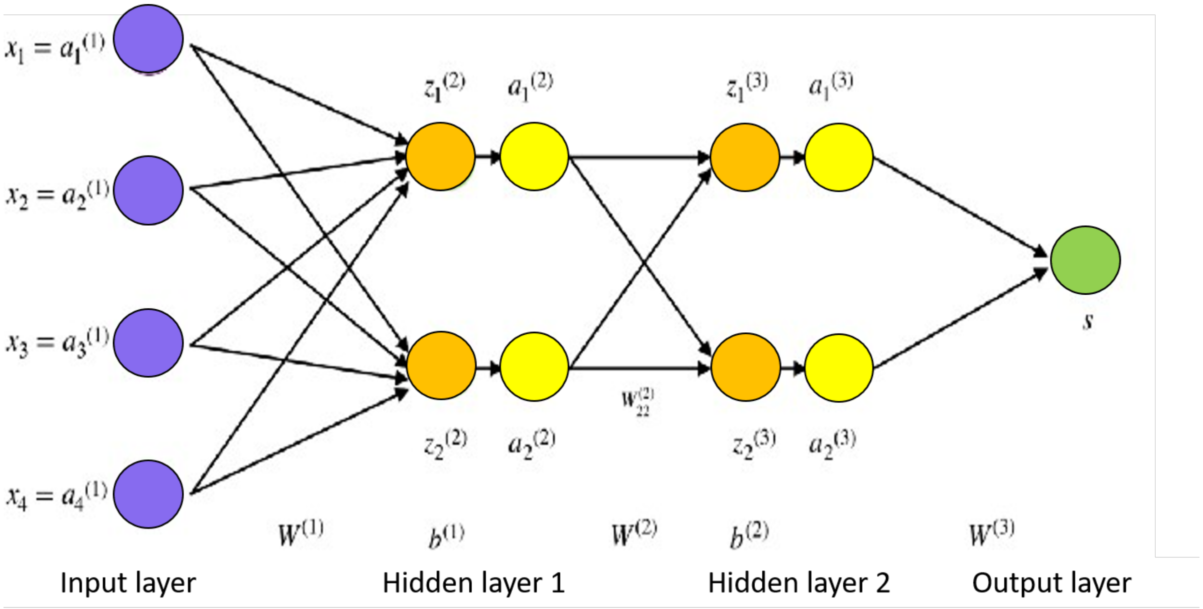


Figure A.2 An example of an FFNN architecture. Composed by an input layer (purple nodes), two hidden layers (green nodes) and one input (blue). The arrows between nodes represent the weights.

The final values at the hidden neurons are computed using $z^l$ (the weighted inputs in layer $l$), and $a^l$ (activations in layer $l$)( see Equations A.2 and A.3. $W^2$ and $W^3$ are the weights in layers 2 and 3 while $b^2$ and $b^3$ are the biases in those layers. Activations $a^2$ and $a^3$ are computed using an activation function $f$. Function $f$ normally is a non-linear function, and allows the network to learn complex patterns in data$^{\mathbf{??}}$.

- **l=2**

$$\begin{aligned} \mathbf{z}^{(\mathbf{2})} &= W^{(1)}x + b^{(1)} \\ \mathbf{a}^{(\mathbf{2})} &= f(z^{(2)}) \end{aligned} \tag{A.2}$$

- **l=3**

$$\begin{aligned} \mathbf{z}^{(\mathbf{3})} &= W^{(2)}a^{(2)} + b^{(2)} \\ \mathbf{a}^{(\mathbf{3})} &= f(z^{(3)}) \end{aligned} \tag{A.3}$$

The final part of the neural network is the output layer which produces the response of the network and it is supposed to be the expected value, in Figure A.2 it is represented as a single neuron and it is calculated as shown in Equation A.4

$$\mathbf{s} = W^{(3)}a^{(3)} \tag{A.4}$$

The last step in a forward pass is to evaluate the response of the output $s$, comparing against the expected value output $y$. This output $y$ is part of the training set $(x,y)$ where $x$ is the input. This evaluation between $s$ and $y$ is done using a cost function (MSE, RMSE, AIC, Cross-entropy)(See Equation A.5.

$$\mathbf{C} = cost(s,y) \tag{A.5}$$

Based on the resulting $C$, the model can adjust its parameters in order to get closer to the expected output $y$. This is when the backpropagation algorithm starts.

