University of Torino

Doctoral School in Life and Health Sciences

PhD Programme in Complex Systems for Life Sciences

Digital Epidemiology: Using Novel Data Streams for Infectious Disease Surveillance, Modelling and Forecasting

---

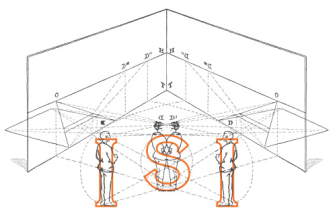
Author:
Daniela Perrotta

Supervisor:
Daniela Paolotti

Co-supervisor:
Michele Tizzoni

Tutor:
Michele Caselle

February 2018



"Begin at the beginning", the King said gravely,
"and go on till you come to the end: then stop."

Lewis Carroll, *Alice in Wonderland*

Abstract

In recent years, traditional epidemiology has been exposed to a new digital revolution driven by the advent of modern communication technologies and the pervasive use of digital devices, that has radically transformed the way people communicate and search for information in real-time through the Web. In this context, non-traditional approaches have emerged in the use of such digital traces generated by human activities on social media, crowdsourced platforms and Internet in general, to monitor global health by exploiting the immediate availability of these data to improve timeliness, and spatial and temporal resolution, thus providing health authorities with an additional and potentially scalable layer of information that may guide the public health decision-making process. In this dissertation, I will present my contributions in the use of such novel digital data streams for monitoring, modelling and predicting the epidemic spreading of infectious diseases. First, I will focus on seasonal influenza that annually affects millions of people around the world with a severity that can vary substantially from year to year. In Italy, a participatory surveillance system called Influweb, part of the European network called Influenzanet, aims at monitoring seasonal influenza activity in a cohort of individuals who self-report their health status through Internet-based surveys, thus providing an additional layer of influenza-related information directly gathered from the general population. The result is a large amount of crowdsourced digital data that can be rapidly analysed for a variety of purposes, including monitoring disease trends, identifying risk factors, but also to complement traditional surveillance data and give early detection of local outbreaks by providing local and timely estimates of the levels of influenza circulating among the population. In this dissertation, I will investigate the representativeness of the participants involved in the data collection and I will evaluate the goodness of the detected epidemiological signal as compared to the traditional surveillance data based on general practitioners' reports. Furthermore, I will show how these different data sources for monitoring influenza epidemics in Italy can be combined by means of simple autoregressive models in order to improve seasonal influenza forecasts by leveraging on the digital component of Influweb having earlier data available. Indeed, real-time forecasts of major influenza indicators, such as peak time and peak intensity, can provide key information for public health interventions, such as resources allocation for influenza prevention and control. Here, I will describe a computational framework for real-time forecast of seasonal influenza based on Influenzanet data, traditional surveillance reports and a dynamical mechanistic model called GLEAM (GLobal Epidemic And Mobility model). Real-time forecasts of seasonal influenza are publicly available at www.fluoutlook.org for several countries and several influenza seasons.

Differently from well-known diseases like seasonal influenza, the emergence of new infectious diseases is continuous, variable and remarkably difficult to predict, such as the Zika outbreak that recently hit the Americas between 2015 and 2016. In this dissertation, I will focus on the epidemic of Zika occurred in Colombia, mainly addressing the role of human mobility in the spreading of the disease by investigating different mobility networks of human movements across the country generated by mobile phone data, mobility models and census data in a metapopulation modelling approach that allows to simulate the spatial and temporal evolution of the epidemic spreading of the disease, accounting for detailed population data and human mobility pattern.

Acknowledgments

I am pleased to acknowledge the role of several people who have been instrumental for the completion of my doctoral research.

First of all, I would like to express my sincere gratitude to my supervisor Daniela Paolotti for the continued support of my PhD study and research, for her patience, motivation and constant encouragement, and for giving me countless opportunities to be exposed to the community of this scientific field, to strengthen my knowledge and increase my experience. In the same way, I would like to thank my co-supervisor Michele Tizzoni for his enthusiasm and willingness to clarify doubts and give suggestions. I am profoundly grateful for the opportunity to do my doctoral programme under their guidance and to learn from their research expertise. They are an extraordinary model for me showing the diligence and dedication to scientific research.

I would like to thank my tutor Michele Caselle for following me during these three years and for being always available to take care of and solve the bureaucratic aspects that allowed me to do my doctoral programme at the ISI Foundation. I also thank all the members of the doctoral school and PhD programme in Complex Systems for Life Sciences for the organization of seminars and events, and always creating the opportunity to interact and learn from other students and researchers.

A special thanks to Alessandro Vespignani who has always pushed me to give my best and has always made himself available despite his busy schedules. I have greatly benefited from his immense knowledge and comprehensive suggestions. I could not have imagined having a better mentor for my PhD study.

I would like to thank all the collaborators with whom I had the chance to work and collaborate at Northeastern University in Boston, with a special thanks to Qian Zhang for her advices, helps and friendship.

I would like to thank the reviewers of my dissertation, Ana Pastore and Stefano Merler, for the time they spent in reading and reviewing my thesis and for their useful comments and suggestions. I am very grateful to have had the opportunity to learn from and interact with them.

I wish to thank all the people from past and present who have passed through the ISI Foundation in these years. I have had the opportunity to meet wonderful people, make friends and share the everyday life in the office and the daily tears and pains of being a PhD student.
I also thank all the administrative staff who helped me solve problems, organize missions and keep everything smooth.

Thanks to Vittoria Colizza and Chiara Poletto for their always kind hospitality at the EPIcx Lab, INSERM, in Paris. A special thanks to Francesco Pinotti, Alexandre Darbon and Caroline Guerrisi for their friendship and for cheering up my stay at the lab.

Thanks to the team of UN Global Pulse in New York for the great opportunity to work in such a prestigious environment, for giving me the chance to be exposed to the world of the United Nations, but more importantly, for making me feel welcomed and accepted like in a big family. I would like to thank all the people with whom I had the chance to work, but also all the friends I had the good fortune to meet. It has been a wonderful experience that radically changed my life.

I am grateful to all the people, colleagues, researchers and friends I had the good fortune to meet at conferences, meetings and schools, for the great memories of the time spent together.

Thanks to many friends, lost, old and new, for joining me in this journey, for their sincere help and encouragements in my way towards what I have gained today.

Thanks to my biggest fan for always being interested and attentive to my work, for believing in me and in my abilities, always present and proud of me as an indispensable source of love, encouragement, support and patience.

Thanks to mum, dad, Davide and Arianna for being my family and for your almost unbelievable support in every moment of my life. You are the most important people in my world and I dedicate this thesis to you.

I am deeply grateful for all the experiences of these years, the wonderful opportunities I had, the people I met and the places I visited. Life changes constantly and puts you to the test many times, bringing you to different directions, towards new paths and incredible adventures. This journey is now over and I am excited about what will come next. The best is yet to come.

Preface

This dissertation is submitted as a requirement for the degree of Doctor of Philosophy in the Doctoral School in Life and Health Sciences at the University of Torino. The research presented here was conducted under the supervision of Dr. Daniela Paolotti and carried out primarily at the Institute for Scientific Interchange (ISI Foundation), in Torino, Italy, between November 2014 and November 2017, plus a three-month internship at United Nations Global Pulse in New York, USA, from June to September 2017. In particular, this dissertation is the result of several works I have carried out in the field of modern computational and digital epidemiology, mainly focused on the use of novel digital data streams for monitoring, modelling and predicting the epidemic spreading of infectious diseases.

Chapter 1 is dedicated to the literature review and description of the state-of-the-art of this scientific field, giving insights on the progress in the study of infectious diseases, from the disease surveillance techniques to the epidemic modelling approaches and the prediction of epidemics.

In Chapter 2, I will focus on an innovative online tool for monitoring seasonal influenza epidemics in Italy as compared to the traditional surveillance system based on general practitioners' reports. This chapter is the result of an ongoing collaboration with the Italian Institute of Public Health (ISS) and is based on the following publication:

- **Daniela Perrotta**, Antonino Bella, Caterina Rizzo, Daniela Paolotti *Participatory Online Surveillance as a Supplementary Tool to Sentinel Doctors for Influenza-Like Illness Surveillance in Italy.* PLoS ONE, *2017*

  In particular, I personally carried out the work presented in this paper, including collecting, processing and analysing the various datasets, and validating and visualising the results, as well as contributing in writing the manuscript.

In Chapter 3, I will focus on the real-time forecasting of seasonal influenza activity by using different forecasting techniques and integrating different data sources, particularly highlighting the benefits of leveraging on novel digital data sources to capture an additional layer of real-time and geo-localized signal. This chapter is based on the following publications:

- John S. Brownstein, Shuyu Chu, Achla Marathe, Madhav V. Marathe, Andre T. Nguyen, Daniela Paolotti, Nicola Perra, **Daniela Perrotta**, Mauricio Santillana, Samarth Swarup, Michele Tizzoni, Alessandro Vespignani, Anil Kumar S. Vullikanti, Mandy L. Wilson, Qian Zhang *Combining Participatory Influenza Surveillance with Modeling and Forecasting: Three Alternative Approaches.* JMIR Public Health and Surveillance, *2017*

  This paper aims at investigating three different participatory disease surveillance systems in the use of modelling, simulation and forecasting. Here I will report only our original contribution based on a computational framework that includes real-time influenza-related data, traditional surveillance reports and a dynamical model for spatial epidemic spreading able to provide long-term predictions of seasonal influenza activity. This methodology

has been previously validated in Ref. [204, 205] of which I am co-author, as a result of an ongoing collaboration with Qian Zhang, Nicola Perra and Alessandro Vespignani at Northeastern University. In this work I personally contributed by exploring the calibration of the model with a different data source, running the simulations, and analysing, validating and visualising the results, as well as contributing in writing the manuscript.

- **Daniela Perrotta**, Michele Tizzoni, Daniela Paolotti *Using Participatory Web-based Surveillance Data to Improve Seasonal Influenza Forecasting in Italy.* Proceeding of the 26th International Conference on World Wide Web (WWW), *2017*

  This paper aims at investigating how traditional surveillance data reported by general practitioners can be combined with digital surveillance data from a participatory Web-based system in order to improve seasonal influenza forecasts in Italy. In particular, I personally carried out the work presented here, including collecting, processing and analysing the two datasets, building and testing the various forecasting models, and analysing and validating the results, as well as writing the manuscript.

In Chapter 4, I will focus on the Zika outbreak occurred in Colombia in 2015-2016, mainly studying the role of human mobility in a metapopulation modelling approach based on real data on population and a detailed description of the epidemiological characteristics of the disease. This work is the result of an ongoing collaboration with Miguel-Luengo Oroz at the United Nations Global Pulse in New York, USA. In this study, I personally conducted the research by performing each step of the work presented here, including the collection, analysis and visualization of the various datasets, as well as building and testing the mobility networks, writing the code of the epidemic model and performing the analysis of the simulations.

In addition, over the course of my PhD program I had the opportunity to collaborate to other projects that are not presented in this dissertation and are listed below:

- Carl Koppeschaar, Vittoria Colizza, Caroline Guerrisi, Clément Turbelin, Jim Duggan, W. John Edmunds, Charlotte Kjels, Ricardo Mexia, Yamir Moreno, Sandro Meloni, Daniela Paolotti, **Daniela Perrotta**, Edward van Straten, Ana O. Franco *Influenzanet: Citizens Among 10 Countries Collaborating to Monitor Influenza in Europe.* JMIR Public Health and Surveillance, *2017*

- Caroline Guerrisi, Clément Turbelin, Thierry Blanchon, Thomas Hanslik, Isabelle Bonmarin, Daniel Levy-Bruhl, **Daniela Perrotta**, Daniela Paolotti, Ronald Smallenburg, Carl Koppeschaar, Ana O. Franco, Ricardo Mexia, W. John Edmunds, Bersabeh Sile, Richard Pebody, Edward van Straten, Sandro Meloni, Yamir Moreno, Jim Duggan, Charlotte Kjels, Vittoria Colizza *Participatory Syndromic Surveillance of Influenza in Europe.* The Journal of Infectious Diseases, *2016*

- Qian Zhang, Nicola Perra, **Daniela Perrotta**, Michele Tizzoni, Daniela Paolotti, Alessandro Vespignani *Forecasting Seasonal Influenza Fusing Digital Indicators and a Mechanistic Disease Model.* Proceeding of the 26th International Conference on World Wide Web (WWW), *2017*

- Qian Zhang, Corrado Gioannini, Daniela Paolotti, Nicola Perra, **Daniela Perrotta**, Marco Quaggiotto, Michele Tizzoni, Alessandro Vespignani *Social Data Mining and Seasonal Influenza Forecasts: The FluOutlook Platform.* ECML PKDD, Springer, *2015*

Daniela Perrotta
December 2017

Contents

List of Figures

List of Tables

Introduction

Infectious diseases have ever been a great concern of humankind and throughout history human beings have served as incidental hosts to many infectious diseases that have resulted in devastating epidemics [98]. In the 14th century, the Black Death carried by rat fleas on different trading routes spread throughout the Mediterranean and Europe claiming a heavy toll of lives killing almost one-fourth of the entire population. In 1918, the special circumstances during the World War I, such as overcrowded camps and hospitals, and soldiers piled in trenches or in transit every day, allowed the spreading of an unusually virulent and deadly flu virus. Perhaps the most lethal pandemic in the history of humankind, the Spanish Flu killed between 20 and 100 million people, more than the number killed in the war itself. In both episodes, human movements across regions and countries promoted the spread of the disease. Nowadays, the growth of the transportation infrastructure, with the airline system being the main and fastest means of transportation, enhanced human mobility and worldwide connection resulting in a larger opportunity for infectious diseases to spread on a large scale more rapidly than ever before. The emergence of new infectious diseases is continuous, variable and remarkably difficult to predict [148], thus feeding an increasing attention and global concern towards the next potential pandemic, mainly wondering when and where it might strike, and whether practitioners and scientists are prepared to respond and prevent the disastrous consequences of wide-spread transmission of a new disease [48]. Indeed, modern life conditions make the scenario of a global pandemic more likely as an increasing share of the planet lives in megacities, humans are encroaching on animal environments and the global trade and travel are constantly growing, thus heightening the likelihood of a sustained person-to-person transmission of pathogens as well as a rapid geographical spread of pathogens in new areas. The last recent global public health threats, such as the 2009 A (H1N1) influenza pandemic, the 2013 MERS-CoV outbreak and the 2016 Zika outbreak in Latin America, are only small glimpses of how quickly a deadly virus can spread, hitting also previously uninfected populations. In fact, infectious disease agents continuously adapt and evolve, resulting in newly emerging viruses (e.g. severe acute respiratory syndrome (SARS)) or re-emerging ones (e.g. West Nile virus), that add up to the other existing diseases, like HIV and influenza, that still remain unresolved threats to human health. Thus, monitoring, modelling and forecasting the spatial and temporal evolution of disease activity in human populations can help in fulfilling the mantra of early detection and early response to stop epidemics in their tracks.

In this context, computational epidemiologic approaches intervene in providing alternative solutions by exploiting the increasing computational and data integration capabilities to develop epidemic models of great complexity and realism. Since experimenting epidemics in-vivo is not a feasible option, epidemic modelling represents the main resort for understanding the disease spreading mechanism, predicting the future course of an epidemic and evaluating control measures and intervention strategies to reduce the overall impact of a disease. In recent years, epidemic models have evolved from simplified compartmental models into data-driven mechanistic approaches focused on large-scale microsimulations for scenario analysis of real infectious disease outbreaks [136]. In such approaches, data-driven epidemic models are able to produce realistic simulations of the global spread of infectious diseases by adopting high-resolution data on populations, human mobility and a detailed description of the socio-demographic interactions

among individuals to mathematically model the disease transmission mechanisms simulating the spatial and temporal evolution of epidemics at the level of single individuals [43, 45]. Given such unprecedented level of detail, data-driven mechanistic models represent a powerful tool to analyse policy making scenario, target efficient control measures, guide public health decision making process and provide quantitative forecasts of real epidemics.

However, since such powerful modelling techniques rely on data, having accurate and readily available data is becoming more and more crucial in order to properly and timely inform data-driven approaches. In recent years, with the advent of modern communication technology and the pervasive use of digital devices, modern epidemiology has been exposed to a new revolution and transformation of the pre-existing practices into digital disease detection techniques and digital warning systems that harness new technologies and novel data streams to monitor the health of populations and predict the future course of an epidemic [183]. With the vast majority of the world getting online, digital records of social interactions can provide sensory information of a potential infection in a certain geographical region, allowing for an early detection of particularly aggressive infection or new emerging diseases among the general population. This is the idea behind digital epidemiology: the fact that the health of a population can be assessed in real-time through digital traces generated by human activities on social media, crowdsourcing platforms and the Internet in general [174]. In this context, a variety of non-traditional approaches have emerged in recent years leveraging on the use of new digital data sources to provide local and timely information about disease outbreaks and related events around the world. Some examples are provided by Twitter [164], Wikipedia [142], Google search [109], but also participatory systems based on the possibility for single individuals to monitor and self-report their own health status through Web-based platforms [201]. Such novel data streams can be useful also to characterize human mobility patterns, necessary to properly inform epidemic models and assess the spatial spreading of infectious diseases, thus allowing for rapid interventions and appropriate control measures [54, 115, 169].

In this dissertation we will describe our original contribution to the field of modern computational and digital epidemiology, mainly in the use of novel digital data streams for monitoring, modelling and predicting the epidemic spreading of infectious diseases. In Chapter 1, we will cover the state-of-the-art of this scientific field, giving insights on the progress in the study of infectious diseases, mainly focusing on the disease monitoring techniques and how they evolved from traditional practices to innovative approaches based on the immediate availability of novel data streams, on the epidemic modelling approaches from simple compartmental models to the more recent data-driven mechanistic models, and eventually on the difficult task of predicting epidemics, trying to harness new technologies to predict the next emerging public health threats. In Chapter 2, we will focus on a participatory Web-based surveillance system present in Italy for monitoring seasonal influenza epidemics in a cohort of individuals who self-report their health status through Internet based surveys. Such novel non-traditional approach to gather self-reported health-related information results in a large amount of crowdsourced digital data that can be rapidly analysed for a variety of purposes, including monitoring disease trends, identifying risk factors, but also to complement traditional surveillance data and give early detection of local outbreaks by providing local and timely estimates of the levels of influenza circulating among the population. In Chapter 3, we will focus on the real-time forecasting of seasonal influenza activity by using different forecasting techniques and integrating different data sources, particularly highlighting the benefits of leveraging on novel digital data sources to capture an additional layer of real-time and geo-localized signal. In Chapter 4, we will focus on the recent Zika outbreak occurred in Latin America in 2015-2016, limiting the study to the epidemic of Zika in Colombia and mainly addressing the role of human mobility in a metapopulation modelling approach based on real data on population and a detailed description of the epidemiological characteristics of the disease, investigating the potential benefits of integrating human mobility patterns provided by mobile phone data.

Chapter 1

Background

> I simply wish that, in a matter which so closely concerns the well-being of the human race, no decision shall be made without all knowledge which a little analysis and calculation can provide.
>
> — Daniel Bernoulli, 1760

Epidemiology is the study of the factors affecting the health of populations in order to develop, implement, and evaluate effective intervention programmes for disease prevention and health promotion. In recent years, traditional epidemiology has been exposed to a new digital revolution driven by the advent of modern communication technologies and the pervasive use of digital devices, that has radically transformed the way people communicate and search for information in real-time through the Web [97]. In this chapter, I will cover the state-of-the-art of modern computational and digital epidemiology, describing the progress in the study of infectious diseases from the point of view of surveillance, modelling and forecasting of epidemics and how they have evolved from traditional practices to innovative approaches based on the immediate availability of novel digital data streams.

1.1 Disease Monitoring

Disease surveillance is an important epidemiological practice for the detection and prevention of the spread of an epidemic. Surveillance is an ongoing, systematic process for the collection, analysis and interpretation of data with the aim of observing patterns of progression and minimizing the harm caused by outbreaks, as well as disseminating results in order to inform, prepare future actions and appropriate interventions, and plan strategies for treatment to reduce overall impact of the disease, including mortality. A key part of modern disease surveillance is the practice of disease case reporting that, with the advent of modern communication technology, has changed dramatically, being transformed from manual record keeping to instant worldwide internet communication. The number of cases can derive from different layers of surveillance, including general practitioners (GPs) counting disease-associated visits, laboratory confirmations, hospitalization admissions, mortality data, depending on the type of disease and the country. Formal reporting of notifiable infectious diseases is a requirement placed upon health care providers by many regional and national governments, and upon national governments by the World Health Organization (WHO) that is the lead agency for coordinating global response to major diseases. Examples of notifiable diseases include diseases preventable by vaccination (e.g. influenza, hepatitis B), sexually transmitted diseases (e.g. HIV, herpes), nosocomial infections (e.g. methicillin-resistant Staphylococcus aureus (MRSA)), foodborne illnesses (e.g. botulism), waterborne diseases (e.g. cholera), contagious diseases caused by airborne particles (e.g. tuberculosis), diseases transmitted by vectors or parasites (e.g. rabies, malaria), and

emerging diseases (e.g. Ebola, Zika). The list of mandatory notifiable diseases may vary over time, according to a country's stage of development and the capacity of its health workforce.

However, traditional disease surveillance systems present practical limitations, mainly due to an heterogeneous population coverage and to considerable delay in disseminating data. In fact, traditional surveillance systems only include those populations who have access to health care or who decide to go to the doctor in the first place, and are often affected by reporting lags due to the time required to collate data and by continuous revision of the numbers initially released. Complementary surveillance systems are thus needed to provide additional and accurate data, but especially available quickly. In this context, a variety of non-traditional approaches have emerged in recent years leveraging on the use of new digital data sources that can provide local and timely information about disease outbreaks and related events around the world [174, 183]. An outstanding example is HealthMap [17], a powerful tool for disease outbreak monitoring and real-time surveillance of emerging public health threats [104, 151]. Founded in 2006, HealthMap acquires data from a variety of freely available digital media sources (e.g. ProMED-mail, Eurosurveillance, Google News, Baidu News) in different languages to obtain a comprehensive view of the current global state of infectious diseases. HealthMap is used as an early detection system and supports situational awareness by providing current, highly local information about outbreaks, used by a variety of organizations including state and local public health agencies. Examples of Twitter-based disease monitoring are provided by EbolaTracking [4] and ZikaTracking [31] that monitor tweets mentioning Ebola- and Zika-related keywords through a machine learning filtering approach with the aim of tracking and providing awareness on ongoing global conversations. But, in general, such phenomenon has particularly grown in the scope of monitoring influenza-like illnesses (ILI) activity that still represents a global public health problem causing high socioeconomic impact and substantial burden. Several digital data sources have been explored for augmenting traditional surveillance systems, such as Twitter data, whose flu-related tweets are used as a proxy for flu activity levels in a population [56, 69, 87, 163, 164, 181], and Wikipedia, whose access logs are used to analyse the amount of Internet traffic on certain influenza-related Wikipedia articles in order to estimate the proportion of the population with ILI symptoms [142]. But the most popular example is certainly Google Flu Trends (GFT). Launched in 2008 to help predict flu epidemics, it was based on the use of flu-related search terms entered in Google's search engine to reveal the presence of flu-like illness in a population [109]. Google Flu Trends popularized the idea of using digital data to derive epidemiological insights, but also demonstrated that this is no easy task when during the 2012-2013 flu season in the Northern Hemisphere GFT drastically overestimated the peak of flu prevalence [60, 131]. This episode serves as a remainder that new disease-tracking techniques based on mining of web data and on social media may complement, but not substitute traditional epidemiological surveillance networks, because, even if they are able to extract useful signals indicating disease activity, they fail in gathering information on the general population and the amount of people who are actually sick.

On the other hand, the advent of modern technology and the increasing community engagement in public health have fostered the birth of participatory surveillance systems based on the possibility for single individuals to monitor and report their own health status through Web-based platforms [201]. In 2009, Influenzanet [21], a European-wide consortium for monitoring influenza-like illness using participatory surveillance systems was established [158]. It was first launched in 2003 in the Netherlands and Belgium as "The Great Influenza Survey" (De Grote Griepmeting [2]) and the system was subsequently adopted by Portugal (Gripenet [11]) in 2005, Italy (Influweb [24]) in 2008, the United Kingdom (Flusurvey [9]) in 2009, Sweden (Halsorapport [16]) in 2011, France (Grippenet [13]) and Spain (Gripenet [12]) in 2012, Ireland (Flusurvey [8]) and Denmark (Influmeter [22]) in 2013 and Switzerland (Grippenet [14]) in 2016. Similar systems were independently implemented in Australia (Flu Tracking [10]) in 2006, Mexico (Reporta [28]) in 2009, the United States (Flu Near You [5]) and Germany (GrippeWeb [15]) in 2011, and Puerto Rico (Salud Boricua [29]) in 2012. Figure 1.1 shows a timeline of the aforementioned

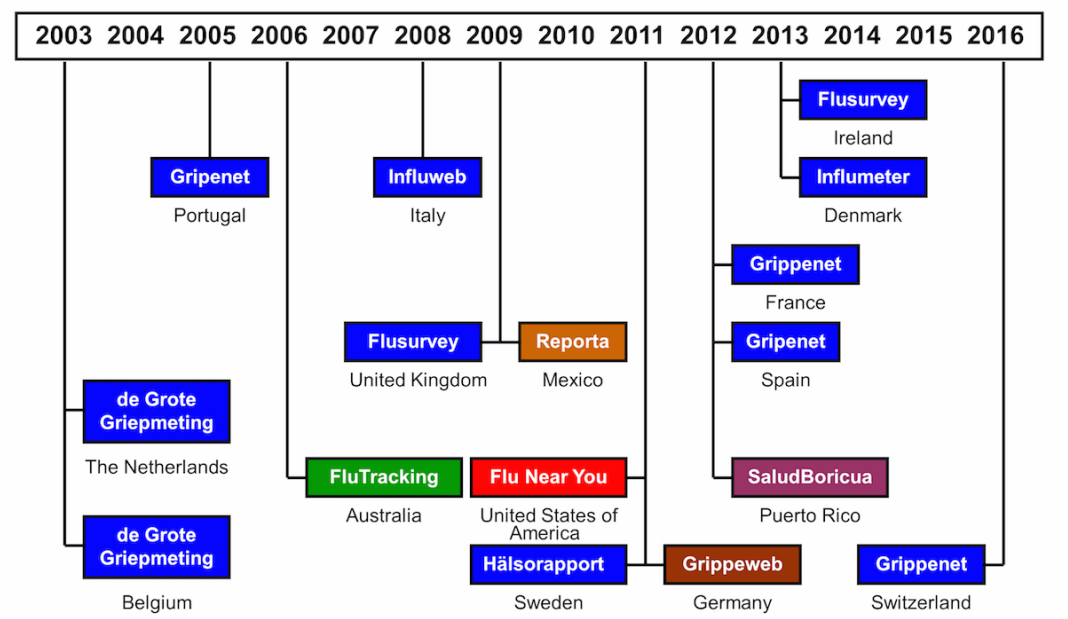


Figure 1.1: Timeline of the participatory surveillance systems for monitoring influenza-like illnesses. The platforms belonging to the Influenzanet network are indicated in blue. Figure reproduced from [125].

participatory surveillance systems for ILI in Europe and worldwide [125]. The result is a large amount of crowdsourced digital data that can complement traditional surveillance data and give early detection of local outbreaks by providing local and timely estimates of the levels of influenza circulating among the population. In particular, participatory surveillance systems have been proven to be accurate and reliable for ILI surveillance, as the detected timing and relative intensities of influenza epidemics are consistent with those reported by general practitioners [86, 88, 158, 167, 192]. Furthermore, participatory surveillance data have been used for a number of purposes, such as to estimate the severity of the 2009 H1N1 influenza pandemic [57, 65, 159], to assess health care usage [165, 189, 192] and to provide relevant information to estimate age-specific influenza attack rates [75, 162, 167], influenza vaccine effectiveness [66, 91, 95, 96] and risk factors for ILI [32, 112, 192]. Challenges to participatory surveillance include the recruitment and retention of participants, the accuracy of self-reported data and the development of nationally representative sample, especially for the populations at risk [39, 42, 64].

In conclusion, despite issues and limitations, digital disease surveillance systems have the potential to support infectious disease monitoring and complement data captured through existing traditional practices. Digital data sources are valuable for detection, monitoring and dissemination of information, facilitating communication during emerging disease infections, but also providing insights of the spatial spread and identifying regions with high prevalence or different demographic groups and communities for the implementation of targeted interventions and control measures, such as delivering specific medicine and distributing vaccines [59, 201]. Moreover, such digital systems can be employed in resource poor regions where supplementary sources (such as high-resolution satellite imagery and syndromic surveillance systems) are needed to fulfill the lack of a strong public health infrastructure [61, 72, 154]. The integration of different sources can improve the surveillance and reduce gaps present in individual sources and systems by monitoring different layers of a population from different perspectives and, if adopted by appropriate public health authorities, the data available through these systems can aid in early detection and response to outbreaks by building digital warning systems designed to track and stop local and global epidemics [35, 47, 183].

1.2 Epidemic Modelling

In recent years, the increasing computational and data integration capabilities have enabled the development of computational epidemic models of great complexity and realism [173]. Given the high human mobility and connection among different parts of the globe, with the airline transportation system being the main and fastest means of transportation, epidemics have the potential to spread on a large scale more rapidly than ever before. On the one hand, this highlights the need for timely and effective surveillance systems capable to detect emerging diseases and support traditional practices of disease monitoring. On the other hand, since experimenting in-vivo epidemics is not a feasible option, modelling approaches are the main resort to understand the spreading mechanism of diseases, predict the future course of an epidemic and evaluate control measures and intervention strategies to reduce the overall impact of a disease. First important steps towards modern epidemiology was made by Daniel Bernoulli, who in 1760 introduced a mathematical method to evaluate the effectiveness of early forms of immunization against smallpox, and by Dr. John Snow, who helped to eradicate the cholera outbreak occurred in London's Soho district in 1854 by identifying and removing the source of the infection, i.e. a public water pump, thus inspiring significant changes and improvement in general public health around the world. In 1927, A. G. McKendrick and W. O. Kermack published the Kermack-McKendrick epidemic model [124] describing the relationship between susceptible, infected and recovered individuals in a population, thus defining the modern mathematical modelling of infectious diseases, which is still ongoing and continuously growing.

Nowadays, disease evolution and contagion processes can be described with a variety of mathematical models of spreading and diffusion processes, ranging from simple compartmental approaches into structured frameworks increasingly focused on the hierarchies and heterogeneities of communities and populations [49]. Figure 1.2 shows the different structures at different scales used in epidemic modelling [37, 49]. Epidemic modelling describes the dynamical evolution of the contagion process within a population that is generally assumed to be divided into different classes (or compartments) depending on the stage of the disease, such as susceptibles (who can contract the infection), infectious (who have contracted the infection and are contagious), and recovered (who have recovered from the disease). Additional compartments can be considered in order to model other possible states of individuals with respect to the disease, for instance immune individuals or individuals exposed to the infection but not yet infectious. This framework can be further extended to take into account vectors for those diseases propagating through contact with an external carrier, such as mosquitoes for malaria or fleas for plague. The temporal evolution of the infection in the population is governed by transitions of individuals from one compartment to another that are generally specified by the disease etiology, such as the transmission rate or the recovery rate. The infection dynamics occurs when individuals get into contact and change their health status according to the disease. In the simplest assumption, individuals within each compartment are considered to be identical and homogeneously mixed, meaning that they can interact homogeneously with each other with random contacts. This is indeed a non-realistic situation as populations are generally structured heterogeneously according to socio-demographic factors, thus a detailed description of social interactions among individuals can help in modelling the epidemic spreading of infectious diseases. Similarly, the introduction of heterogeneous connectivity patterns in multiscale models helps in better describing population dynamics based on the spatial structure of the environment, transportation infrastructures and human movement patterns. In particular, metapopulation models describe spatially structured interacting subpopulations, such as cities, urban areas, or any defined geographical regions, accounting for the possibility that people move between different locations. Figure 1.3 schematizes a metapopulation modelling approach in which the system is composed of a heterogeneous network of subpopulations (or patches) connected by migration processes of individuals [81]. Each subpopulation contains a population of individuals classified according to their health status with respect to the disease (e.g. susceptible, infected, removed) as the

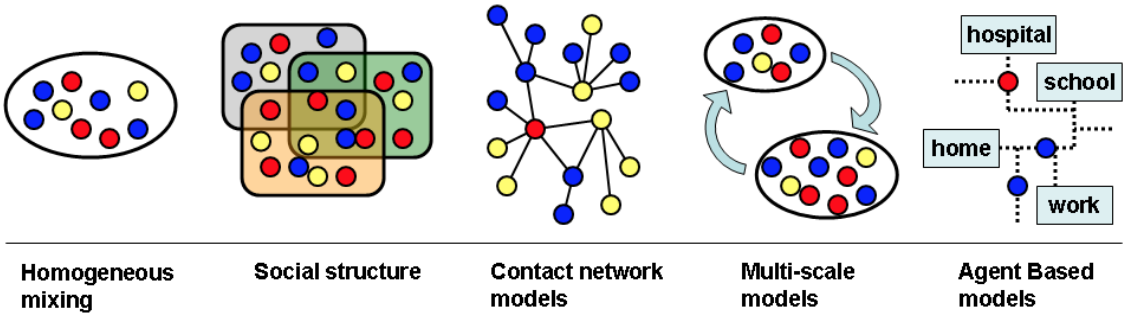


Figure 1.2: Structures at different scales used in epidemic modelling where circles represent individuals and colors indicate a specific stage of the disease. From left to right: homogeneous mixing, in which individuals are assumed to interact homogeneously with each other at random; social structure, where people are classified according to demographic information (age, gender, etc.); contact network models, in which the detailed network of social interactions between individuals provide the possible virus propagation paths; multiscale models which consider subpopulations coupled by movements of individuals, while homogeneous mixing is assumed on the lower scale; agent-based models which recreate the movements and interactions of any single individual on a very detailed scale (a schematic representation of a part of a city is shown). Figure reproduced from [49].

usual compartmental framework with the assumption of homogeneous mixing. The interaction among subpopulations is the result of the movement of individuals from one subpopulation to another on the network of connections among subpopulations. A simplified modelling approach assumes a Markovian process to approximate origin-destination mobility [49, 170]. In this case, the movement of individuals at each time step is given according to a matrix p_{ij} expressing the probability for an individual in the subpopulation i to travel to the subpopulation j. Thus, individuals are not labelled according to their original subpopulation and at each time step the same traveling probability applies to all individuals in the subpopulation without having memory of their origin. Such approach is widely used for very large populations when the traffic w_{ij} among subpopulations is known, by stating that $p_{ij} \sim w_{ij}/N_j$, where N_j is the number of individuals in subpopulation j. Several modelling approaches to the large-scale spreading of infectious diseases use this mobility process based on transportation networks for which it is now possible to obtain detailed data [77, 79]. On the other hand, even complicated mechanistic patterns can be accounted in a non-Markovian travelling process that allows individuals of subpopulation i to travel to destination j and come back at a constant rate as the usual commuting process. Here the subpopulations of the metapopulation system are coupled through detailed rate of traveling/commuting, thus defining the mixing subpopulation N_{ij} denoting the number of individuals of the subpopulation i present in the subpopulation j. In this case, the effective couplings result in a force of infection generated by the infectious individuals in subpopulation j on the individuals in subpopulation i [49, 122].

Lastly, agent-based models represent the class of most realistic epidemic models, in which the agents are the individuals described on a very detailed scale and the dynamics of interaction among agents is based on the socio-demographic structure of the population. Consequently, the infection spreads from one agent to another by direct contact, that may take place within households members, workplace colleagues or school and so on, thus resembling an actual situation occurring during an epidemic. Agent-based models provide a very rich data scenario and allow to make projections for policy makers using population specific socio-demographic features of the population [113, 143, 144], but the computational cost and most importantly the need for very detailed input data might be a limitation for their use. On the other hand, the structured metapopulation models are less detailed but fairly scalable and can be conveniently used to provide worldwide scenarios and patterns throughout Monte Carlo techniques based on the analysis of thousands of stochastic realizations exploring a multidimensional parameter space [43, 46, 77, 82]. An example of data-driven epidemic model is provided by the Global Epidemic

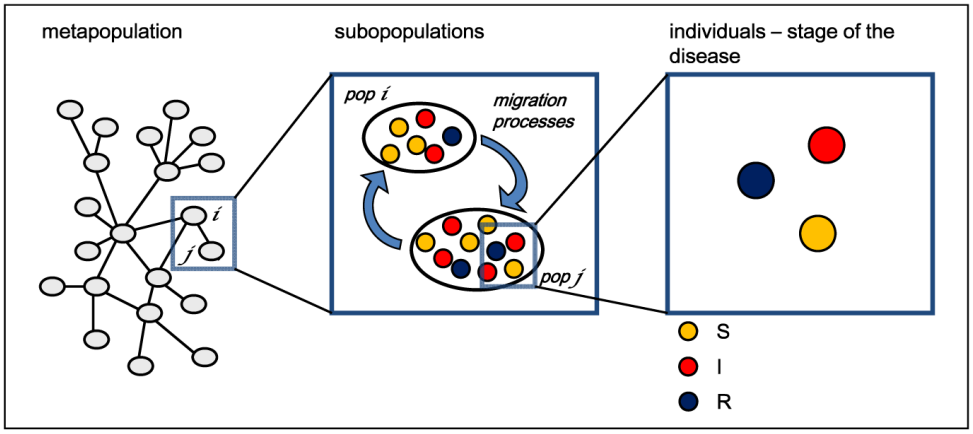


Figure 1.3: Schematic representation of a metapopulation model. The system consists of a heterogeneous network of interacting subpopulations (or patches) connected by migration processes. Individuals within each subpopulation are classified according to their health status (e.g. susceptible, infected, removed) and can move among subpopulations on the network of connections. Figure reproduced from [81].

and Mobility model (GLEAM) [43, 45] that is a stochastic generative model able to produce realistic simulations on the global spread of infectious diseases by integrating high-resolution data on populations, human mobility and a detailed description of the disease of interest as well as a great flexibility in integrating new data sources and adopting new techniques. GLEAM supports policy-making and emergency planning by developing epidemic models and scenario analysis, allowing to model containment and mitigation strategies and provide quantitative projections that better informs the analysis of their potential impact. More details about GLEAM can be found in Section 3.2.2.1.

