Advanced Statistical Approaches for the Global Analysis of Influenza Virus Circulation

Francesco Bonacina

École doctorale de Sciences Mathématiques de Paris Centre

Laboratoire de Probabilités, Statistique et Modélisation - UMR 8001, Sorbonne Université

Equipe Surveillance et modélisation des maladies transmissibles, IPLESP UMR-S 1136, INSERM, Sorbonne Université

DOCTORAL THESIS in Applied Mathematics

author

Francesco Bonacina

Advanced Statistical Approaches for the Global Analysis of Influenza Virus Circulation

---

supervisors

Olivier Lopez, Chiara Poletto, Maud Thomas

Thesis defended on **July 1, 2024**, before a jury composed of:

Pierre-Yves Böelle	Sorbonne Université	President
Julien Chiquet	Université Paris-Saclay, AgroParisTech INRAE	Referee
Giorgio Guzzetta	Fondazione Bruno Kessler	Referee
Olivier Lopez	ENSAE - IP Paris	Supervisor
Chiara Poletto	Università di Padova	Supervisor
Ganna Rozhnova	University Medical Center Utrecht	Examiner
Maud Thomas	Sorbonne Université	Invited member

Approches statistiques avancées pour l'analyse globale de la circulation des virus de la grippe

Francesco Bonacina

Abstract

De multiples types et sous-types de virus de la grippe co-circulent dans le monde, avec une dynamique caractérisée par des épidémies annuelles et des changements exceptionnels dus à des événements épidémiologiques majeurs. Cette thèse développe des outils statistiques pour étudier certains aspects clés de cette dynamique ponctuée, proposant des approches non conventionnelles en épidémiologie. Les analyses sont basées sur les données de FluNet, un jeu de données fourni par l'Organisation mondiale de la santé qui comprend des comptages hebdomadaires d'échantillons de grippe provenant de plus de 150 pays, catégorisés par type et sous-type. Les deux premiers projets de recherche inclus dans la thèse sont axés sur l'application, tandis que la troisième étude est orientée vers la théorie, bien qu'elle comprenne une application aux données de surveillance de la grippe humaine. La première étude examine le déclin de la grippe pendant la pandémie COVID-19, en évaluant l'ampleur du déclin et en utilisant des techniques basées sur des arbres de régression pour identifier les facteurs associés à ce déclin au niveau des pays. La deuxième étude examine la dynamique couplée des (sous-)types de grippe, en se concentrant sur leur abondance relative dans chaque pays et chaque année, par le biais de l'analyse des données de composition. Elle démontre l'évolution du mélange des (sous-)types au cours de la pandémie COVID-19 et développe des algorithmes de prévision probabiliste pour prédire la composition des (sous-)types un an à l'avance. La troisième étude formule un modèle de copule conditionnelle pour décrire les dépendances de données multivariées nettes de certaines covariables. La consistance asymptotique du modèle est ensuite étudiée. Enfin, le modèle est utilisé pour classer les pays et les années caractérisés par des dépendances similaires dans les proportions relatives des (sous-)types de grippe.

Advanced Statistical Approaches for the Global Analysis of Influenza Virus Circulation

Francesco BONACINA

Abstract

The mitigation of human Influenza remains a challenge due to the complexities characterizing its spread. Multiple types and subtypes of influenza viruses co-circulate globally, with a dynamic characterized by annual epidemics and occasional shifts due to major epidemiological events. This thesis develops statistical tools to study some key aspects of influenza spatiotemporal ecological dynamics, proposing unconventional approaches in epidemiology. The analyses are based on data from FluNet, a comprehensive dataset provided by the World Health Organization that includes weekly counts of influenza samples from over 150 countries, categorized by type and subtype. The first two research projects included in the thesis have an applied focus, while the third study is theoretically oriented, although it includes an application to influenza surveillance data. The first study examines the decline of influenza during the COVID-19 pandemic, assessing the magnitude of the decline by country globally and using regression tree-based techniques to identify country-level factors associated with the decline. The second study examines the coupled dynamics of influenza (sub)types, focusing on their relative abundance across countries and years through the lens of Compositional Data Analysis. It provides evidence of the changes in (sub)type mixing during the COVID-19 pandemic and develops probabilistic forecasting algorithms to predict (sub)type composition one year in advance. The third study formulates a conditional copula model to describe the dependencies of multivariate data conditionally upon certain covariates. The asymptotic consistency of the model is then investigated. Finally, the model is used to classify countries and years characterized by similar dependencies in the relative abundances of influenza (sub)types.