The backpropagation algorithm aims to minimize the cost function by adjusting the weights and biases of the network. The level of adjustment is determined by the gradients of the cost function with respect to those parameters. The gradient calculation shows how much parameter x needs to change to minimize $C$.

The gradient used for a single weight $(w_{jk})^l$ is shown in Equation A.6, similarly the gradient for the bias $b$ is shown in Equation A.7.

$$\frac{\partial \mathbf{C}}{\partial \mathbf{w^l_{jk}}} = \frac{\partial C}{\partial z^l_j} a^{l-1}_k \tag{A.6}$$

$$\frac{\partial \mathbf{C}}{\partial \mathbf{b^l_j}} = \frac{\partial C}{\partial z^l_j} 1 \tag{A.7}$$

The gradients provide information to optimize the weights and bias of the network, by using the gradient descent algorithm, which backpropagates the adjustments through the network.

$while\ (termination\ condition\ not\ met)$

$$\mathbf{w} := w - \varepsilon \tfrac{\partial C}{\partial w}$$

$$\mathbf{b} := b - \varepsilon \tfrac{\partial C}{\partial b} \tag{A.8}$$

$end$

Where:

- Initial values of $w$ and $b$ are randomly generated.

- Epsilon ($\varepsilon$) is the learning rate. It determines the gradient's influence.
- $w$ and $b$ are matrix representations of the weights and biases.
- Derivative of $C$ in $w$ or $b$ can be calculated using partial derivatives of $C$ in the individual weights or biases.
- Termination condition is met once the cost function is minimized.

This description tries to give a general idea of the composition of FFNNs, their internal parts and interactions, for a deeper understanding other sources should be reviewed.

# Appendix B

# Discarded Methods and Metrics

## B.1 Smoothing techniques

### B.1.1 Moving averages (MA)

Widely used as smoothing techniques of signals and time series, the moving averages are known as trend-following or lagging indicators because they are based on past values of the variables. There are several variations for this technique: Simple moving average, exponential moving average, triangular moving average, weighted moving average and modified moving average. Mostly, the difference relies in the weight coefficients given to most recent values. MAs can predict or estimate the immediate next value from a series.

#### Simple moving averages (SMA)

This is the simplest form of the moving averages technique, and it works by calculating the average of the $n$ last points of the series. These $n$ last values are known as *sliding window of size n*, the value of $n$ can be selected to enhance the fitting of the SMA. Let $S_t : t = 1, \ldots m$, so

$$\mathbf{SMA}(S,n) = \sum_{i=1}^{n} \frac{S_{m-i}}{n} \tag{B.1}$$

where $(t \geq n)$ and $(n \leq m)$.

The main disadvantage with the simple moving average is that the previous values used for the average are considered equally weighted, giving the same weight to the oldest used value and the most recent used value. It tends to behave better with small window sizes.

### Weighted moving averages (WMA)

This variation, uses the same idea of the simple MA but adds a vector of weights of the same size than the window used, meaning that a weight will be applied to each value of the window depending on its position (See Equation B.2), normally, most recent values should have bigger weight than older values in the window. The sum of weights should add up to 1. By defining $S_t : t = 1, \dots, m$ and $W_i : i = 1, \dots, n$, we have:

$$\mathbf{WMA}(S,W) = \sum_{i=1}^{n} \frac{S_{m-i} * W_{n-i}}{n} \tag{B.2}$$

where $(t \geq n)$ , $\sum_{i=1}^{n} W_i = 1$ , $W_i \geq 0$ , $(n < m)$ and $(n > 0)$

There are two parameters to adjust in this method, the window size $n$ and the array of weights $W$, depending on the problem to be solved they should be properly adjusted. By adding weights to the averages there is more control over the influence of older and newer values for the estimation.