It is clear that the key issue in such modelling approaches is the accuracy in the description of each dynamics component and in the data used to inform the model. Population data and mobility patterns certainly represent a fundamental aspect as the epidemic spreading of infectious diseases is strongly influenced by the amount of people living in each region and moving to other regions. Indeed, a detailed description of human mobility is important for characterizing and forecasting the spatial and temporal spread of infectious diseases [172] both at global [43, 77, 132] and national scale [68, 100, 143]. Recently, the last global public health threats, such as the 2009 A (H1N1) influenza pandemic, the 2013 MERS-CoV outbreak and the 2016 Zika outbreak in Latin America, highlighted the urgent need for accurate human mobility data to properly inform epidemic models and assess the risk of importation from the affected areas to the rest of the world, thus allowing for rapid interventions and appropriate control measures [54, 115, 169, 196]. In developed countries, mobility data is usually available and easily accessible from official sources for airline transportation, train trips, or commuting, but such datasets may be not updated for several years or aggregated at a wide geographical resolution, often limiting the potential impact of many studies. Also, this information may be inadequate or completely unavailable in developing countries. Thus, mobility models come in help by inferring local and global human mobility flows on a synthetic network whose connections and the corresponding intensity represent the flow of people among different regions. The most extensively used models for the estimation of trip distribution are the traditional gravity model [43], based on Newton's law of gravity, that assumes that the number of trips is related to the population at origin and destination and to decrease with the distance, and the more recent radiation model [182], that instead is inspired by the theory of intervening opportunities and considers human movements as diffusion processes that depend on the population distribution over the space. The gravity law and the radiation law have been widely tested and compared showing advantages and limitations and giving superiority to either of the approaches depending on the specific geographical setting and modelling assumptions [133, 134, 140, 203]. The limited generalizability of mobility models,

whose use can be hindered in the absence of good calibration data, led to the study of the large volumes of digital traces left by humans over the Internet allowing for a better understanding of mobility processes. Moreover, the continuous growth of the transport infrastructure and the fast evolution of mobility patterns allow people to travel more, thus rapidly changing travel patterns with important consequences for epidemic spreading and planning. Several datasets coming from different geo-located data sources have been analysed to characterize human mobility patterns. Some examples include credit card transactions [116], Twitter data [117, 135, 147], Foursquare data [153] or Flickr data [50]. In particular, several works have investigated the mobility flows obtained from mobile phone data to study the temporal and spatial information on humans physical displacements given by calls among mobile phone users and extract origin-destination matrix of the amount of people moving among different locations [34, 62, 110, 156]. Also, the use of mobile phone data is useful for low-income countries where the penetration is increasing and fully justify the use of CDRs data to estimate mobility flows. The availability of human mobility data at such high resolution has impacted several research fields, ranging from urban planning to social sciences [36, 38, 63, 76], but more important their application to the spatial epidemiology of infectious diseases [51, 101, 157, 190, 197, 198, 199, 200].

In conclusion, all these different structures of mathematical models have been developed to study the dynamic properties of disease transmission, characterize spatio-temporal spreading of infectious disease, determine the biological characteristics of specific pathogens, and analyse historical transmission behaviour during past events. Given that epidemic models must rely heavily on the realism put into the models that strongly impact the mechanisms of diseases propagation, they might be used to pose important questions about the underlying mechanisms of infection spread, possible means of control of the disease or epidemic prediction of real outbreaks.

1.3 Predicting epidemics

Throughout history, human beings have served as incidental hosts to many infectious diseases, often vector-borne or zoonotic in origin, which have resulted in devastating epidemics [98, 148]. Even though diseases like HIV and avian influenza remain unresolved threats to human health, increasingly attention and global concern is shifting to the next potential pandemic, trying to harness new technologies to predict when and where the next pandemic might strike and prevent the disastrous consequences of wide-spread transmission of a new disease [48]. Indeed, modern conditions make the scenario of a global pandemic more likely as humans are encroaching on animal environments raising chances for pathogens to adapt from animals to people, and an increasing share of the planet lives in megacities heightening the likelihood of person-to-person transmission of pathogens. The outbreaks related to avian influenza, the emergence of severe acute respiratory syndrome (SARS) and Middle Eastern respiratory syndrome (MERS), and the recent outbreaks of Ebola and Zika, to name a few, profoundly illustrate how an infection can spread worldwide in a very rapid fashion, also affecting previously uninfected populations. Thus, monitoring and forecasting the evolution of disease activity in human populations can help in early detection of newly emerging viruses (e.g. SARS) or re-emerging ones (e.g. West Nile virus), as well as complement traditional surveillance practices and support decision makers in designing effective interventions and allocating resources to mitigate their impact.

Influenza, one of the most common infectious diseases, is a highly contagious airborne disease that occurs in seasonal epidemics, but occasionally novel influenza A strains arise and may evade existing antibody immunity, thus giving rise to potential severe outbreaks. For example, the 1918 pandemic caused around 20-40 million deaths, while pandemics in 1957 and 1968 involved many infections but fewer deaths than in the 1918 pandemic. Real-time and accurate forecasts of major influenza indicators, such as peak time and peak intensity, can provide key information for appropriately preparing for and responding to influenza epidemics and pandemics [53]. In the literature, there are several studies advancing flu forecasting efforts mainly utilizing novel sources of digital surveillance [73], such as Google [109, 176], Yahoo [171], Twitter

[56, 69, 87, 129, 163, 164, 181, 204, 205], Wikipedia [108, 120, 142] and Web-based participatory surveillance systems [58, 168, 175, 204]. Also, to encourage development and innovation in influenza forecasting, every year since 2013 CDC organizes a flu forecasting challenge, asking researchers to develop cost effective methods to predict flu activity and to come up with their best predictions for the timing, peak, and intensity of the season [53]. After a three-year collaboration between CDC Influenza Division and external research groups, CDC launched in 2016 a dedicated website called "FluSight" to house the weekly influenza activity forecasts provided by the various research teams involved. Moreover, the more recent availability of virologic surveillance data and genetic sequence databases has fostered the birth of innovative studies incorporating the evolutionary change of influenza viruses [93, 150]. Despite increasing effort in identifying new methodology, often based on novel data streams, to provide a statistical framework capable of accurate estimation of flu prevalence in a population, our ability to predict the timing, duration and magnitude of local seasonal outbreaks of influenza remains limited.

On the other hand, some vector-borne diseases, such as dengue, chikungunya, and West Nile virus, are emerging in countries where they were unknown previously because of globalization of travel and trade and environmental challenges, such as climate change. The spatial dynamics of human infectious diseases are determined by the mobility of individuals who carry a disease into previously uninfected populations. Analogously, human migration and mobility mediate a large number of bioinvasions, defined as the introduction of previously unknown organisms in ecosystems. A clear example is provided by the new development of Zika virus, that was first identified in 1952 and then undergone a mutation in the South American outbreak of 2015 [178], now able to produce serious defects in newborns from infected mothers and other neurological complications, such as the Guillain-Barré syndrome. Response to and containment of an infectious disease outbreak can be greatly improved if health care response and outbreak control measures can be focused to areas predicted to be at the highest risk of experiencing new outbreaks. Accurate models of the geographic distribution of epidemic risk could significantly enhance the population-level effects of interventions implemented to control the spread of transmittable diseases. Thus, mapping is essential in spatial epidemiology [118, 119] in order to enhance our knowledge on the global geographic distribution of infectious agents and their spatial limits and better understand their risk of transmission, that up to date still remains incomplete. In this context, a large database has been recently compiled for mosquitoes *Aedes aegypti* and *Aedes albopictus* that are vectors for several globally important viral human diseases, such as dengue, chikungunya, yellow fever and Zika. Global maps of the predicted distributions of both species have been computed by coupling their global distribution with relevant environmental variables [126, 127]. Dengue is the most prevalent human arboviral infection causing approximately 100 million new annual infections in more than 120 endemic countries, mostly in the tropics and sub-tropics, but more recently also introduced to Europe [177]. Chikungunya has caused over 2.5 million infections over the past decade mainly occurred in Africa and Asia, but recently emerged in the Americas and Europe, thus posing new challenges to health systems and receiving considerable public health attention as the virus spreads into new areas, infecting naive populations and consequently causing large outbreaks. Similarly, yellow fever infections were significantly reduced due to large-scale vector control and vaccination programmes developed more than 70 years ago, but the virus still causes a significant disease burden in tropical and subtropical areas in South America and Africa. Given the public health impact and increasing concerns of these diseases due to their rapid geographical spread in new areas, understanding the distributions of their shared vectors can enable more efficient measures for disease control. In fact, these diseases can only persist where their vectors are present, thus mapping the global distribution of these vectors and determining their geographic limits is essential for public health planning, but strongly hindered by a continuous expansion fuelled by increased global trade and travel. Other previous works have been carried out in this context, such as mapping the global distributions of the dominant vectors of malaria [184, 185, 186] in order to improve efforts to understand the spatial epidemiology of associated arboviruses, and to predict how these could

change in the future. However, predicting spatial transmission routes of epidemics has proven to be remarkably difficult, due to the importance of long-distance transmission events, limited data on population mobility, unknown population immunity levels, low sensitivity and specificity of case reports, limited access to accurate and spatiotemporally resolved case data and a general stochasticity in the outbreak propagation.

In recent years, mathematical and computational approaches to the study of epidemics have been increasingly relevant in providing quantitative forecasts and scenario analysis of real infectious disease outbreaks [137]. Epidemic models have evolved into data-driven computational approaches focused on large-scale microsimulations aimed to explore the feasibility of reliable epidemic forecasts and spreading scenario analysis, and provide information at very detailed spatial resolutions. Data-driven approaches can generate results at unprecedented level of detail, and have been used successfully in the analysis and forecast of real epidemics [43, 44, 145], and policy making scenario analysis [55, 80, 99, 138]. Moreover, mechanistic and mathematical approaches aid not only in the response to particular diseases, but also in illuminating basic epidemiologic principles and important parameters that dictate whether a novel (or existing) pathogen can be controlled [103]. However, paremeterising such models is often difficult in real-time, when information on behavioural changes, interventions and routes of transmission are not readily available [106]. Indeed, forecasting the course of disease spread is a difficult task, particularly in the response to an emerging disease threat, but still remains a major goal of the disease-modelling community and has become increasingly important to help guide critical decision-making during infectious disease outbreaks. Since disease reporting is often delayed and initially inaccurate, forecasting techniques include not only projections into the future, but also nowcasting of incidence based on earlier available information, like Twitter, Google or participatory systems [109, 164, 168, 175].

However, much more needs to be done to integrate such forecasting modelling techniques into existing practices in public health. Epidemic forecasts are rarely evaluated during or after the event, and it has not been established what the best metrics for assessment are. As forecasts become a routine part of the toolkit in public health, standards for evaluation of performance will be important for assessing quality and improving credibility of mathematical models, and for elucidating difficulties and trade-offs when aiming to make the most useful and reliable forecasts [107]. As the progress made in weather forecasting over the past 60 years, infectious diseases forecasting is now in the process of steadily improving the accuracy and reliability of predictions. The hope is that one day researchers will forecast disease outbreaks in the same way meteorologists forecast the weather, but with the advantage of being able to prevent and stop them with effective control measures and prevention strategies. But it is clear that many basic theoretical questions are still open, such as understanding how the complex nature of the real world affects our predictive capabilities in computational epidemiology or the fundamental limits on epidemic evolution predictability with computational modelling. Understanding the mechanisms influencing the epidemic spreading of infectious diseases, characterizing the spatio-temporal transmission and diffusion of diseases, harnessing new technologies to predict the next emerging public health threats, investigating the behaviour of people during a disease outbreak, all these aspects tackle the development of a modern approach to computational and digital epidemiology.

Chapter 2

Monitoring the activity of seasonal influenza in Italy

> When we think of the major threats to our national security, the first to come to mind are nuclear proliferation, rogue states and global terrorism. But another kind of threat lurks beyond our shores, one from nature, not humans – an avian flu pandemic.
>
> — Barack Obama, 2005

The pervasive use of digital communication technologies for public health and the increasing community engagement in public health have fostered the birth of a variety of non-traditional approaches to provide additional data sources for monitoring influenza epidemics directly among the general population. In this context, next to national public health surveillance infrastructures, participatory surveillance systems have emerged in different parts of the globe with the aim of monitoring the influenza activity by directly involving self-selected volunteers among the general population reporting their health status through Internet based surveys.

In this chapter, I will focus on the two surveillance systems present in Italy for monitoring seasonal influenza: the traditional surveillance system called Influnet [23] based on general practitioners reports, and the participatory Web-based surveillance system called Influweb [24]. The aim of this work is to describe the two data sources, investigate the representativeness of the two populations compared to the Italian general population, evaluate weekly incidence rates and estimate age-specific influenza attack rates. As a result of a collaboration with the Italian Institute of Public Health (ISS) [23], this chapter is based on Ref. [167] in which I personally contributed as the first author by performing the cleaning process, analysis, validation and visualisation of datasets, as well as writing the manuscript.

2.1 Introduction

Seasonal influenza is an acute contagious respiratory illness caused by viruses that can be easily transmitted from person to person. Influenza viruses circulate worldwide causing annual epidemics with highest activity during winter seasons in temperate regions, resulting in about 3-5 million cases of severe illness and about 250-500 thousand deaths around the world [19]. The rates and severity of epidemics can vary substantially from year to year due to several factors, including the types, subtypes and strains of circulating viruses, and the level of protective antibodies in the population. In fact, influenza viruses undergo high mutation rates and frequent genetic re-assortment [188] that cause a continuous formation of newly emerging or re-emerging influenza viruses and thus a subsequent lack of immunisation in the population. Influenza viruses

can cause zoonotic infections and adapt to humans resulting in sustained transmission and emergence of novel viruses leading then to seasonal viruses and sometime to pandemics, such as the Spanish Flu between 1918 and 1920 or the H1N1 influenza pandemic in 2009. Since influenza viruses constantly change, the seasonal influenza vaccines are updated and administered annually to provide the necessary protection to the population, especially to high-risk groups, such as children, the elderly, health care workers, and people who have chronic illnesses, for whom the influenza vaccine is recommended by the World Health Organization (WHO) [19]. Indeed, influenza is a global public health problem causing high general practice consultation rates, increased hospital admissions, excess deaths, and high absenteeism in schools and workplaces, including in health workers. Its high socioeconomic impact and burden is not just limited to the industrialized world but extends to low- and middle-income countries, encompassing multiple dimensions such as direct costs to the health service and households and indirect costs because of productivity losses, as well as broadly affecting the overall economy [90].

Nationally organized networks of general practitioners (GPs) constitute the basis of public health surveillance, reporting the weekly number of patients visited with influenza-like illness (ILI) or acute respiratory infection (ARI) in selected healthcare facilities (sentinels). Some countries also report virological information from a subset of patients, influenza-confirmed hospitalizations, or mortality data (see Figure 2.1a). The aim of collating data from different layers of surveillance is to better assess the intensity and spread of influenza, identify trends and risk groups, and inform actions to reduce the influenza-associated burden. On the other hand, traditional surveillance techniques are notorious for severe time lags of up to two weeks due to the time required to collect, collate and distribute data, and typically the numbers initially released are continuously revised and updated as more data are recorded throughout the influenza season. This means that by the time the data is available, the information is already 1 or 2 weeks old and subject to change in the following published bulletin. Moreover, these traditional ways of collecting epidemiological data only includes those people who have access to the health care system and subsequently seek health care treatment for their illness, but they typically cannot account for a variable (depending on age, gender or other characteristics) proportion of individuals who instead do not seek health care assistance. Furthermore, there is lack of spatial resolution and uniform standards in clinical definitions, which may vary considerably between countries and even between reporters.

Consequently, next to national public health surveillance infrastructures, a variety of non-traditional approaches have emerged in recent years [174] in order to overcome some issues and limitations of traditional disease surveillance approaches and provide additional data sources for monitoring influenza epidemics [112]. The pervasive use of digital communication technologies for public health [183] and the increasing community engagement in public health have fostered the birth of surveillance systems based on the possibility for single individuals to monitor and report their own health status through Web-based platforms [201]. The result is a large amount of crowdsourced digital data that can be rapidly analysed for a variety of purposes, including monitoring disease trends, identifying risk factors, but also to complement traditional surveillance data and give early detection of local outbreaks by providing local and timely estimates of the levels of influenza circulating among the population. Participatory surveillance systems for seasonal influenza are currently running in 13 countries around the world and collect, aggregate and communicate data in real time during the course of influenza seasons. Specifically, the systems that are currently online are: Influenzanet, a network of Web-platforms running in eleven European countries [21], FluNearYou in the United States [5] and FluTracking in Australia [10]. In particular, the Influenzanet participatory surveillance system was established in Europe in 2009 and included 5 countries with prior Web-based participatory surveillance experience: first launched in the Netherlands and Belgium as "The Great Influenza Survey" (De Grote Griepmeting [2]) in the 2003-2004 influenza season, then implemented in Portugal (Gripenet [11]) in 2005, Italy (Influweb [24]) in 2008 and the United Kingdom (Flusurvey [9]) in 2009. At the outset, the Influenzanet platform was not homogeneous across countries because of historical develop-

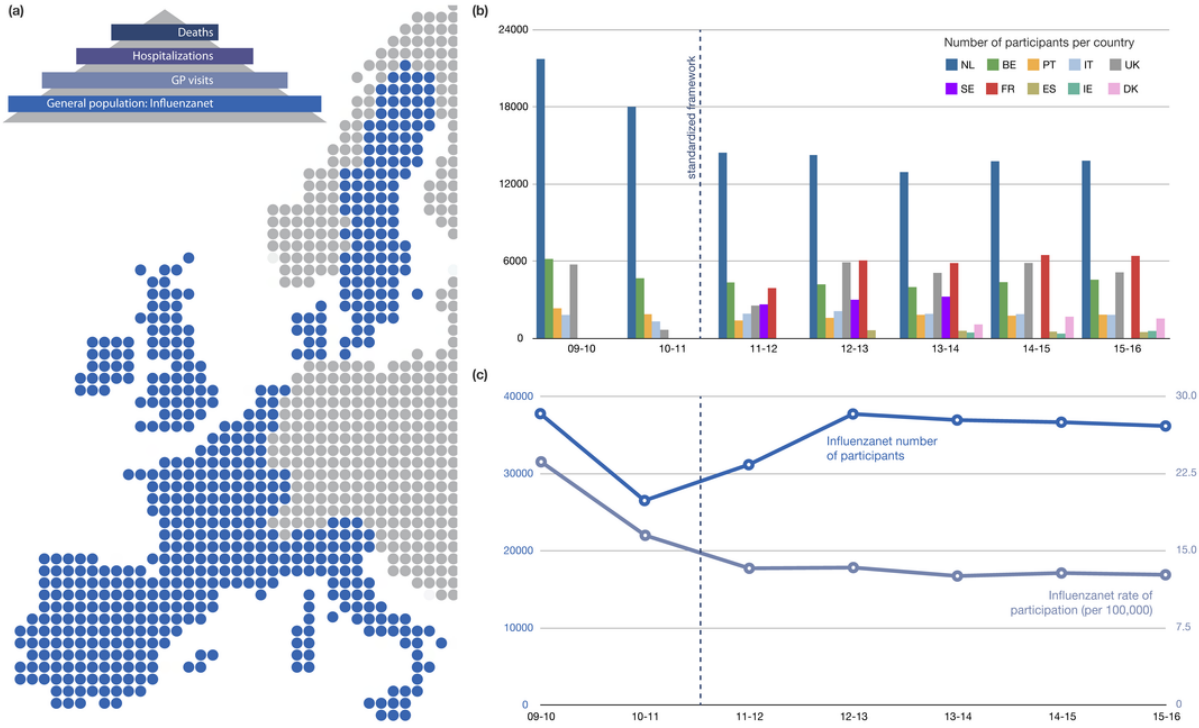


Figure 2.1: Influenzanet participatory surveillance system. a) ILI monitoring scheme illustrating different layers of surveillance used by public health authorities. Influenzanet represents an additional layer to monitor ILI in the general population and it is now present in 11 European countries as represented in the map. b) Number of Influenzanet participants per country per season. c) Total number of Influenzanet participants per season (left axis) and rate of Influenzanet participation per season (right axis) expressed as the number of participants per 100,000 individuals of the total population of Influenzanet countries. The dashed vertical line indicates the standardized framework introduced from the 2011-2012 season.

ments leading to the project [158], but since the 2011-2012 influenza season, the Influenzanet platforms share a common and standardized data collection approach throughout the European countries involved. Subsequently, the system was adopted by Sweden (Halsorapport [16]) in 2011, France (Grippenet [13]) and Spain (Gripenet [12]) in 2012, Ireland (Flusurvey [8]) and Denmark (Influmeter [22]) in 2013, and Switzerland (Grippenet [14]) in 2016. Figure 2.1a shows the European countries involved in the Influenzanet platform (Switzerland joined later) [112]. In each country, the platform is coordinated by a team of local researchers from University, Research Institution or Public Health Institution and consists of a website where individuals can register and have access to a personal account where they can insert and update their data. Participation by country varies considerably [64], with averages over all seasons ranging from 1.2 individuals per 100,000 population, for Spain, to almost 100 individuals per 100,000 population for the Netherlands, notably the most successful example within Influenzanet, as shown in Figures 2.1b and 2.1c .

Over the course of the years, data collected through such participatory systems have been used for a variety of purposes, including estimating the severity of the 2009 H1N1 influenza pandemic [57, 65, 159], assessing health care usage [165, 189, 192] and providing relevant information to estimate age-specific influenza attack rates [75, 162, 167], influenza vaccine effectiveness [66, 91, 95, 96, 125] and risk factors for ILI [32, 112, 192]. Furthermore, participatory surveillance systems have been proven to be accurate and reliable for ILI surveillance, as the detected timing and relative intensities of influenza epidemics are consistent with those reported by general practitioners [86, 88, 158, 167, 192].

In this chapter, we will focus on the two surveillance systems that are present in Italy for monitoring seasonal influenza epidemics: the traditional surveillance system called Influnet [23] that is based on general practitioners reports, and the participatory Web-based surveillance system, Influweb [24], that is part of the Influenzanet platform. Since 2012 data collected by Influweb have been experimentally adopted by the Italian National Institute of Health as an additional source of data about the circulation of influenza-like illness among the general population by collecting in a single weekly bulletin called FluNews [6] all information gathered by the various epidemiological surveillance systems monitoring seasonal influenza in Italy, including the traditional surveillance from Influnet, the participatory surveillance from Influweb, the syndromic surveillance of access to Emergency Rooms and the monitoring of more severe influenza-related cases. In particular, this chapter is based on Ref. [167] in which we evaluated the performance of Influweb in combination with traditional surveillance as a tool for improving the national surveillance of influenza-like illness in Italy during the first three years of such integrated approach for flu surveillance, namely from 2012-2013 to 2014-2015. Here, we will describe the systems, summarizing advantages and limitations and mainly focusing on the representativeness of individuals with respect to the Italian general population, but also investigate the quality and accuracy of the epidemiological signal that can be extracted from Influweb compared to the ILI incidence reported by Influnet, that represents our ground truth. Results showed that, despite the existing participation biases, the ILI incidence detected by Influweb is able to capture both the timing and the relative intensity of seasonal flu activity in Italy, thus justifying its use in further works related to real-time influenza forecasts, as we will tackle in Chapter 3.

2.2 Dataset

In this section, we describe the two data sources for influenza-like illnesses (ILI) in Italy, highlighting their main characteristics and explaining how they have been used in our analysis.

2.2.1 Traditional surveillance system

Surveillance is an ongoing, systematic collection, analysis, interpretation, and dissemination of data regarding the diffusion of the disease for use in public health action and program planning and evaluation to reduce morbidity and mortality and to improve health among the general population. In particular, for influenza surveillance, in most of the developed countries a national network of general practitioners traditionally report the weekly number of cases with influenza-like illness and collect samples from a subset of patients for virological confirmation. In Italy, such surveillance system for influenza syndromes is called Influnet [23] and it is coordinated by the Italian National Institute of Health, with the support of the Ministry of Health. The system is based on a network of about 1,000 volunteering physicians and pediatricians, evenly distributed to represent all the Italian regions, covering about 2% of the Italian population. Practitioners share a common operational protocol to report weekly ILI cases, defined according to EU case definition, as well as seasonal influenza vaccine uptake. Influnet analyses the data and estimates the national and regional incidence rates by age groups (0-4, 5-14, 15-64, >64 years), compiling and publishing a weekly report during the winter season, generally from week 42 to week 17 of the calendar year, to evaluate the duration and intensity of the influenza epidemic. Therefore, by the time the ILI incidence data are published, they are already at least one week old, and typically new reports provide only a first estimate of the weekly ILI incidence which is then continuously updated in the following weeks as more data from GPs are recorded. Moreover, traditional surveillance can sometimes over- or underestimate the true burden of the epidemic, depending on what ILI fraction corresponds to real influenza cases and on the reporting rates of sentinel doctors [142]. However, traditional surveillance data are considered a highly reliable indicator of influenza activity and ILI incidence data reported by Influnet represent the ground truth for this study.

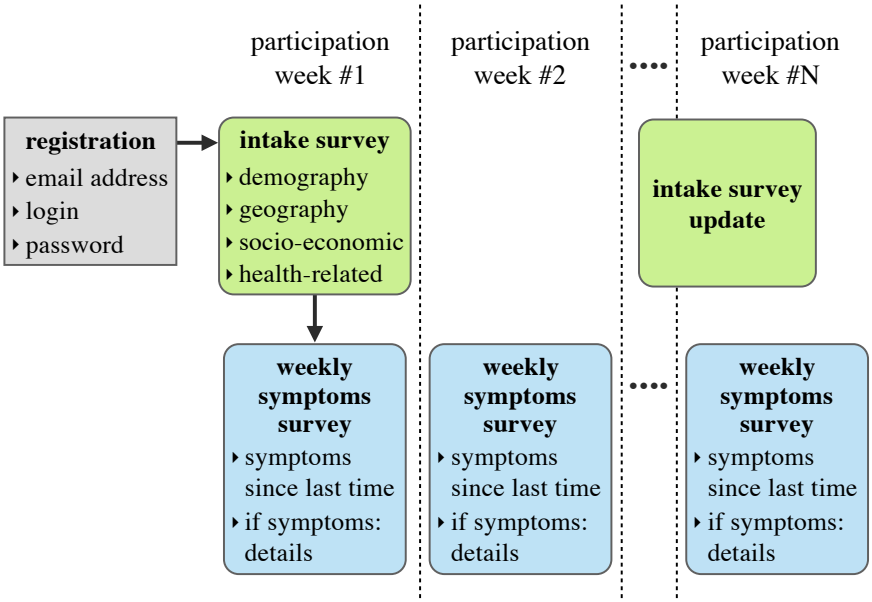


Figure 2.2: Schematic representation of the registration process to the platform and the data collection through the compilation of two types of surveys: the intake survey aims to collect generic information about participants, while the symptoms survey aims to collect data on the episodes of illness and the consequent behaviour of participants. See Appendix A for more details on the questionnaires.

2.2.2 Web-based surveillance system

Influweb [24] is the Italian Web-based surveillance system that monitors the seasonal influenza activity in Italy since 2008. Influweb is part of the Influenzanet [21] platform and is coordinated by ISI Foundation in Torino, Italy. The data collection generally runs from October/November to April/May, allowing for flexibility (e.g. in 2016 the surveillance was extended through the summer to monitor cases of Zika virus infection) and it is usually disseminated among the general population at the beginning of each influenza season through a number of press releases, general media campaigns, specific dissemination events (e.g. science fairs) or word of mouth. Participation is voluntary and anonymous, and open to all individuals living in Italy. To join the network, individuals register on their national platform and complete an intake survey covering demographic, geographic, socioeconomic and health questions, including age, gender, household size and composition, location of home and workplace, education level, occupation, vaccination status for the previous and the present influenza season, the presence of a chronic disease, a possible pregnancy and other issues. The intake survey can be then updated throughout the season to account for changes (e.g. vaccination, pregnancy, etc.). Users can also create accounts on behalf of other members of their family or household and report for individuals, such as children or elderly people, who are unable to navigate the Internet. Participants receive a weekly e-mail newsletter as remainder to fill in a symptoms survey in which they are asked whether since the last time they visited the platform they experienced any general, respiratory or gastrointestinal symptom from a list of 20 symptoms. If symptoms are reported, further questions are asked to assess the syndrome (e.g. sudden onset of symptoms and body temperature) and participant behaviour (e.g. changes in the daily routine, health-seeking behaviour and medicine uptake, including painkillers or antipyretics, cough medications, antivirals, and antibiotics). Figure 2.2 schematizes the processes of registration to the platform and compiling of the weekly symptoms surveys. Intake and symptoms questionnaires are reported in Appendix A.

In general, participants must fulfill some conditions to be considered in the data analysis as *active participants* in order to avoid having a variable and biased sample [105], mainly due to the possibility for volunteers to join the platform throughout the influenza season or stop reporting during the influenza season. Inclusion criteria to select participants may vary and depend on the

specific aim of the study [42, 64, 192, 193]. Here we define as active participants those users who joined the data collection by completing the intake survey at least once since their registration to the system, and a minimum of two symptoms surveys with a frequency of at least one every three weeks, on average [42]. In the following, we will equally refer to Influweb population or Influweb active population to indicate the population of active participants. If a participant completed the intake survey multiple times, we consider the report that is most recent according to the influenza season under study. Symptoms surveys are weekly filtered, keeping only the last symptoms survey filled within a certain week, thus assuming that it corresponds to the most updated health status of the participant. Moreover, since information are provided and manually inserted by participants, they might be affected by inaccuracy in self-reporting, due to misunderstandings and wrong interpretations of the questions or as the result of a deliberate action. For this reason, collected data are first processed and cleaned in order to check for mistakes or misreporting (e.g. a date of birth in the future or a non-valid zip code).

2.3 Methods

In this study we included data collected by Influnet and Influweb during three influenza seasons, namely the 2012-2013, 2013-2014 and 2014-2015 influenza season, limiting the analysis to the time window of full overlap of both systems, that is from week 47 to week 14 of the calendar year. Aim of this work is to describe the two data sources in terms of demographic indicators by investigating the representativeness of the sample compared to the Italian general population, evaluate the ILI incidence rates extracted from Influweb in comparison to the ILI signal detected by sentinel doctors, and estimate age-specific influenza attack rates by using self-reported symptoms from Influweb and official surveillance data reported by Influnet.

Demographic analysis Influnet provides only information about patients aggregated for age groups (0-4, 5-14, 15-64, >64 years) and regions, and no other details are provided. Influweb, instead, collects information at a higher resolution level, including age, gender, zip code of residence and household composition, thus allowing for a detailed comparison with the Italian general population. Influweb data are mapped from zip code resolution to region level to allow the comparison with Influnet data and national data.

For this analysis we used demographic data of the Italian general population at 1st January of 2013, 2014 and 2015, respectively, provided by the National Institute for Statistics Studies (ISTAT) [27] as well as publicly available geographical data and shapefiles at the level of NUTS2 regions. Maps are generated by manipulating the shapefiles with the *Basemap* library available in Python. Summary statistics given here comprise simple counts and percentages. We used χ^2-test for non-continuous variables, and non-parametric test (Mann-Whitney U test and Kruskal-Wallis test) to compare distributions.

ILI incidence Weekly ILI incidence can be extracted from Influweb data by dividing the number of ILI cases by the number of active participants per week. The flexibility of the collected data allow for assessing the ILI syndromes by applying different case definitions. Here we use the standard case definition provided by the European Center for Disease and Control (ECDC) and also adopted by Influnet, that defines an ILI case as the sudden onset of symptoms with one or more systemic symptoms (fever or feverishness, malaise, headache, myalgia) plus one or more respiratory symptoms (cough, sore throat, shortness of breath) [18]. Only participants who reported symptoms within 15 days from the onset of the symptoms are included and, in order to avoid double-counting of a single ILI episode, individuals who fit the ILI definition for two consecutive weeks are considered ILI only during the first week, while in the few cases where individuals reported ILI for three consecutive weeks, only the second episode is discarded.

To quantify the performance of the weekly ILI incidence detected by Influweb, we compute the Pearson correlation and the cross-correlation between the two time series. In particular, for

this latter measure, we first smoothed the Influweb time series using a simple moving average technique with a time window of three weeks. To this aim we used the moving average function $ma()$ of the R *forecast* package and cross-correlation function $ccf()$ of the R *stats* package.

Attack Rate The influenza attack rate is defined as the cumulative incidence of influenza virus infections and its annual global estimate corresponds to 5-10% in adults and 20-30% in children [19]. It represents an important measure of the rate of infections in at risk population, but it is a difficult parameter to obtain. Here we estimate age-specific influenza attack rates by using self-reported symptoms from Influweb active participants and official surveillance data reported by Influnet. Similar studies have been carried out in The Netherlands [162] with data from the national Influenzanet platform called Grote Griepmeting and in the United States [75] with data from the Flu Near You platform. Here we adopt the same methodology developed in Ref. [162] based on the use of weekly rates of self-reported ILI (defined as self-reported fever and cough/sore throat) among participants in the Influweb cohort in combination with an external separate dataset, extracted from the Hong Kong household studies [84, 85, 130], consisting of symptoms reported by influenza-positive individuals in households where clinical cases of influenza had been observed. The influenza attack rates are estimated by using the inference method described in Ref. [162], and, in particular, equation 3 in Ref. [162]. Briefly, it requires the assessment of baseline ILI rates during periods of low influenza activity and during periods of active influenza circulation as reported by the traditional surveillance system. Such rates are then converted to influenza attack rates via estimates of the probability $P(ILI|Flu)$ of self-reported ILI for influenza cases. This probability is estimated using data from PCR-positive individuals in Hong Kong household studies [84, 85]. The influenza attack rate AR_{flu} for a given season between calendar week $t = i$ and week N is estimated as:

$$AR_{flu} = \frac{\sum_{t=i}^{N}(ILI(t) - Base)}{P(ILI|Flu) - Base} \tag{2.1}$$

where the numerator represents the excess ILI rate (above the baseline) during the period of active influenza circulation, while the denominator is the excess probability of reporting ILI for influenza cases compared to non-influenza cases ($P(ILI|Flu) - Base$), from which the influenza attack rate is estimated. Posterior samples for each of the quantities in eq. 2.1 (e.g. ILI(t)) are independently extracted to get a posterior sample of estimates for AR_{flu}, for which the mean and the 95% credible intervals are reported. More details can be found in Ref. [162].

In particular, we used data on self-reported symptoms among Influweb active participants during the 2012-2013, 2013-2014 and 2014-2015 influenza seasons. For each season, we defined four age-group specific "main cohorts" (ages $\leq$24, 25-44, 45-64, and $\geq$65 years) as the set of persons in each age group who 1) filled out a report by a "cutoff week", that is the calendar week preceding the week in which the epidemic threshold is crossed and 2) number of completed reports during at least 50% of the weeks from the date of their first report through week 14. The cutoff weeks have been gathered from the official surveillance data reported by Influnet and correspond to weeks 2012-51, 2013-52 and 2014-51, respectively for the three influenza seasons.

2.4 Results

A total of 2,127 unique individuals actively participated during one or more of the three influenza seasons under study. In particular, 35.68% participated for one season, 22.99% for two seasons and 41.33% participated for all three seasons. Table 2.1 reports the number of registered and active participants as well as the average number of symptoms surveys completed during the three seasons under study. Every year, Influweb is able to attract new participants, with 446 new individuals during the 2013-2014 influenza season, and 233 new individuals during the 2014-2015 influenza season, thus continuously increasing the total number of individuals registered to the

Table 2.1: Participation to Influweb during the 2012-2013, 2013-2014 and 2014-2015 influenza seasons.

season	no. registered individuals	no. active participants	% active sample	no. active in country (per 100,000)	average symptoms surveys (95% CI)
2012-2013	3,041	1,452	47.75%	2.43%	10.4 (10.0, 10.7)
2013-2014	3,487	1,458	41.81%	2.4%	11.9 (11.5, 12.2)
2014-2015	3,720	1,464	39.35%	2.41%	12.4 (12.1, 12.7)

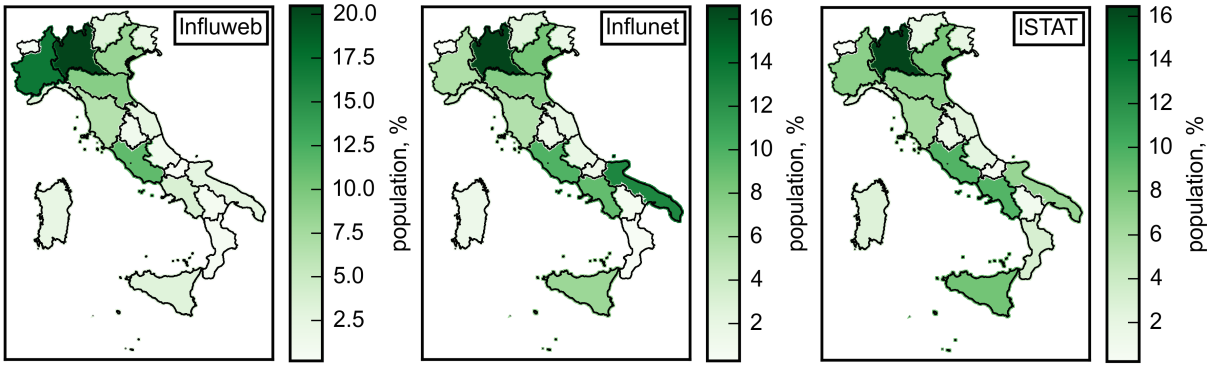


Figure 2.3: Geographic distributions of the Influweb active participants (left), the Influnet sample (middle) and the Italian general population (right) at the level of NUTS2 regions during the 2014-2015 influenza season. The colour code indicates the proportion of individuals living in each region.

system. However, participation to data collection remained quite constant with a total of 1,452, 1,458 and 1,464 active participants during the 2012-2013, 2013-2014 and 2014-2015 influenza seasons, respectively, representing about 2.4% (per 100,000 individuals) of the Italian general population. On average, active participants completed from 10 to 13 weekly symptoms surveys every season, corresponding to about one survey every two weeks.

Referring to the 2014-2015 influenza season, approximately 78% of the active participants had a single membership account, while 22% belonged to a multiple account with at least two active participants. In particular, 11% of the multiple accounts had at least one participants aged less than 14 years and 9% had at least one participant aged over 65 years. Among the sample of active participants, about 49% never updated the intake survey, about 46% updated it twice, and about 5% updated it at least three times during the 2014-2015 season.

Despite the small sample, Influweb showed a good coverage of all the Italian regions, with an average of 2.41% active participants per 100,000 individuals, and an active participation rate per region that varies between 0.4 per 100,000 individuals (Calabria) to 5.6 per 100,000 individuals (Piemonte). The geographic distributions of the Influweb active participants, the Influnet sample and the Italian general population were not statistically different (p=0.886 for 2012-2013, p=0.925 for 2013-2014, p=0.923 for 2014-2015). Corresponding maps at the level of NUTS2 regions are displayed in Figure 2.3 for the 2014-2015 influenza season. Maps for the 2012-2013 and 2013-2014 seasons are not reported here because of their similarity.

We further analysed the Influweb population in terms of age, gender and household composition in comparison to the Italian general population, as shown in Table 2.2. Such level of detail is not available for the traditional surveillance system that reports only data for some specific age groups. During the 2012-2013 influenza season participants were statistically representative of the Italian general population in terms of age (p=0.226), whereas during the 2013-2014 and

Table 2.2: Age, gender and household size of the Influweb population and the Italian general population (IT pop) during the 2012-2013, 2013-2014 and 2014-2015 influenza seasons.

Season	Age		Gender		Household Size	
	Influweb, avg. age (95% CI)	IT pop, avg. age	Influweb, %male, %female	IT pop, %male, %female	Influweb, avg. household size (95% CI)	IT pop, avg. household size
2012-2013	43.8 (42.9, 44.7)	43.5	57.8, 42.2	48.4, 51.6	2.6 (2.5, 2.7)	2.4
2013-2014	45.4 (44.4, 46.3)	43.7	59.1, 40.9	48.5, 51.5	2.8 (2.6, 2.9)	2.4
2014-2015	45.7 (44.8, 46.7)	43.9	58.2, 41.8	48.5, 51.5	2.9 (2.7, 3.1)	2.4

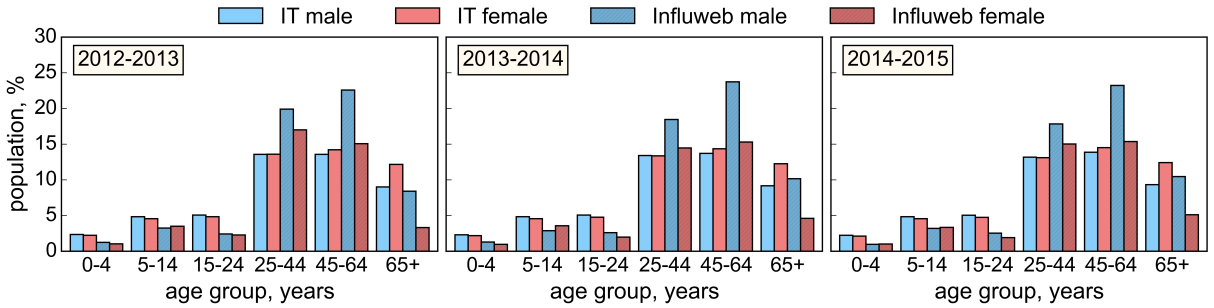


Figure 2.4: Age and gender distribution of the Influweb population and the Italian general population during the 2012-2013, 2013-2014 and 2014-2015 influenza seasons.