Acknowledgements

Saint Antoine, SUMO team - Jussieu, LPSM
Between these two labs, this manuscript was written, between these two labs, I spent a considerable part of my Parisian years, surrounded by fantastic colleagues and engaged in stimulating research activities.

I would like to start by thanking my supervisors. Thank you Chiara, it is mainly thanks to you that I ended up in Paris for this adventure. Thank you for your passion, your patience, and your precision. Thank you for teaching me the intricacies of research. For pushing me to look for interesting results even where the analyses seemed to be inconclusive, for pushing me to explain the analyses I did down to the finest details, adjusting each paragraph word by word, for pushing me to explore all the palettes of matplotlib: I never imagined I could test so many different nuances for one scatterplot. Thank you Maud for welcoming me to LPSM. Thank you for your energy, helpfulness and pragmatism. Your advice has often helped me not to get lost in a glass of water. Thank you also for supporting me in my teaching activities by suggesting courses to take. This part of the work was particularly enriching for me. Thank you Olivier for your optimism, for being able to reassure me about the choices to be made when the simulations never seemed to work. I am also grateful to Pierre-Yves, who was always available to discuss the details of each statistical model with patience and enviable pedagogical skills.

Thank you to the jury members who agreed to evaluate my research. Thank you Giorgio and Julien for your time and your wise comments. Thank you Ganna for coming from the Netherlands.

Thanks to all my colleagues in Saint Antoine and Jussieu. To all the PhD students, post-docs, researchers, and professors who welcomed me when I arrived, and to all those with whom we shared at least part of this journey. I will try to write down your names here, but I am sure there are more to be named. Thank you Pourya, Benjamin, Marika, Chiare, Elisabetta, Laura, Giulia, Davide, Albano, Francesco, Anthony, Mattia, Giulia, Lucille, Wen, Ilona, Ingrid, Claudio, Boxuan, Younjung, Jonathan, Suprabhath, Federico, Eugenio, Raphaelle, Vittoria. Thank you Grâce, Alice, Ludovic, Adeline, Romain, Nina, Antoine, Paul, Barbara, Sara, Antonio, Camila, Miguel, Lucas, Alexis, Iqraa, Arianne, Gloria, Thibault, Joseph, Marine, Nicklas, Stéphane, Anna, Claire, Jean-Patrick, Arnaud, Maxime, Antoine. With some of you, I was lucky enough to share moments outside the lab. Here are just a few of the highlights of Parisian social life: the Cinéermitage, les aperitifs aux petites arènes, the parties at Camila's coloc, et le poulet DG servi avec un grand sourire.

Rue Montera, Lausanne, Bergamo
Viva Rue Montera, le meilleur coin de Paris ! Merci Chachou et Flo de m'avoir accueillie à Paris et de m'avoir fait sentir comme chez moi. Brocolis surgelés, yaourts maison, courses matinales et petits déjeuners partagés sont devenus au fil des années des routines précieuses dont je me souviendrai avec nostalgie. Si j'ai appris à parler français, ou du moins à me débrouiller, c'est en grande partie grâce à vous. Si j'ai ensuite perdu une partie de mon français, c'est surtout grâce à mes colocataires suivants. Qui, pourtant, ont été aussi fantastiques et pleins de qualité! Merci Chiara pour ta spontanéité et les pensées partagées du dimanche soir. Merci Julius pour ton esprit d'aventure et ton enthousiasme qui font que tout semble facile, même la construction d'un four à pizza. Merci Daniele pour avoir toujours maintenu le niveau des glucides assez élevé, comme il faut. Merci à Raph pour ta bonne humeur. Merci à Lollo et Matteo, les voisins parfaits, dont la porte est toujours ouverte.