### Triangular moving averages(TMA)

The TMA is a double-smoothed SMA. It is an average of data calculated over a window of time, where the central values have more weight (See Equation B.3). it is commonly used with prices, but can also be used in any time series. TMA smoothes data series, which is useful in very irregular data sets. This method will react to fluctuations in the time series even slower than the SMA, this could be problematic if the intention is to quickly adapt to data changes. It is beneficial in specific cases, if the data moves back and forth in a range, the TMA will not respond immediately to the variations, thus giving a more stable signal, that points to a not changed trend. It would take a more sustained move in data to cause TMA to change directions. Let $\{S_t : t = 1, \dots, m\}$ and $l = \lceil \frac{n+1}{2} \rceil$, then:

$$\mathbf{TMA}(S,l) = SMA(SMA(S,l),l) \tag{B.3}$$

where $(t \geq n)$ and $(n < m)$ and $(n > 0)$

### Exponential moving averages (EMA)

The EMA is similar to the SMA and the WMA, it uses weights to accent most recent data, the difference is that it automatically adjusts the smoothing constant, EMA is able to adjust its sensitivity, allowing it to adjust better in more variating series. Nevertheless, as EMA responds so quickly to changes in the series it could point to a new trend forming in the series when in reality there is just a spike in data, and could be affected significantly by outliers in most recent values. Let $\{S_t : t = 1, \ldots, m$ and $k = \frac{2}{n+1}$, then

$$\mathbf{EMA_t}(S,n) = kS_t + ((1-k)\mathbf{EMA_{t-1}}(S,n)) \tag{B.4}$$

where $(t > n)$ and $(n < m)$ and $(n > 0)$

As it can be seen in the Equation B.4 the previous EMA is needed to calculate the new one, in this case, we need to calculate the first EMA (See Equation B.5 ), the most used way to calculate the first value is to make use of the SMA, as shown below:

$$\mathbf{EMA_t}(S,n) = SMA_t(S,n) \tag{B.5}$$

where $(t = n)$

### Modified moving averages (MMA)

Also known as Running Moving Average (RMA) or Smoothed Moving Average (SMMA), it is another variant of the SMA, where the new values are calculated by adding the new price, and then subtracting the last average from the resulting sum, the difference is the new point. As the previous value of the MMA is needed to estimate the most recent, the first value is calculated using the SMA, just as in the EMA. As $S_t : t = 1, \ldots, m$ and $k = \frac{2}{n+1}$

$$\mathbf{MMA_t}(S,n) = \mathbf{MMA_{t-1}}(S,n) + \frac{S_t - \mathbf{MMA_{t-1}}(S,n)}{n} \tag{B.6}$$

where $(t > n)$ , $(n < m)$ and $(n > 0)$

## B.1.2 Exponentially Weighted Moving Averages

Exponentially Weighted Moving Averages (EWMA), also known as Exponential Smoothing (ES), is a common way to produce smoothed time series. ES focuses in exponentially decrease the weight of observations as they get older. There are different variations of ES where various smoothing parameters must be estimated to better fit the smoothing of the time series. ES is commonly used in the analysis of financial time series or signal processing.

### Single exponential smoothing (SES)

The simplest case of exponential smoothing. This method is better when using data with no seasonal or trend pattern. The idea is to give greater weights to more recent observations than older ones. The forecasts are calculated using weighted averages that decrease exponentially as observations come from further in the past. The difference between some MAs that also use weights and the ES is that the latter uses exponential weights in their calculations. The calculations for the SES are shown below: let $S_t : t = 1, \ldots, m$, so

$$\mathbf{SES_t}(S,\alpha) = \alpha\mathbf{SES_t}(S,\alpha) + (1-\alpha)\mathbf{SES_{t-1}}(S,\alpha) \tag{B.7}$$

where $(0 < \alpha \leq 1)$
The $\alpha$ value is a parameter to be set, and as shown in Equation B.7 its value must be between 0 and 1. $\alpha$ is known as the smoothing constant. Single exponential smoothing, by definition, has problems to follow data when a trend is present.