2014-2015 influenza seasons participants were found to be older than the general population ($p < 10^{-3}$). A comparison between the age groups distributions of the two surveillance systems showed that in the Influweb population the young adults and adults age groups are overrepresented, while school age children are underrepresented, whereas in the Influnet sample, patients aged less than 25 years are slightly overrepresented, while adults aged between 25 and 64 years are slightly underrepresented, as displayed in Figure 2.5. In addition, Figure 2.4 shows the age and gender distributions of the Influweb active participants compared to the Italian general population for the influenza seasons under study. A larger proportion of male individuals participated to Influweb, in contrast with the gender distribution of the Italian general population ($p < 10^{-4}$ for all seasons). Influweb participants were found to live in larger families with respect to the Italian general population as the distributions of the number of household's members were statistically different (p=0.024 for 2012-2013, $p < 10^{-4}$ for 2013-2014 and 2014-2015).

Table 2.3 shows the values for the vaccination coverage as reported by the Influweb active participants and as collected by the Influnet sentinel doctors. On average, about 16% of the Italian population got the seasonal influenza vaccine according to the data collected by Influnet, while about 14.5% of the active participants reported to have got the vaccine. Vaccination coverage detected by Influweb was statistically representative for the 2012-2013 (p=0.722) and the 2013-2014 (p=0.235) influenza seasons, whereas during the 2014-2015 Influweb detected a larger proportion of vaccinated people ($p < 10^{-2}$). Moreover, if we limit the analysis only to people older than 65 years old, corresponding to a high-risk group for which the seasonal influenza vaccine is recommended by WHO, the vaccination coverage was larger in the Influweb participants for the 2012-2013 and 2013-2014 influenza seasons ($p < 10^{-4}$), whereas it was statistically representative in the 2014-2015 (p=0.06).

The weekly incidence of ILI cases among Influweb active participants were found to correlate

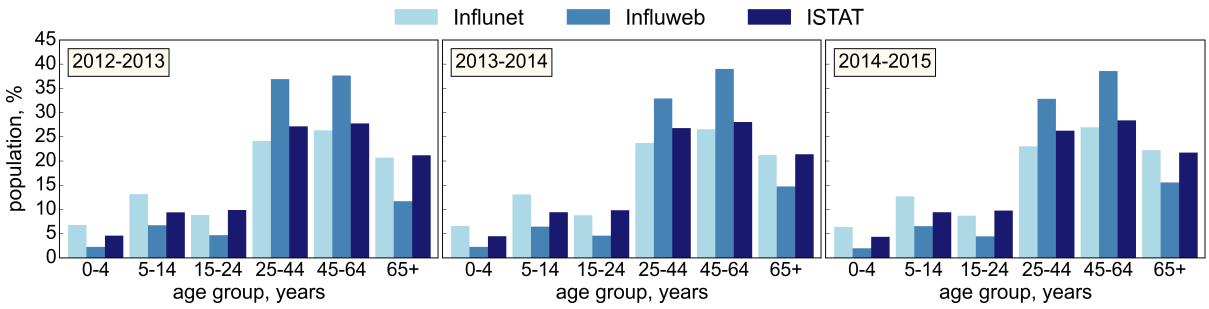


Figure 2.5: Age distribution of the Influweb population (blue), the Influnet population (light blue) and the Italian general population (dark blue) during the 2012-2013, 2013-2014 and 2014-2015 influenza seasons.

Table 2.3: Vaccination coverage for the Influweb population and the Influnet population, as well as limited to the age group of individuals aged over 65 years, during the 2012-2013, 2013-2014 and 2014-2015 influenza seasons.

Season	Vaccination Coverage			
	Influweb pop.	Influnet pop.	Influweb pop. (aged>65 years)	Influnet pop. (aged>65 years)
2012-2013	15.15%	14.79%	32.3%	53.0%
2013-2014	16.46%	15.31%	37.9%	53.9%
2014-2015	16.19%	13.35%	40.9%	47.6%

well with the weekly incidence detected by the Influnet sentinel doctors for the three seasons under study. In particular, the Pearson correlation coefficient for 2012-2013 is 0.605 ($p < 10^{-2}$), for 2013-2014 is 0.472 (p <0.05) and for 2014-2015 is 0.699 ($p < 10^{-3}$). Figure 2.6 shows the weekly incidence curves of Influweb and Influnet, which are reported on different scales for ease of comparison in order to highlight that the two curves are consistent in both timing and relative magnitude for the whole duration of each of the three influenza seasons under exam. In the early phase of the epidemic we can observe an initial overestimation of the Influweb weekly incidence, mainly due to a smaller sample of participants involved in the data collection in the first weeks. As already mentioned, the data collection period generally starts at the beginning of the influenza season in November, through a first e-mail reminder to the participants already enrolled in the system and a number of press releases to attract new volunteers among the general population. Therefore, the system needs a few weeks to reach a stable cohort of participants and early data are usually noisy. In general, to address this issue, data are filtered according to a certain fraction of the number of active participants present at the peak during that influenza season or during the previous influenza season, depending on whether the analysis is performed in real-time or retrospectively [168]. In this study, we chose not to discard any data points in order to present the entire analysis for the same time window, but, for instance, only removing the data point at week 47 would be sufficient to obtain better results in terms of correlation, i.e. 0.706 for 2012-2013, 0.616 for 2013-2014 and 0.678 for 2014-2015.

Furthermore, we analysed the cross-correlation between the smoothed time series of Influweb data (3-weeks centered moving average) and the Influnet time series, as shown in Figure 2.7. The maximum level of cross-correlation is found at lag of one week (ρ=0.813 for 2012-2013, ρ=0.724 for 2013-2014, ρ=0.795 for 2014-2015).

When participants report symptoms, they are also asked to answer to some follow up questions regarding healthcare-seeking behaviour and whether they changed their daily routine due to the illness (see *Dataset*). The fraction of participants who reported to have consulted a general practitioner after an episode of ILI remained fairly constant during each season, with an average of about 36%. Among this subset, the largest fraction of participants visiting their fam-

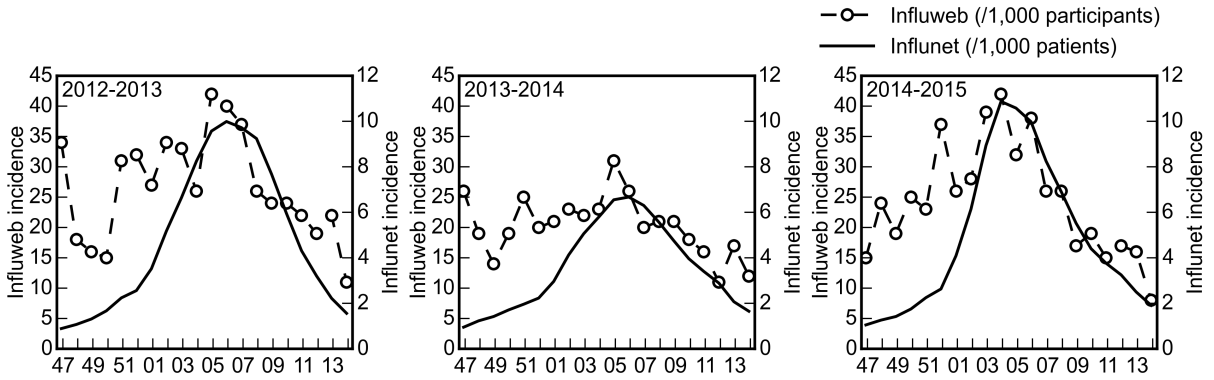


Figure 2.6: Weekly ILI incidence rates as extracted from Influweb (left axis) and reported by Influnet (right axis) during the 2012-2013, 2013-2014 and 2014-2015 influenza seasons. Incidence values are intended for 1,000 participants for Influweb, while for 1,000 patients for Influnet.

Table 2.4: Estimated age-specific influenza attack rate for the period (in weeks) of high incidence, that is when the incidence reported by Influnet is above the epidemic threshold, during the three influenza seasons under study.

Season, Calendar Weeks	Estimated Attack Rate % (95% CI)		
	Age Groups, years		
	25-44	45-64	65+
2012-2013, 52-12	16.06 (0, 39.92)	18.72 (0, 38.95)	-
2013-2014, 01-12	10.75 (0, 31.3)	4.04 (0, 18.42)	12.04 (0, 28.4)
2014-2015, 52-13	16.2 (0, 41.39)	19.43 (0, 38.73)	15.32 (0, 38.72)

ily doctor corresponded to children up to 14 years old, followed by elderly, while young adults aged between 15 and 24 years old were less likely to visit a healthcare provider for their illness. The corresponding age distribution is shown in Figure 2.8a and highlights the differences among age groups in healthcare consultation. Moreover, children were more likely to change their daily routine during an episode of ILI, staying at home from school, as shown in Figure 2.8b.

Estimated influenza attack rates in the different Influweb cohorts during the 2012-2013, 2013-2014 and 2014-2015 seasons are reported in Table 2.4. The age group of individuals younger than 24 years were excluded from this analysis because of its small size. For the same reason, during the 2012-2013 season it was not possible to estimate an attack rate for the age group 65+ because of the small size of the sample. Sensitivity analysis with respect to the choice of the baseline period in equation 2.1 is presented in Table 2.5.

2.5 Discussion

A participatory surveillance system called Influweb has been implemented since 2008 in Italy as a tool to collect influenza-like illness data among the general population. In this work we aimed at showing the utility and reliability of Influweb data as an additional layer of surveillance data to support and complement traditional surveillance data based on general practitioners.

A total of 2,127 individuals actively participated during one or more of the three influenza seasons under study, thus representing only a small fraction of the total population living in Italy (approximately 60 millions). Among the total active participants about 38% participated for all three seasons, thus meaning that the sample is quite stable from one year to the other, with a consistent fraction of participants who continue to be motivated to participate over several years. Moreover, Influweb is capable of attracting new participants every year, with an average of 1,458 active participants per season among which about 60% are new to the system.

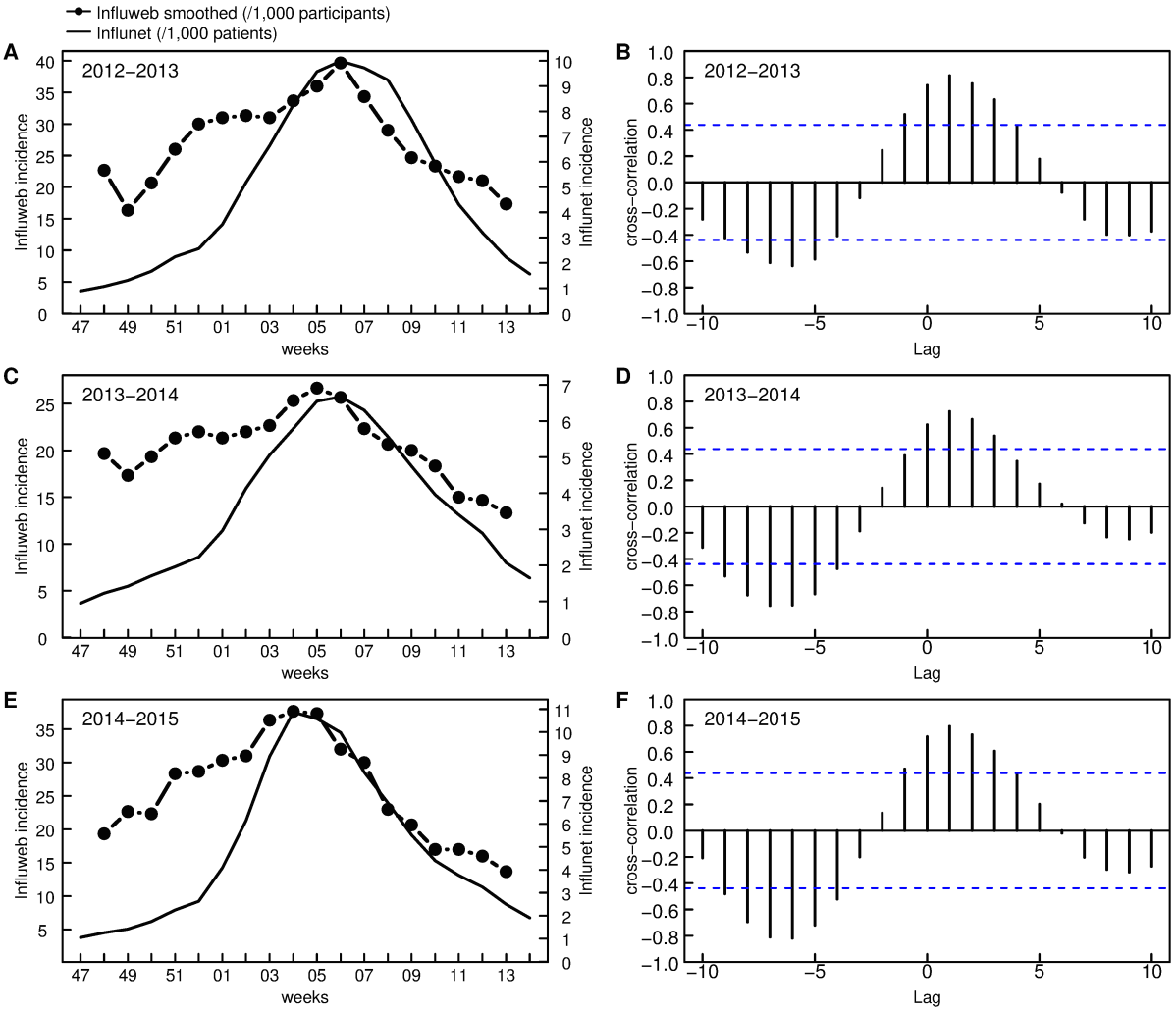


Figure 2.7: ILI incidence analysis for the three influenza seasons under study. Subplots (A), (C), (E) show the smoothed Influweb incidence curve (left axis) and the Infunet incidence curve (right axis). Subplots (B), (D), (F) show the cross-correlation between the two time series as a function of the lag (weeks).

The geographical distribution of the Influweb active participants covers all the Italian regions and reflects well the heterogeneous distribution of the population in the various regions of Italy (see Figure 2.3). However, a higher number of active participants is consistently observed in the region named Piemonte, which hosts the Institution conducting the Influweb project, likely reflecting a more powerful effect of communication campaigns at the local level.

The distribution of age is statistically different from the Italian general population. The young adults and adults age groups are overrepresented, while school age children are underrepresented. Underrepresentation in the groups between 0 and 25 years old may be due to the impossibility to access the Internet in an unsupervised way for the youngest children and to a lack of interest in influenza or health-related topics for teenagers and people in their 20's, as previously pointed out in [64]. The system already incorporates the possibility of adding multiple users to an account managed by a single participant who is supposed to facilitate the input of data for individuals who cannot or are not familiar with Internet tools. The results for the 2014-2015 season showed that 11% of the multiple accounts had at least one participant aged less than 14 years and 9% had at least one participants aged over 65 years. Interestingly, elderly participants, corresponding to individuals aged more than 65 years old, are well represented, thus probably meaning that the familiarity with computers and the usage of the Web is increas-

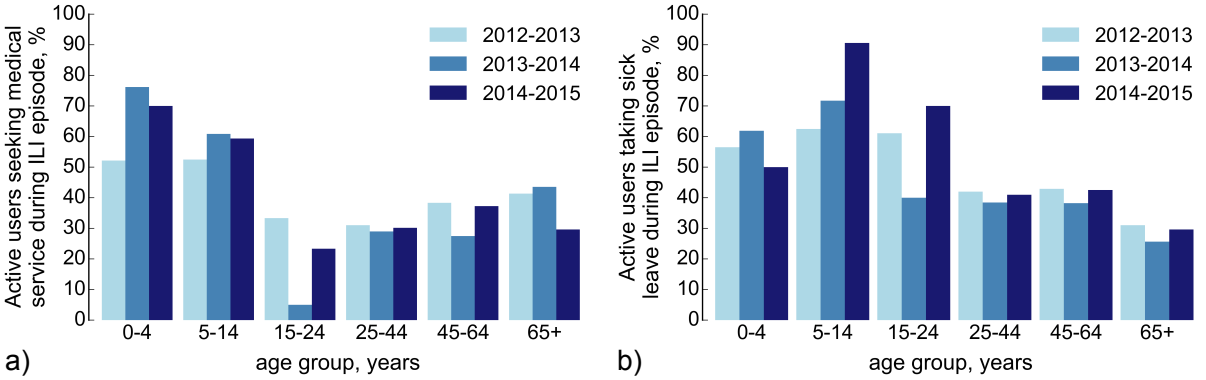


Figure 2.8: Age distribution of the proportion of Influweb active participants, who during an episode of ILI (a) sought medical assistance for their illness, and (b) changed their daily routine staying off from school or work, during the 2012-2013, 2013-2014 and 2014-2015 influenza seasons.

Table 2.5: Sensitivity analysis for the age-specific influenza attack rate with respect to the choice of the baseline period (in weeks), during the three influenza seasons under study.

Season	Baseline Weeks	Estimated Attack Rate % (95% CI)		
		Age Groups, years		
		25-44	45-64	65+
2012-2013	47-51	16.06 (0, 39.92)	18.72 (0, 38.95)	-
	46-51	10.76 (0, 35.49)	18.7 (0, 38.96)	-
	46-50	8.89 (0, 35.67)	12.76 (0, 35.59)	-
	47-50	15.7 (0, 41.15)	12.77 (0, 35.56)	-
2013-2014	47-52	10.75 (0, 31.3)	4.04 (0, 18.42)	12.04 (0, 28.4)
	46-52	10.8 (0, 30.72)	6.63 (0, 20.17)	13.0 (0, 28.92)
	47-51	12.47 (0, 33.57)	4.48 (0, 19.35)	10.5 (0, 27.56)
	46-51	11.05 (0, 31.77)	4.99 (0, 19.27)	11.65 (0, 28.17)
2014-2015	47-51	16.2 (0, 41.39)	19.43 (0, 38.73)	15.32 (0, 38.72)
	46-51	19.23 (0, 42.41)	17.51 (0, 36.22)	19.18 (0, 41.48)
	46-50	17.82 (0, 42.3)	16.66 (0, 36.43)	24.3 (0.63, 45.98)
	47-50	12.37 (0, 39.02)	20.14 (0, 40.3)	21.46 (0, 44.53)

ing even among age groups that used to be considered harder to reach through the Web. On the other hand, while for the Influweb sample there is an underrepresentation of children aged less than 25, this age group is well covered (if not slightly overrepresented) in the Influnet sample. This is a clear example of the kind of complementarity that the two systems can achieve.

The percentage of male participants is larger than the percentage of female participants, in contrast with the gender distribution in the Italian general population. This might be due to the fact that traditionally in Italy the familiarity with Web technologies is more typical of the male gender with a larger fraction of men (65%) accessing the Internet compared to women (55.8%), despite the fact that in general women are usually more active and involved on websites and forums dealing with health-related content [105]. This is also reflected by the very low rates of participation of elderly women.

Influweb participants live in larger households than the general population, in agreement with previous findings from other Influenzanet platforms [64].

Vaccination coverage is statistically representative of the national coverage in Italy during two seasons (2012-2013 and 2013-2014), whereas Influweb detected a larger vaccination coverage in Italy during the 2014-2015 influenza season. However, both systems detected the slight decrease in the percentage of vaccination in the 2014-2015 influenza season. Looking at the 65+ age class,

i.e. the age group who is most at risk of complications due to influenza-like illness, Influweb found a smaller proportion of vaccinated people in comparison with official data collected by the Ministry of Health [20] during the 2012-2013 and the 2013-2014 influenza seasons. This might be a consequence of the fact that this age class is not well represented in the Influweb cohort. On the other hand, as the sample size of this age group increased over the course of the years, in the 2014-2015 the vaccination coverage was statistically representative. Potentially, data from Influweb could also be used to estimate vaccine effectiveness during the influenza season [91].

Figures 2.6 and 2.7 show that the ILI incidence calculated from Influweb and from Influnet correlate well: the two curves are consistent in both timing and relative magnitude for the whole duration of each of the three influenza seasons under exam. This is in line with what has been observed previously with other Web-based platforms for influenza surveillance in other countries [105]. Moreover, as shown in Figure 2.7 panel B, D and F, Influweb detected the peak incidence one week earlier than traditional surveillance, thus suggesting that Influweb might be able to detect temporal variations in incidence rates in advance with respect to the sentinel doctors surveillance. This might be explained by the fact that most people generally do not seek healthcare assistance on the first day they feel sick, while sentinel doctors report the day of the visit as being the first day of illness, thus causing a slight delay in reporting.

The portion of volunteers seeking medical assistance when experiencing ILI remained fairly constant during each season, with an average of about 36%. However, differences among age groups in healthcare consultation for individuals with ILI symptoms were evident, showing that children are more likely to be visited by a doctor with respect to young adults.

Most participants do not stay at home or change their routine when they experience an episode of ILI. Overall, children are more likely to stay at home from school, as expected, while elderly participants rarely report to have changed their daily routine during an ILI episode. This might be due to the fact that the largest part of participants aged over 65 years are retired and only 9.3% still have a paid employment.

Self-reported symptoms collected by the Influweb platform have also been used to calculate influenza attack rates during the 2012-2013, 2013-2014 and 2014-2015 influenza seasons. A good agreement is observed between the weekly incidence of ILI in the Influweb population and attack rates measured by the inference framework (see *Methods*). In fact, during the 2013-2014 season ILI attack rates were smaller compared to the other seasons, as the 2013-2014 influenza season was milder than the others. In the 2012-2013 influenza season, it was not possible to estimate an attack rate for the age group 65+ because of the small size of the sample, but this shortcoming was fixed in the 2013-2014 and 2014-2015 seasons as the sample size increased. For the same reason, we restricted the estimation of influenza attack rates to the participants aged over 25 years, but future estimates can be performed even for the age group 0-24, provided that the corresponding cohort size would be sufficiently large. Overall, as already discussed in previous papers [75, 162], despite limitations of the inference method and of data gathered through the Influweb participatory system, the results show that it is possible to have real-time estimation of influenza attack rates among the general population during each influenza season. Such a feature is specific of the Web-based surveillance systems detecting ILI cases directly from the general population.

2.6 Conclusion

Since 2012 data collected by the Web-platform Influweb have been experimentally adopted by the Italian National Institute of Health as an additional source of data about the circulation of influenza-like illness among the general population. As a result, a weekly bulletin called FluNews [6] is being published during the influenza season since 2012 as a single collector of all information gathered by the various epidemiological surveillance systems monitoring seasonal influenza in Italy, including the GPs surveillance from Influnet, the participatory surveillance from Influweb, the syndromic surveillance of access to Emergency Rooms and the monitoring of

more severe influenza-related cases.

In this study, we presented the results of the first three seasons of such integrated approach for flu surveillance in order to assess the contribution that an online system can provide to the traditional influenza-like illness surveillance system. Here we demonstrated that despite the sample of individuals involved in the data collection is limited and not representative of the general population, the detailed information provided by participants enables to estimate weekly ILI incidence rates and age-specific influenza attack rates in good agreement with data reported by the traditional surveillance. Among the advantages and strengths of Influweb we denote the real-time component that allows to extract the current level of influenza circulating among the population, while traditional surveillance data are usually reported with at least one week lag and numbers initially reported are subject to continuous revision throughout the influenza season. In addition, Influweb allows to gather information directly from the general population and not only from medically attended ILI. In fact, we showed that only a fraction of the Influweb participants reported to have consulted a healthcare provider for their illness, corresponding to about 36% of the ILI cases.

On the other hand, recruiting and maintaining participants are the main challenges, while limitations are mainly due to the self-selected sample, potential misreporting, and lack of validation by a physician or by virological testing. To improve the representativeness of the sample, targeted strategies for communication informed by the results of this study can be used to increase participation rates in Italy, which are indeed lower than in other European countries in which Influenzanet is implemented. However, with the large majority of participants willing to contribute to additional studies beyond ILI, Influweb may become in the near future a powerful system that, once adjusted for sample biases, can offer a timely tool to measure the epidemiological status, opinion, or behaviour of the general population with regard to different indicators and diseases. Moreover, results showed here allow to further use Influweb collected data to provide real-time forecasts on seasonal influenza, as we will see in the next Chapter.

Chapter 3

Real-time forecasts of seasonal influenza epidemics

> It is quite probable that influenza will continue to be prevalent all over the world for some years to come. May we hope that etiological and epidemiological work will furnish us with more competent methods for prevention and delimation before the world is visited by another pandemic.
>
> — Hans Zinsser, 1922

Seasonal influenza is an acute contagious respiratory illness that annually produces about 3 to 5 million cases of severe illness and about 250 to 500 thousand deaths around the world [19]. Monitoring and forecasting the evolution of influenza activity can help in early detection and response in order to minimize the impact of potentially devastating epidemics. In this chapter, I will focus on the real-time forecasts of seasonal influenza by using different forecasting techniques and integrating different data sources, particularly highlighting the benefits of leveraging on novel digital data sources to capture an additional layer of real-time and geo-localized signal. In the first section, I will describe a computational framework previously validated in Ref. [204, 205], based on real-time influenza-related data, traditional surveillance reports and a dynamical mechanistic model called GLEAM (GLobal Epidemic And Mobility model) able to provide short-term predictions of seasonal influenza activity. This section is based on Ref. [58] in which I personally contributed by exploring the calibration of the model with Influenzanet data, running the simulations, and analysing, validating and visualising the results.

In the second section, I will report our work about the way traditional surveillance data reported by general practitioners can be combined with digital surveillance data from Influweb, a participatory Web-based system described in Chapter 2, in order to improve seasonal influenza forecasts in Italy. This section is based on Ref. [168] in which I personally carried out the work, including collecting, processing and analysing the two datasets, building and testing the various forecasting models, and analysing and validating the results, as well as writing the manuscript.

3.1 Introduction

Seasonal epidemics of influenza affect millions of people every year, causing high general practice consultation rates, increased hospital admissions and excess deaths, and consequent socio-economic impact and burden [90]. Seasonal influenza epidemics typically occur in fall/winter seasons in the Northern Hemisphere, annually varying in terms of timing and intensity. In fact, once the influenza season starts, it is not an easy task to predict its spatial and temporal evolution as the infection intensity and geographical spread can change dramatically year by year.

As an example, Figure 3.1 shows the influenza-like illness (ILI) consultation rates per 100,000 individuals as reported by the European Centre for Disease Prevention and Control (ECDC) in Belgium, Italy and Spain during some historical seasons. Indeed, both timing and intensity of influenza epidemics are country- and season-specific. Moreover, in contrast to seasonal influenza, novel influenza A strains capable of sustained person-to-person transmission arise occasionally and may give rise to pandemic outbreaks if population lack pre-existing antibody immunity. For example, the 1918 pandemic caused around 20-40 million deaths, while pandemics in 1957 and 1968 involved many infections but fewer deaths than in the 1918 pandemic. Thus, monitoring and forecasting the evolution of influenza activity in populations can help in early detection of novel circulating viruses with pandemic potential or particularly virulent influenza seasons as well as complement traditional surveillance practices and support decision makers in designing effective interventions and allocating resources to mitigate their impact. Real-time and accurate forecasts of major influenza indicators, such as peak time and peak intensity, can provide key information for preparing for and responding appropriately to influenza epidemics [53].

The first popular example of early warning system for predicting flu epidemics was Google Flu Trends (GFT). Launched in 2008, it was based on the idea of using flu-related search terms entered in Google's search engine to estimate influenza-like illness activity in a population rather than traditional statistical predictive analysis [109]. However, significant discrepancies between GFT's flu estimates and those measured by the Centers for Disease Control (CDC) in subsequent years demonstrated the hard challenge of such digital disease detection systems [60]. Despite increasing effort in identifying new methodology, often based on novel data streams, to provide a statistical framework capable of accurate estimation of flu prevalence in a population, our ability to predict the timing, duration and magnitude of local seasonal outbreaks of influenza remains limited. In the literature there are several studies advancing flu forecasting efforts [73], such as mining Twitter to uncover flu-related tweets and Wikipedia access logs to analyse the amount of Internet traffic on certain influenza-related Wikipedia articles to be used as a proxy for flu activity levels in a population [56, 69, 87, 142, 163, 164, 181]. To encourage development and innovation in influenza forecasting, the CDC organized a challenge during the 2013-2014 influenza season in United States, mainly asking researchers to utilize novel sources of digital surveillance data to develop cost effective methods to predict flu activity [53]. A team at Columbia University won the competition with a mathematical model that incorporates GFT data as well as CDC's ILI data and is weekly calibrated and optimized in order to produce an accurate and reliable forecast [179, 180]. Moreover, this team presented their forecasts similarly to weather forecast, thus helping the communication to both public health officials and the public. After a three-year collaboration, CDC launched in 2016 a dedicated website called "FluSight" to house the weekly influenza activity forecasts provided by the various research teams involved. Indeed, much more needs to be done to integrate such digital epidemiology modelling approach with existing practices in public health. As the progress made in weather forecasting over the past 60 years, infectious diseases forecasting is now in the process of steadily improving the accuracy and reliability of predictions and increasingly investment are now concentrated to connect predictions to public health decision making.

In this chapter, we summarize our efforts and main findings in forecasting seasonal influenza by using different forecasting techniques and integrating different data sources, particularly highlighting the benefits of leveraging on digital data sources to capture an additional layer of real-time and geo-localized signal. The chapter is subdivided in two sections. The first section is based on Ref. [58] in which three different participatory disease surveillance systems have been investigated in the use of modelling, simulation and forecasting. Here we report only our original contribution based on a computational framework based on real-time influenza-related data, traditional surveillance reports and a dynamical model for spatial epidemic spreading, called GLEAM (GLobal Epidemic And Mobility model), which is able to produce realistic simulations of the global spread of infectious diseases by combining high-resolution data on populations and human mobility with stochastic mathematical models of disease transmission

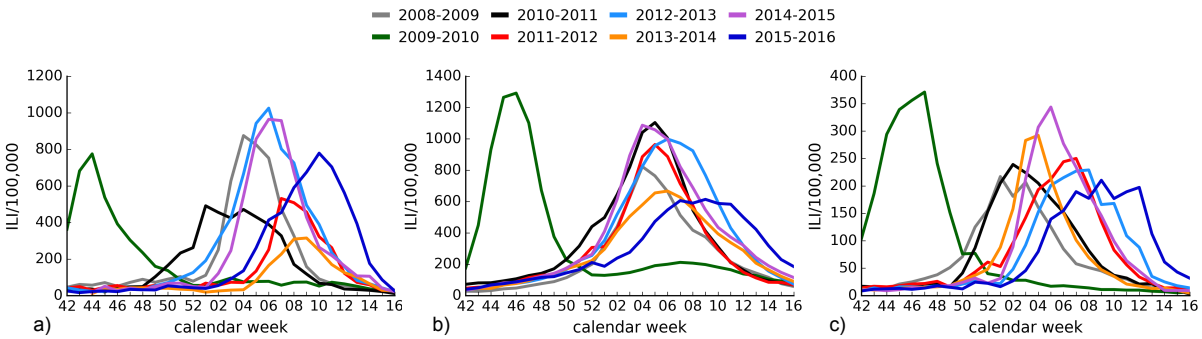


Figure 3.1: The influenza-like illness (ILI) consultation rates per 100,000 individuals as reported by the ECDC during selected influenza seasons in (a) Belgium, (b) Italy, and (c) Spain.

simulating the spatial and temporal evolution of epidemics at the level of single individuals. Such methodology has been previously validated in Ref. [204, 205] with different calibrations of the model through microblogging data from Twitter platform as well as participatory surveillance data from Influenzanet platforms. Results derived from this methodology have been made publicly available on an website, called FluOutlook (`fluoutlook.org`) [7], launched in 2014 as a joint collaboration between ISI Foundation in Torino, Italy, and Northeastern University in Boston, USA. FluOutlook is an online platform exposing real-time seasonal influenza forecasts obtained from our computational framework calibrated both with Twitter data and Influenzanet data as well as from some standard regression statistical models [204]. For each country and each season, FluOutlook reports in real-time the influenza intensity projected up to four weeks into the future, thus providing a description of the seasonal influenza progression that could be used by public health agency to guide their decision making process, as well as to compare and assess the performance of different forecast approaches.

The second section of this chapter is based on Ref. [168] in which we investigated how traditional surveillance data reported by general practitioners (GPs) can be combined by means of linear autoregressive models with digital surveillance data from the participatory Web-based system called Influweb, thoroughly described in Chapter 2, in order to improve seasonal influenza forecasts in Italy. Here we address the main issues affecting traditional surveillance systems (i.e. reporting lags and continuous revision of data) and we show how to exploit one of the main advantages of Influweb of having earlier data available.

Indeed, statistical and mechanistic models present very different features and level of knowledge on the disease and the biological mechanisms that drive the infection dynamics [136]. On the one hand, statistical models such as regression models generally ignore any details of the disease transmission, being based on historical data to provide short terms predictions of the ongoing season. Time series forecasting techniques can range from simple linear autoregressive models based on a combination of past values of the variable to be predicted and then increasingly add more details and accuracy, including seasonality. On the other hand, the mechanistic models are structured to make explicit assumptions about the biological mechanisms that drive infection dynamics, thus allowing for a better characterization of the disease, including the explicit estimation of key epidemiological parameters relevant to the public health decision-making that cannot be achieved with statistical models that do not consider the disease dynamic. However, in both approaches we were able to highlight the added value provided by integrating a digital real-time participatory component into seasonal influenza forecasting models.

3.2 Using a mechanistic model to forecast seasonal influenza

In this section, we present our work in forecasting seasonal influenza through a computational framework that combines real-time influenza-related data, traditional surveillance reports and a dynamical model for spatial epidemic spreading, called GLEAM (GLobal Epidemic And Mobility model), capable of providing short-term predictions of seasonal influenza activity by producing realistic simulations of the spatial and temporal evolution of the disease progression at the level of single individuals.

This section is based on Ref. [58] in which three different participatory disease surveillance systems, namely WISDM (Widely Internet-Sourced Distributed Monitoring), Influenzanet, and Flu Near You (FNY), have been investigated in the use of modelling, simulation, and forecasting. Here we report only our original contribution focused on the aforementioned computational framework calibrated with Influenzanet data for the 2015-2016 influenza season for six countries involved in the Influenzanet network (i.e. Belgium, Denmark, Italy, the Netherlands, Spain, and the United Kingdom). The methodology presented here has been previously validated in Ref. [204, 205] in which the model has been calibrated through microblogging data from Twitter platform as well as participatory surveillance data from Influenzanet platform.

3.2.1 Dataset

In the case of the seasonal influenza, we generally lack surveillance data at a high spatial resolution as traditional surveillance systems usually aggregate and report influenza incidence rates at the level of country or regions. In order to inform the model about the number of initial infections in each geographical area, we use a source of geo-localized data at a high spatial granularity that allows to estimate the relative incidence of influenza activity across regions at any given point in time. Such dataset corresponds to the participatory surveillance data from Influenzanet platforms. Launched in 2008, Influenzanet is a network of Web-based platforms for participatory surveillance of ILI present in 10 European countries [21]. Real-time information on population health is collected through weekly Internet-based surveys in which volunteers self-report their health status and consequent behaviour. The unique advantage provided by the self-reported information collected by the Influenzanet platforms consists in their high resolution both in time (daily collected) and space (at postal code level). Furthermore, Influenzanet data represent a high-specificity ground truth for the ILI incidence that cannot be obtained with any other source of information. More details about Influenzanet can be found in Chapter 2.

Users reporting a case of ILI are determined by applying the ECDC case definition [18] to the set of symptoms reported by the participants. In particular, the ECDC defines an ILI case as the sudden onset of symptoms with one or more systemic symptoms (fever or feverishness, malaise, headache, myalgia) plus one or more respiratory symptoms (cough, sore throat, shortness of breath). The weekly ILI incidence is determined by dividing the number of ILI cases by the number of active participants N in week w and postal code p, as $ILI_{p,w}/N_{p,w}$. When retrospectively analysing these data, inclusion criteria to determine active participants may vary and depend on the specific aim of the study [42, 64, 192, 193]. Here, as we are interested in extracting the number of infections early on in the first weeks of the influenza season, we consider as active those participants who completed at least one Influenzanet symptoms questionnaire during the period of interest. The full symptoms questionnaire can be found in Appendix A.

3.2.2 Methods

In order to provide real-time forecasts of seasonal influenza we combine digital indicators, surveillance reports and a mechanistic modelling approach in a four stages computational framework as presented in Figure 3.4. The mechanistic model, called GLEAM (Global Epidemic And Mobility model), is described in the following section, then follows the description of the methodology implemented at each stage of the framework.

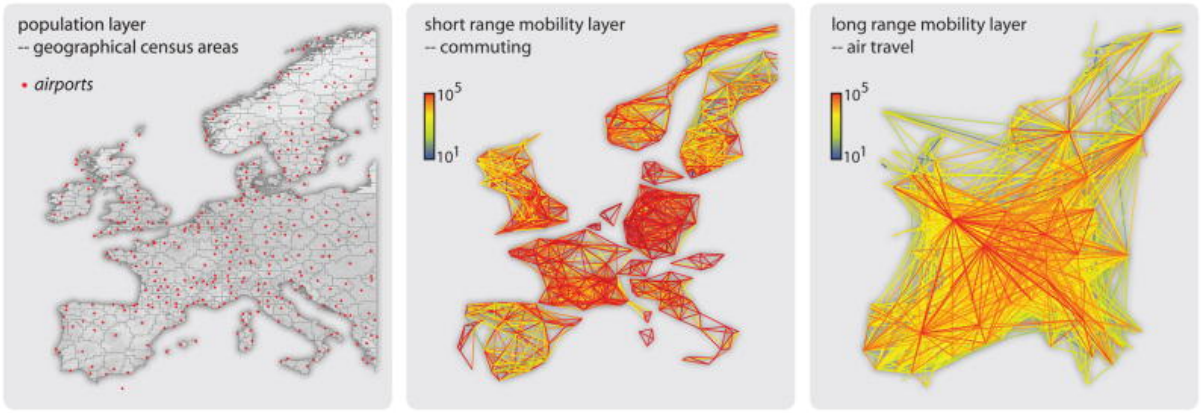


Figure 3.2: Population layer and mobility layer in GLEAM. The world surface is represented in a grid-like partition whose cells are assigned to the closest airport, thus defining the geographical census areas (i.e. subpopulations) of the metapopulation model. Subpopulations are connected by two mobility networks, the short range commuting network and the long range air travel network. Figure reproduced from [45].