### Double exponential smoothing (DES)

To improve the behavior of the SES, a second constant is added. The constant $\beta$ is the second constant, which should be chosen in conjunction with $\alpha$. The use of these two constants and their two equations is known as the Double exponential smoothing or Exponential trend method. $\beta$ is known as the level constant. These two equations are to calculate trend (or slope) and level. The slope is the well known algebraic slope $m = \frac{\Delta y}{\Delta x}$. The level is the same value that was used in the SES equation, but this time it is just part of the whole calculation, as shown below

$$\begin{aligned}
\ell_\mathbf{t} &= \alpha S_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) && level \\
\mathbf{b_t} &= \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1} && trend \\
\mathbf{DES_{t+1}}(S,\alpha,\beta) &= \ell_t + b_t && forecast
\end{aligned} \tag{B.8}$$

where $(0 < \alpha \leq 1)$ and $(0 < \beta \leq 1)$ As with simple exponential smoothing, the level equation in Equation B.8 shows that $\ell_t$ is a weighted average of observation $y_t$ and the within-sample one-step-ahead forecast for time $t$, here given by $(\ell_{t-1} + b_{t-1})$.The trend equation shows that $b_t$ is a weighted average of the estimated trend at time $t$ based on $(\ell_t - \ell_{t-1})$ and $b_{t-1}$, the previous estimate of the trend.

The Equation B.8 can estimate not only the immediate next value but $h$ steps ahead, as shown in Equation B.9.

$$\mathbf{DES_{t+h}}(S,\alpha,\beta) = \ell_t + hb_t \tag{B.9}$$

Equations B.8 and B.9 show the additive method of the DES, the multiplicative is very similar, but uses multiplication instead of addition for the calculations, as shown below

$$\begin{aligned}
&\ell_{\mathbf{t}} = \alpha S_t + (1-\alpha)(\ell_{\mathbf{t-1}}\mathbf{b_{t-1}}) && level \\
&\mathbf{b_t} = \beta\frac{\ell_t}{\ell_{t-1}} + (1-\beta)b_{t-1} && trend \\
&\mathbf{DES_{t+1}}(S,\alpha,\beta) = \ell_t b_t && forecast
\end{aligned} \tag{B.10}$$

where $(0 < \alpha \leq 1)$ and $(0 < \beta \leq 1)$

**Triple exponential smoothing (Holt-Winters seasonal method)**

Between 1957 and 1960, Holt and Winters enhanced the DES method to capture seasonality. The Holt-Winters seasonal method comprises the forecast equation and three smoothing equations; one for the level $\ell_t$, one for trend $b_t$, and one for the seasonal component denoted by $\varsigma$, with smoothing parameters $\alpha$, $\beta$ and $\gamma$.

$L$ will denote the period of the seasonality, the number of seasons in a year, for example. For example, for quarterly data $L = 4$, and for monthly data $L = 12$. The seasonal component ($\varsigma$) is an additional deviation from $level + trend$ that repeats itself at the same offset time into the season (See Equation). There is a seasonal component for every point in a season, if $L = 12$, there are 12 seasonal components.

As mentioned in the previous methods, there is an additive and a multiplicative version of the method. Both variations are shown below.

**Holt-Winters additive:**

$$\begin{aligned}
&\ell_{\mathbf{t}} = \alpha(S_t - \varsigma_{t-L}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) && level \\
&\mathbf{b_t} = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1} && trend \\
&\varsigma_{\mathbf{t}} = \gamma(S_t - \ell_t) + (1-\gamma)\varsigma_{t-L} && seasonal \\
&\mathbf{DES_{t+h}}(S,\alpha,\beta,\gamma) = \ell_t + hb_t + \varsigma_{t-L+1+(m-1)\%L} && forecast
\end{aligned} \tag{B.11}$$

where $(0 < \alpha \leq 1)$ , $(0 < \beta \leq 1)$ and $(0 \leq \gamma \leq 1)$

**Holt-Winters multiplicative:**

$$
\begin{aligned}
\ell_{\mathbf{t}} &= \alpha\frac{S_t}{\varsigma_{t-L}} + (1-\alpha)(\ell_{t-1}+b_{t-1}) && \textit{level}\\
\mathbf{b_t} &= \beta(\ell_t-\ell_{t-1}) + (1-\beta)b_{t-1} && \textit{trend}\\
\varsigma_{\mathbf{t}} &= \gamma\frac{S_t}{\ell_{t-1}+b_{t-1}} + (1-\gamma)\varsigma_{t-L} && \textit{seasonal}\\
\mathbf{DES_{t+h}}(S,\alpha,\beta,\gamma) &= (\ell_t+hb_t)\varsigma_{t-L+1+(m-1)\%L} && \textit{forecast} \qquad \text{(B.12)}
\end{aligned}
$$

where $(0 < \alpha \leq 1)$ , $(0 < \beta \leq 1)$ and $(0 \leq \gamma \leq 1)$

The multiplicative method is recommended for situations where the time series shows a trend, a tendency to repeat its behavior increasing its values by some factor. Whereas, the additive version is suitable for time series with no detectable trend, as the ARI data. This method was tested in several scenarios and even gave good results but it was surpassed by the sum of sines smoothing technique. An example of the results obtained with this method are shown in Figure B.1.

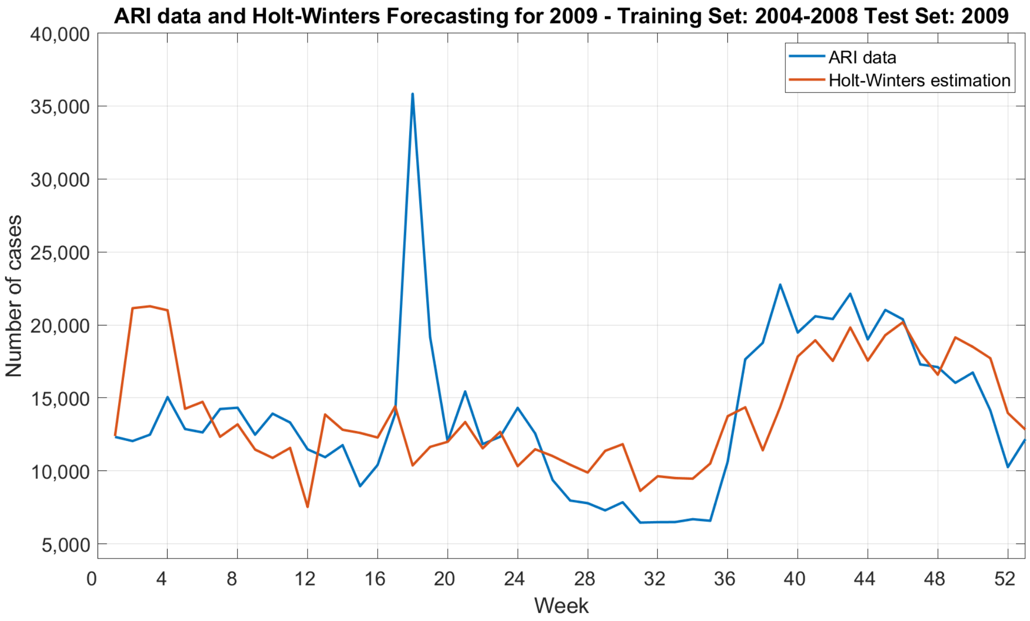


Figure B.1 Example of the test made using the Holt-Winters smoothing

## B.2 Stock market technique: Directional Movement Indicator and Average Directional Index (DMI-ADX)

This method was designed originally for commodities and daily prices; it can also be applied to stocks and to identify trends in the financial market. It helps in the decision making for financial investments minimizing risks by identifying trends.

Proposed by Wilder [102], the Directional Movement is a system that identifies if the market is trending before providing signals for trading the trend. It is composed by three variables the ADX, +DI and -DI, these values are calculated from the time series to be analyzed. The Plus Directional Indicator (+DI) and Minus Directional Indicator (-DI) are derived from smoothed averages of these differences, and measure trend direction over time. Together, they are known as the Direct Movement Indicator (DMI).