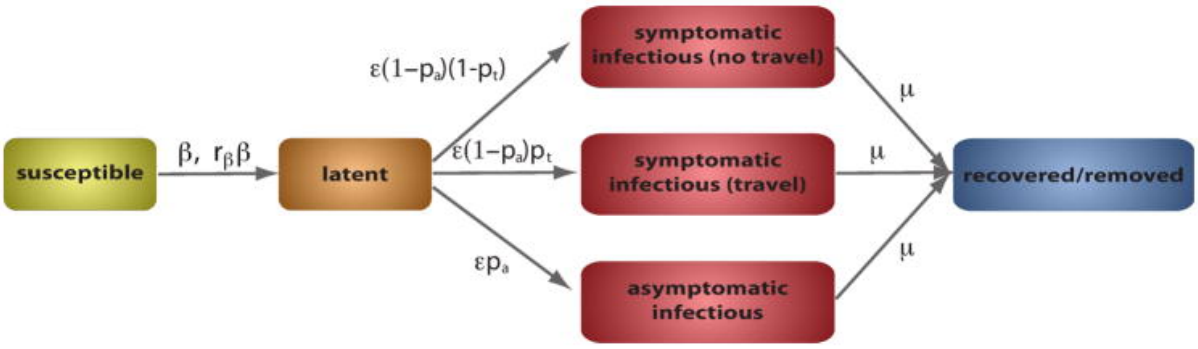


Figure 3.3: Compartmental structure of the epidemic model. Figure reproduced from [45].

3.2.2.1 The Global Epidemic and Mobility Model (GLEAM)

The Global Epidemic and Mobility model (GLEAM) [43, 45] is a stochastic generative model able to produce realistic simulations on the global spread of infectious diseases by combining real-world data on populations and human mobility and a detailed description of the specific characteristics and epidemic parameters of the disease with stochastic mathematical models of infectious dynamics simulating the spatio-temporal evolution of epidemics at the level of single individuals. GLEAM supports policy-making and emergency planning by developing epidemic models and scenario analysis able to gauge the actual threat of highly pathogenic diseases and minimize the impact of potentially devastating epidemics. In the event of epidemic outbreaks public health authorities can choose to activate prevention and emergency response policies, such as vaccination, travel restrictions, or school and business closures, but the socio-economic cost of such programmes can be high and their impact on the epidemic hard to determine. In this context, GLEAM allows the modelling of containment and mitigation strategies providing quantitative projections that better informs the analysis of their likely impact. GLEAM has been thoroughly tested and validated against historical outbreaks, including the 2002-2003 SARS epidemic [78], the 2009 H1N1 influenza pandemic [40, 41, 44, 46, 191], but also in a real-time fashion during the more recent outbreaks of Ebola in West Africa [161] and Zika in Latina America [206].

GLEAM is based on a metapopulation approach [123] whose subpopulations are defined through a Voronoi tessellation of the Earth surface that is represented in a grid-like partition

where each cell is approximately 25x25 km and corresponds to an estimated population value. Population data are obtained from the websites of the Gridded Population of the World and the Global Urban-Rural Mapping projects, run by the Socioeconomic Data and Application Center (SEDAC) of Columbia University. Each cell is assigned to the closest airport from the International Air Transport Association (IATA) [26] database, thus generating the 3,253 subpopulations of the metapopulation model, as shown in Figure 3.2. Groups of airports located close to each other (e.g. London is served by six airports, New York City by three and so on) are manually identified and assigned to the same subpopulation, hereafter called *basin*. Basins are connected by a network of human travel fluxes corresponding to transportation infrastructures and mobility patterns. In particular, GLEAM integrates two different types of mobility processes: a long-range mobility given by air travel data and a short-range mobility given by commuting data (see Figure 3.2). The flight network is derived from the worldwide booking datasets from the Official Airline Guide (OAG) database that contains more than 3,800 commercial airports in about 230 countries, and includes over 4,000,000 connections representing the estimated bookings between any two of these airports for each month. The commuting patterns are collected from the national statistics offices of more than 40 countries in five continents, covering more than 78,000 administrative regions. On the other hand, for those countries where real data are not available, synthetic commuting data can be generated by applying some mobility models, such as the gravity model [43] and the radiation model [182]. The mobility networks exhibit important variability both in the number of people travelling on each connection and in the total number of travellers per geographical census area, capturing the irregular network structure that affects the local and global diffusion of infections among subpopulations.

The infection dynamics occurs within each subpopulation and is governed by a disease-specific compartmental model in which individuals are classified according to their health status with respect to the disease, such as infected, susceptible, immune, etc., and can change health status according to the characteristics of the disease (incubation times, the proportion of asymptomatic yet infectious individuals, mortality rates and immunity, etc.) coupled with any prevention and intervention measures. The spread of the infection occurs through interactions between individuals belonging to the same subpopulation, while the infection is transmitted between subpopulations when people commute to work or school, or travel longer distances on national and international flights. The epidemic evolution is modelled using an individual dynamic where transitions are mathematically defined by chain binomial and multinomial processes [114] to preserve the discrete and stochastic nature of the processes. The model is fully stochastic and from any nominally identical initialization (initial conditions and disease model) generates an ensemble of possible epidemic evolution for epidemic observables, such as newly generated cases.

In the application to seasonal influenza, the disease dynamics is modelled with a Susceptible-Latent-Infectious-Recovered (SLIR) compartmental scheme, typical of ILI. Individuals can occupy one of these four discrete disease states at each discrete time step: susceptible individuals, S, who lack immunity against the infection, latent individuals, L, who are infectious but do not display the symptoms yet, infectious individuals who can transmit the infection, and removed/recovered individuals, R, who no longer have the infection. In addition, infectious individuals are further subdivided into 3 compartments: asymptomatic individuals, I^a, who do not display any symptoms of their illness, symptomatic individuals who can be allowed to travel, I^t, or not, I^{nt}. The condition of homogeneous mixing is assumed within each subpopulation, meaning that each individual of each subpopulation can interact and have contact with all the other individuals in a random way, without any specific characteristics related to their social network. Figure 3.3 shows the compartmental structure along the transitions between compartments. Susceptible individuals can acquire the infection through the interactions with infectious individuals either in their home subpopulation or in their neighbouring subpopulations on the mobility network. In particular, a susceptible individual in contact with a symptomatic or asymptomatic infectious person contracts the infection at a rate β or $r_\beta\beta$, respectively, and

enters the latent compartment. This distinction is due to the fact that an asymptomatic person is less infectious than a symptomatic one and in this case the transmission rate has to be reduced by a factor r_β=0.5 [46, 102, 138]. In addition, a fraction r of the population is considered to be not susceptible to the disease because of the residual immunity from previous seasons or vaccination. After the latency period ϵ^{-1}, each latent individual becomes infectious at a rate ϵ=1.5, entering the symptomatic compartments with probability $1 - p_a$ or the asymptomatic compartment with probability p_a. To reflect the changes of human traveling behaviour after the onset of symptoms, infected individuals with symptoms are further divided into two categories: those who can travel (I^t) with probability p_t=0.5 [46, 138], and those who are travel-restricted (I^{nt}) with probability $1 - p_t$. All the infected individuals permanently recover at a rate μ after the average infectious period, μ^{-1}, entering the recovered compartment, R.

The basic reproduction number R_0, defined as the average number of secondary cases generated by an infected individual in a fully susceptible population [37] for seasonal influenza might change from season to season according to the residual immunity r in a population. Thus, the effective reproduction number is defined to take into account this variability and takes the form $R^{\text{eff}} = (1 - r)R_0$, where $R_0 = (r_\beta p_a + (1 - p_a))\beta/\mu$.

3.2.2.2 Computational Framework

The computational framework consists of 4 main components: input, mechanistic modelling, model selection and output (Figure 3.4). In the following we detail the methodology implemented at each stage of the framework.

In the input component we estimate the initial conditions needed to inform the model about the number of initial infections so that it is able to numerically generate the epidemic progression by explicitly simulating the transmission. The number of infected individuals extracted from Influenzanet (see *Dataset*) in a given week w are mapped into the corresponding GLEAM basins k and used as seeds to initialize the simulations, as :

$$I^C_{k,w} = \left(\frac{ILI_{k,w}}{N_{k,w}}\right) \alpha_k Y \tag{3.1}$$

The coefficient α_k is the ratio of census population to the total number of Influenzanet participants that are estimated to live in the subpopulation k, thus taking into account the heterogeneous coverage of Influenzanet in a country [64, 147]. Y is a free parameter necessary to fine tune the rescaled number of ILI extracted from Influenzanet compared to the actual number of infected individuals, while w represents the starting week in the simulation, hereafter called *seeding week*. Since the ILI incidence rates reported by traditional surveillance systems are quite noisy at the early stages of the influenza season, the seeding week is shifted forward in time until the first week for which the ILI incidence crosses the epidemic threshold, that is specific for each country and seasonally evaluated [194].

In the second component, such initial conditions are used as seeds to initialize the simulations that generate an ensemble of possible epidemic evolution of the disease. GLEAM performs a Latin hypercube sampling of a 4-dimensional phase space defined by the vector $\overrightarrow{\theta} = r \times \mu \times R_0 \times Y$. The variable r describes the fraction of the population not susceptible to the disease due to the residual immunity and the fraction of vaccinated population. At the beginning of the influenza season, there is no explicit data about the fraction of population immunized due to exposure to the virus during the previous season or due to vaccination campaign, but in general this quantity varies from 25% to 45%, thus we consider $r \in [0.0, 0.45]$. The inverse of the recovery rate μ defines the infectious period that for influenza typically varies from 2 to 5 days, thus we consider $\mu \in [0.2, 0.5]$. The parameter β defines the disease transmission rate, and together with r and μ, is determined by the effective reproduction number, R^{eff}, that typically varies from 0.9 to 2.1 for seasonal influenza [71]. Here we consider $R^{\text{eff}} \in [0.8, 3.0]$. The parameter Y is a tuning parameter regulating the number of generated infected individuals

in each basin k. We consider $Y \in [10^{-6}, 10]$. Each range of these parameters is sampled with different resolutions resulting in a phase space formed by a total of 58,000 sampling points. For each sampled point, the model generates a statistical ensemble of 500 identically initialized Monte Carlo simulations providing for each subpopulation k the number of new flu cases in time $G_k(t : r, \mu, R_0, Y)$, among other indicators. Here, for sake of simplicity, we consider only the country level aggregation $G(t) = \sum_k G_k(t)$.

Since the simulated epidemic profiles $G(t)$ produce a large number of infected individuals, we rescale them according to the average of the ILI surveillance data from the last ten influenza seasons, excluding the season 2009-2010 because of the H1N1 pandemic. In particular, the rescaling considers both the average peak intensities and off-season average of the ILI surveillance data with variations in the rescaling factor of the order of the standard deviation.

Among these rescaled epidemic profiles, we use a statistical inference approach based on the Akaike Information Criterion (AIC) [33] to select only those models that are more likely according to the goodness of fit with the ILI surveillance data. In particular, given a set of candidate models, AIC provides an estimation of the information lost when the selected model is used to represent the real data. For each candidate model i, the AIC value is defined as:

$$AIC_i = N\, ln(\frac{RSS}{N}) + 2K \tag{3.2}$$

where N is the number of fitting points, K is the number of parameters, and RSS is the residual sum of squares of the vertical distances of the fitting points from the curve. The model associated with the smallest AIC value, indicated with AIC_{min}, minimizes the information loss to approximate the real data and represents the best model. For each candidate model i, we can define $\Delta AIC_i = AIC_i - AIC_{min}$, which is a measure that allows for an immediate ranking of each model with respect to the best model. According to some rules of thumb, models having $\Delta AIC_i < 2$ have substantial evidence to support their validity against data, values between 3 and 7 indicate that the model has considerably less support, whereas models with $\Delta AIC_i > 10$ indicates that they are very unlikely.

In our application, we select models with the minimum loss of information and the maximum likelihood with respect to the real data within a certain *training window*, shown in yellow in Figure 3.4(D). For each training window $x_0, x_1, ..., x_{T-1}$, this selected ensemble of models is then used to forecast the epidemic in the following weeks $T, T+1, ...$, and to estimate peak time and peak intensity of the influenza season. Such model selection technique is repeated week by week as new data from the ILI surveillance system is available, thus being more and more stringent as the season progresses and reducing uncertainties and stabilizing the forecast.

3.2.3 Results

In this section, we report the results obtained for the calibration of the model with Influenzanet data for the 2015-2016 influenza season for six countries of the Influenzanet network (i.e. Belgium, Denmark, Italy, the Netherlands, Spain, and the United Kingdom) [58]. Regarding the traditional surveillance data, we use the weekly ILI consultation rate per 100,000 individuals reported by ECDC.

Forecasts are provided along the entire season by considering x-week lead predictions (x-*wlp* for short) where $x \in [1, 4]$, thus predicting up to four weeks ahead of the release of the new surveillance report, but actually predicting three weeks into the future. Results for 1-*wlp* correspond to the nowcasting task, i.e. inferring the incidence value that the traditional surveillance systems will report in the following week. Figure 3.5 illustrates the values of the forecast along the entire season for 1-week, 2-week, 3-week, and 4-week lead predictions for the calibration with Influenzanet data. For all the countries under study the 95% confidence interval inferred from the selected ensemble of models is able to forecast the range of the entire epidemic profile. In fact, the empirical observations (i.e. the ground truth of the traditional surveillance data represented as black dots in the figure) lay within the confidence intervals for

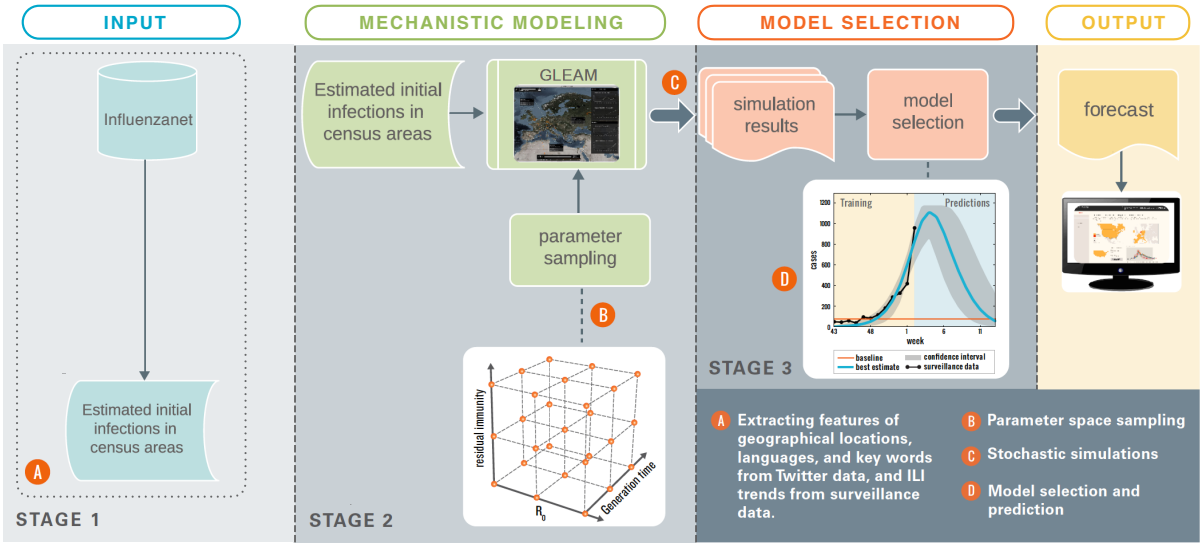


Figure 3.4: Computational forecast framework. Figure reproduced and updated from [205].

most of the weeks. To quantify the simulation's forecast performance, we compute the Pearson correlation and the mean absolute percentage error (MAPE) between each predicted time series and official surveillance time series. Moreover, in order to deepen the analysis of the accuracy of the model we report weekly forecasts for one of the main indicators of the influenza season, i.e. the peak week. Here we define the peak week accuracy as the percentage of the selected ensemble of simulations providing predictions within one week for peak time. In particular, Table 3.1 reports the values of the Pearson correlation and the mean absolute percentage error. As expected, the correlation generally decreases and the MAPE increases as the epidemic profiles are made by forecasts considering larger lead. Furthermore, the 2015-2016 influenza season was very mild in Europe and our method provides more reliable predictions with one or two weeks lead, mainly due to the fact that the dynamic pattern of the epidemic and the surveillance data signals are affected by large relative fluctuations. It is worth remarking that the ILI rate reported by official surveillance systems compounds together with the flu several other pathogens like rhinovirus and respiratory syncytial virus, thus the model should be generalized to multiple pathogens in order to improve the accuracy also at the very beginning and end of the season. In the case of Influenzanet, the quality of the forecasts is also affected by the noise of the corresponding ILI incidence curves, mainly due to the level of participation reached in each country. More information can be found in Chapter 2.

Table 3.2 shows the accuracy in predicting the peak week for the various countries under study. The peak accuracy of our predictions is consistently high for short-term predictions, but as already pointed out, the 2015-2016 influenza season was very peculiar and mild in Europe with an unusually late peak, and the forecasting accuracy depends on the severity of the season under consideration. Indeed, the large fluctuations in the surveillance data and the confounding factors introduced by other ILI pathogens that are predominant in mild influenza seasons reverberate in the ensemble selection process inducing unstable predictions.

3.2.4 Discussion

In this study, we presented a computational framework to forecast the unfolding of seasonal influenza combining digital ILI indicators, official surveillance data, and a data-driven mechanistic epidemiological model. This methodology has been previously validated adopting two different calibrations of the model: microblogging data from Twitter platform [205] and participatory surveillance data from Influenzanet platforms [58]. In particular, here we evaluated

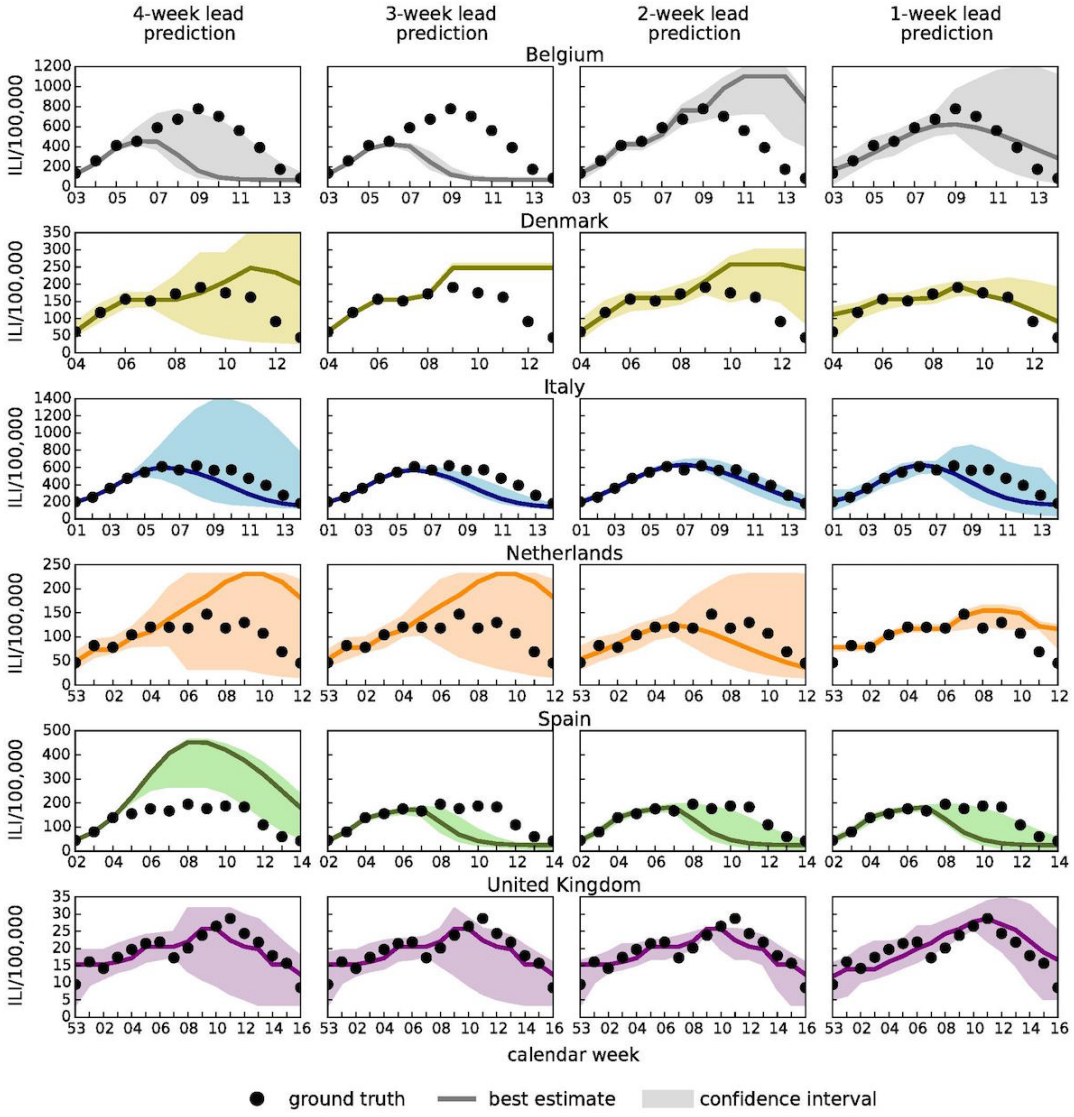


Figure 3.5: Influenzanet-based forecasts for the 2015-2016 influenza season in Belgium, Denmark, Italy, the Netherlands, Spain, and the United Kingdom, considering 4-week, 3-week, 2-week, and 1-week lead predictions. The best estimation (solid line) and the 95% confidence interval (shadow area) are shown together with official surveillance data (black dots).

its performance during the 2015-2016 influenza season in Belgium, Denmark, Italy, the Netherlands, Spain, and the United Kingdom. Remarkably, the predictions are in good agreement with real observations and the framework is able to provide forecasts that stabilize and become more accurate as the season progresses. Overall, from these analyses we can conclude that our methodology provides reliable predictions often with significant lead times. The quality of the forecasts is related to the intensity of the epidemic and for the 2015-2016 influenza season, that was particularly mild, we observed that the fluctuations in the signal generates a larger uncertainty in the predictions. The importance of accurate forecasts is certainly higher during severe influenza season and it is encouraging to observe that the forecast framework performs better exactly when it is most needed. Furthermore, our approach presents a number of advances with respect to previous work. In fact, we adopt an individual based data-driven epidemiological model in which the census areas are coupled by real mobility data, thus being able to capture the geospatial spreading patterns in a country. Moreover, we make use of digital ILI indicators only to initialize the epidemiological model with the result that the predictions are just indirectly influenced by such indicators, and the dependence from their availability is limited to

Table 3.1: Pearson correlations and mean absolute percentage errors (MAPE) as obtained by comparing the forecast results and the traditional surveillance data along the entire season in each country under study. Significant correlations (i.e. p$< 10^{-2}$) are indicated with *.

country	Pearson correlation				MAPE			
	4-wlp	3-wlp	2-wlp	1-wlp	4-wlp	3-wlp	2-wlp	1-wlp
Belgium	0.283	0.193	0.236	0.922*	43	46	149	39
Denmark	0.236	0.255	0.199	0.965*	60	75	76	25
Italy	0.879*	0.809*	0.973*	0.812*	15	20	6	18
Netherlands	0.387	0.387	0.763*	0.632	69	69	18	31
Spain	0.791*	0.494	0.521	0.521	124	35	34	34
United Kingdom	0.845*	0.845*	0.845*	0.810*	15	15	15	19

Table 3.2: Peak week accuracy defined as the percentage of the selected ensemble of epidemic profiles providing predictions within one week for peak time.

country	Peak week accuracy			
	4-wlp	3-wlp	2-wlp	1-wlp
Belgium	17	0	5	56
Denmark	12	0	84	74
Italy	56	0	100	61
Netherlands	59	60	34	83
Spain	100	83	97	100
United Kingdom	9	3	68	100

just few weeks in the early stages of the influenza season. Also, we are not limited to a specific source of digital ILI indicators as we demonstrated that the same methodology can adopt different data, being able to be extended to any other geo-localized data sources [205]. Traditional surveillance data are usually affected by reporting lags of at least one week and by weekly revision of the numbers initially released and generally lack a fine geographical resolution needed to inform high-resolution dynamical models such as GLEAM. Thus, geo-localized and real-time influenza-related indicators, such as data from social media and participatory systems, represent an important component in such modelling approach. Finally the presented framework can be easily refined with multiple data sources, such as weather data information, specific contact matrices and school calendars.

3.3 Using participatory Web-based surveillance data to improve seasonal influenza forecasting in Italy

In this section, we will describe the second approach we developed in the scope of real-time forescasting of seasonal influenza in Italy. This work is based on Ref. [168] in which we investigated how two data sources deriving from two different monitoring systems can be combined to improve seasonal influenza forecasts in Italy. The first one is Influnet [23], the national surveillance system for influenza syndromes in Italy, which is coordinated by the Italian National Institute of Health (ISS). The second one is Influweb [24], a Web-based participatory surveillance system, part of the Influenzanet network [21], that monitors ILI activity in Italy since 2008. Both systems have already been described in Caption 2, here we will briefly explaine how they have been used in this study.

Our approach is based on linear autoregressive models that integrate ILI prevalence reported by the traditional surveillance system and detected by the Influweb platform to generate predictions in a real-time fashion up to four weeks in advance. We retrospectively evaluate the predictive ability of our forecasting models over the course of four influenza seasons in Italy, from 2012-2013 to 2015-2016, for each of the four weekly time horizons.

3.3.1 Dataset

In this section, we describe the data collection for the two surveillance systems, highlighting their main characteristics and explaining how they have been used in our analysis. More details about these platforms can be found in Chapter 2.

3.3.1.1 Influnet

Influenza activity in Italy is officially monitored by the Italian National Institute of Health, "Istituto Superiore di Sanità", through a system called Influnet [23]. ILI incidence data reported by Influnet represent the ground truth for this study. The Influnet system collects data from a network of sentinel General Practitioners (GPs) and compiles a weekly report in which the national and regional incidence rates by age group are published during the winter season, generally from week 42 to week 17 of the calendar year. The system covers about 2% of the Italian population. Influnet data are published with at least one-week lag and typically new reports provide a first estimate of the weekly ILI incidence which is then updated in the following weeks as more data from sentinel GPs are recorded.

In this study, we distinguished between final *revised* reports, i.e. ILI data that are no longer being revised and are available only at the end of the influenza season, and weekly *unrevised* reports, i.e. ILI data that are actually available with one-week lag and subject to weekly revision until the end of the influenza season. We collected the Influnet revised and unrevised reports for five influenza seasons, from 2011-2012 to 2015-2016, from week 47 to week 17. Reports for weeks 2011-52 and 2013-51 were not available. Weekly reports released on week *WW* of the year *YYYY* are available at the following URLs: `http://www.iss.it/binary/iflu/cont/Influnet_YYYY_WW.pdf`. Final revised reports are available at the following URLs: `http://www.iss.it/flue/index.php?lang=1&anno=2016&tipo=13`. All reports were accessed and downloaded on September 1, 2016.

3.3.1.2 Influweb

Influweb collects weekly symptoms reports in Italy with the aid of self-selected volunteers from the general population [192]. Generally, the data collection is active from November to May during each influenza season. Participants, upon registration, are invited to provide some general background information (e.g. postal code of residence, their birth month and year, their vaccination status, their household composition etc.; further details about the background questions

Table 3.3: Participation to Influweb during the five influenza seasons under study.

Season	Number of registered users	Average number of weekly active users	Total number of symptoms surveys
2011-2012	2,270	1,085	14,681
2012-2013	3,057	1,165	15,789
2013-2014	3,513	1,244	18,471
2014-2015	3,753	1,290	20,224
2015-2016	4,053	1,224	20,823

are available upon request). Once they are enrolled in the system they are invited by means of a weekly e-mail reminder to report whether or not they experienced respiratory, gastrointestinal and systemic symptoms (see the full list in [64]).

ILI cases among participants are determined by applying the ECDC case definition [18] to the set of symptoms reported by volunteers. Accordingly, ILI is defined as the acute onset (within a few hours) of symptoms, including at least one among fever or feverishness, malaise, headache and myalgia and at least one among cough, sore throat and shortness of breath. The day of symptoms onset determines the day of ILI onset. Weekly ILI incidence is determined by dividing the number of ILI onsets by the number of active participants per week. When retrospectively analysing these data, inclusion criteria to determine active participants may vary and depend on the specific aim of the study [42, 64, 192, 193]. Here, we consider as active those participants who filled the background survey and at least two symptoms surveys, to avoid sporadic participation [64, 193]. ILI cases detected in the first symptoms survey were not taken into account because of a potential correlation between symptom presence and willingness to join the web-platform [32, 57, 193]. Participants are considered active for two weeks around the week of reporting.

The fact that volunteers are self-selected is a well known issue that usually affects participatory systems and causes the sample to be non-representative of the general population. In the case of the Influenzanet platforms, self-selection biases have been discussed and quantified in a previous work [64] in which all the relevant aspects for epidemiological analyses, including geography, mobility, demographic, socio-economic and health indicators, have been extensively explored and compared with national statistics. More recently, the representativeness of the users sample has been thoroughly investigated by restricting the analysis only to the Italian Web-platform [167]. In particular, the representativeness of the sample has been examined in terms of age, gender and geographic distribution with respect to the general Italian population and to the Influnet population sample during three influenza seasons, from 2012-2013 to 2014-2015. Results have shown that, despite the existing participation biases, the epidemiological signal extracted from Influweb correlates well with the ILI signal detected by the Influnet sentinel network, capturing both the timing and the relative intensity of seasonal flu activity in Italy. More details can be found in Chapter 2. In addition, differently from Influnet, Influweb data are available in real-time and the system can detect ILI cases from the general population and not only from medically attended patients.

Here, we have used Influweb data from five influenza seasons, from 2011-2012 to 2015-2016. The data collection period generally starts at the beginning of the influenza season in November, through a first e-mail reminder to the participants already enrolled in the system and a number of press releases to attract new volunteers among the general population. Thus, the system needs some weeks to reach a stable cohort of participants and early data are usually noisy and must be discarded to avoid biases due to a variable sample of reporting participants. To address this issue, for each influenza season we have selected a starting week according to a threshold on the number of active participants of at least 25% of the total number of participants in the previous season. Accordingly, the influenza seasons under study start on weeks 2011-47, 2012-48, 2013-47, 2014-47 and 2015-47. We have also tested different thresholds and 25% has been found as the optimal value to remove the initial noise and obtain the largest number of data points at

the same time, for all the seasons under study.

In Table 3.3, we report the number of registered users, the average number of weekly active users and the total number of symptoms surveys for all influenza seasons considered in this work from the starting week until the last week of active GP surveillance.

Data collected from the Influweb platform are gathered according to the Italian regulations on privacy which states that only aggregated and anonymized data can be published and shared. Raw data are available upon request from third parties wishing to conduct scientific research and upon discussion with other members of the Influenzanet Scientific Committee. Weekly incidence data aggregated at country level for the current influenza season (2016-2017) are publicly available at the following URLs (in Italian): `http://www.epicentro.iss.it/problemi/influenza/flunews.asp`. Weekly incidence data used in this study are available at the following URLs: `https://www.influweb.it/it/dati/`.

3.3.2 Methods

In this section we describe three models we have used to produce seasonal influenza forecasts: one model is based on data from GP surveillance reports only, while the other two combine data from both traditional influenza surveillance and Web-based participatory surveillance. We also present evaluation metrics we used to quantify the forecasting accuracy of the three models over the four influenza seasons under study.

Following the same methodology adopted in previous works [131, 164, 175], we considered as our reference a baseline model that considers only ILI data from the traditional GP surveillance Influnet. The baseline model consists of a linear autoregressive model with three weekly lagged components (AR3) as independent variables and takes the form:

$$y_{w+k} = \alpha_1 \widetilde{y}_{w-1} + \alpha_2 \widetilde{y}_{w-2} + \alpha_3 \widetilde{y}_{w-3} \tag{3.3}$$

where y_w denotes the ILI incidence value at week w and y indicates the value reported by the *revised* reports, while $\widetilde{y}$ indicates the *unrevised* report that is most recent at the time of the forecast.

According to the value of k, a distinction can be made between *nowcasting* ($k = 0$), i.e. inferring the present ILI incidence value that Influnet will report in the following week, and *forecasting* ($k > 0$), i.e. predicting the ILI incidence value in k weeks. ILI predictions generated with the baseline model are then contrasted with those produced by the forecasting models that integrate data from participatory surveillance, to assess their added value.

As mentioned above, one advantage of the participatory system Influweb is that data are available in real-time, not just for the previous week as for Influnet. Thus, we used a forecasting model that integrates the real-time ILI signal detected by Influweb into the epidemiological signal of Influnet. First, we produced ILI predictions by including the previous three weeks of ILI incidence reported by Influnet and the Influweb signal for the current week in a linear autoregressive exogenous (ARX) model, as described in [164]. This model, hereafter called ARX_{1w}, takes the following form:

$$y_{w+k} = \sum_{i=1}^{3} \alpha_i \widetilde{y}_{w-i} + \gamma_0 \widetilde{z}_w \tag{3.4}$$

Then, we further extended the ARX model by adding other three weekly lagged terms of the Influweb ILI incidence, (hereafter called ARX_{4w}), as follows:

$$y_{w+k} = \sum_{i=1}^{3} \alpha_i \widetilde{y}_{w-i} + \sum_{j=0}^{3} \gamma_j \widetilde{z}_{w-j} \tag{3.5}$$

In the previous models, the regression coefficients α_i and γ_i are estimated separately for different values of k by a least-squares regression.

We considered the 2011-2012 influenza season as a training set for all the models and the subsequent four seasons, from 2012-2013 to 2015-2016, for validation purposes. The goal was to use all available information, at a given point in time, to produce accurate predictions of the ILI incidence one, two, three and four weeks ahead of the release of Influnet reports, effectively predicting ILI three weeks into the future. To this aim, the training set was dynamically increased to include all available information at a given week. In other words, for our first prediction on week 2012-48, the training set included 23 weeks of historical data, for the second prediction on week 2012-49, the training set consisted of 24 weeks, including the values at week 2012-48, and so on for all subsequent weeks.

For each forecasting model we report four different evaluation metrics: Pearson correlation, mean absolute error (MAE), root mean squared error (RMSE) and mean absolute percentage error (MAPE). The definitions of all evaluation metrics are given below, conforming to the following notation: x_i indicates the predicted value at time i, y_i corresponds to the ground truth value at time i, and, finally, $\overline{x}$ and $\overline{y}$ denote the average of the values $\{x_i\}$ and $\{y_i\}$, respectively.

- The Pearson's Correlation is a measure of the linear dependence between two variables and is defined as:

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{3.6}$$

- The Mean Absolute Error is a measure of the average of the absolute errors between predicted and true values and is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \, y_i - x_i \, | \tag{3.7}$$

- The Root Mean Squared Error (RMSE) is a measure of the difference between predicted and true values and is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \tag{3.8}$$

- The Mean Absolute Percentage Error (MAPE) is a measure of prediction accuracy of a forecasting method and is defined as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - x_i|}{y_i} \times 100 \tag{3.9}$$

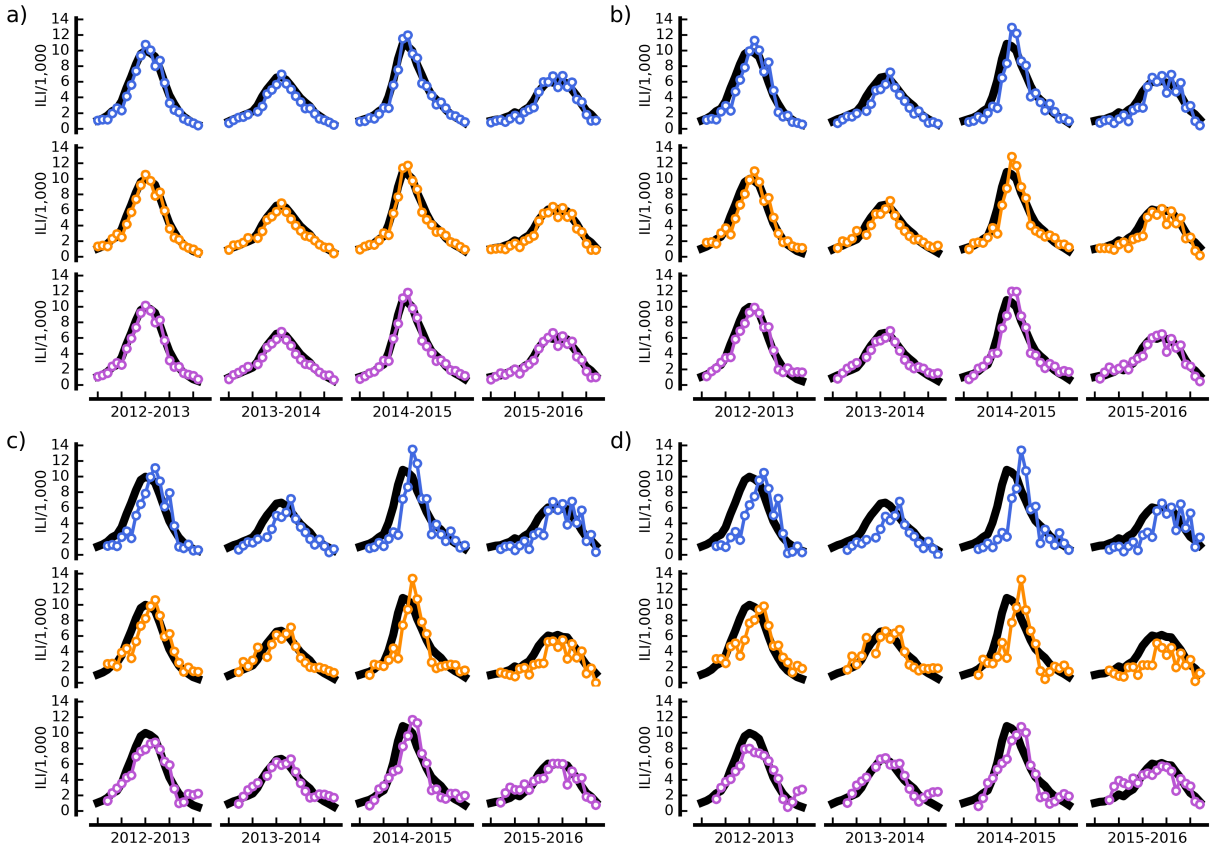
These similarity metrics were calculated for the time period from the beginning of the 2012-2013 influenza season, week 2012-48, to the end of the 2015-2016 influenza season, week 2016-16.

Moreover, for each forecasting model we evaluate both the timing and magnitude of the peaks in the influenza seasons with respect to the ground truth. In particular, for each value of k, we report the number of weeks lag between the predictions and the ground truth and the percent error (PE) around the peak, that is a measure of the discrepancy between predicted and true values and is defined as:

$$PE = \frac{|y_i - x_i|}{y_i} \times 100 \tag{3.10}$$

The peak analysis is performed separately for each of the four influenza seasons, but for simplicity's sake, we report only the minimum and maximum number of weeks lag and the average together with the minimum and maximum percent error around the peak intensity.

Figure 3.6: Comparison between the ground truth (black) and the forecasting models: baseline model (blue), ARX_{1w} model (orange) and ARX_{4w} model (purple), for the four time horizons: a) k=0; b) k=1; c) k=2; d) k=3.



3.3.3 Results

Figure 3.6 shows the ILI predictions, while Figure 3.7 shows the weekly errors associated to each model. Table 3.4 displays all the values of the evaluation metrics and the peak analysis obtained by comparing the baseline model and the forecasting models ARX_{1w} and ARX_{4w} for the four weekly time horizons, corresponding to $k = 0, 1, 2, 3$, against the ground truth. The Pearson's correlation increases with respect to the baseline model of about 0.2% and 0.3%, respectively for the ARX_{1w} and the ARX_{4w} models, for $k = 0$; 0.8% and 2.5% for $k = 1$; 3.4% and 9.9% for $k = 2$; 10.9% and 29.9% for $k = 3$. The MAE decreases with respect to the baseline model of about 6.7% and 13.3%, respectively for the ARX_{1w} and the ARX_{4w} models, for $k = 0$; 12.0% and 24.1% for $k = 1$; 18.6% and 34.9% for $k = 2$; 21.1% and 43.3% for $k = 3$. The root mean squared error (RMSE) decreases with respect to the baseline model of about 9.0% and 15.2%, respectively for the ARX_{1w} and the ARX_{4w} models, for $k = 0$; 13.8% and 28.8% for $k = 1$; 17.2% and 37.8% for $k = 2$; 20.1% and 46.2% for $k = 3$. The mean absolute percentage error (MAPE) increases with respect to the baseline model of about 0.2% and 1.5%, respectively for the ARX_{1w} and the ARX_{4w} models, for $k = 0$; 8.6% and 14.5% for $k = 1$; for $k = 2$, the MAPE decreases of about 3.2% for the ARX_{1w} model, while it increases of about 2.5% for the ARX_{4w} model; finally, for $k = 3$, the MAPE decreases of about 4.4% and 5.1%, respectively for the ARX_{1w} and the ARX_{4w} models.