A DMI crossover generates the bullish (tendency to increase) and bearish (tendency to decrease) signals. When the DMI+ crosses above the DMI-, a bullish signal is identified. When the DMI- crosses below the DMI+ a bearish signal is identified. The ADX does not identify the direction of a trend by itself; it only identifies the degree of strength within a trending market. Initially, DMI assumes that the positive trend (+DI) is basically defined by the maximum price, when the actual maximum price is greater than the previous one, and the negative trend (-DI) is calculated by the minimum price. The average that predicts a positive or negative trend is defined by normalizing all the average values, higher and lower of the actual measurement (True Range, also developed by Wilder)(see EquationB.13). Three price values are required for each period measured, a maximum value of the period, a minimum value and a close value of the period.

$$\begin{aligned} +\mathbf{DI} &= \tfrac{+DM_n}{TR_n} & -\mathbf{DI} &= \tfrac{+DM_n}{TR_n} \\ +\mathbf{DM} &= H_t - H_{t-1} & -\mathbf{DM} &= L_t - L_{t-1} \\ \mathbf{CL} &= C_t - C_{t-1} & \mathbf{TR} &= T_h - T_l \end{aligned} \tag{B.13}$$

Where:

$+\mathbf{DI}$ = Positive DI

$-\mathbf{DI}$ = Negative DI

$+\mathbf{DM_n}$ = Current updated mobile average of +DM

$-\mathbf{DM_n}$ = Current updated mobile average of -DM

$+\mathbf{DM}$ = Current value of the +DI

$\mathbf{H_t}$ = Current maximum value

$\mathbf{H_{t-1}}$ = Previous maximum value

$\mathbf{TRn}$ = Current modified mobile average of TR

$\mathbf{TR}$ = True Range

$\mathbf{T_h}$ = Highest price compared to the maximum current value or the previous value

$\mathbf{T_l}$ = Lowest current value or previous value

The Average Directional Index (ADX) is derived from the smoothed averages of the difference between +DI and -DI, and measures the strength of the trend over time, regardless of the direction of the trend. It provides an additional prediction about the intensity by evaluating the normalized difference between +DI and -DI. It is obtained by calculating the average of m days for +DI and -DI (Equation B.14). ADX values range from 0 to 100 where 0 would mean there is no trend in data, and 100 would be an extremely strong trend (See Table B.1).

$$\mathbf{ADX} = 100 \times \frac{(+DI - -DI)}{(+DI + -DI)} \tag{B.14}$$

The ADX, +DI and -DI calculations involve a lot of smoothing, consequently it takes around 150 periods of data to get true ADX values, one of the techniques used is the exponential moving average, the calculation starts with the sum of the first m periods Together the DMI and ADX can provide the direction an strength of a trend.

Some issues when using this technique are that it is based on Moving Averages (MA), this means that it reacts slowly to low moving series, it is known as a lagging indicator. Some crossover between DMI indicators can happen too often, giving false indications. It is strongly recommended to use ADX when other reliable indicators are supporting them.

The algorithm has been enhanced to adjust itself to the ranges of the input data, meaning that the algorithms needs to define what range of values are low, medium and high to work properly and be sensitive to the input data. After preparing the algorithm, and after some tests and tuning the algorithm with the ARI data, it showed to be a good predictor of the increasing and decreasing tendencies of the ARI totals series, and particularly it became a good predictor of start of the end of the infection season (see Figure B.2).

Table B.1 Definition of trend strengths by ADX range of values in finances

| ADX value | Trend Strength |
|---|---|
| 0 - 25 | Absent or weak trend |
| 25 - 50 | Strong trend |
| 50 - 75 | Very strong trend |
| 75 - 100 | Extremely strong trend |

## B.3 Accuracy Metrics

### B.3.1 Spearman Correlation

In statistics, the Spearman Correlation (Rank-Order Correlation), is a non-parametric measure for ranked monotonic correlation between two variables, different to the Pearson Correlation which assesses linear relationships. Basically it consists in calculating the Pearson Correlation to the ranking of the variables. This correlation was considered because of the less restrictive nature of the monotonicity, and was discarded after finding the values of the correlation did not reflected the behavior of the data as well as the Pearson Correlation.

### B.3.2 Mean Absolute Error (MAE)

MAE is a measure of errors of two observations expressing the same phenomenon, including predicted versus observed. It is the arithmetic average of the absolute errors. It is scale dependant and can not be used to compare series with different scales. This metric does not pose any enhancement to other similar metrics, its scale dependency makes it difficult to use it to compare with other solutions proposed in the state of the art.

### B.3.3 Mean Square Error (MSE)

The MSE of an estimator is a very common statistical measure of the quality of an estimator. By definition, it is always non-negative, and closer to zero values are desired. Compared to the Root Mean Squared Error (RMSE) the MSE has several disadvantages, i.e. it does not make difference between small or big errors, while the RMSE gives higher weights to bigger errors. The MSE was replaced by the RMSE as it has many other desirable properties.

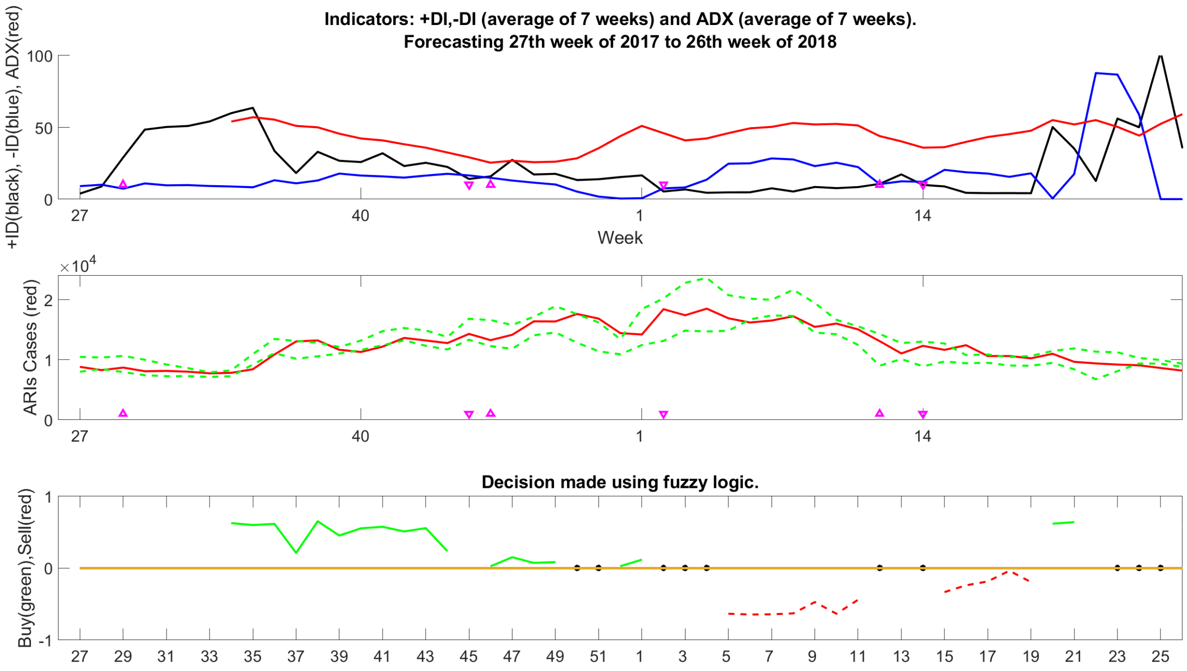


Figure B.2 Use of the DMI-ADX indicators on the ARI data, for 2017 and 2018.

### B.3.4 Normalized Mean Square Error (NMSE)

Used as a function to determinate the goodness of fit of an estimator, the NMSE is the MSE normalized by signal power. After several tests this metric was replaced by the RMSPE to facilitate result comparisons with other methods.

### B.3.5 Coefficient of determination ($R^2$)

$R^2$ is defined as the proportion of the variance in the dependent variable that is predictable from the independent variables. Used commonly in statistical models prediction of future outcomes. It gives a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.This metric should not be used to validate a model but rather after the model is validated and to compare it with another validated model.

### B.3.6 Aikaike Information Criterion (AIC)

AIC is used to measure quality of statistical models for a given dataset. AIC provides a means for model selection by estimating the relative amount of information lost by a given model: the less information lost the higher quality of the model. This metric was used in many test performed but it was visually clear that the best results of the AIC where not better than the other metrics used in this research.