As observed in Figure 3.6, the predictions curves produced by the baseline model undergo a slight shift forward. This trend is then corrected by incorporating a source of data that integrates the current value of the ILI incidence and contributes to readjust the predictions and partially remove the shift.

Table 3.4: Similarity metrics and peak analysis of the forecasting models with respect to the ground truth. The best performing model per metric is bold faced.

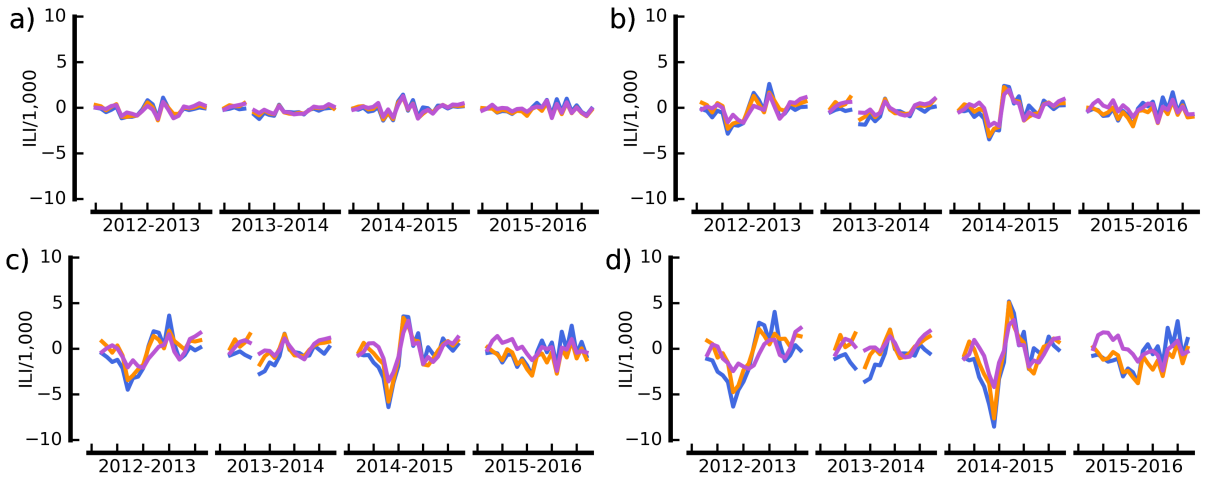
		Similarity Metrics				Peak Analysis	
	Model	CORR	MAE (ILI/1,000)	RMSE (ILI/1,000)	MAPE (%)	weeks lag [Min, Max]	PE (%) Mean [Min, Max]
	baseline	0.981	0.45	0.5955	**12.93**	[0, 1]	8.6 [4.8, 10.9]
$k = 0$	ARX_{1w}	0.983	0.42	0.5419	12.95	[0, 1]	5.5 [3.5, 7.7]
	ARX_{4w}	**0.984**	**0.39**	**0.5052**	13.13	[0, 1]	5.6 [1.9, 9.0]
	baseline	0.934	0.83	1.1174	**22.39**	[1, 2]	13.7 [8.9, 19.3]
$k = 1$	ARX_{1w}	0.942	0.73	0.9635	24.31	[1, 2]	9.3 [0.8, 18.1]
	ARX_{4w}	**0.957**	**0.63**	**0.7952**	25.63	[0, 1]	5.5 [0.7, 10.6]
	baseline	0.835	1.29	1.7497	33.79	[2, 3]	13.8 [7.7, 24.3]
$k = 2$	ARX_{1w}	0.864	1.05	1.4488	**32.72**	[2, 3]	11.7 [6.5, 23.0]
	ARX_{4w}	**0.918**	**0.84**	**1.0885**	34.62	[0, 2]	5.2 [0.3, 12.0]
	baseline	0.677	1.80	2.4287	45.23	[0, 3]	9.7 [2.3, 23.2]
$k = 3$	ARX_{1w}	0.751	1.42	1.9404	43.23	[0, 3]	10.7 [1.7, 22.1]
	ARX_{4w}	**0.879**	**1.02**	**1.3058**	**42.93**	[0, 3]	7.2 [0.6, 19.7]

In the nowcasting task ($k = 0$) the three forecasting models show a comparable performance in predicting the peak with 0 or 1-week lag. The ARX_{1w} model predicts the peak with the same range of weeks lag with respect to the baseline model, but with a lower percent error around the peak, for all the four weekly time horizons. The ARX_{4w} model predicts the peak with a greater accuracy with respect to the baseline model and the ARX_{1w} model for $k = 1, 2$ and with a lower percent error around the peak, for all the four weekly time horizons. As expected, the ability of the forecasting models to capture the peaks in the influenza seasons decays as the time horizon increases, as shown in Figure 3.7.

3.3.4 Discussion

In this study, we have investigated how real-time forecasts of seasonal influenza activity in Italy can be improved by integrating surveillance data from the GP based monitoring system, called Influnet, with data from a Web-based participatory platform, called Influweb. Indeed, ILI incidence estimates produced by traditional surveillance systems such as Influnet undergo weekly revisions during the influenza season as more data from sentinel doctors are collected and official reports are usually released with at least one-week lag. On the other hand, participatory surveillance data such as those produced by Influweb are available as soon as participants report their health status. To produce accurate seasonal influenza forecasts, we have leveraged on this real-time component and built two forecasting models incorporating the weekly lagged terms of the ILI incidence of both surveillance systems in a linear autoregressive exogenous (ARX) model. One is the ARX_{1w} model, which combines incidence data from GP surveillance with a single term of the Influweb incidence for the current week. The second is the ARX_{4w} model which integrates incidence data from traditional surveillance with four weekly lagged terms of the Influweb incidence. We have found that both models outperform the predictions based on GP surveillance data only. The ARX_{4w} model achieves the best performance by predicting the unfolding of the influenza epidemic four weeks in advance and showing the best accuracy in terms of lowest MAE and RMSE and highest correlation with the ground truth across all time horizons. These results highlight the added value provided by the integration of a digital

Figure 3.7: Errors associated with each forecasting models, baseline model (blue), ARX_{1w} model (orange) and ARX_{4w} model (purple), are displayed for the four time horizons: a) k=0; b) k=1; c) k=2; d) k=3.



real-time participatory component into seasonal influenza forecasting models. For sensitivity analysis, we also tested the performance of a linear autoregressive model with three weekly lagged components (AR3) based only on Influweb ILI incidence. This model resulted to be less accurate than the baseline model (results not presented) showing that, although ILI retrospective estimates tracked by Influweb correlate well with the ground truth data, a model relying only upon Web-based surveillance data is not able to capture the main indicators of the epidemic season.

Our study relies on the assumption that ILI incidence reported by GP surveillance represents the best available measure of influenza activity in Italy. Such assumption may not be completely accurate, as GP surveillance can sometimes over- or underestimate the true burden of the epidemic, depending on what ILI fraction corresponds to real influenza cases and on the reporting rates of sentinel medical doctors [141]. However, sentinel incidence data is considered a highly reliable indicator of influenza activity and adjusting sentinel data for under-ascertainment would require additional information, such as consultancy rates by age groups or reporting rates by GPs, which is usually not available.

The advantages of an integrated framework have been explored also in previous works [142, 164, 175] in which it has been shown that, at least in the United States, the most effective approaches aimed at improving influenza forecasts combine information from multiple flu predictors extracted from various Web data sources such as Twitter, Google Trends, Wikipedia, including data from participatory surveillance. In countries where the penetration of social media is low or Natural Language Processing algorithms do not reach the same accuracy as for the English language and such abundance of different data sources may not be harnessed with equal efficacy, participatory surveillance systems remain not only an important tool to measure the levels of flu activity but also a data source that, combined with the traditional GP surveillance, can be used to provide real-time accurate forecasts.

In the next future, we will extend the present work to include all the countries in which the Influenzanet project has been deployed so far. Another promising extension of our study lies in the integration of other Web sources, such as Twitter or Wikipedia, into our framework. While the majority of the existing studies are focused on the United States and the English language, it would be of interest to understand how much real-time flu forecasts can be improved in countries like Italy, where the penetration of Twitter is more limited, and how this compares to the results obtained by assimilating participatory surveillance data into predictive models.

Traditional surveillance of seasonal influenza is generally affected by reporting lags of at least one week and by continuous revisions of the numbers initially released. As a consequence, influenza forecasts are often limited by the time required to collect new and accurate data. On the

other hand, the availability of novel data streams for disease detection can help in overcoming these issues by capturing an additional surveillance signal that can be used to complement data collected by public health agencies. In this study, we investigate how combining both traditional and participatory Web-based surveillance data can provide accurate predictions for seasonal influenza in real-time fashion. To this aim, we use two data sources available in Italy from two different monitoring systems: traditional surveillance data based on sentinel doctors reports and digital surveillance data deriving from a participatory system that monitors the influenza activity through Internet-based surveys. We integrate such digital component in a linear autoregressive exogenous (ARX) model based on traditional surveillance data and evaluate its predictive ability over the course of four influenza seasons in Italy, from 2012-2013 to 2015-2016, for each of the four weekly time horizons. Our results show that by using data extracted from a Web-based participatory surveillance system, which are usually available one week in advance with respect to traditional surveillance, it is possible to obtain accurate weekly predictions of influenza activity at national level up to four weeks in advance. Compared to a model that is only based on data from sentinel doctors, our approach significantly improves real-time forecasts of influenza activity, by increasing the Pearson's correlation up to 30% and by reducing the Mean Absolute Error up to 43% for the four weekly time horizons.

Seasonal influenza epidemics occur annually during winter in temperate regions, resulting in around 3-5 million cases of severe illness and 250,000-500,000 deaths worldwide each year [19]. Accurate influenza incidence predictions are needed to estimate in advance, rapidly and reliably, the number of influenza cases and to properly prepare for and respond to the unfolding of the influenza epidemics. In developed and developing countries, national syndromic (i.e. based on observed symptoms) surveillance systems monitor levels of influenza-like illness (ILI) cases among the general population by gathering information from physicians, known as sentinel doctors, who record the number of people seeking medical attention and presenting ILI symptoms. These traditional surveillance systems for seasonal influenza are generally affected by reporting lags of at least one week and by continuous revision of the numbers initially released [67]. Thus, influenza forecasts are often limited by the time required to collect new, accurate data.

The pervasive use of digital communication technologies for public health [174] and the increasing community engagement in public health have fostered the birth of surveillance systems based on the possibility for single individuals to monitor and report their own health status through Web-based platforms [201]. These so-called participatory surveillance systems aim at capturing seasonal influenza activity directly from the general population through Internet-based surveys. Participatory surveillance systems for seasonal influenza are currently running in 13 countries worldwide and collect, aggregate and communicate data in real time during the course of every influenza season. Specifically, the systems that are currently online are: Influenzanet, a network of Web platforms running in eleven European countries [21], FluNearYou in the United States [74, 86, 187] and FluTracking in Australia [65, 66, 88, 89, 160]. Participatory surveillance systems have proven to be accurate and reliable for ILI surveillance, as the detected timing and relative intensities of influenza epidemics are consistent with those reported by sentinel doctors [86, 88, 158, 167, 192]. Furthermore, it has been shown that participatory surveillance systems can also provide relevant information to estimate age-specific influenza attack rates [75, 162, 167] and influenza vaccine effectiveness [66, 96], to assess health care usage [189], and to estimate risk factors for ILI [32].

Previous studies have developed a range of model-inference systems to generate real-time influenza forecasts [180], some of them exploiting the availability of large scale Web data from search queries and social media, such as Google [109, 176], Yahoo [171], Twitter [56, 69, 87, 129, 163, 164, 181, 204], Wikipedia [108, 120, 142] and also by including participatory surveillance data in their models [175]. However, the large majority of published studies have focused on forecasting the influenza activity in the United States only. This is mainly due to the availability of advanced Natural Language Processing algorithms for the English language and the easy access to ILI data provided by the Centers for Disease Control and Prevention (CDC),

also in the context of the "Forecasting influenza season challenge" [53]. To what extent similar forecasting approaches can be extended to non-native English speaking countries remains largely unexplored, with a few exceptions [204].

In this study, we combine traditional surveillance data and Web-based participatory surveillance data to improve forecasts of seasonal influenza activity in an European country, Italy, for which such a forecasting approach has never been tested. To this aim, we use two influenza activity monitoring systems. The first one is Influnet [23], the national surveillance system for influenza syndromes in Italy, which is coordinated by the Italian National Institute of Health (ISS). The second one is Influweb [24], a Web-based participatory surveillance system, part of the Influenzanet network, that monitors ILI activity in Italy since 2008.

Our approach is based on linear autoregressive models that integrate ILI prevalence for the current week as reported from the Influweb platform to generate predictions in a real-time fashion up to four weeks ahead of the release of Influnet report. We retrospectively evaluate the predictive ability of our forecasting models over the course of four influenza seasons in Italy, from 2012-2013 to 2015-2016, for each of the four weekly time horizons.

Chapter 4

Modelling the epidemic spreading of Zika using mobile phone data

> Not a single year passes without [which]...
> we can tell the world: here is a new disease!
>
> — Rudolf Virchow, 1867

A large outbreak of Zika started in Brazil in July 2015 recently hit a total of 48 countries and territories in the Americas causing up to the end of 2016 more than 700 thousand cases, among which about 25% were laboratory confirmed, resulting in a cumulative incidence rate of about 71 cases/100,000 population, strongly affected by underreporting due to the clinical similarities of mild illness symptoms associated with Zika, asymptomatic cases, limited sentinel sites, and medically unattended cases. In this chapter, I will focus on the epidemic of Zika occurred in Colombia in 2015-2016, mainly studying the role of human mobility in a metapopulation modelling approach based on real data on population and a detailed description of the epidemiological characteristics of the disease, as well as informed by real data on the number of cases occurred in the early stage of the epidemic. In particular, I will investigate different mobility networks, mainly focusing on the potential benefits of integrating human mobility patterns given by mobile phone data, as well as human mobility generated by mobility models (i.e. the gravity model and the radiation model) and census commuting data. In this study I personally carried out the collection, analysis and visualization of the data, as well as generating the mobility networks and writing the code of the epidemic model and performing the analysis of the simulations.

The research presented here is the result of a collaboration with United Nations Global Pulse in New York, USA, where I spent three months from June to August 2017 for a summer internship in the Data Science group. Global Pulse was launched in 2009 as an innovation initiative of the United Nations Secretary-General with the mission to accelerate discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action.

4.1 Introduction

The Zika virus (ZIKV) is a mosquito-borne flavivirus that was first identified in the Zika Forest of Uganda in 1947 in monkeys and later in humans in 1952. ZIKV is primarily transmitted by infected *Aedes* mosquitoes [70, 111], which are also responsible for the transmission of dengue, chikungunya and yellow fever. However, other ways of transmission have been reported, such as sexual and perinatal transmission [52, 92, 146, 202] and blood transmission through blood transfusion [149]. It generally causes mild symptoms, such as fever, skin rashes, conjunctivitis, muscle and joint pain, malaise, and headache, although almost 80% of infections are asymptomatic [94]. Being a mild disease, no specific treatment is required, but only get plenty of rest,

drink enough fluids, and treat pain and fever with common medicines.

Until recently, ZIKV was considered a neglected tropical disease with only local outbreaks, typically accompanied by mild illness. Underreporting, due to the clinical similarities of (mild) illness symptoms associated with Zika, dengue, and chikungunya infections might also account for previous Zika outbreaks being overlooked. The first large outbreak of disease caused by Zika infection was reported in 2007 in the Pacific Island of Yap in the Federated States of Micronesia with an estimate of 73% of Yap residents infected with Zika virus. In 2013-2014 outbreaks occurred in 4 other groups of Pacific islands (French Polynesia, Easter Island, the Cook Islands, and New Caledonia) with thousands of suspected infections in French Polynesia and possible associations between Zika virus and congenital malformations and severe neurological and autoimmune complications. Lastly, during the recent outbreak in Latin America started in Brazil in July 2015, WHO declared a Public Health Emergency of International Concern (PHEIC) [155] on February 2016 (and that lasted for nearly 10 months) due to the increased incidence of neurological complications, including microcephaly in newborns and Guillain-Barré syndrome, associated to ZIKV infections. The epidemic reached a total of 48 countries and territories in the Americas causing up to the end of 2016 more than 700 thousand cases, of which about 25% were laboratory confirmed.

In this study we focus on the Zika outbreak occurred in Colombia that, following Brazil, was the second earliest country that experienced a large-scale ZIKV epidemic in Latin America [1]. From week 2015-32 to week 2016-30, when health officials declared the end of the Zika epidemic in Colombia, the Colombian Instituto Nacional de Salud (INS) [25] reported a total of 100,888 cases (about 9% were laboratory confirmed) corresponding to a cumulative incidence rate of 214 cases/100,000 population. In particular, here we investigate the role of human mobility in the epidemic spreading of Zika in Colombia, with a special focus on the benefits of integrating human mobility patterns provided by Call Detail Records (CDRs) in an epidemic modelling approach. Recently, the latest global public health threats have highlighted the urgent need for accurate human mobility data to properly inform epidemic models and assess the spatial spreading of diseases and the risk of importation from the affected areas to the rest of the world, thus allowing for rapid interventions and appropriate control measures [54, 115, 169, 196]. In developed countries, mobility data is usually available and easily accessible from official sources for airline transportation, train trips, or commuting, but such datasets may be not updated for several years or aggregated at a wide geographical resolution, often limiting the potential impact of many studies. Moreover, this information may be inadequate or completely unavailable in developing countries. Mobility models, such as the gravity model [43] and the radiation model [182], come in help by inferring local and global human mobility flows on a synthetic network whose connections and the corresponding intensity represent the flow of people among different regions. On the other hand, the limited generalizability of mobility models whose use can be hindered in the absence of good calibration data, led to the study of the digital traces left by humans over the Internet (e.g. Twitter, Flickr, Foursquare) and the footprints left by mobile phone users whenever they make a call or send an SMS. In particular, CDRs data carry temporal and spatial information on the position of mobile phone users at the level of tower signal cells and can provide a proxy for humans physical displacements over time that may be used as a primary data source for early response in case of outbreak. The availability of human mobility data at such high resolution has impacted several research fields, ranging from urban planning to social sciences [36, 38, 63, 76], but more important their application to the spatial epidemiology of infectious diseases [51, 101, 157, 190, 197, 198, 199, 200].

In this study, we develop a metapopulation model directly informed with the reported Zika cases in Colombia in order to analyse the spatiotemporal spread of the epidemic, accounting for detailed population data and human mobility pattern. Next to mobile phone data, we consider different synthetic networks of human movements across the country as well as census commuting data in order to investigate and evaluate the metapopulation epidemic outcomes obtained by integrating different mobility sources.

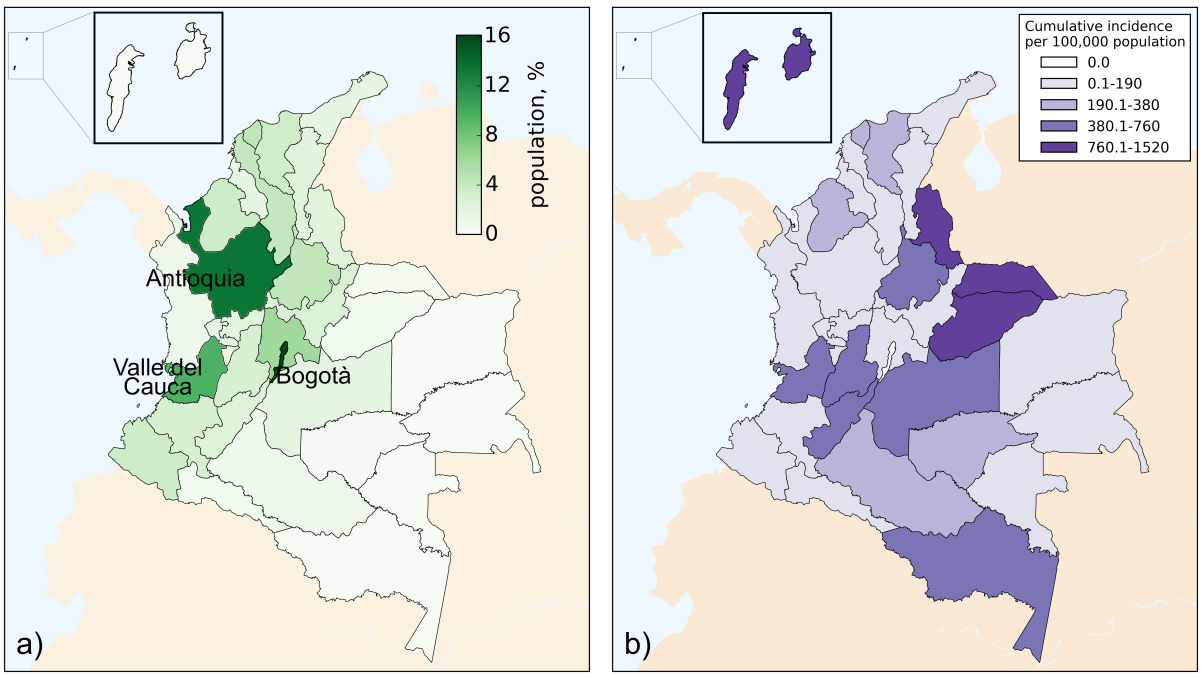


Figure 4.1: a) Population distribution: Bogotá, Antioquia and Valle del Cauca are the most populated departments. b) Cumulative incidence rate per 100,000 population from week 2015-32 to week 2016-30.

4.2 Dataset

In this section, we give a description of Colombia in terms of geography, population and the epidemiological situation of the Zika outbreak. We then describe the mobile phone data, highlighting their main characteristics and explaining how they have been used in our analysis.

Population

Colombia is situated in the northwest of South America and is bordered to the northwest by Panama; to the east by Venezuela and Brazil; to the south by Ecuador and Peru. Archipelago of San Andres, Providencia and Santa Catalina consists of two island groups about 775 km northwest of mainland Colombia. Colombia is divided into 32 Departments and 1 Federal District, which is the country's capital, Bogotá. Departments are formed by a grouping of municipalities, 1,120 in total. The urban centres are mostly located in the highlands of the Andes mountains, thus most people live in the western portion of the country, while the southern and eastern portions are mostly sparsely inhabited with tropical rainforest, large livestock farms, oil and gas production facilities, small farming communities and indigenous tribes.

Population data have been obtained from WorldPop [30] that provides estimates of population counts and densities for multiple years per 100x100m grid cells through integrating census, survey, satellite and GIS datasets in a flexible machine-learning framework. In this study, we used the dataset lastly updated in 2015 by projecting population estimates on the administrative resolution of departments (Figure 4.1). About 47 million people live in Colombia according to this dataset, mostly distributed in the departments of Bogotá (about 16%), Antioquia (about 13%) and Valle del Cauca (about 10%), as reported in Table 4.1.

Zika case data

Although preliminary monitoring began in Colombia after the recognition of the outbreak in Brazil, the Colombian Instituto Nacional de Salud (INS) [25] began official surveillance for ZIKV

Table 4.1: Administrative subdivision of Colombia in 33 departments with the population, the cumulative number of Zika cases and the incidence rate reported from week 2015-32 to week 2016-30 by the Instituto Nacional de Salud (INS). Departments are sorted according to the population size.

Code	Department	Population (%)	Zika cases (confirmed)	Incidence Rate (cases/100,000)
11	Santafe De Bogotá D.C.	7,572,580 (16.07)	0 (0)	0
05	Antioquia	6,350,420 (13.48)	2,393 (335)	38
76	Valle del Cauca	4,525,666 (9.61)	25,341 (895)	560
25	Cundinamarca	2,840,969 (6.03)	5,212 (317)	183
08	Atlantico	2,090,062 (4.44)	6,627 (359)	317
68	Santander	2,065,317 (4.38)	9,570 (443)	463
13	Bolivar	1,970,779 (4.18)	1,885 (242)	96
52	Narino	1,690,071 (3.59)	67 (20)	4
23	Cordoba	1,655,242 (3.51)	3,195 (253)	193
47	Magdalena	1,619,426 (3.44)	3,233 (295)	200
73	Tolima	1,362,998 (2.89)	6,904 (822)	507
19	Cauca	1,362,202 (2.89)	300 (34)	22
15	Boyaca	1,300,493 (2.76)	346 (88)	27
41	Huila	1,153,011 (2.45)	6,767 (915)	587
54	Norte de Santander	1,112,279 (2.36)	10,112 (1,521)	909
20	Cesar	1,015,579 (2.16)	1,558 (245)	153
17	Caldas	994,946 (2.11)	277 (74)	28
66	Cisaralda	953,055 (2.02)	1,296 (130)	136
50	Meta	921,498 (1.96)	3,995 (580)	434
44	La Guajira	902,996 (1.92)	712 (95)	79
70	Sucre	851,034 (1.81)	1,610 (107)	189
63	Quindio	555,222 (1.18)	377 (24)	68
27	Choco	479,515 (1.02)	50 (5)	10
18	Caqueta	469,704 (1.00)	1,144 (234)	244
85	Casanare	347,198 (0.74)	3,827 (280)	1,102
86	Putumayo	324,807 (0.69)	510 (110)	157
81	Arauca	230,173 (0.49)	1,820 (191)	791
95	Guaviare	105,759 (0.22)	206 (15)	195
88	Archipelago of San Andres, Providencia and Santa Catalina	73,965 (0.16)	1,124 (66)	1,520
99	Vichada	68,010 (0.14)	75 (5)	110
91	Amazonas	65,264 (0.14)	329 (28)	504
97	Vaupes	42,915 (0.09)	13 (0)	30
94	Guainia	35,059 (0.07)	13 (3)	37
	Colombia	4,7108,214 (100)	100,888 (8,731)	214

in August 2015. In early October 2015, a Zika outbreak was declared after the first cluster of laboratory-confirmed cases was identified in nine patients from northern Colombia. After that report, the INS retrospectively identified Zika virus in an archived serum sample from July 2015. In July 2016, INS declared the end of the Zika epidemic in Colombia.

In this study we use the weekly epidemiological reports published from week 2015-32 to week 2016-30 and available at the following URL: `http://www.ins.gov.co/boletin-epidemiologico/Paginas/default.aspx`. All reports were accessed and downloaded on June 30, 2017. Each re-

Table 4.2: Basic properties of the weekly OD matrices at the municipality level generated from mobile phone data. Table includes the weekly number of nodes, links and total volume of trips, and some statistics on the distribution of trips T_{ij} and active phones n_i, including minimum, maximum and average values.

week	nodes	links	volume	$T_{ij_{min}}$	$T_{ij_{max}}$	$\overline{T}_{ij}$	$n_{i_{min}}$	$n_{i_{max}}$	$\overline{n}_i$
2013-49	786	41,805	4,005,184	1	174,714	96	1	489,267	5,060
2013-50	787	46,096	4,210,148	1	175,590	91	1	504,840	5,266
2013-51	824	49,401	4,394,688	1	187,247	89	2	517,384	5,311
2013-52	831	55,410	4,415,700	1	171,634	80	3	487,953	5,504
2014-01	839	63,194	4,730,996	1	147,393	75	6	430,362	5,877
2014-02	848	62,587	4,503,580	1	159,557	72	1	486,339	5,383
2014-03	849	54,934	4,352,551	1	159,989	79	2	499,834	5,150
2014-04	854	51,607	4,255,056	1	158,437	82	1	505,074	5,013
2014-05	855	49,283	3,595,741	1	136,644	73	1	492,084	4,669
2014-06	856	48,882	3,660,297	1	141,324	75	1	506,978	4,786
2014-07	856	48,735	4,360,665	1	168,967	89	1	521,226	5,061
2014-08	858	48,793	4,303,000	1	166,948	88	1	523,650	5,065
2014-09	858	49,420	4,513,878	1	174,699	91	1	544,196	5,278
2014-10	858	47,818	4,542,935	1	180,103	95	1	545,182	5,214
2014-11	859	48,055	4,453,554	1	170,364	93	1	542,875	5,172
2014-12	858	50,117	4,510,703	1	166,130	90	0	540,254	5,278
2014-13	859	49,564	4,283,952	1	159,565	86	0	536,457	4,996
2014-14	859	48,187	4,412,165	1	168,483	92	1	544,497	5,093
2014-15	860	52,640	4,612,189	1	166,435	88	1	547,129	5,250
2014-16	860	63,826	4,500,126	1	117,738	71	3	497,451	5,456
2014-17	860	55,510	4,623,062	1	165,102	83	2	545,608	5,153
2014-18	860	46,636	4,372,846	1	157,824	94	2	540,645	5,077
2014-19	861	47,853	4,739,432	1	172,058	99	2	559,030	5,265
2014-20	861	50,360	4,686,350	1	163,605	93	2	557,085	5,257
2014-21	863	48,998	4,580,521	1	161,257	93	4	549,841	5,130

port includes the number of suspected and confirmed Zika cases per each department. Table 4.1 reports the number of cases, including the laboratory confirmed, and the cumulative incidence rate per each department, also shown in Figure 4.1. A total of 100,888 cases have been reported in Colombia corresponding to an incidence rate of 214 cases/100,000 population and about 9% of the cases were laboratory confirmed through RT-PCR assay. No cases have been reported in Bogotá that is situated at an average altitude of 2,640 metres above sea level with an average monthly air temperature of 18°C, thus not favouring the presence of mosquitoes. Valle del Cauca experienced the highest number of cases (25,341), whereas Archipelago of San Andres, Providencia and Santa Catalina had the highest cumulative incidence rate (1,520 cases/100,000). However, these numbers underestimate the total effect of the ZKV outbreak, since they do not account for asymptomatic infection or unreported clinical illness.

Mobile Phone data

Aggregated and encripted Origin-Destination (OD) matrix has been provided by a telecommunication operator in Colombia. The network consists of about 4 thousand cell towers distributed in the territory of Colombia accordingly to the population density (Figure 4.2a). Cellphone coverage is heterogeneous across municipalities with cell towers mostly concentrated in Bogotá (about 18% of the total number of cell towers), while 258 out of 1,120 municipalities have no cell towers (Figure 4.2b). Our dataset consists of weekly OD matrices containing the number of trips

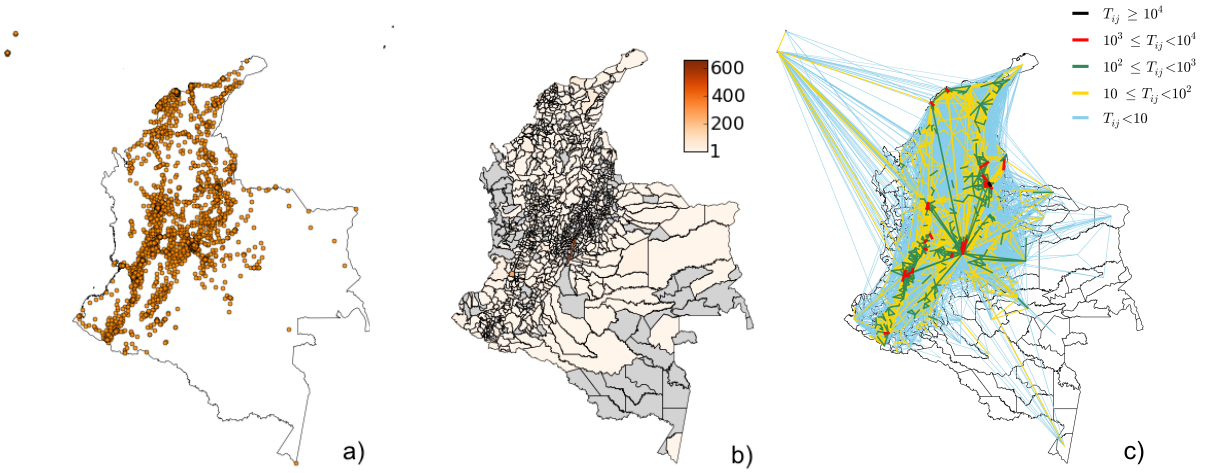


Figure 4.2: a) Location of cell towers in Colombia. b) Number of cell towers per municipality. The 258 municipalities with no cell towers are shown in grey. c) Average number of trips $\overline{T}_{ij}$ among municipalities.

T_{ij} from municipality i to municipality j that occurred in a week for a six-month period, from 2013-12-02 to 2014-05-19 (25 weeks in total). Moreover, we are also provided with the number of active phones n_i per municipality. OD matrices have been previously generated from more than two billion encripted and anonymized calls made by around seven million phone numbers in Colombia. More details on this process and the topology of the network can be found in Ref. [83]. Briefly, each phone number is associated to a cell tower whenever it makes a call, so that a movement can be identified when the same phone number is handled by two different cell towers in two consecutive calls, thus reconstructing the mobility network of human trips aggregated then at the level of municipality. In this process each phone number is encrypted and anonymized and only phone numbers which originated and received at least six calls during the observation period have been considered in order to drop foreign phones and all special phone numbers that are likely to be not associated with an actual person (e.g. call centers).

Table 4.2 shows the characteristics of the weekly data, including the number of nodes (i.e. municipalities), links and volume of trips, as well as the statistics of the weekly distribution of the number of trips and active phones per municipality. The structure of the networks varies week by week with the number of municipality that ranges from 786 to 863, the number of links connecting municipalities that ranges from 41,805 to 63,826 and the volume of trips that ranges from 3,595,741 to 4,739,432. The OD matrices are constructed to account at least 1 trip along a link during a week. The average number of trips ranges from 71 to 99. Differently, the number of active phones ranges from 0 to a maximum of 559,030 with an average that ranges from 4,669 to 5,877. On average, trips mostly interest the western portion of the country, including links with the islands and few links towards the south, as shown in Figure 4.2c.

4.3 Materials and Methods

In this study we use a metapopulation modelling approach [123] to perform numerical simulations of the epidemic spreading of Zika in Colombia at the geographical resolution of departments. Metapopulation models describe spatially structured interacting subpopulations, such as city locations, urban areas, or any defined geographical regions [81]. Individuals within each subpopulation are classified according to their health status with respect to the disease, such as infected, susceptible, immune, etc., and the compartment dynamics accounts for the possibility that individuals in the same location may get into contact and change their health status according to the infection dynamics. The interaction among subpopulations is the result of the movement of individuals from one subpopulation to another.

Here we consider the 33 departments of Colombia as the subpopulations of the metapopulation model that interact according to different human mobility patterns. The migration process among subpopulations is modelled with a Markovian dynamics, representing individuals who are indistinguishable regarding their travel pattern, so that at each time step the same traveling probability applies to all individuals without having memory of their origin [79, 80, 81]. Since this study focuses on identifying the benefits of using CDRs data, we first create a mobility network generated from mobile phone data and then we test it against different mobility networks through a detailed comparison of datasets and analysis of outcomes of the epidemic model. In particular, we create a mobility network based on the census data from the National Institute of Statistics [3] that represents our benchmark as it comprises the entire population of the country and its mobility features. On the other hand, mobile phone data are usually affected by the sampling bias corresponding to the operator's coverage and to the selection of the subset available for the analysis (it only represents a fraction of the total population) and by the algorithm used to identify movements. Next to census data, population movements can be described by mobility models determining the daily commuting fluxes across the country. Here we create two synthetic mobility networks by applying the gravity model [43], based on Newton's law of gravity, that assumes that the number of trips is related to the population at origin and destination and to decrease with the distance, and the more recent radiation model [182], that instead is inspired by the theory of intervening opportunities and considers human movements as diffusion processes that depend on the population distribution over the space.

In the next sections follows a description of the mobility networks and the epidemic model used to simulate the ZIKV transmission dynamic.

4.3.1 Mobility Networks

In this section, we describe the various mobility networks and how they have been generated at the geographical resolution of departments. In particular, we model the migration process with a Markovian process in which the memory of the origin of each traveller is lost and this requires a symmetric mobility network, i.e. the strength of the connection from i to j is the same of the one from j to i. Consequently, for each mobility network under study (that are asymmetric by definition) we will compute the average of w_{ij} and w_{ji} in order to obtain $w_{ij} = w_{ji}$. Moreover, for each of these networks we consider only those connections having $w_{ij} \geq 1$, allowing at least one person daily travelling on that link.

4.3.1.1 Mobile Phone Network

The weekly OD matrices generated from CDRs data in Colombia contains the number of trips T_{ij} from municipality i to municipality j occurred in a week. To generate the mobility network of flows w_{ij} of people travelling among departments we have to collapse our 25 weekly OD matrices in a single mobility network by accounting for the weekly fluctuations in the network and aggregating at the level of departments. Moreover, to obtain an estimate of the number of people moving from one location to another we rescale the number of trips T_{ij} on the number of active phones n_i in location i and on the population N_i living in i. Figure 4.3 shows the weekly number of active phones n_i within each department. Large weekly fluctuations can be observed in some departments, such as department D18 with a number of active phones that varies from a minimum of 0 to a maximum of 15,325 active phones per week, or department D86 (min=239; max=8,980) or department 99 (min=921; max=2,186). In conjunction with the celebration of the New Year at week 2014-01, almost all departments present a peak in the number of active phones, while department D11, that is the country's capital Bogotá, shows a decrease in the number of active phones, thus highlighting the outward movement of people travelling from the most populated region to other locations in the occasion of the New Year.

Since we do not know *a priori* the effect of such variability in the number of active phones on the final mobility network, we calculate the flows w_{ij} with two different rescaling approaches:

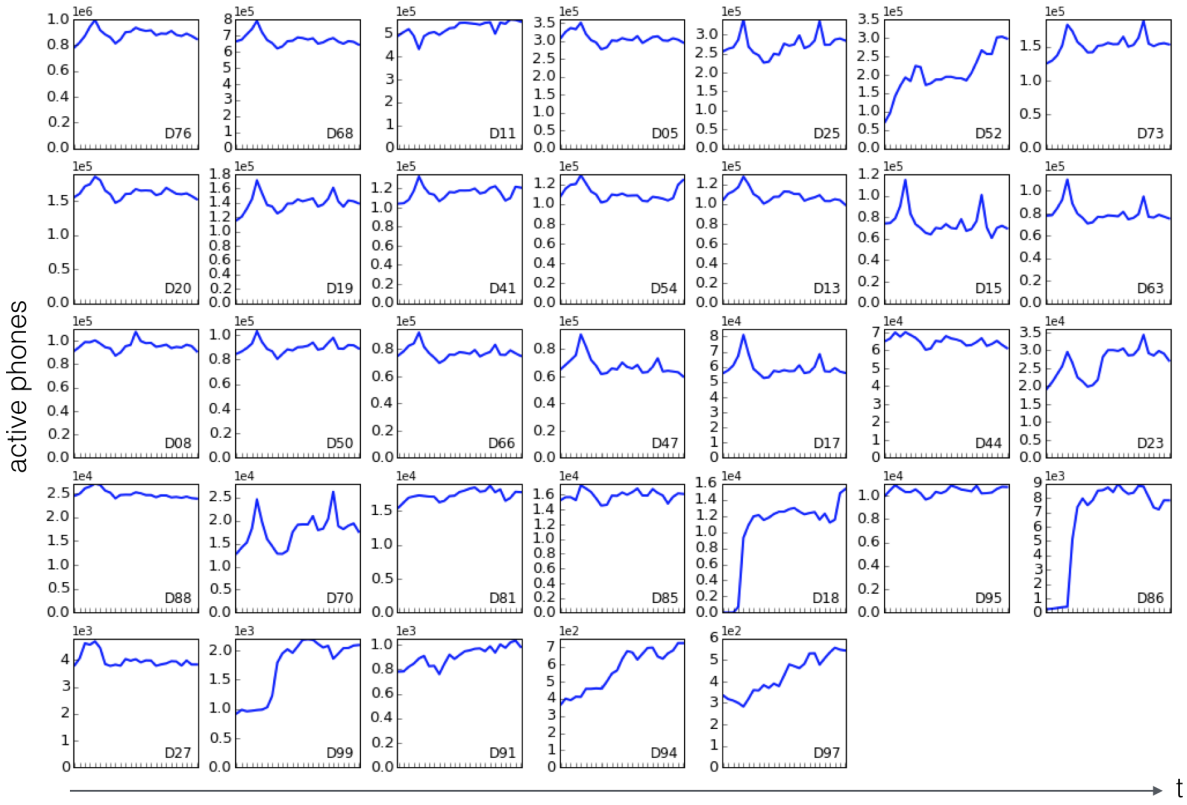


Figure 4.3: Weekly number of active phones within each department D in Colombia (subplots are sorted according to the maximum number of active phones).

1. In the first approach, flows are computed by rescaling the weekly average number of trips $\overline{T}_{ij}$ on the weekly average number of active phones $\overline{n}_i$ and on the population N_i, as follows: $\widetilde{w_{ij}} = \frac{\overline{T}_{ij} * N_i}{\overline{n}_i}$. Flows obtained with this method are indicated hereafter with $\widetilde{w_{ij}}$ and the network is labelled as *"MP1"*.

2. In the second approach, first we rescale each weekly number of trips T_{ij}^k at week k on the number of active phones n_i^k at week k and on the population N_i, as follows: $w_{ij}^k = \frac{T_{ij}^k * N_i}{n_i^k}$, where k=2013-49, ..., 2014-21, then we compute the median of the rescaled weekly flows. Flows obtained with this method are indicated hereafter with $\overline{w_{ij}}$ and the network is labelled as *"MP2"*.

In both methods flows are made symmetric (i.e. $w_{ij} = w_{ji}$) and only links having $w_{ij} \geq 1$ are considered. Moreover, since the OD matrices have been constructed to include movements occurred in a week, here we divided the flows by a factor 7 in order to get the daily flows.

4.3.1.2 Census Commuting Network

Commuting data for Colombia is extracted from the 2005 Colombian census of the National Institute of Statistics [3]. Even though this dataset is not recent and mobility might be changed in the meantime, we will consider it as our benchmark since it represents the entire population of the country and its mobility features. Data is collected at the level of municipality and the matrix of census commuting flows is asymmetric, i.e. the strength of the connection from municipality m_1 to municipality m_2 is not the same of the one from m_2 to m_1. Thus, here we aggregate at the geographical resolution of departments and compute the average of w_{ij} and w_{ji}

Table 4.3: Parameters of the gravity law as obtained by applying a multivariate analysis to census data.

Parameter	Estimate	Standard Error	p-value
α	0.63	0.05	$< 10^{-17}$
γ	0.37	0.04	$< 10^{-17}$
β	1.94	0.09	$< 10^{-17}$
C	3.25	2.86	0.3

removing then those connections having $w_{ij} < 1$. In the following we will indicate with w_{ij}^C the census fluxes of commuters from departments i to department j.

4.3.1.3 Synthetic Mobility Networks

Next to census data and mobile phone data, we generate human mobility fluxes across departments by creating two fully connected synthetic networks whose weights of edges connecting nodes are computed with the gravity model [43] and the radiation model [182], respectively.

The gravity model assumes that the commuting flow w_{ij} between location i with population N_i and location j with population N_j takes the following form:

$$w_{ij} = C\frac{N_i^\alpha N_j^\gamma}{f(d_{ij})} \tag{4.1}$$

where C is a proportionality constant, α and γ tune the dependence with respect to each location size, and $f(d_{ij})$ is a distance dependent functional form. Gravity laws usually consider power or exponential laws for the behaviour of $f(d_{ij})$. The results reported in the literature are variable and generally depend on the geographical resolution used in the study. We tested both power and exponential decay functions of the distance and we found that the best fit is obtained by using a power function $f(d_{ij}) = d_{ij}^\beta$. In particular, the regression coefficients are $R^2 = 0.579$ for a power function and $R^2 = 0.499$ for an exponential function.

The gravity law of equation 4.1 has 4 free parameters: the exponents of the population sizes, α and γ, the exponent of the distance, β, and the proportionality constant, C. A multivariate regression analysis is applied to obtain the values of the parameters that better fit our data as well as an estimation of their statistical significance. The values estimated for α, γ, β and C are reported in Table 4.3 along their p-values. All p-values are significant, except for the value of the intercept that shows a p-value>0.05, therefore we cannot reject the null hypothesis that the coefficient is equal to 0, thus meaning $C = 1$. These estimated parameters fitted to the census commuting data allows to apply the gravity law and generate the human mobility fluxes among departments, hereafter called w_{ij}^G.

In the case of the radiation model, instead, the flows w_{ij} of people travelling between different locations are obtained with the following expression:

$$w_{ij} = T_i\frac{N_iN_j}{(N_i + s_{ij})(N_i + N_j + s_{ij})},\ \ T_i = \sum_{i\neq j} w_{ij} \tag{4.2}$$

where N_i is the population living at origin i, N_j is the population living at destination j, s_{ij} is the total population living in a circle of radius d_{ij} centered at i, excluding the populations of origin and destination locations. Differently from the gravity model, the radiation model is parameter-free, i.e. it does not require regression analysis or fit on existing data. However, it requires the information on the total number of residents T_i who commute in each administrative unit. Given these quantities, the radiation model yields the mobility fluxes w_{ij}^R for each pair of departments of Colombia.

For each mobility network we analyse the structural properties as well as a detailed comparison of the distributions of flows and distances, for which also a two-sample Kolmogorov-Smirnov test is performed in order to asses if they derive from the same underlying distribution. Restricting the analysis to the topological intersection of the census network and the other mobility networks, we compute different statistical tests and similarity measures, including the Spearman's correlation coefficient, the Jaccard index, the cosine similarity and the common part of commuters (CPC) in order to evaluate the correlation and similarity of both links and weights of the networks.

- The Spearman's rank correlation coefficient is a nonparametric measure to assess monotonic relationships (whether linear or not) between two variables. It takes values from -1 (decreasing monotonic trend) to +1 (increasing monotonic trend).

- The Jaccard index is a measure of similarity that quantifies the fraction of connections in common between two networks. In particular, given a network A with elements A_{ij} and a network B with elements B_{ij}, the Jaccard index measures the proportion of shared links between A and B relative to the total number of links connected to A or B, i.e. the size of the intersection divided by the size of the union of the two sets, defined as follows:

$$J(A_{ij}, B_{ij}) = \frac{|A_{ij} \cap B_{ij}|}{|A_{ij} \cup B_{ij}|}, \quad J(A_{ij}, B_{ij}) \in [0, 1] \tag{4.3}$$

  Values of J close to 1 mean that the two networks under study share a large fraction of links, while values close to 0 mean that the two networks present very different structures.

- The cosine similarity is a measure of similarity that takes into account both links and weights shared by two networks. In particular, given a network A with weights $w_{ij,A}$ and a network B with weights $w_{ij,B}$, the cosine similarity is defined as follows:

$$\sigma(w_{ij,A}, w_{ij,B}) = \frac{\sum_j w_{ij,A} w_{ij,B}}{\sqrt{\sum_j w_{ij,A}^2 \sum_j w_{ij,B}^2}}, \quad \sigma(w_{ij,A}, w_{ij,B}) \in [0, 1] \tag{4.4}$$

  It takes values from 0, i.e. the sets of links are completely different in the two configurations, to 1, i.e. the two networks share the same set of links with same weight.

- The common part of commuters (CPC) is a measure of similarity that compares the modelled weights $\widetilde{w}_{ij}$ against the real weights w_{ij} according to the following expression:

$$CPC(w, \widetilde{w}) = \frac{2 \sum_{i,j} min(w_{ij}, \widetilde{w}_{ij})}{\sum_{i,j} w_{ij} + \sum_{i,j} \widetilde{w}_{ij}} \tag{4.5}$$

  This indicator is based on the Sørensen-Dice similarity coefficient that, given two sets A and B, is defined as $s(A, B) = 2|A \cap B|/(|A| + |B|)$. Values of $CPC(w, \widetilde{w})$ varies from 0, when no agreement is found, to 1, when the two networks are identical.

4.3.2 Epidemic Model

Following the same methodology adopted in previous works [128, 206], to study the spatiotemporal ZIKV spread we used a compartmental mathematical model simulating vector-borne transmission. Figure 4.4 describes the compartmental classification used to simulate ZIKV transmission dynamics. Humans can occupy one of four compartments: susceptible individuals S_H who lack immunity against the infection, exposed individuals E_H who have acquired the infection but are not yet infectious, infected individuals I_H who can transmit the infection (and may or

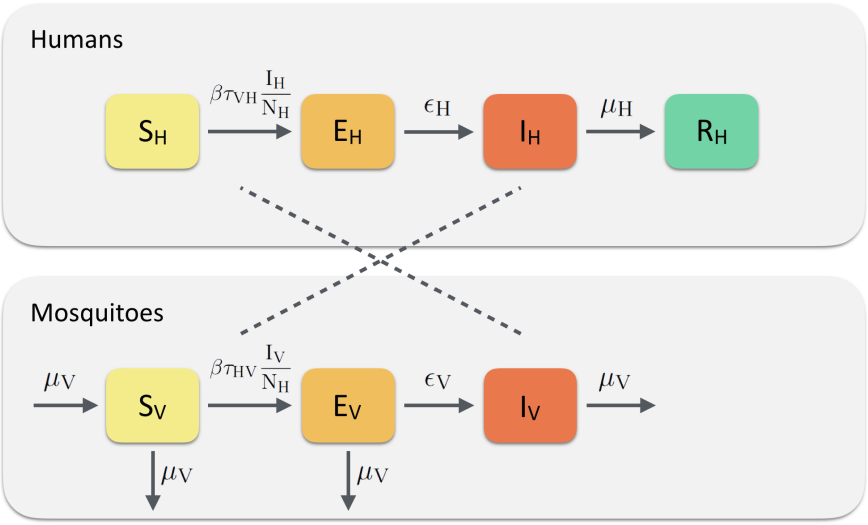


Figure 4.4: Human-vector transmission model for ZIKV infection. The transmission dynamics can occur when a susceptible human (S_H) is bitten by an infectious mosquito (I_V) or, viceversa, when a susceptible mosquito (S_V) bites an infectious human (I_H).

may not display symptoms), and removed individuals R_H who no longer have the infection and are immune to further ZIKV infection. As the mean human lifespan is much longer than the outbreak duration, we omitted human births and deaths and we considered the total human population size to be constant, i.e. $S_H + E_H + I_H + R_H = N_H$. Susceptible humans transition to the exposed compartment occurs under the vector-to-human force of infection that follows the usual mass-action law and is given by the expression $\lambda_{VH} = \beta\tau_{VH}\frac{I_H}{N_H}$. Exposed individuals become infectious at a rate ϵ_H, which is inversely proportional to the mean intrinsic latent period of the infection, ϵ_H^{-1}. Infectious individuals then recover from the disease at a rate μ_H, which is inversely proportional to the mean infectious period, μ_H^{-1}.

The mosquito vector population is described by the number of susceptible S_V, exposed E_V, and infectious mosquitoes I_V. The number of mosquitoes per person is expressed by k, so that $k = \frac{N_V}{N_H}$. The human-to-vector force of infection λ_{HV} governing the transition rate from susceptible to exposed individuals is given by $\lambda_{HV} = \beta\tau_{HV}\frac{I_V}{N_H}$. Exposed mosquitoes transition to the infectious class occurs at a rate ϵ_V, which is inversely proportional to the mean extrinsic latent period in the mosquito population, ϵ_V^{-1}. Mosquitoes in each compartment die at a rate μ_V, that is inversely proportional to the mosquito lifespan, μ_V^{-1}.

The dynamics is discrete and stochastic, and individuals are assumed to be homogeneously mixed within each subpopulation. Subpopulations are coupled by weighted links representing the mobility fluxes between two locations, thus defining the metapopulation structure of the model [123]. No other type of movement is considered. The transmission model is fully stochastic. The coupled population equations describing the epidemic time evolution read as follows:

$$S_{t+1}^H = S_t^H - \Delta_{S^H \longrightarrow E^H} \tag{4.6}$$
$$E_{t+1}^H = E_t^H + \Delta_{S^H \longrightarrow E^H} - \Delta_{E^H \longrightarrow I^H} \tag{4.7}$$
$$I_{t+1}^H = I_t^H + \Delta_{E^H \longrightarrow I^H} - \Delta_{I^H \longrightarrow R^H} \tag{4.8}$$
$$R_{t+1}^H = R_t^H + \Delta_{I^H \longrightarrow R^H}, \tag{4.9}$$

$$S_{t+1}^V = S_t^V - \Delta_{S^V \longrightarrow E^V} + \Delta_{E^V \longrightarrow S^V} + \Delta_{I^V \longrightarrow S^V} \tag{4.10}$$
$$E_{t+1}^V = E_t^V + \Delta_{S^V \longrightarrow E^V} - \Delta_{E^H \longrightarrow S^V} - \Delta_{E^V \longrightarrow I^V} \tag{4.11}$$
$$I_{t+1}^V = I_t^V + \Delta_{E^V \longrightarrow I^V} - \Delta_{I^V \longrightarrow S^V}. \tag{4.12}$$

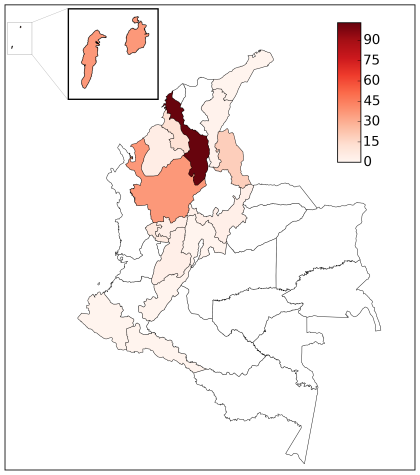


Figure 4.5: Initial condition at week 2015-42.

Parameter		Value	Ref.
transmission prob.	τ_{VH} τ_{HV}	0.399, 0.504, 0.591	-
mosquito biting rate	β	0.7	[121]
intrinsic latent period	ϵ_H^{-1}	7 days	[206]
human infectious period	μ_H^{-1}	5 days	[121]
extrinsic latent period	ϵ_V^{-1}	8 days	[206]
mosquito lifespan	μ_V^{-1}	7.8 days	[128]
mosquito per person	k	2	[121]
reproduction number	R_0	3.0, 4.8, 6.6	[152]

Table 4.4: Summary of the epidemiological parameters.

Each term $\Delta_{X\longrightarrow Y}$ represents the number of human or vector individuals transitioning from state X to state Y. Transitions are calculated according to chain binomial processes $\Delta_{X\longrightarrow Y} = Binomial(X, p_{X\longrightarrow Y})$ considering a number of trials equal to the number of individuals X in the compartments and a transition probability $p_{X\longrightarrow Y}$ determined by the force of infection and the average lifetime of individuals in each compartment. Each transition is a memoryless discrete stochastic transition process.

The basic reproduction number R_0 can be expressed as the standard approach of Ref. [123]:

$$R_0 = \frac{\epsilon_V}{\mu_H \mu_V (\epsilon_V + \mu_V)} k\beta^2 \tau_{VH}\tau_{HV} \tag{4.13}$$

For each epidemic scenario, defined by certain initial conditions and a set of parameters, we simulate 1,000 stochastic realizations using discrete time steps of one full day over a period of 1 year, thus resembling the actual duration of the epidemic of Zika in Colombia. The model is fully stochastic and from any nominally identical initialization generates an ensemble of possible epidemics, as described by newly generated infections in each subpopulation. Initial conditions are given by the number of Zika cases in each department at a certain week as reported by the Instituto Nacional de Salud in Colombia. Among each set of simulations, we then compute the median and the 95% confidence interval (CI), and we weekly aggregate the results in order to make them comparable with the surveillance case data.

As initial phase of this study, we investigate the role of human mobility in the spreading of ZIKV in Colombia in a very simple setting in which we first neglect the potential correlation of ZIKV transmission dynamics with environmental, climatic and socio-economic factors. To this aim, the study is limited to some epidemic scenarios in order to evaluate the variability in the epidemic outcomes and test the performances of the various mobility networks. In particular, we inform the model about the number of initial infections by considering the number of Zika cases reported in the early stage of the epidemic in Colombia, taking into account that Zika cases reported over that period might not be accurate as the outbreak was just declared and a potential underreporting affected the collected data due to asymptomatic cases, limited sentinel sites, and cases not seeking treatment (see *Dataset*). Thus, here we assume that reported cases correspond to 1% of the actual number of Zika cases and we consider the number of Zika cases reported at week 2015-42 (18-24 October, 2015) as initial conditions to run the epidemic simulations (see Figure 4.5). At this stage, such choice is arbitrary and further investigations will be needed to better understand how results may be affected by using different set of initial conditions.

The epidemiological parameters that intervene in the model are taken from the literature, in particular from articles presenting similar compartmental models for ZIKV transmission dynam-

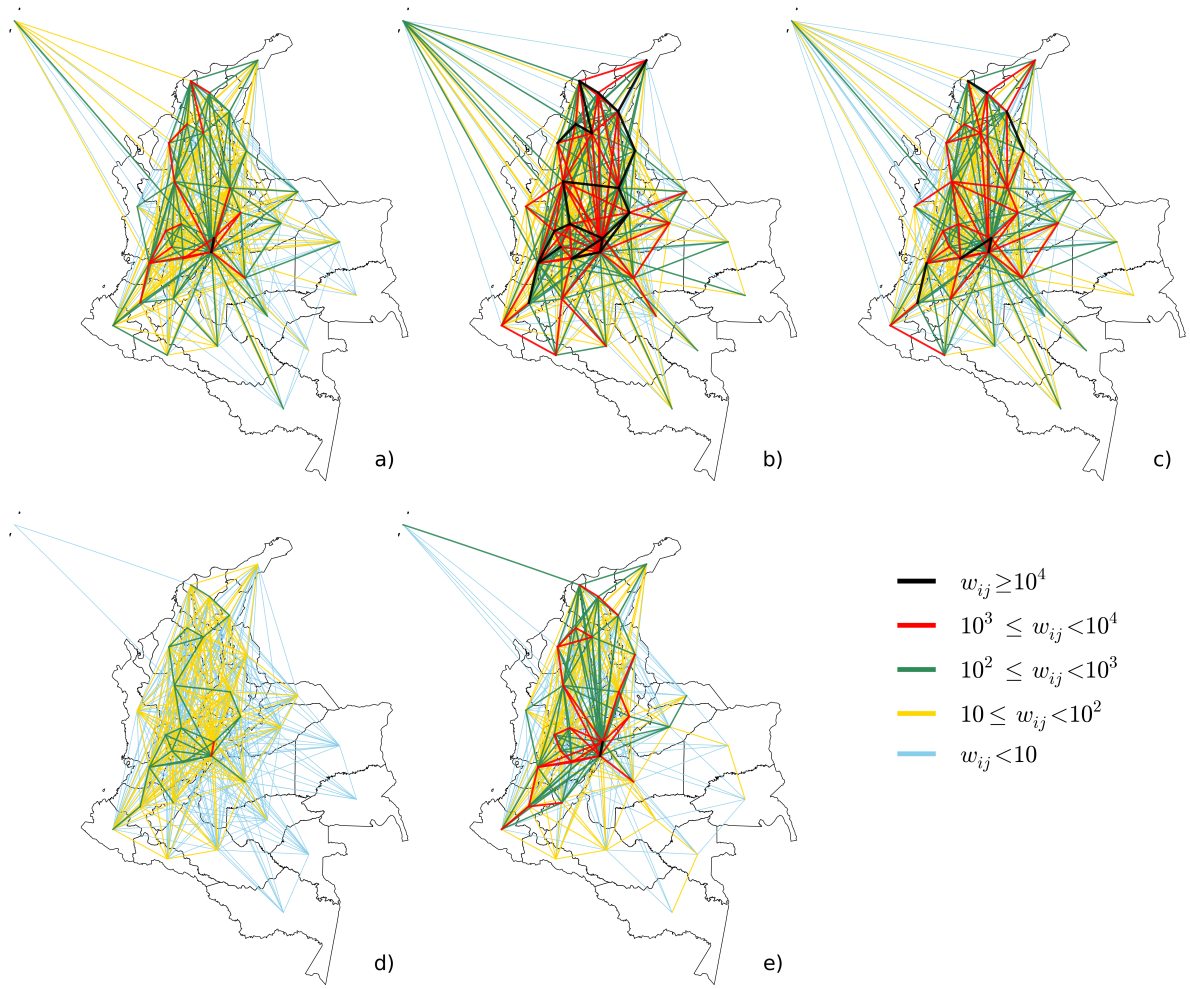


Figure 4.6: Flows w_{ij} of people travelling among departments as obtained for the census network (a), the mobile phone networks MP1 (b) and MP2 (c), the gravity network (d) and the radiation network (e). Colours refer to different ranges of values.

ics [121, 128, 139, 166, 206]. Given that many parameters characterizing ZIKV are surrounded by uncertainty, here we consider the average values of the ranges reported in the literature, as shown in Table 4.4. In particular, the basic reproduction number R_0 for Colombia was estimated to range from 3.0 to 6.6 [152], thus here we set three epidemic scenarios according to the minimum, mean and maximum value of this range (i.e. R_0=3.0, 4.8, 6.6). Given R_0, we estimate the value of the probabilities of transmission τ_{VH} and τ_{HV}, as reported in Table 4.4. Moreover, as in a previous study [121], we assume an average of 2 mosquitoes per person, i.e. the population of susceptible mosquitoes in each subpopulation at the beginning of the simulations is twice the population of humans. Spatio-temporal variability in the number of mosquitoes is not taken into account, but will be further explored in the future.

4.4 Results

In this section we describe the topological structure of the various mobility networks by comparing their statistical properties against the census network that represents our benchmark. Then, we compare the metapopulation epidemic outcomes obtained integrating these different mobility sources into the three scenarios (see *Methods*).

Table 4.5: Basic properties of each mobility network under study, including the number of nodes, links and links shared with the census network, and the total volume of travellers without considering self-loops.

network	#nodes	#link	# shared link (%)	volume
census	33	734	-	298,962
MP1	33	820	690 (94)	1,577,198
MP2	33	802	684 (93)	688,850
gravity	33	814	684 (93)	25,354
radiation	33	666	590 (80)	277,872

Table 4.6: Statistics on the distribution of flows and distances for each mobility network under study. Columns report the minimum (*min*), maximum (*max*), mean (*mean*), 95% confidence interval (95% CI), median (*median*) and standard deviation (*std*) of the flows w_{ij} and the distances d_{ij}, respectively. Distances are reported in kilometers.

network	$w_{ij_{min}}$	$w_{ij_{max}}$	$w_{ij_{mean}}$	95% CI	$w_{ij_{median}}$	$w_{ij_{std}}$
census	1	82,311	407	(1; 1,583)	22	4,311
MP1	1	255,949	1,923	(2; 12,140)	49	13,191
MP2	1	106,033	859	(1; 7,264)	23	5,597
gravity	1	2,044	31	(1; 184)	8	111
radiation	1	57,205	417	(1; 2,087)	18	3,173
network	$d_{ij_{min}}$	$d_{ij_{max}}$	$d_{ij_{mean}}$	95% CI	$d_{ij_{median}}$	$d_{ij_{std}}$
census	57	1,320	485	(113; 1,082)	446	256
MP1	57	1,422	512	(116; 1,135)	479	271
MP2	57	1,422	507	(116; 1,135)	470	270
gravity	57	1,247	484	(116; 1,006)	462	235
radiation	57	1,250	450	(110; 1,045)	406	238

4.4.1 Statistical comparison of mobility networks

Final mobility networks are displayed in Figure 4.6. All mobility networks under study share the same number of nodes, thus meaning that all 33 Colombian departments are covered by all datasets, with variations in the number of links and total volume of travellers. The census commuting network for Colombia includes 298,962 people travelling among departments through 734 weighted connections. The mobile phone networks MP1 and MP2 are formed by 820 links and 1,577,198 travellers, and 802 links and 688,850 travellers, respectively. For the mobility models we found 25,354 people moving along 814 links and 277,872 people moving along 666 links for the gravity and the radiation model, respectively. A summary of the basic statistics of the networks is reported in Table 4.5. In particular, the number of links shared with the census network varies from 80% for the radiation network to 94% for the mobile phone network MP1, thus reflecting the Jaccard index that ranges from 0.72 for the radiation network to 0.80 for both mobile phone networks (see Table 4.7).

For each mobility network, we further analyse the travel fluxes w_{ij} as well as the travelled distances d_{ij}, this latter calculated from the coordinates of the department's centroid by applying the Vincenty's formula [195]. Figure 4.7 shows the probability density distributions of flows and distances (reported in kilometers), whereas Table 4.6 reports the statistics on such distributions. The largest number of travellers is observed between Cundinamarca (D25) and country's capital Bogotá (D11) for all mobility networks, with significant variations in the weights of the link. In particular, there are 82,311 travellers for the census network, 255,949 and 106,033 for the mobile

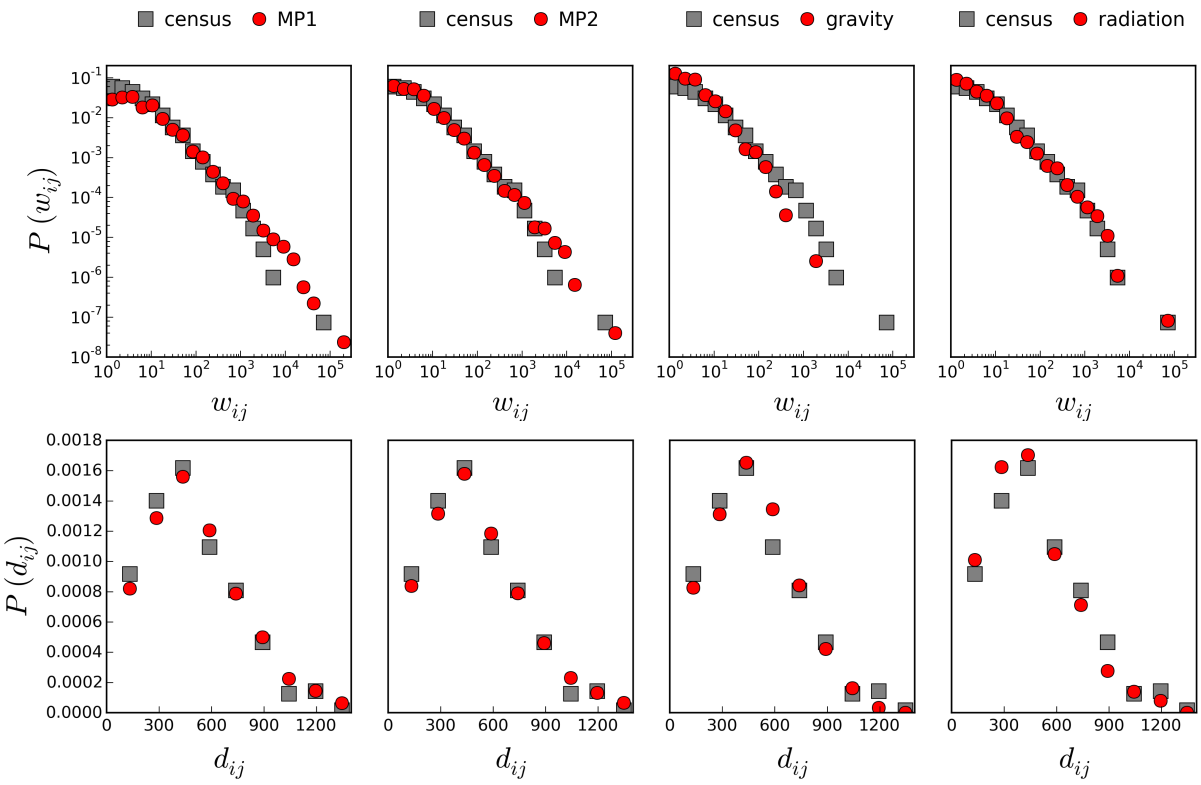


Figure 4.7: Probability density distributions of the weights w_{ij} (top) and the distances d_{ij} (bottom) for the census network and the various mobility networks. Distances are reported in kilometers.

phone networks, 2,044 for the gravity network and 57,205 for the radiation network. In general, both mobile phone networks show large weights, whereas the gravity network presents very small flows, with an average of 31 people travelling through a link. A two-sample Kolmogorov-Smirnov test performed between the census network and the various mobility networks showed that only the flows of the radiation network and the census network derive from the same underlying distribution (p=0.1, while $p < 10^{-3}$ for the remaining mobility networks). Furthermore, a deeper analysis of the spatial relative differences in the incoming/outgoing traffic at the level of departments, confirmed the similarity between the census network and the radiation network, presenting 16 out of 33 departments in the interval of the national traffic value (grey coloured in Figure 4.9) and only few departments strongly underestimated, located in less populated areas in the southern and eastern portion of the country. Larger discrepancies are instead observed for the other mobility networks with the gravity network being strongly underestimated and the mobile phone networks strongly overestimated, as shown in Figure 4.9.

On the other hand, travelled distances are quite similar across the mobility networks, with a minimum of 57 km between Cundinamarca (D25) and Bogotá (D11) and an average daily distance that ranges from 450 to 512 km. The largest daily distances correspond to 1,320 km between Huila (D41) and Archipelago of San Andres, Providencia and Santa Catalina (D88) for the census network, 1,422 km between Archipelago of San Andres, Providencia and Santa Catalina (D88) and Meta (D50) for the mobile phone networks, 1,247 km between Narino (D52) and La Guajira (D44) for the gravity network, and 1,250 km between Archipelago of San Andres, Providencia and Santa Catalina (D88) and Bogotá (D11) for the radiation network. Note that in a fully connected network the overall largest distance would be of 1,947 km between Archipelago of San Andres, Providencia and Santa Catalina (D88) and Amazonas (D91) or, if we restrict only to mainland, 1,441 km between La Guajira (D44) and Amazonas (D91). A two-sample Kolmogorov-Smirnov test performed between the census network and the other

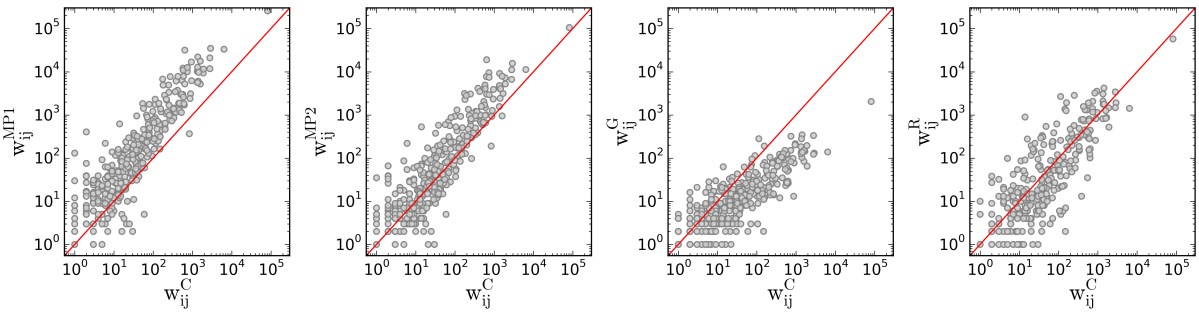


Figure 4.8: Comparison of the weights w_{ij} in the various mobility networks and the weights w_{ij}^C in the census networks. Grey points are scatter plot for each connection. The red line is given by $y = x$.

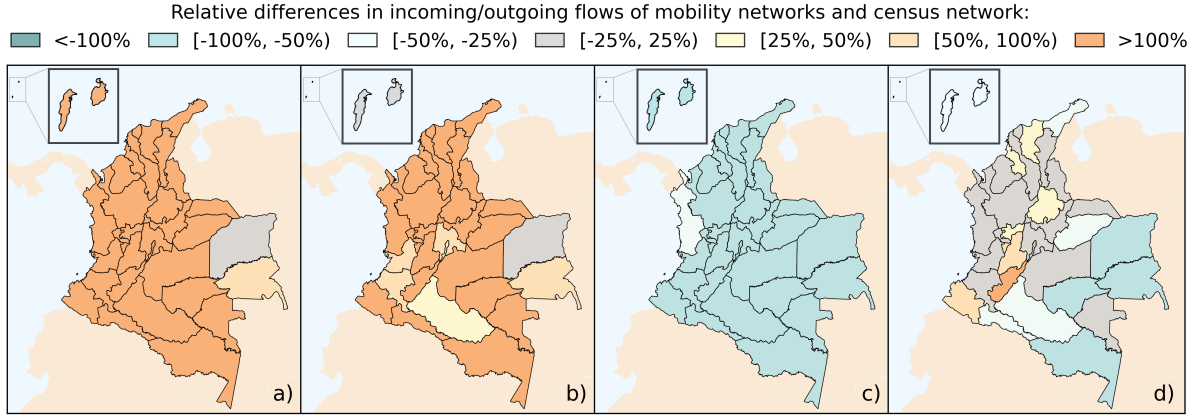


Figure 4.9: Spatial relative difference in the incoming/outgoing traffic between the census network and the various mobility networks at the level of departments. The colour code indicates the relative difference between the census traffic and the corresponding traffic of the mobile phone networks MP1(a) and MP2 (b), the gravity network (c) and the radiation network (d). Since the networks are symmetric, the incoming traffic $T_j = \sum_i w_{ij}$ is equal to the outgoing traffic $T_i = \sum_j w_{ij}$.

mobility networks confirmed that the corresponding distributions of distances derive from the same underlying distribution (all $p > 0.05$).

Restricting our analysis to the topological intersection of the mobility networks and the census network, fluxes are found to be larger than the census ones, except for the gravity network (Figure 4.8). A side-by-side weight comparison on each link shows high correlation between datasets, as reported in Table 4.7. Strong correlations are also obtained for the total number of travellers leaving or entering a given department (outgoing and incoming flows are the same because the networks are symmetric). In particular, the correlation is lower for the gravity network and higher for the radiation network, thus confirming the results found for the spatial relative differences in the incoming/outgoing traffic (see 4.9). The cosine similarity measured between the census network and the various mobility networks ranges from 0.91 for the gravity network to 0.99 for the radiation network, as reported in Table 4.7. Furthermore, we tested the various mobility networks by analysing the performance for different flows subsets organized by distance and destination population, using the common part of commuters (CPC) that compares the modelled flows $\widetilde{w}_{ij}$ against the real flows w_{ij} given by census data. Figure 4.10 shows the goodness of fit in a grid of (distance, population) ranges, where each cell represents a subset of pairs of origin-destination links filtered by distance and by destination population. In general, the gravity network shows the worst performance with low values of similarity with the census flows. The performance increases in the mobile phone network MP1 with increasing values of

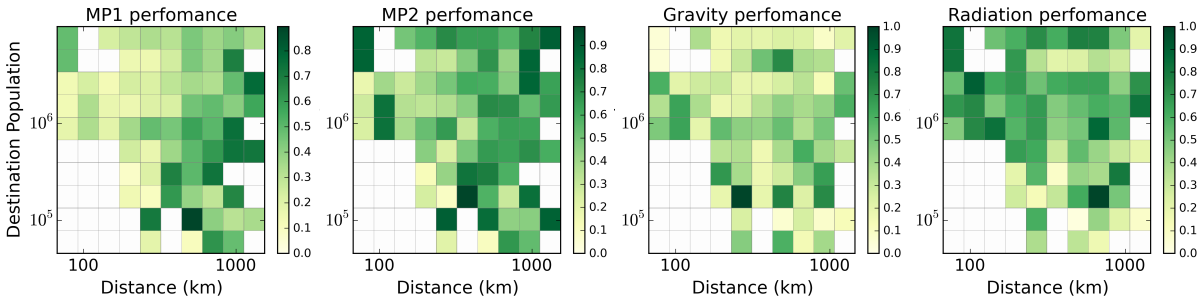


Figure 4.10: Performance of the mobility networks against the census network according to the common part of commuters (CPC) in a grid of (distance, population) subsets. The colour code indicates the level of similarity between the predicted and real flows, from low (light yellow) to high similarity (dark green).

Table 4.7: Values of the Spearman's coefficient, the Jaccard index and the cosine similarity as obtained from the various mobility networks compared to the census network. The Spearman's coefficient is measured on both weights w_{ij} and outgoing flows $\sum_i w_{ij}$.

Mobility network	Spearman's coefficient		Jaccard Index	Cosine Similarity
	w_{ij}	$\sum_i w_{ij}$		
MP1	0.88	0.91	0.80	0.97
MP2	0.86	0.87	0.80	0.96
gravity	0.79	0.90	0.79	0.91
radiation	0.77	0.96	0.73	0.99

the distances. Best performances are presented by the mobile phone network MP2 for high populated departments and long distances, and by the radiation network for high populated departments and short distances.

4.4.2 Comparison of epidemic outcomes

Figure 4.11 shows the simulated epidemic profiles with the 95% confidence interval (CI) aggregated at the country level in the three epidemic scenarios based on the different values of the basic reproduction number R_0. Peak weeks and Pearson correlations are reported in Table 4.8. In general, the same behaviour is observed in all three scenarios in terms of peak timing and peak intensity, with the mobile phone network MP1 being the one with the highest number of infected individuals and presenting an early peak, followed then by the mobile phone network MP2, then the census network, then the radiation network, and eventually the gravity network. In particular, the census network and the radiation network present very similar epidemic profiles, due to the similarity we already observed in the structure of the mobility networks.

The epidemic of Zika in Colombia reached the highest number of reported cases at week 2016-05. The first scenario, corresponding to R_0=3.0, presents a slow infection dynamics that causes the curves to be wider with a late peak, in general after week 2016-12. Values of the Pearson correlation ranges from 0.557 for the gravity network to 0.66 for the mobile phone network MP1. In the second scenario (R_0=4.8), both mobile phone networks are able to capture the peak week within their 95% CI and the correlation is generally high for all mobility networks, ranging from 0.758 for the mobile phone network MP1 to 0.802 for the radiation network. In the third scenario, the high value of R_0 allows for a faster infection dynamics that causes more narrow curves with early peaks. In this case, the mobile phone network MP1 peaks at week 2016-03 and has the lowest value of correlation of 0.39, while the gravity network captures the peak week with a correlation of 0.667.

A deeper analysis of the simulated epidemic profiles at the level of departments is shown

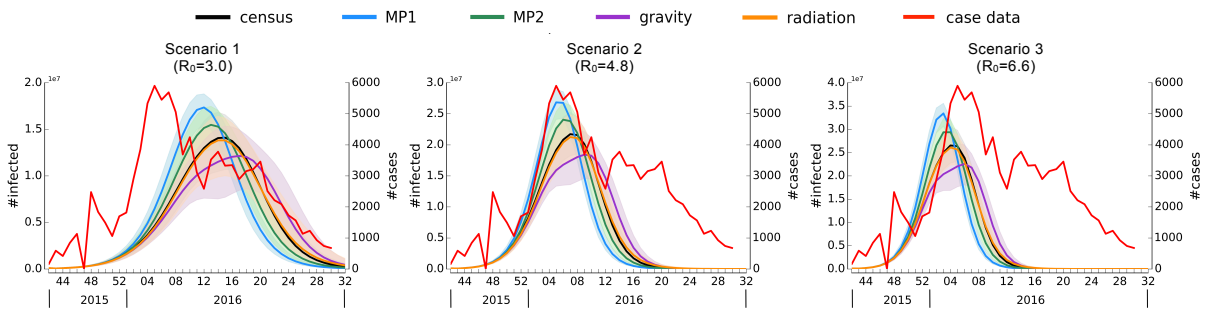


Figure 4.11: The simulated epidemic profiles and the official case data aggregated at the country level. Solid lines correspond to the median, while shadow areas correspond to 95% CI.

Table 4.8: Peak weeks and Pearson correlation values as obtained from the simulated epidemic profiles compared to the official case data aggregated at the country level. All p-values are significant ($p < 10^{-4}$).

Mobility network	Scenario 1 (R_0=3.0) peak week (95% CI)	CORR	Scenario 2 (R_0=4.8) peak week (95% CI)	CORR	Scenario 3 (R_0=6.6) peak week (95% CI)	CORR
census	2016-15 [2016-14, 2016-16]	0.60	2016-07 [2016-07, 2016-08]	0.79	2016-04 [2016-04, 2016-05]	0.61
MP1	2016-12 [2016-12, 2016-12]	0.66	2016-05 [2016-05, 2016-06]	0.76	2016-03 [2016-03, 2016-03]	0.39
MP2	2016-13 [2016-13, 2016-13]	0.63	2016-06 [2016-06, 2016-07]	0.79	2016-03 [2016-03, 2016-04]	0.52
gravity	2016-17 [2016-15, 2016-22]	0.56	2016-08 [2016-08, 2016-09]	0.78	2016-05 [2016-05, 2016-06]	0.67
radiation	2016-14 [2016-14, 2016-15]	0.59	2016-07 [2016-07, 2016-08]	0.80	2016-04 [2016-04, 2016-05]	0.61

in Figure 4.12, limiting the study to the first ten departments according to the highest value of cumulative incidence rate during the Zika outbreak in Colombia, namely Valle del Cauca (D76), Norte de Santander (D54), Santander (D68), Tolima (D73), Huila (D41), Atlantico (D08), Cundinamarca (D25), Meta (D50), Casanare (D85), and Magdalena (D47). Since the weekly numbers of Zika cases reported at the level of departments are very noisy, we apply a smoothing technique based on a 3-weeks centered moving average. At such geographical resolution, the temporal evolution in different departments shows strong heterogeneity with non-well defined peaks, that most of the time are very broad or present multiple spikes during the epidemic, thus being very difficult to predict. As for the results aggregated at the country level, also in this case we observe that the mobile phone network MP1 tends to hasten the infections dynamics with an early peak, followed then by the mobile phone network MP2, while the census network and the radiation network present very similar epidemic profiles, that often coincide, and finally the gravity network with a late peak. In some departments, such as Valle del Cauca (D76) and Santander (D76), each mobility network presents a well-distinct simulated epidemic profile in all three epidemic scenarios, while in other departments, such as Tolima (D73) and Atlantico (D08), all simulated epidemic profiles almost coincide in all three epidemic scenarios. As the value of R_0 increases, the simulated curves become more narrow, presenting early peaks, thus reflecting the more rapid dynamics of infection induced by increased transmissibility between humans and mosquitoes.

The correlation measured between each epidemic profile and the official case data is very heterogeneous, as reported in Table 4.9, and at this stage we do not observe any epidemic

Table 4.9: Values of the Pearson correlation computed between the curve of the official case data and each epidemic profile as obtained by integrating the various mobility networks under study. Significant correlations (i.e. p-value< 0.01) are indicated with $*$. Departments are sorted according to the cumulative incidence rate reported during the period from week 2015-32 to week 2016-30.

		Pearson correlation				
Dep.	scenario	census	MP1	MP2	gravity	radiation
Valle del Cauca (D76)	scenario 1	0.78*	0.62*	0.87*	0.85*	0.72*
	scenario 2	0.53*	0.63*	0.24	0.45*	0.58*
	scenario 3	0.16	0.30	-0.18	0.04	0.22
Norte de Santander (D54)	scenario 1	-0.39	-0.29	0.25	-0.01	-0.27
	scenario 2	0.40	0.40	0.98*	0.88*	0.51*
	scenario 3	0.88*	0.89*	0.79*	0.91*	0.94*
Santander (D68)	scenario 1	0.34	0.30	0.41*	0.36	0.39
	scenario 2	-0.34	-0.35	-0.30	-0.33	-0.32
	scenario 3	-0.52*	-0.52*	-0.50*	-0.51*	-0.51*
Tolima (D73)	scenario 1	0.45*	0.47*	0.44*	0.33	0.38
	scenario 2	0.90*	0.90*	0.91*	0.91*	0.92*
	scenario 3	0.52*	0.51*	0.58*	0.63*	0.57*
Huila (D41)	scenario 1	-0.14	-0.14	-0.01	-0.10	-0.20
	scenario 2	0.74*	0.73*	0.84*	0.77*	0.66*
	scenario 3	0.96*	0.96*	0.93*	0.95*	0.97*
Atlantico (D08)	scenario 1	-0.10	-0.47*	0.42*	0.23	-0.21
	scenario 2	0.82*	0.32	0.89*	0.89*	0.75*
	scenario 3	0.83*	0.79*	0.66*	0.74*	0.84*
Cundinamarca (D25)	scenario 1	-0.26	-0.34	0.09	-0.10	-0.18
	scenario 2	0.46*	0.43*	0.63*	0.56*	0.52*
	scenario 3	0.63*	0.64*	0.53*	0.61*	0.63*
Meta (D50)	scenario 1	0.55*	0.44*	0.86*	0.78*	0.42*
	scenario 2	0.69*	0.72*	0.39	0.58*	0.73*
	scenario 3	0.43*	0.46*	-0.06	0.22	0.51*
Casanare (D85)	scenario 1	0.87*	0.90*	0.58*	0.78*	0.90*
	scenario 2	0.19	0.23	0.02	0.12	0.25
	scenario 3	-0.03	0.02	-0.23	-0.11	0.02
Magdalena (D47)	scenario 1	-0.37	-0.49*	0.21	0.02	-0.40
	scenario 2	0.52*	0.34	0.80*	0.77*	0.48*
	scenario 3	0.75*	0.73*	0.65*	0.70*	0.75*

scenario that is better than the others. In particular, among the ten departments we considered in this analysis, the census network outperforms only in one department (i.e. Magdalena (D47), the gravity network outperforms in two departments (i.e. Tolima (D73) and Atlantico (D08)), the mobile phone network MP1 outperforms in four departments (i.e. Valle del Cauca (D76), Tolima (D73), Cundinamarca (D25) and Casanare (D85)), the radiation network outperforms in six departments (i.e. Norte de Santader (D54), Tolima (D73), Huila (D41), Meta (D50), Casanare (D85) and Magdalena (D47)), the mobile phone network MP2 outperforms in eight departments (i.e. Valle del Cauca (D76), Norte de Santader (D54), Santader (D68), Huila (D41), Atlantico (D08), Cundinamarca (D25), Meta (D50) and Magdalena (D47)).

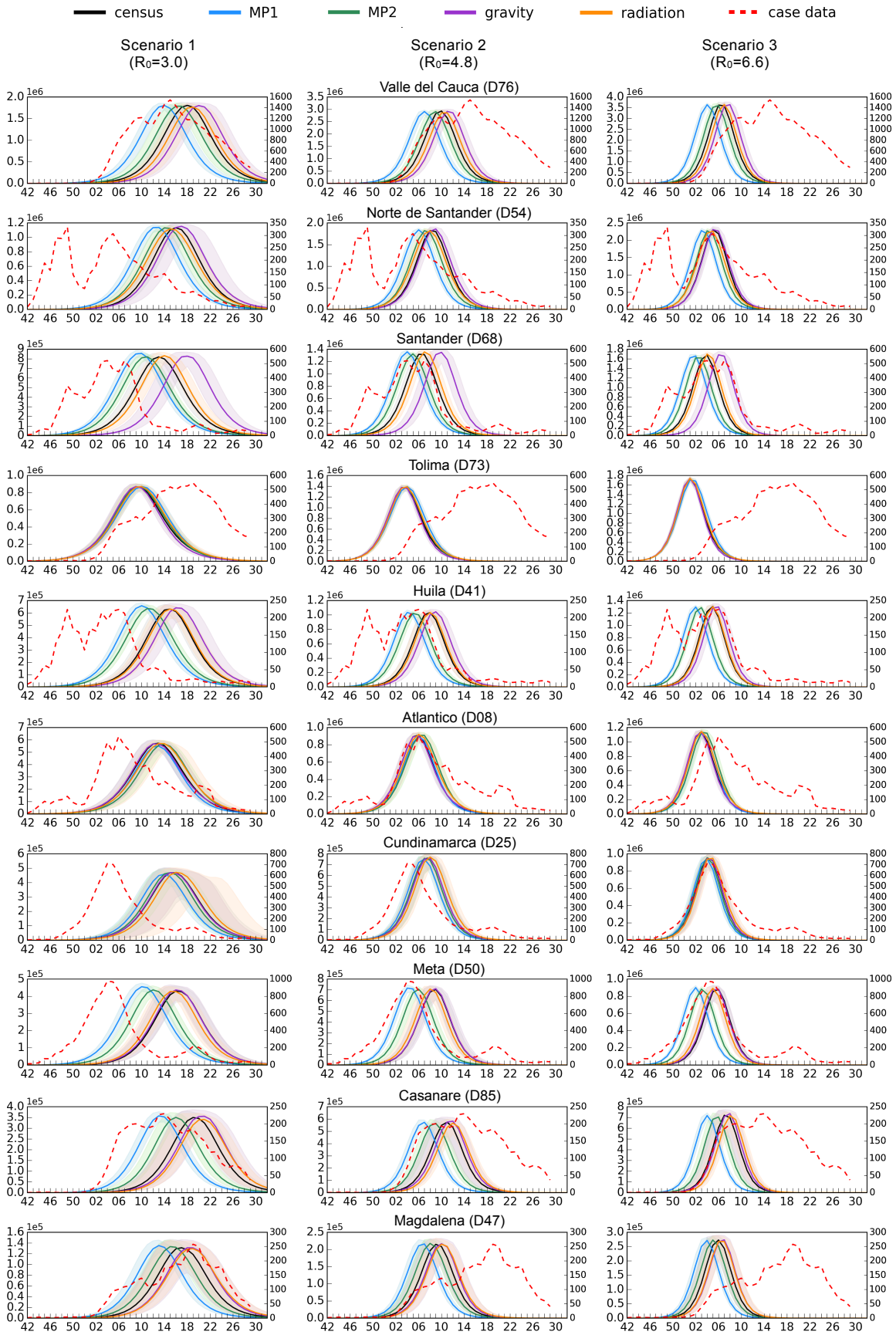


Figure 4.12: Weekly number of infectious individuals as obtained from the metapopulation model by integrating the various mobility networks under study. Solid lines correspond to the median, while shadow areas correspond to 95% CI for each department of Colombia. Departments are sorted according to the cumulative incidence rate reported during the period from week 2015-32 to week 2016-30.

4.5 Discussion and Future Work

In this study we focused on the Zika outbreak occurred in Colombia in 2015-2016 by developing a metapopulation model based on real data on population and human mobility, and directly informed with the number of cases reported in the early stage of the epidemic. Aim of this work is to investigate the role of human mobility given by different mobility networks, including human mobility patterns provided by mobile phone data, as well as generated by mobility models (i.e. the gravity model and the radiation model) and census commuting data. Here we thoroughly analysed the various mobility networks and we undertook the initial step towards the investigation of the spatiotemporal spread of the epidemic as obtained by integrating different mobility sources into the metapopulation modelling approach.

In particular, we found that all final mobility networks under study share the same number of nodes (i.e. departments) with variations in the number of weighted connections. The radiation network shares a smaller number of links with the census network (i.e. 80% with a Jaccard index of 0.73) compared to the remaining networks (i.e. 93-94% with a Jaccard index of 0.8). In general, fluxes are found to be larger than the census ones, except for the gravity network that has very small flows, with an average of 31 people travelling through a link. Consequently, the incoming/outgoing traffic at the level of departments is strongly underestimated by the gravity network, while is strongly overestimated by both mobile phone networks. The radiation network shows the best performance in terms of similarity with the census traffic with 16 out of 33 departments in the interval of the national traffic values and only few departments strongly underestimated, located in less populated areas in the southern and eastern portion of the country. This is also confirmed by the Spearman correlation computed on the incoming/outgoing traffic, that is lower (=0.87) for the gravity network and higher (=0.96) for the radiation network. Indeed, from this analysis of the structural and fluxes properties of the mobility networks under study, the radiation network resulted to be the best mobility network in terms of performance and similarity with the census network, that represents our benchmark. Also, travelled distances are quite similar across the mobility networks, with a minimum of 57 km and an average daily distance that ranges from 450 to 512 km. The migration process among departments has been modelled with a Markovian dynamics because of the long distances travelled on a daily basis. However, human mobility might be split into a short-range and a long-range mobility in order to integrate a non-Markovian dynamics that allows individuals to travel to a destination and come back at a constant rate as it usually happens for the daily commuting.

To study the spatiotemporal ZIKV spread we adopted a compartmental mathematical model simulating the vector-borne transmission that characterizes Zika, whose dynamics of infection occur in the interaction between humans and mosquitoes, i.e. when a human is bitten by an infectious mosquito or when a mosquito bites an infectious human. Then, the possibility for individuals to move from one location to another promotes the epidemic spreading of the disease. In particular, the subpopulations of the metapopulation model correspond to the 33 Colombian departments. As initial stage of this study, we investigated the role of human mobility in the spreading of ZIKV in Colombia in a very simple setting in which we first neglect the potential correlation of ZIKV transmission dynamics with environmental, climatic and socio-economic factors. To this aim, the study is limited to three epidemic scenarios based on different values of the basic reproduction number R_0 in order to evaluate the variability in the epidemic outcomes and test the performance of the various mobility networks. In general, we observed that both mobile phone networks present early peaks in the epidemic, explained by their larger fluxes connecting different locations. The census network and the radiation network instead show very similar epidemic profiles, that often coincide, thus reflecting the similarity we found in the structure of the two networks. Lastly the gravity network shows late peaks due to its small fluxes as compared to the other mobility networks.

The epidemic of Zika in Colombia reached the highest number of reported cases at week 2016-05 and in the second scenario, corresponding to R_0=4.8, both mobile phone networks were

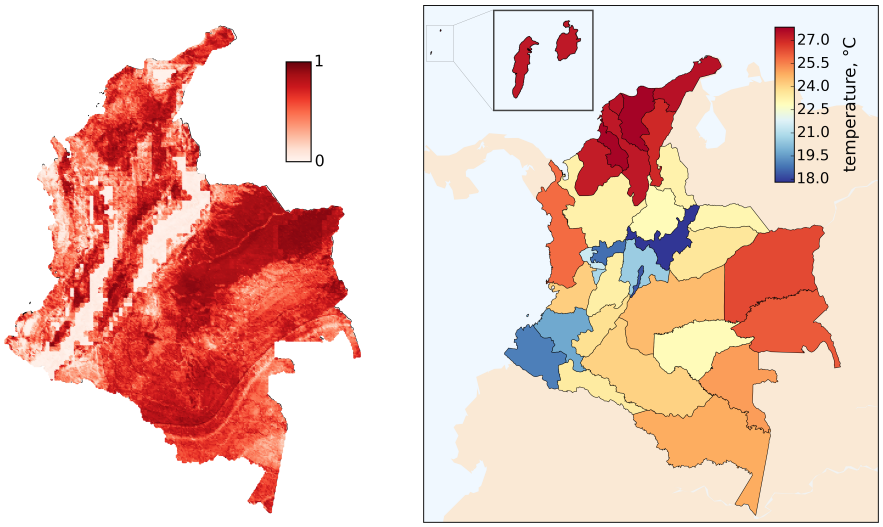


Figure 4.13: Left: Probability distribution of mosquitoes in Colombia. Right: Average monthly air temperature in Colombia.

able to capture the peak week within their 95% CI and the correlation was generally high for all mobility networks, ranging from 0.758 for the mobile phone network MP1 to 0.802 for the radiation network.

However, at this stage of the work, the results found for these three epidemic scenarios are not sufficient to assess the performance of the various mobility networks. Also, at the resolution of departments, the temporal evolution of the number of reported cases shows strong heterogeneity with non-well defined peaks, that most of the time are very broad or present multiple spikes during the epidemic, thus being very difficult to predict, especially for the strong assumptions of geographical homogeneity we made in our modelling approach. Indeed, we made several assumptions and simplifications in the study of the spatiotemporal spreading of Zika. First of all, the epidemiological parameters that intervene in the model are taken from the literature as the average values of the ranges reported in similar studies. For some specific parameters, such the transmission probability and the mosquito lifespan, there exists some polynomial expressions describing their temperature dependence, that might be used to modulate the spatiotemporal variability of these parameters [206]. On the other hand, Markov Chain Monte Carlo (MCMC) methods can be used to estimate the key model parameters involved in the infection dynamics by calibrating the model on the official surveillance data. In the next future, we plan to calibrate the model using a Metropolis-Hastings algorithm as well as a deeper exploration of possible ranges of the epidemiological parameters to understand which factors affect more the model.

Moreover, the underreporting of Zika cases is here assumed to correspond to 1% and the abundance of mosquitoes is considered to be homogeneous in the country, corresponding to 2 mosquitoes per person (i.e. the population of mosquito is twice the population of humans). This latter is indeed a strong assumption as we know that the abundance of mosquitoes is associated with strong spatial heterogeneity, driven by variability and seasonality due to the temperature dependence modulating the vector competence. Many locations, such as those at high elevation, are not at risk for autochthonous ZIKV transmission simply because the vector is absent as we observed for the country's capital Bogotá where no cases have been reported. In other locations the vector may be present but sustained transmission may not be possible because of environmental factors that affect the vector's population dynamics, such as temperature or precipitation. Housing conditions, availability of air conditioning, and socioeconomic factors also contribute significantly in determining the fraction of the population likely exposed to the vector. Indeed, for a vector-borne disease like Zika, additional layers of information concerning the distribution of the vector and the consequent exposed populations due to socio-economic factors,

are extremely important in order to model a more realistic spatiotemporal transmissibility of the disease. In the future we will integrate several other fundamental datasets at a high spatial resolution, including the following:

- **Temperature** The global air temperature dataset (available at `climate.geog.udel.edu/climate/`) contains average monthly air temperature interpolated to a 0.5 by 0.5 degree grid resolution (centered on 0.25 degree). To match the spatial resolution of departments, grid nodes are associated to each department according to the position (coordinates). If none nodes fall within a department (e.g. islands), the nearest node is associated to it. A department contains n measurements of the temperature for each month m, $T_1^m, .., T_n^m$, thus we consider the average temperature as $\overline{T}^m = \sum_{i=1}^{n} T_i^m$, for m=January, ..., December. Daily average temperatures are linearly interpolated from monthly averages. Figure 4.13 shows the average air temperature in Colombia.

- **Distribution of mosquitoes** The global *A. aegypti* and *A. albopictus* distribution database provides uncertainty estimates for the vector's distribution at a spatial resolution of 5 km x 5 km [126, 127]. The probability distribution of mosquitoes in Colombia is shown in Figure 4.13. Colombia is characterized by the presence of the Andes mountains, where the probability of having mosquitoes is almost zero.

- **Socio-economic factor** Dataset from G-Econ Project, Yale University (available at `http://gecon.yale.edu/colombia`) contains the per capita Gross Domestic Product (GDP, computed at purchasing power parity (PPP) exchange rates) at a 1 degree longitude by 1 degree latitude resolution (approximately 100 km by 100 km).

Human mobility is certainly one of the main factors that affect the epidemic spreading of an infectious disease, but in this case the correlation of ZIKV transmission dynamics with environmental, climatic and socio-economic factors play a fundamental role [128, 139, 206]. These extra datasets must be included in the model in order to assess and justify the epidemic outcomes obtained from the various mobility networks.

Conclusion

In this dissertation we explored how novel digital data streams can be integrated into the monitoring, modelling and forecasting of the epidemic spreading of infectious diseases. The pervasive use of digital communication technologies, including electronic devices and smartphones, has dramatically changed the way people communicate and search for information in real-time through the Web [97], generating a large volume of digital traces left by human activities on social media, crowdsourced platforms and Internet in general [174]. Modern epidemiology has thus been exposed to a new digital revolution and transformation of the pre-existing practices into digital disease detection techniques and digital warning systems that harness new technologies and novel data streams to monitor the health of populations and predict the future course of an epidemic.

The first part of this work is dedicated to an innovative online tool called Influweb [24] that since 2008 aims at monitoring seasonal influenza activity in Italy by directly involving self-selected volunteers among the general population reporting their health status through Web-based surveys. In this way, Influweb is able to provide an additional layer of surveillance to the traditional existing practices based on general practitioners' reports. We thoroughly investigated the representativeness of the population monitored by Influweb in terms of demographic and geographic indicators, as well as the epidemiological signal that can be extracted from Influweb collected data, and we found that, despite the limited sample and the existing participation biases, Influweb is a powerful tool to estimate weekly incidence rates as compared to the traditional surveillance data, with additional advantages provided by a real-time component and a higher spatial resolution [167]. Indeed, recruiting and maintaining participants are the main challenges, but targeted strategies informed by the results of our study can be implemented to increase participation rates in Italy. With the vast majority of people getting online and the increasing community engagement in public health, such participatory Web-based systems may become in the near future powerful tools to measure health status, opinion, or behaviour of the general population with regard to different indicators and diseases.

In the second part of this dissertation, we demonstrated how such digital crowdsourced data from participatory systems can be used in the scope of real-time forecasting of seasonal influenza epidemics by using different forecasting techniques, from simple linear autoregressive models to a dynamical mechanistic model called GLEAM (GLobal Epidemic And Mobility model), which is able to produce realistic simulations of the global spread of infectious diseases by combining high-resolution data on populations and human mobility with stochastic mathematical models of disease transmission simulating the spatial and temporal evolution of epidemics at the level of single individuals. In particular, in the first approach we used a previously validated computational framework [204, 205] based on real-time influenza-related data, traditional surveillance reports and stochastic Monte Carlo simulations computed by GLEAM, to provide long-term predictions of seasonal influenza activity, as well as key indicators, such as the peak timing and peak intensity, up to four weeks in advance. The novel component of this methodology consists in the calibration of the model with the initial infections estimated from various participatory Web-based systems available in Europe under the umbrella of Influenzanet [21], including Bel-

gium, Denmark, Italy, the Netherlands, Spain, and the United Kingdom [58]. In the second approach, we investigated how traditional surveillance data reported by general practitioners can be combined with digital surveillance data from the participatory Web-based system Influweb, in order to improve seasonal influenza forecasts in Italy [168]. We used some standard autoregressive models to address the main issues affecting traditional surveillance systems (i.e. reporting lags and continuous revision of data throughout an influenza season) and to exploit one of the main advantages of Influweb of having earlier data available. Indeed, statistical and mechanistic models present very different features and level of knowledge on the disease and the biological mechanisms that drive the infection dynamics [136]. However, in both approaches we were able to highlight the added value provided by integrating a digital real-time participatory component into seasonal influenza forecasting models.

In the third and last part of this work, we focused on the recent Zika outbreak occurred in Colombia in 2015-2016, mainly addressing the role of human mobility in the epidemic spreading of the disease. The continuous growth of the transportation infrastructure and the high human mobility and connection among different parts of the globe lead to a larger opportunity for infectious diseases to spread on a large scale more rapidly than ever before. Accurate human mobility data are needed to properly inform epidemic models and assess the spatial spreading of the disease and the risk of importation from the affected areas to the rest of the world, thus allowing for rapid interventions and appropriate control measures [54, 115, 169, 196]. Here we investigated different mobility networks, mainly focusing on the potential benefits of integrating human mobility patterns provided by mobile phone data, as well as human mobility generated by mobility models (i.e. the gravity model and the radiation model) and census commuting data.

Indeed, much more needs to be done to integrate such digital disease monitoring techniques and computational modelling and forecasting approaches into existing practices in public health and the efforts presented in this dissertation are only few glimpses of what can be done to help guide this process. In our case, since 2012 data collected by Influweb have been adopted by the Italian National Institute of Health [23] as an additional source of data about the circulation of influenza-like illness among the general population by collecting in a single weekly bulletin, called FluNews [6], all information gathered by the various epidemiological surveillance systems monitoring seasonal influenza in Italy. Moreover, in 2014 we launched FluOutlook (`fluoutlook.org`) [7], an online platform exposing real-time seasonal influenza forecasts projected up to four weeks into the future, thus providing a description of the seasonal influenza progression that could be used by public health agency to guide their decision making process, as well as to compare and assess the performance of different forecast approaches.

In the event of epidemic outbreaks public health authorities can choose to activate prevention and emergency response policies, such as vaccination, travel restrictions, or school and business closures, but the socioeconomic cost of such programmes can be high and their impact on the epidemic hard to determine. In this context, a high-resolution mechanistic model like GLEAM, allows the modelling and scenario analysis of containment and mitigation strategies providing quantitative projections that better informs the analysis of their likely impact. Furthermore, it can be easily refined with multiple data sources, from weather data information, to specific contact matrices and genetic sequence data, in order to include an increasing level of details and achieve better performances in the modelling and response to real epidemic situation. Luckily, we can now witness public health authorities more and more willing to adapt to this ongoing evolution and revolution, and to adopt these novel powerful tools for prevention strategies and emergency response planning. Certainly digital epidemiology represents a good premise in these challenges, thanks to the great improvements in the speed, scope, and focus of information available for public health purposes.

Appendix A

Influenzanet questionnaires

A.1 Intake Questionnaire

Intake Q0:
For whom are you filling this survey in? (If you are filling in the survey on behalf of someone else, then make sure that you have the consent of that person to do so)
Q0.1 - Yes
Q0.2 - No

Intake Q1:
What is your gender?
Q1.1 - Male
Q1.2 - Female

Intake Q2:
What is your date of birth (month and year)?
Q2.1 - XX/XXXX

Intake Q3:
What is your home postal code?
Q3.1 - XXXX

Intake Q4:
What is your main activity?
Q4.1 - Paid employment, full-time
Q4.2 - Paid employment, part-time
Q4.3 - Self-employed (businessman, farmer, tradesman, etc.)
Q4.4 - Attending daycare/school/college/university
Q4.5 - Home-maker (e.g. housewife)
Q4.6 - Unemployed
Q4.7 - Long-term sick-leave or parental leave
Q4.8 - Retired
Q4.9 - Other

Intake Q4b: (follow-up question to Q4.1 and Q4.2)
What is the first part of your school/college/workplace postal code (where you spend the majority of your working/studying time)?
Q4b.1 - XXXX
Q4b.2 - I don't know/can't remember

Q4b.3 - Not applicable (e.g. don't have a fixed workplace)

Intake Q4c: (follow-up question to Q4.1 and Q4.2)
Which of the following descriptions most closely matches with your main occupation?
Q4c.1 - Professional (e.g. manager, doctor, teacher, nurse, engineer)
Q4c.2 - Office work (e.g. admin, finance assistant, receptionist, etc.)
Q4c.3 - Retail, sales, catering and hospitality and leisure (e.g. shop assistant, waiter, bar-staff, gym instructor etc.)
Q4c.4 - Skilled manual worker (e.g. mechanic, electrician, technician)
Q4c.5 - Other manual work (e.g. cleaning, security, driver)
Q4c.6 - Other [free text here to allow users to write their occupation]

Intake Q4d: (only if age>16)
What is the highest level of formal education qualification that you have? (If you are still in education, then tick this box with the appropriate highest level that you have already achieved)
Q4d.1 - I have no formal qualifications
Q4d.2 - GCSE's, O'levels, CSEs or equivalent
Q4d.3 - A-Levels or equivalent (e.g. Highers, NVQ Level3, BTEC)
Q4d.4 - Bachelor's Degree (BA, BSc) or equivalent (e.g. HND, NVQ Level 4)
Q4d.5 - Higher Degree or equivalent (e.g. Masters Degree, PGCE, PhD, Medical Doctorate, Advanced Professional Awards)
Q4d.6 - I am still in education

Intake Q5:
Except people you meet on public contact, do you have contact with any of the following during the course of a typical day? (Select all options that apply, if any)
Q5.1 - More than 10 children or teenagers over the course of the day
Q5.2 - More than 10 people aged over 65 over the course of day
Q5.3 - Patients
Q5.4 - Groups of people (more than 10 individuals at any one time)
Q5.5 - None of the above

Intake Q6:
INCLUDING YOU, how many people in each of the following age groups live in your household? (Insert number for each age groups)
Q6.1 - 0-4 years
Q6.2 - 5-18 years
Q6.3 - 19-44 years
Q6.4 - 45-64 years
Q6.5 - 65+ years

Intake Q6b: (if any in household are aged 0-18, including participant)
How many of the children in your household go to school or day-care?
Drop-down menu

Intake Q7:
What is your main means of transportation?
Q7.1 - Walking
Q7.2 - Bike
Q7.3 - Motorbike/scooter
Q7.4 - Car
Q7.5 - Public transportation (bus, train, tube, etc.)

Q7.6 - Other

Intake Q7b:
On a normal day, how much time do you spend on public transport (bus, train, tube, etc.)?
Q7b.1 - No time at all
Q7b.2 - 0-30 minutes
Q7b.3 - 30 minutes - 1.5 hours
Q7b.4 - 1.5 hours - 4 hours
Q7b.5 - Over 4 hours

Intake Q8:
How often do you have common colds or flu-like diseases?
Q8.1 - Never
Q8.2 - Once or twice a year
Q8.3 - Between 3 and 5 times a year
Q8.4 - Between 6 and 10 times a year
Q8.5 - More than 10 times a year
Q8.6 - I don't know

Intake Q9:
Have you received a flu vaccine this autumn/winter season?
Q9.1 - Yes
Q9.2 - No
Q9.3 - I don't know/can't remember

Intake Q9b: (follow-up question to Q9.1)
When were you vaccinated against flu this season?
Q9b.1 - XX/XX/XXXX
Q9b.2 - I don't know/can't remember

Intake Q9c: (follow-up question to Q9.1)
What were your reasons for getting a seasonal influenza vaccination this year? (Select all options that apply)
Q9c.1 - I belong to a risk group (e.g. pregnant, over 65, underlying health condition, etc.)
Q9c.2 - Vaccination decreases my risk of getting influenza
Q9c.3 - Vaccination decreases the risk of spreading influenza to others
Q9c.4 - My doctor recommended it
Q9c.5 - It was recommended in my workplace/school
Q9c.6 - The vaccine was readily available and vaccine administration was convenient o The vaccine was free (no cost)
Q9c.7 - I don't want to miss work/school
Q9c.8 - I always get the vaccine
Q9c.9 - Other reason(s)

Intake Q9d: (follow-up question to Q9.2)
What were your reasons for NOT getting a seasonal influenza vaccination this year? (Select all options that apply)
Q9d.1 - I am planning to be vaccinated, but haven't been yet
Q9d.2 - I haven't been offered the vaccine
Q9d.3 - I don't belong to a risk group
Q9d.4 - It is better to build your own natural immunity against influenza
Q9d.5 - I doubt that the influenza vaccine is effective
Q9d.6 - Influenza is a minor illness

Q9d.7 - I don't think that I am likely to get influenza
Q9d.8 - I believe that influenza vaccine can cause influenza
Q9d.9 - I am worried that the vaccine is not safe or will cause illness or other adverse events
Q9d.10 - I don't like having vaccinations
Q9d.11 - The vaccine is not readily available to me
Q9d.12 - The vaccine is not free of charge
Q9d.13 - No particular reason
Q9d.14 - Although my doctor recommended a vaccine, I did not get one
Q9d.15 - Other reason(s)

Intake Q10:
Did you receive a flu vaccine during the last autumn/winter season?
Q10.1 - Yes
Q10.2 - No
Q10.3 - I don't know/can't remember

Intake Q11:
Do you take regular medication for any of the following medical conditions? (Select all options that apply)
Q11.1 - No
Q11.2 - Asthma
Q11.3 - Diabetes
Q11.4 - Chronic lung disorder besides asthma e.g. COPD, emphysema, or other disorders that affect your breathing
Q11.5 - Heart disorder
Q11.6 - Kidney disorder
Q11.7 - An immunocompromising condition from treatment or illness including splenectomy, organ transplant, acquired immune deficiency, cancer treatment

Intake Q12: (Only to women aged between 15 and 50)
Are you currently pregnant?
Q12.1 - Yes
Q12.2 - No
Q12.3 - Don't know/would rather not answer

Intake Q12b: (follow-up question to Q12.1)
Which trimester of the pregnancy are you in?
Q12b.1 - First trimester (week 1-12)
Q12b.2 - Second trimester (week 13-28)
Q12b.3 - Third trimester (week 29-delivery)
Q12b.4 - Don't know/would rather not answer

Intake Q13:
Do you smoke tobacco?
Q13.1 - No
Q13.2 - Yes, occasionally
Q13.3 - Yes, daily, fewer than 10 times a day
Q13.4 - Yes, daily, 10 or more times a day
Q13.5 - Don't know/would rather not answer

Intake Q14:
Do you have one of the following allergies that can cause respiratory symptoms? (Select all options that apply)

Q14.1 - Hay fever
Q14.2 - Allergy against house dust mite
Q14.3 - Allergy against domestic animals or pets
Q14.4 - Other allergies that cause respiratory symptoms (e.g. sneezing, runny eyes)
Q14.5 - I do not have an allergy that causes respiratory symptoms

Intake Q15:
Do you follow a special diet? (Select all options that apply)
Q15.1 - No special diet
Q15.2 - Vegetarian
Q15.3 - Veganism
Q15.4 - Low-calorie
Q15.5 - Other

Intake Q16:
Do you have pets at home? (Select all options that apply)
Q16.1 - No
Q16.2 - Yes, one or more dogs
Q16.3 - Yes, one or more cats
Q16.4 - Yes, one or more birds
Q16.5 - Yes, one ore more other animals

A.2 Symptoms Questionnaire

If you are filling this in on behalf of someone else, please answer all the questions as if you are that person.

Weekly Q1:
Have you had any of the following symptoms since your last visit (or in the past week, if this is your first visit)? (Select all options that apply)
Q1.1 - No symptoms
Q1.2 - Fever
Q1.3 - Chills
Q1.4 - Runny or blocked nose
Q1.5 - Sneezing
Q1.6 - Sore throat
Q1.7 - Cough
Q1.8 - Shortness of breath
Q1.9 - Headache
Q1.10 - Muscle/joint pain
Q1.11 - Chest pain
Q1.12 - Feeling tired or exhausted (malaise)
Q1.13 - Loss of appetite
Q1.14 - Coloured sputum/phlegm
Q1.15 - Watery, bloodshot eyes
Q1.16 - Nausea
Q1.17 - Vomiting
Q1.18 - Diarrhoea
Q1.19 - Stomach ache
Q1.20 - Other

Weekly Q2: (If the participant was still ill on their last visit and has reported symptoms this time)
On *DATE OF LAST VISIT* you reported that you were still ill with symptoms that began on *DATE OF FIRST SYMPTOMS REPORTED PREVIOUSLY*. Are the symptoms you reported today part of the same bout of illness?
Q2.1 - Yes
Q2.2 - No
Q2.3 - I don't know/can't remember

Weekly Q3: (if symptoms)
When did the first symptoms appear?
Q3.1 - Choose date XX/XX/XXXX
Q3.2 - I don't know/can't remember

Weekly Q4: (if symptoms)
When did your symptoms end?
Q4.1 - Choose date XX/XX/XXXX
Q4.2 - I don't know/can't remember
Q4.3 - I am still ill

Weekly Q5: (if symptoms)
Did your symptoms develop suddenly over a few hours?
Q5.1 - Yes
Q5.2 - No
Q5.3 - I don't know/can't remember

Weekly Q6 (if fever)
When did your fever begin?
Q6.1 - Choose date XX/XX/XXXX
Q6.2 - I don't know/can't remember

Weekly Q6a: (if fever)
Did your fever develop suddenly over a few hours?
Q6a.1 - Yes
Q6a.2 - No
Q6a.3 - Don't know

Weekly Q6b: (if symptoms)
Did you take your temperature?
Q6b.1 - Yes
Q6b.2 - No
Q6b.3 - I don't know

Weekly Q6c: (if symptoms and if took temperature)
What was your highest temperature measured?
Q6c.1 - Below 37°C
Q6c.2 - 37° - 37.4°C
Q6c.3 - 37.5° - 37.9°C
Q6c.4 - 38° - 38.9°C
Q6c.5 - 39° - 39.9°C
Q6c.6 - 40°C or more
Q6c.7 - I don't know/can't remember

Weekly Q7: (if symptoms)
Because of your symptoms, did you VISIT (see face to face) any of medical services? (Select all options that apply)
Q7.1 - No
Q7.2 - GP or GP's practice nurse
Q7.3 - Hospital admission
Q7.4 - Hospital accident & emergency department/out of hours service o Other medical services
Q7.5 - No, but I have an appointment scheduled

Weekly Q7b: (if symptoms)
How soon after your symptoms appeared did you visit this medical service?
Q7b.1 - Same day
Q7b.2 - 1 day
Q7b.3 - 2 days
Q7b.4 - 3 days
Q7b.5 - 4 days
Q7b.6 - 5-7 days
Q7b.7 - More than 7 days
Q7b.8 - I don't know/can't remember

Weekly Q8: (if symptoms)
Because of your symptoms, did you contact via telephone or internet any of the following? (Select all options that apply)
Q8.1 - No
Q8.2 - GP, spoke to receptionist only
Q8.3 - GP, spoke to doctor or nurse
Q8.4 - Other

Weekly Q8b: (if symptoms)
How soon after your symptoms appeared did you contact via telephone or internet any of the following?
Q8b.1 - Same day
Q8b.2 - 1 day
Q8b.3 - 2 days
Q8b.4 - 3 days
Q8b.5 - 4 days
Q8b.6 - 5-7 days
Q8b.7 - More than 7 days
Q8b.8 - I don't know/can't remember

Weekly Q9: (if symptoms)
Did you take medication for these symptoms? (Select all options that apply)
Q9.1 - No medication
Q9.2 - Pain killers (e.g. paracetamol, ibuprofen, aspirin, etc.)
Q9.3 - Cough medication (e.g. expectorants)
Q9.4 - Antivirals (Tamiflu, Relenza)
Q9.5 - Antibiotics
Q9.6 - Other
Q9.7 - I don't know/can't remember

Weekly Q9b: (follow-up question to Q9.4)
How long after the beginning of your symptoms did you start taking antiviral medication?

Q9b.1 - Same day (within 24 hours)
Q9b.2 - 1 day later
Q9b.3 - 2 days later
Q9b.4 - 3 days later
Q9b.5 - 4 days later
Q9b.6 - 5-7 days later
Q9b.7 - More than 7 days later
Q9b.8 - I don't know/can't remember

Weekly Q10: (if symptoms)
Did you change your daily routine because of your illness?
Q10.1 - No
Q10.2 - Yes, but I did not take time off work/school
Q10.3 - Yes, I took time off work/school

Weekly Q10b: (follow-up question to Q10.3)
Are you still off work/school?
Q10b.1 - Yes
Q10b.2 - No
Q10b.3 - Other (e.g. I wouldn't usually be at work/school today anyway)

Weekly Q10c: (follow-up question to Q10.3)
How have you been off work/school for?
Q10c.1 - 1 day
Q10c.2 - 2 days
Q10c.3 - 3 days
Q10c.4 - 4 days
Q10c.5 - 5 days
Q10c.6 - 6 to 10 days
Q10c.7 - 11 to 15 days
Q10c.8 - More than 15 days

Weekly Q11: (if symptoms)
What do you think is causing your symptoms?
Q11.1 - Flu or flu-like illness o Common cold
Q11.2 - Allergy/hay fever
Q11.3 - Asthma
Q11.4 - Gastroenteritis/gastric flu
Q11.5 - Other

Bibliography

[1] Cumulative Zika cases reported by countries and territories in the Americas up to 29 December 2016. Available at: http://www.paho.org/hq/index.php?option=com_docman&task=doc_view&Itemid=270&gid=37582&lang=en (accessed on October 31, 2017).

[2] De Grote Griepmeting (Holland and Belgium). Available at: https://www.degrotegriepmeting.nl (accessed on October 31, 2017).

[3] Departamento Administrativo Nacional de Estadística (DANE). Available at: http://www.dane.gov.co (accessed on October 31, 2017).

[4] EbolaTracking. Available at: http://ebolatracking.org/ (accessed on October 31, 2017).

[5] Flu Near You (USA). Available at: https://flunearyou.org/ (accessed on October 31, 2017).

[6] FluNews. Available at: http://www.epicentro.iss.it/problemi/influenza/FluNews.asp. (accessed on June 7, 2016).

[7] FluOutlook, an Epidemic Forecasting Observatory. Available at: http://fluoutlook.org/ (accessed on October 31, 2017).

[8] Flusurvey (Ireland). Available at: https://flusurvey.ie/ (accessed on October 31, 2017).

[9] Flusurvey (United Kingdom). Available at: https://flusurvey.org.uk/ (accessed on October 31, 2017).

[10] FluTracking (Australia). Available at: http://www.flutracking.net/ (accessed on October 31, 2017).

[11] Gripenet (Portugal). Available at: http://www.gripenet.pt (accessed on October 31, 2017).

[12] Gripenet (Spain). Available at: https://www.gripenet.es/ (accessed on October 31, 2017).

[13] Grippenet (France). Available at: https://www.grippenet.fr/ (accessed on October 31, 2017).

[14] Grippenet (Switzerland). Available at: http://www.grippenet.ch/ (accessed on October 31, 2017).

[15] GrippeWeb (Germany). Available at: https://grippeweb.rki.de (accessed on October 31, 2017).

[16] Halsorapport (Sweden). Available at: https://www.halsorapport.se/sv/ (accessed on October 31, 2017).

[17] HealthMap. Available at: http://www.healthmap.org (accessed on October 31, 2017).

[18] Influenza case definition. Available at: http://ecdc.europa.eu/en/healthtopics/influenza/surveillance/Pages/influenza_case_definitions.aspx.

[19] Influenza (Seasonal), Fact Sheet Number 211. Available at: http://www.who.int/mediacentre/factsheets/fs211/en/index.html.

[20] Influenza Vaccine Coverage. Ministry of Health. 2016. Available at: http://www.salute.gov.it/portale/documentazione/p6_2_8_3_1.jsp?lingua=italiano&id=19 (accessed on December 6, 2016).

[21] Influenzanet. Available at: https://www.influenzanet.eu (accessed on October 31, 2017).

[22] Influmeter (Denmark). Available at: http://www.influmeter.dk (accessed on October 31, 2017).

[23] Influnet website. Available at: http://www.iss.it/flue/index.php?lang=1.

[24] Influweb (Italy). Available at: https://www.influweb.it (accessed on October 31, 2017).

[25] Instituto Nacional de Salud (INS). Available at: www.ins.gov.co (accessed on October 31, 2017).

[26] International Air Transport Association (IATA), available at: http://www.iata.org.

[27] National Institute for Statistics Studies (ISTAT). Available at: dati.istat.it (accessed on October, 2017).

[28] Reporta (Mexico). Available at: http://reporta.c3.org.mx (accessed on October 31, 2017).

[29] Salud Boricua (Puerto Rico). Available at: https://saludboricua.org/ (accessed on October 31, 2017).

[30] WorldPop. Available at: worldpop.org.uk (accessed on 31 October, 2017).

[31] ZikaTracking. Available at: http://zikatracking.org (accessed on October 31, 2017).

[32] A. J. Adler, K. T. Eames, S. Funk, and W. J. Edmunds. Incidence and risk factors for influenza-like-illness in the UK: online surveillance using Flusurvey. *BMC infectious diseases*, 14(1):1, 2014.

[33] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[34] L. Alexander, S. Jiang, M. Murga, and M. C. González. Origin–Destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58:240–250, 2015.

[35] B. M. Althouse, S. V. Scarpino, L. A. Meyers, J. W. Ayers, M. Bargsten, J. Baumbach, J. S. Brownstein, L. Castro, H. Clapham, D. A. Cummings, et al. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Science*, 4(1):1, 2015.

[36] A. Amini, K. Kung, C. Kang, S. Sobolevsky, and C. Ratti. The impact of social segregation on human mobility in developing and industrialized regions. *EPJ Data Science*, 3(1):6, 2014.

[37] R. M. Anderson, R. M. May, and B. Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992.

[38] P. Bajardi, M. Delfino, A. Panisson, G. Petri, and M. Tizzoni. Unveiling patterns of international communities in a global city using mobile phone data. *EPJ Data Science*, 4(1):3, 2015.

[39] P. Bajardi, D. Paolotti, A. Vespignani, K. Eames, S. Funk, W. J. Edmunds, C. Turbelin, M. Debin, V. Colizza, R. Smallenburg, et al. Association between recruitment methods and attrition in Internet-based studies. *PloS one*, 9(12):e114925, 2014.

[40] P. Bajardi, C. Poletto, D. Balcan, H. Hu, B. Gonçalves, J. J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, et al. Modeling vaccination campaigns and the Fall/Winter 2009 activity of the new A (H1N1) influenza in the Northern Hemisphere. *Emerging Health Threats Journal*, 2(1):7093, 2009.

[41] P. Bajardi, C. Poletto, J. J. Ramasco, M. Tizzoni, V. Colizza, and A. Vespignani. Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. *PloS one*, 6(1):e16591, 2011.

[42] P. Bajardi, A. Vespignani, S. Funk, K. T. Eames, W. J. Edmunds, C. Turbelin, M. Debin, V. Colizza, R. Smallenburg, C. E. Koppeschaar, et al. Determinants of follow-up participation in the Internet-based European influenza surveillance platform Influenzanet. *Journal of medical Internet research*, 16(3):e78, 2014.

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

[44] D. Balcan, V. Colizza, A. C. Singer, C. Chouaid, H. Hu, B. Gonçalves, P. Bajardi, C. Poletto, J. J. Ramasco, N. Perra, et al. Modeling the critical care demand and antibiotics resources needed during the Fall 2009 wave of influenza A (H1N1) pandemic. *PLoS currents*, 1, 2009.

[45] D. Balcan, B. Gonçalves, H. Hu, J. J. Ramasco, V. Colizza, and A. Vespignani. Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *Journal of computational science*, 1(3):132–145, 2010.

[46] D. Balcan, H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, et al. Seasonal transmission potential and activity peaks of the new influenza A (H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC medicine*, 7(1):45, 2009.

[47] S. Bansal, G. Chowell, L. Simonsen, A. Vespignani, and C. Viboud. Big data for infectious disease surveillance and modeling. *The Journal of Infectious Diseases*, 214(suppl_4):S375–S379, 2016.

[48] E. Barclay. Predicting the next pandemic, 2008.

[49] A. Barrat, M. Barthelemy, and A. Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.

[50] M. G. Beiró, A. Panisson, M. Tizzoni, and C. Cattuto. Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Science*, 5(1):30, 2016.

[51] L. Bengtsson, J. Gaudart, X. Lu, S. Moore, E. Wetter, K. Sallah, S. Rebaudet, and R. Piarroux. Using mobile phone data to predict the spatial spread of cholera. *Scientific reports*, 5, 2015.

[52] M. Besnard, S. Lastere, A. Teissier, V. Cao-Lormeau, D. Musso, et al. Evidence of perinatal transmission of Zika virus, French Polynesia, December 2013 and February 2014. *Euro surveill*, 19(13):20751, 2014.

[53] M. Biggerstaff, D. Alper, M. Dredze, S. Fox, I. C.-H. Fung, K. S. Hickmann, B. Lewis, R. Rosenfeld, J. Shaman, M.-H. Tsou, et al. Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. *BMC Infectious Diseases*, 16(1):357, 2016.

[54] I. I. Bogoch, O. J. Brady, M. Kraemer, M. German, M. I. Creatore, M. A. Kulkarni, J. S. Brownstein, S. R. Mekaru, S. I. Hay, E. Groot, et al. Anticipating the international spread of Zika virus from Brazil. *Lancet (London, England)*, 387(10016):335–336, 2016.

[55] D. Brockmann and D. Helbing. The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342(6164):1337–1342, 2013.

[56] D. A. Broniatowski, M. J. Paul, and M. Dredze. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12):e83672, 2013.

[57] E. Brooks-Pollock, N. Tilston, W. J. Edmunds, and K. T. Eames. Using an online survey of healthcare-seeking behaviour to estimate the magnitude and severity of the 2009 H1N1v influenza epidemic in England. *BMC infectious diseases*, 11(1):1, 2011.

[58] J. S. Brownstein, S. Chu, A. Marathe, M. V. Marathe, A. T. Nguyen, D. Paolotti, N. Perra, D. Perrotta, M. Santillana, S. Swarup, et al. Combining participatory influenza surveillance with modeling and forecasting: Three alternative approaches. *JMIR Public Health and Surveillance*, 3(4):e83, 2017.

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

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

[61] P. Butler, N. Ramakrishnan, E. O. Nsoesie, J. S. Brownstein, et al. Satellite imagery analysis: What can hospital parking lots tell us about a disease outbreak? 2014.

[62] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area. *IEEE Pervasive Computing*, 99, 2011.

[63] F. Calabrese, Z. Smoreda, V. D. Blondel, and C. Ratti. Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PloS one*, 6(7):e20814, 2011.

[64] P. Cantarelli, M. Debin, C. Turbelin, C. Poletto, T. Blanchon, A. Falchi, T. Hanslik, I. Bonmarin, D. Levy-Bruhl, A. Micheletti, et al. The representativeness of a European multi-center network for influenza-like-illness participatory surveillance. *BMC public health*, 14(1):1, 2014.

[65] S. J. Carlson, C. B. Dalton, D. N. Durrheim, J. Fejsa, et al. Online FluTracking survey of influenza-like illness during pandemic (H1N1) 2009, Australia. *Emerg Infect Dis*, 16(12):1960–1962, 2010.

[66] S. J. Carlson, D. N. Durrheim, and C. B. Dalton. FluTracking provides a measure of field influenza vaccine effectiveness, Australia, 2007–2009. *Vaccine*, 28(42):6809–6810, 2010.

[67] P. Chakraborty, P. Khadivi, B. Lewis, A. Mahendiran, J. Chen, P. Butler, E. O. Nsoesie, S. R. Mekaru, J. S. Brownstein, M. V. Marathe, et al. Forecasting a moving target: Ensemble models for ILI case count predictions. In *SDM*, pages 262–270. SIAM, 2014.

[68] D. L. Chao, M. E. Halloran, V. J. Obenchain, and I. M. Longini Jr. FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS computational biology*, 6(1):e1000656, 2010.

[69] L. Chen, K. T. Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash. Flu gone viral: Syndromic surveillance of flu on Twitter using temporal topic models. In *2014 IEEE International Conference on Data Mining*, pages 755–760. IEEE, 2014.

[70] T. Chouin-Carneiro, A. Vega-Rua, M. Vazeille, A. Yebakima, R. Girod, D. Goindin, M. Dupont-Rouzeyrol, R. Lourenço-de Oliveira, and A.-B. Failloux. Differential Susceptibilities of Aedes aegypti and Aedes albopictus from the Americas to Zika Virus. *PLoS neglected tropical diseases*, 10(3):e0004543, 2016.

[71] G. Chowell, M. Miller, and C. Viboud. Seasonal influenza in the United States, France, and Australia: transmission and prospects for control. *Epidemiology & Infection*, 136(6):852–864, 2008.

[72] J.-P. Chretien, H. S. Burkom, E. R. Sedyaningsih, R. P. Larasati, A. G. Lescano, C. C. Mundaca, D. L. Blazes, C. V. Munayco, J. S. Coberly, R. J. Ashar, et al. Syndromic surveillance: adapting innovations to developing settings. *PLoS medicine*, 5(3):e72, 2008.

[73] J.-P. Chretien, D. George, J. Shaman, R. A. Chitale, and F. E. McKenzie. Influenza forecasting in human populations: a scoping review. *PloS one*, 9(4):e94130, 2014.

[74] R. Chunara, S. Aman, M. Smolinski, and J. S. Brownstein. Flu Near You: An online self-reported influenza surveillance system in the USA. *Online Journal of Public Health Informatics*, 5(1), 2012.

[75] R. Chunara, E. Goldstein, O. Patterson-Lomba, and J. S. Brownstein. Estimating influenza attack rates in the United States using a participatory cohort. *Scientific reports*, 5:9540, 2015.

[76] S. Çolak, A. Lima, and M. C. González. Understanding congested travel in urban areas. *Nature communications*, 7, 2016.

[77] V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, and A. Vespignani. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS medicine*, 4(1):e13, 2007.

[78] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani. Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study. *BMC medicine*, 5(1):34, 2007.

[79] V. Colizza, R. Pastor-Satorras, and A. Vespignani. Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nature Physics*, 3(4):276–282, 2007.

[80] V. Colizza and A. Vespignani. Invasion threshold in heterogeneous metapopulation networks. *Physical review letters*, 99(14):148701, 2007.

[81] V. Colizza and A. Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *Journal of theoretical biology*, 251(3):450–467, 2008.

[82] B. S. Cooper, R. J. Pitman, W. J. Edmunds, and N. J. Gay. Delaying the international spread of pandemic influenza. *PLoS medicine*, 3(6):e212, 2006.

[83] M. Coscia and R. Hausmann. Evidence that calls-based and mobility networks are isomorphic. *PloS one*, 10(12):e0145091, 2015.

[84] B. J. Cowling, K.-H. Chan, V. J. Fang, C. K. Cheng, R. O. Fung, W. Wai, J. Sin, W. H. Seto, R. Yung, D. W. Chu, et al. Facemasks and hand hygiene to prevent influenza transmission in households: A cluster randomized trial. *Annals of internal medicine*, 151(7):437–446, 2009.

[85] B. J. Cowling, K. H. Chan, V. J. Fang, L. L. Lau, H. C. So, R. O. Fung, E. S. Ma, A. S. Kwong, C.-W. Chan, W. W. Tsui, et al. Comparative epidemiology of pandemic and seasonal influenza A in households. *New England journal of medicine*, 362(23):2175–2184, 2010.

[86] A. W. Crawley. Flu Near You: Comparing crowdsourced reports of influenza-like illness to the CDC outpatient influenza-like illness surveillance network, October 2012 to March 2014. In *2014 CSTE Annual Conference*. Cste, 2014.

[87] A. Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122. ACM, 2010.

[88] C. B. Dalton, S. J. Carlson, M. T. Butler, E. Elvidge, and D. N. Durrheim. Building influenza surveillance pyramids in near real time, Australia. *Emerging infectious diseases*, 19(11):1863, 2013.

[89] C. B. Dalton, S. J. Carlson, L. McCallum, M. T. Butler, J. Fejsa, E. Elvidge, and D. N. Durrheim. FluTracking weekly online community survey of influenza-like illness: 2013 and 2014. *Communicable diseases intelligence quarterly report*, 39(3):E361, 2015.

[90] N. de Francisco, M. Donadel, M. Jit, and R. Hutubessy. A systematic review of the social and economic burden of influenza in low-and middle-income countries. *Vaccine*, 33(48):6537–6544, 2015.

[91] M. Debin, V. Colizza, T. Blanchon, T. Hanslik, C. Turbelin, and A. Falchi. Effectiveness of 2012–2013 influenza vaccine against influenza-like illness in general population: Estimation in a French web-based cohort. *Human vaccines & immunotherapeutics*, 10(3):536–543, 2014.

[92] E. D'Ortenzio, S. Matheron, X. de Lamballerie, B. Hubert, G. Piorkowski, M. Maquart, D. Descamps, F. Damond, Y. Yazdanpanah, and I. Leparc-Goffart. Evidence of sexual transmission of Zika virus. *New England Journal of Medicine*, 374(22):2195–2198, 2016.

[93] X. Du, A. A. King, R. J. Woods, and M. Pascual. Evolution-informed forecasting of seasonal influenza A (H3N2). *Science Translational Medicine*, 9(413):eaan5325, 2017.

[94] M. R. Duffy, T.-H. Chen, W. T. Hancock, A. M. Powers, J. L. Kool, R. S. Lanciotti, M. Pretrick, M. Marfel, S. Holzbauer, C. Dubray, et al. Zika virus outbreak on Yap Island, federated states of Micronesia. *New England Journal of Medicine*, 360(24):2536–2543, 2009.

[95] K. T. Eames, N. L. Tilston, E. Brooks-Pollock, and W. J. Edmunds. Measured dynamic social contact patterns explain the spread of H1N1 influenza. *PLoS computational biology*, 8(3):e1002425, 2012.

[96] W. J. Edmunds and S. Funk. Using the Internet to estimate influenza vaccine effectiveness. *Expert review of vaccines*, 11(9):1027–1029, 2012.

[97] G. Eysenbach. Infodemiology and infosurveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *Journal of medical Internet research*, 11(1), 2009.

[98] A. S. Fauci and D. M. Morens. The perpetual challenge of infectious diseases. *New England Journal of Medicine*, 366(5):454–461, 2012.

[99] N. M. Ferguson, D. A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056):209–214, 2005.

[100] N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448–452, 2006.

[101] F. Finger, T. Genolet, L. Mari, G. C. de Magny, N. M. Manga, A. Rinaldo, and E. Bertuzzo. Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. *Proceedings of the National Academy of Sciences*, 113(23):6421–6426, 2016.

[102] C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, et al. Pandemic potential of a strain of influenza A (H1N1): early findings. *Science*, 324(5934):1557–1561, 2009.

[103] C. Fraser, S. Riley, R. M. Anderson, and N. M. Ferguson. Factors that make an infectious disease outbreak controllable. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16):6146–6151, 2004.

[104] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein. 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, 2008.

[105] I. Friesema, C. Koppeschaar, G. Donker, F. Dijkstra, S. Van Noort, R. Smallenburg, W. Van der Hoek, and M. Van der Sande. Internet-based monitoring of influenza-like illness in the general population: experience of five influenza seasons in The Netherlands. *Vaccine*, 27(45):6353–6357, 2009.

[106] S. Funk, A. Camacho, A. J. Kucharski, R. M. Eggo, and W. J. Edmunds. Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. *Epidemics*, 2016.

[107] S. Funk, A. Camacho, A. J. Kucharski, R. Lowe, R. M. Eggo, and W. J. Edmunds. Assessing the performance of real-time epidemic forecasts. *bioRxiv*, page 177451, 2017.

[108] N. Generous, G. Fairchild, A. Deshpande, S. Y. Del Valle, and R. Priedhorsky. Global disease monitoring and forecasting with Wikipedia. *PLoS Comput Biol*, 10(11):e1003892, 2014.

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

[110] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[111] G. Grard, M. Caron, I. M. Mombo, D. Nkoghe, S. M. Ondo, D. Jiolle, D. Fontenille, C. Paupy, and E. M. Leroy. Zika virus in Gabon (Central Africa)–2007: a new threat from Aedes albopictus? *PLoS neglected tropical diseases*, 8(2):e2681, 2014.

[112] C. Guerrisi, C. Turbelin, T. Blanchon, T. Hanslik, I. Bonmarin, D. Levy-Bruhl, D. Perrotta, D. Paolotti, R. Smallenburg, C. Koppeschaar, et al. Participatory syndromic surveillance of influenza in Europe. *Journal of Infectious Diseases*, 214(suppl 4):S386–S392, 2016.

[113] M. E. Halloran, N. M. Ferguson, S. Eubank, I. M. Longini, D. A. Cummings, B. Lewis, S. Xu, C. Fraser, A. Vullikanti, T. C. Germann, et al. Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences*, 105(12):4639–4644, 2008.

[114] M. E. Halloran, I. M. Longini, and C. J. Struchiner. Binomial and stochastic transmission models. *Design and Analysis of Vaccine Studies*, pages 63–84, 2010.

[115] M. E. Halloran, A. Vespignani, N. Bharti, L. R. Feldstein, K. Alexander, M. Ferrari, J. Shaman, J. M. Drake, T. Porco, J. N. Eisenberg, et al. Ebola: mobility data. *Science*, 346(6208):433–433, 2014.

[116] S. Hasan, C. M. Schneider, S. V. Ukkusuri, and M. C. González. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151(1-2):304–318, 2013.

[117] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014.

[118] S. I. Hay, K. E. Battle, D. M. Pigott, D. L. Smith, C. L. Moyes, S. Bhatt, J. S. Brownstein, N. Collier, M. F. Myers, D. B. George, et al. Global mapping of infectious disease. *Phil. Trans. R. Soc. B*, 368(1614):20120250, 2013.

[119] S. I. Hay, A. Graham, and D. J. Rogers. *Global mapping of infectious diseases: methods, examples and emerging applications*, volume 62. Academic Press, 2006.

[120] K. S. Hickmann, G. Fairchild, R. Priedhorsky, N. Generous, J. M. Hyman, A. Deshpande, and S. Y. Del Valle. Forecasting the 2013–2014 influenza season using Wikipedia. *PLoS Comput Biol*, 11(5):e1004239, 2015.

[121] M. A. Johansson, A. M. Powers, N. Pesik, N. J. Cohen, and J. E. Staples. Nowcasting the spread of chikungunya virus in the Americas. *PloS one*, 9(8):e104915, 2014.

[122] M. J. Keeling. Metapopulation moments: coupling, stochasticity and persistence. *Journal of Animal Ecology*, 69(5):725–736, 2000.

[123] M. J. Keeling and P. Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2008.

[124] W. Kermack and A. McKendrick. A contribution to the mathematical theory of epidemics. *P Roy Soc Long A Mat.*, 115:700–721, 1927.

[125] C. E. Koppeschaar, V. Colizza, C. Guerrisi, C. Turbelin, J. Duggan, W. J. Edmunds, C. Kjelsø, R. Mexia, Y. Moreno, S. Meloni, et al. Influenzanet: Citizens among 10 countries collaborating to monitor influenza in Europe. *JMIR Public Health and Surveillance*, 3(3), 2017.

[126] M. U. Kraemer, M. E. Sinka, K. A. Duda, A. Mylne, F. M. Shearer, O. J. Brady, J. P. Messina, C. M. Barker, C. G. Moore, R. G. Carvalho, et al. The global compendium of Aedes aegypti and Ae. albopictus occurrence. *Scientific Data*, 2:sdata201535, 2015.

[127] M. U. Kraemer, M. E. Sinka, K. A. Duda, A. Q. Mylne, F. M. Shearer, C. M. Barker, C. G. Moore, R. G. Carvalho, G. E. Coelho, W. Van Bortel, et al. The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. *Elife*, 4:e08347, 2015.

[128] A. J. Kucharski, S. Funk, R. M. Eggo, H.-P. Mallet, W. J. Edmunds, and E. J. Nilles. Transmission dynamics of Zika virus in island populations: a modelling analysis of the 2013–14 French Polynesia outbreak. *PLoS neglected tropical diseases*, 10(5):e0004726, 2016.

[129] V. Lampos and N. Cristianini. Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):72, 2012.

[130] L. L. Lau, B. J. Cowling, V. J. Fang, K.-H. Chan, E. H. Lau, M. Lipsitch, C. K. Cheng, P. M. Houck, T. M. Uyeki, J. M. Peiris, et al. Viral shedding and clinical illness in naturally acquired influenza virus infections. *The Journal of infectious diseases*, 201(10):1509–1516, 2010.

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

[132] P. Lemey, A. Rambaut, T. Bedford, N. Faria, F. Bielejec, G. Baele, C. A. Russell, D. J. Smith, O. G. Pybus, D. Brockmann, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS pathogens*, 10(2):e1003932, 2014.

[133] M. Lenormand, A. Bassolas, and J. J. Ramasco. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography*, 51:158–169, 2016.

[134] M. Lenormand, S. Huet, F. Gargiulo, and G. Deffuant. A universal model of commuting networks. *PloS one*, 7(10):e45985, 2012.

[135] M. Lenormand, M. Picornell, O. G. Cantú-Ros, A. Tugores, T. Louail, R. Herranz, M. Barthelemy, E. Frias-Martinez, and J. J. Ramasco. Cross-checking different sources of mobility information. *PLoS One*, 9(8):e105184, 2014.

[136] J. Lessler and D. A. Cummings. Mechanistic models of infectious disease and their impact on public health. *American journal of epidemiology*, 183(5):415–422, 2016.

[137] E. T. Lofgren, M. E. Halloran, C. M. Rivers, J. M. Drake, T. C. Porco, B. Lewis, W. Yang, A. Vespignani, J. Shaman, J. N. Eisenberg, et al. Opinion: Mathematical models: A key tool for outbreak response. *Proceedings of the National Academy of Sciences*, 111(51):18095–18096, 2014.

[138] I. M. Longini, A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D. A. Cummings, and M. E. Halloran. Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087, 2005.

[139] G. Marini, G. Guzzetta, R. Rosà, and S. Merler. First outbreak of Zika virus in the continental United States: a modelling analysis. *Eurosurveillance*, 22(37), 2017.

[140] A. P. Masucci, J. Serras, A. Johansson, and M. Batty. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E*, 88(2):022812, 2013.

[141] S. A. McDonald, A. M. Presanis, D. D. Angelis, W. van der Hoek, M. Hooiveld, G. Donker, and M. E. Kretzschmar. An evidence synthesis approach to estimating the incidence of seasonal influenza in the Netherlands. *Influenza and Other Respiratory Viruses*, 8(1):33–41, nov 2013.

[142] D. J. McIver and J. S. Brownstein. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. *PLoS Comput Biol*, 10(4):e1003581, 2014.

[143] S. Merler and M. Ajelli. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proceedings of the Royal Society of London B: Biological Sciences*, 277(1681):557–565, 2010.

[144] S. Merler, M. Ajelli, L. Fumanelli, M. F. Gomes, A. P. y Piontti, L. Rossi, D. L. Chao, I. M. Longini, M. E. Halloran, and A. Vespignani. Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. *The Lancet Infectious Diseases*, 15(2):204–211, 2015.

[145] S. Merler, M. Ajelli, A. Pugliese, and N. M. Ferguson. Determinants of the spatiotemporal dynamics of the 2009 H1N1 pandemic in Europe: implications for real-time modelling. *PLoS computational biology*, 7(9):e1002205, 2011.

[146] J. Mlakar, M. Korva, N. Tul, M. Popović, M. Poljšak-Prijatelj, J. Mraz, M. Kolenc, K. Resman Rus, T. Vesnaver Vipotnik, V. Fabjan Vodušek, et al. Zika virus associated with microcephaly. *N Engl J Med*, 2016(374):951–958, 2016.

[147] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. The Twitter of babel: Mapping world languages through microblogging platforms. *PloS one*, 8(4):e61981, 2013.

[148] D. M. Morens, G. K. Folkers, and A. S. Fauci. Emerging infections: a perpetual challenge. *The Lancet infectious diseases*, 8(11):710–719, 2008.

[149] D. Musso, T. Nhan, E. Robin, C. Roche, D. Bierlaire, K. Zisou, A. Shan Yan, V. Cao-Lormeau, and J. Broult. Potential for Zika virus transmission through blood transfusion demonstrated during an outbreak in French Polynesia, November 2013 to February 2014. *Euro Surveill*, 19(14):20761, 2014.

[150] M. I. Nelson, L. Simonsen, C. Viboud, M. A. Miller, and E. C. Holmes. Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. *PLoS pathogens*, 3(9):e131, 2007.

[151] R. Nelson. HealthMap: the future of infectious diseases surveillance? *The Lancet Infectious Diseases*, 8(10):596, 2008.

[152] H. Nishiura, K. Mizumoto, W. E. Villamil-Gómez, and A. J. Rodríguez-Morales. Preliminary estimation of the basic reproduction number of Zika virus infection during Colombia epidemic, 2015–2016. *Travel medicine and infectious disease*, 14(3):274–276, 2016.

[153] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027, 2012.

[154] E. O. Nsoesie, P. Butler, N. Ramakrishnan, S. R. Mekaru, and J. S. Brownstein. Monitoring disease trends using hospital traffic data from high resolution satellite imagery: A feasibility study. *Scientific reports*, 5, 2015.

[155] W. H. Organization et al. WHO Director-General summarizes the outcome of the Emergency Committee regarding clusters of microcephaly and Guillain-Barré syndrome. *Saudi medical journal*, 37(3):334, 2016.

[156] V. Palchykov, M. Mitrović, H.-H. Jo, J. Saramäki, and R. K. Pan. Inferring human mobility using communication patterns. *Scientific reports*, 4, 2014.

[157] C. Panigutti, M. Tizzoni, P. Bajardi, Z. Smoreda, and V. Colizza. Assessing the use of mobile phone data to describe recurrent mobility patterns in spatial epidemic models. *Royal Society Open Science*, 4(5):160950, 2017.

[158] D. Paolotti, A. Carnahan, V. Colizza, K. Eames, J. Edmunds, G. Gomes, C. Koppeschaar, M. Rehn, R. Smallenburg, C. Turbelin, et al. Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience. *Clinical Microbiology and Infection*, 20(1):17–21, 2014.

[159] D. Paolotti, C. Gioannini, V. Colizza, and A. Vespignani. Internet-based monitoring system for influenza-like illness: H1N1 surveillance in Italy. In *Proceedings of the 3rd International ICST Conference on Electronic Healthcare for the 21st century. Casablanca*, pages 13–15, 2010.

[160] A. Parrella, C. B. Dalton, R. Pearce, J. C. Litt, and N. Stocks. ASPREN surveillance system for influenza-like illness: A comparison with FluTracking and the National Notifiable Diseases Surveillance System. *Australian family physician*, 38(11):932, 2009.

[161] A. Pastore-Piontti, Q. Zhang, M. F. Gomes, L. Rossi, C. Poletto, V. Colizza, D. L. Chao, I. M. Longini, M. E. Halloran, and A. Vespignani. Real-time assessment of the international spreading risk associated with the 2014 West African Ebola outbreak. In *Mathematical and Statistical Modeling for Emerging and Re-emerging Infectious Diseases*, pages 39–56. Springer, 2016.

[162] O. Patterson-Lomba, S. Van Noort, B. J. Cowling, J. Wallinga, M. G. M. Gomes, M. Lipsitch, and E. Goldstein. Utilizing syndromic surveillance data for estimating levels of influenza circulation. *American journal of epidemiology*, 179(11):1394–1401, 2014.

[163] M. J. Paul and M. Dredze. You are what you Tweet: Analyzing Twitter for public health. *ICWSM*, 20:265–272, 2011.

[164] M. J. Paul, M. Dredze, and D. Broniatowski. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*, 2014.

[165] M. Peppa, W. J. Edmunds, and S. Funk. Disease severity determines health-seeking behaviour amongst individuals with influenza-like illness in an internet-based cohort. *BMC infectious diseases*, 17(1):238, 2017.

[166] T. A. Perkins, A. S. Siraj, C. W. Ruktanonchai, M. U. Kraemer, and A. J. Tatem. Model-based projections of Zika virus infections in childbearing women in the Americas. *Nature microbiology*, 1:16126, 2016.

[167] D. Perrotta, A. Bella, C. Rizzo, and D. Paolotti. Participatory online surveillance as a supplementary tool to sentinel doctors for influenza-like illness surveillance in Italy. *PloS one*, 12(1):e0169801, 2017.

[168] D. Perrotta, M. Tizzoni, and D. Paolotti. Using participatory Web-based surveillance data to improve seasonal influenza forecasting in Italy. In *Proceedings of the 26th International Conference on World Wide Web*, pages 303–310. International World Wide Web Conferences Steering Committee, 2017.

[169] C. Poletto, C. Pelat, D. Levy-Bruhl, Y. Yazdanpanah, P. Boelle, and V. Colizza. Assessment of the Middle East respiratory syndrome coronavirus (MERS-CoV) epidemic in the Middle East and risk of international spread using a novel maximum likelihood analysis approach. *Eurosurveillance*, 19(23):3, 2014.

[170] C. Poletto, M. Tizzoni, and V. Colizza. Heterogeneous length of stay of hosts' movements and spatial epidemic spread. *Scientific reports*, 2, 2012.

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

[172] O. G. Pybus, A. J. Tatem, and P. Lemey. Virus evolution and transmission in an ever more connected world. In *Proc. R. Soc. B*, volume 282, page 20142878. The Royal Society, 2015.

[173] S. Riley. Large-scale spatial-transmission models of infectious disease. *Science*, 316(5829):1298–1301, 2007.

[174] M. Salathe, L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, et al. Digital epidemiology. *PLoS Comput Biol*, 8(7):e1002616, 2012.

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

[176] S. V. Scarpino, N. B. Dimitrov, and L. A. Meyers. Optimizing provider recruitment for influenza surveillance networks. *PLoS Comput Biol*, 8(4):e1002472, 2012.

[177] F. Schaffner and A. Mathis. Dengue and dengue vectors in the WHO European region: past, present, and scenarios for the future. *The Lancet Infectious Diseases*, 14(12):1271–1280, 2014.

[178] L. Schuler-Faccini. Possible association between Zika virus infection and microcephaly—Brazil, 2015. *MMWR. Morbidity and mortality weekly report*, 65, 2016.

[179] J. Shaman and A. Karspeck. Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, 109(50):20425–20430, 2012.

[180] J. Shaman, A. Karspeck, W. Yang, J. Tamerius, and M. Lipsitch. Real-time influenza forecasts during the 2012–2013 season. *Nature communications*, 4, 2013.

[181] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS one*, 6(5):e19467, 2011.

[182] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási. A universal model for mobility and migration patterns. *arXiv preprint arXiv:1111.0586*, 2011.

[183] L. Simonsen, J. R. Gog, D. Olson, and C. Viboud. Infectious disease surveillance in the big data era: Towards faster and locally relevant systems. *The Journal of Infectious Diseases*, 214(suppl_4):S380–S385, 2016.

[184] M. E. Sinka, M. J. Bangs, S. Manguin, T. Chareonviriyaphap, A. P. Patil, W. H. Temperley, P. W. Gething, I. R. Elyazar, C. W. Kabaria, R. E. Harbach, et al. The dominant Anopheles vectors of human malaria in the Asia-Pacific region: occurrence data, distribution maps and bionomic précis. *Parasites & vectors*, 4(1):89, 2011.

[185] M. E. Sinka, M. J. Bangs, S. Manguin, M. Coetzee, C. M. Mbogo, J. Hemingway, A. P. Patil, W. H. Temperley, P. W. Gething, C. W. Kabaria, et al. The dominant Anopheles vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. *Parasites & vectors*, 3(1):117, 2010.

[186] M. E. Sinka, Y. Rubio-Palis, S. Manguin, A. P. Patil, W. H. Temperley, P. W. Gething, T. Van Boeckel, C. W. Kabaria, R. E. Harbach, and S. I. Hay. The dominant Anopheles vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic précis. *Parasites & vectors*, 3(1):72, 2010.

[187] M. S. Smolinski, A. W. Crawley, K. Baltrusaitis, R. Chunara, J. M. Olsen, O. Wójcik, M. Santillana, A. Nguyen, and J. S. Brownstein. Flu Near You: crowdsourced symptom reporting spanning 2 influenza seasons. *American journal of public health*, 105(10):2124–2130, 2015.

[188] J. K. Taubenberger and J. C. Kash. Influenza virus evolution, host adaptation, and pandemic formation. *Cell host & microbe*, 7(6):440–451, 2010.

[189] N. L. Tilston, K. T. Eames, D. Paolotti, T. Ealden, and W. J. Edmunds. Internet-based surveillance of influenza-like-illness in the UK during the 2009 H1N1 influenza pandemic. *BMC Public Health*, 10(1):650, 2010.

[190] M. Tizzoni, P. Bajardi, A. Decuyper, G. K. K. King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. González, and V. Colizza. On the use of human mobility proxies for modeling epidemics. *PLoS computational biology*, 10(7):e1003716, 2014.

[191] M. Tizzoni, P. Bajardi, C. Poletto, J. J. Ramasco, D. Balcan, B. Gonçalves, N. Perra, V. Colizza, and A. Vespignani. Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm. *BMC medicine*, 10(1):165, 2012.

[192] S. P. van Noort, C. T. Codeço, C. E. Koppeschaar, M. Van Ranst, D. Paolotti, and M. G. M. Gomes. Ten-year performance of influenzanet: ILI time series, risks, vaccine effects, and care-seeking behaviour. *Epidemics*, 13:28–36, 2015.

[193] Y. Vandendijck, C. Faes, and N. Hens. Eight years of the Great Influenza Survey to monitor influenza-like illness in Flanders. *PLoS One*, 8(5):e64156, 2013.

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

[195] T. Vincenty. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. *Survey review*, 23(176):88–93, 1975.

[196] A. Wesolowski, C. O. Buckee, L. Bengtsson, E. Wetter, X. Lu, and A. J. Tatem. Commentary: containing the Ebola outbreak-the potential and challenge of mobile network data. *PLoS currents*, 6, 2014.

[197] A. Wesolowski, C. O. Buckee, K. Engø-Monsen, and C. Metcalf. Connecting mobility to infectious diseases: the promise and limits of mobile phone data. *The Journal of Infectious Diseases*, 214(suppl_4):S414–S420, 2016.

[198] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

[199] A. Wesolowski, C. Metcalf, N. Eagle, J. Kombich, B. T. Grenfell, O. N. Bjørnstad, J. Lessler, A. J. Tatem, and C. O. Buckee. Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *Proceedings of the National Academy of Sciences*, 112(35):11114–11119, 2015.

[200] A. Wesolowski, T. Qureshi, M. F. Boni, P. R. Sundsøy, M. A. Johansson, S. B. Rasheed, K. Engø-Monsen, and C. O. Buckee. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences*, 112(38):11887–11892, 2015.

[201] O. P. Wojcik, J. S. Brownstein, R. Chunara, and M. A. Johansson. Public health for the people: participatory infectious disease surveillance in the digital age. *Emerging themes in epidemiology*, 11(1):1, 2014.

[202] L. Yakob, A. Kucharski, S. Hue, and W. J. Edmunds. Low risk of a sexually-transmitted Zika virus outbreak. *The Lancet infectious diseases*, 16(10):1100–1102, 2016.

[203] Y. Yang, C. Herrera, N. Eagle, and M. C. González. Limits of predictability in commuting flows in the absence of data for calibration. *Scientific reports*, 4, 2014.

[204] Q. Zhang, C. Gioannini, D. Paolotti, N. Perra, D. Perrotta, M. Quaggiotto, M. Tizzoni, and A. Vespignani. Social data mining and seasonal influenza forecasts: The FluOutlook platform. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 237–240. Springer, 2015.

[205] Q. Zhang, N. Perra, D. Perrotta, M. Tizzoni, D. Paolotti, and A. Vespignani. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of the 26th International Conference on World Wide Web*, pages 311–319. International World Wide Web Conferences Steering Committee, 2017.

[206] Q. Zhang, K. Sun, M. Chinazzi, A. P. y Piontti, N. E. Dean, D. P. Rojas, S. Merler, D. Mistry, P. Poletti, L. Rossi, et al. Spread of Zika virus in the Americas. *Proceedings of the National Academy of Sciences*, 114(22):E4334–E4343, 2017.