During these years I was lucky to have a second coloc in Lausanne. Thank you Matte, Marc, Jad, Abe, with you I always felt at home.

Grazie alla mia famiglia, in italia e in giro per il mondo. Grazie mamma e papà per la vostra disponibilità ad assecondare i programmi più improbabili, fatti di arrivi notturni con amici al seguito e di ripartenze di corsa, inseguendo qualche treno. A Sara per quando vieni a prendermi in stazione con il tandem, a Luca, con cui ci siamo ritrovati sulla Barre Des Ecrins.

Grazie Natha, che hai fatto parte di tutte queste case, e di chissà quali altre ancora in futuro.

Le Vélo Volant
Je ne peux pas oublier le Vélo Volant et toutes les belles personnes que j'y ai rencontrées : Adrien, Rebecca, Caroline, Thérèse, Marie, Wenbo, Léo, Sinan, Julien, Julius, Benjamin, Mireia, Saul et d'autres encore. Merci pour votre générosité, votre bonne humeur et votre disponibilité à l'improvisation. Merci pour les après-midi à l'atelier, les goûters partagés, Brassens, les soirées au Club Des Poètes, les balades à vélo en Bretagne, sur la Loire, à Fontainebleau.

Le mur d'escalade, le chapiteau, la coloc, les aperos, les cineclubs, les voyages
Merci aux copain.e.s de l'escalade et du cirque. À la communauté italienne, accueillante, festive, un peu nostalgique. À toutes les personnes précieuses que j'ai rencontrées au fil des années à Paris, mais que je n'ai pas encore nommées : Tibo, Tilman, Maéva, Leo, Alice, Etta, Xhafer. Infine, grazie agli amici e amiche di sempre, da Bergamo e da Torino, a voi che siete passati di qui in questi anni e a voi che mi aspettate ogni volta che torno a casa.

La fin de mon doctorat correspond également à la fin de mon aventure parisienne, au moins pour l'instant. J'aurais aimé faire beaucoup plus de choses avec chacun d'entre vous, mais je pense quand même d'avoir vécu des années intenses, et je suis reconnaissante du temps que nous avons passé ensemble.

Contents

List of Figures

List of Tables

Introduction

This thesis examines the global circulation of human influenza viruses. It is motivated by the need to better understand the spatio-temporal patterns of influenza spread during periods of stable activity and how these patterns were disrupted during the COVID-19 pandemic.

Influenza viruses cause respiratory disease and infect millions of people worldwide each year. Their circulation is characterized by temporal patterns that show annual cycles, with seasonalities related to latitudes and climatic regions [1, 2, 3]. However, these regularities can be altered by exceptional epidemiological events. In the last decades, sporadic influenza pandemics occurred due to the emergence of a new variant that managed to spread rapidly even outside the usual periods of influenza activity [4]. Moreover, a significant perturbation of the influenza circulation was observed during the COVID-19 pandemic [5, 6, 7].

Anticipating the onset, peak time and severity of an epidemic season is essential. This remains a challenge due to the multiplicity of factors involved. In particular, the characteristics of influenza epidemics crucially depend on viral composition [8, 9]. Specifically, there exist three main influenza (sub)types (A\H1N1, A\H3N2, and B) that co-circulate. They exhibit different temporal trends [10] and have different impacts on different age groups [11, 12, 13, 14]. The composition of the (sub)types therefore strongly influences the burden of the epidemic season. A better understanding of these aspects would allow optimization of vaccine allocation and health system preparedness. Beyond seasonal influenza, epidemic anticipation is all the more important when the seasonal pattern is perturbed by exogenous events, as was the case when new variants emerged, or during the COVID-19 pandemic. In this case, uncertainties about the altered drivers of propagation (e.g. altered human behavior, altered susceptibility to infection) and possible future scenarios require maximum effort to understand the ongoing epidemiological situation and its possible developments.

A major challenge in epidemic anticipation is that epidemics in different countries are interdependent due to the continuous import and export of influenza viruses from one country to another [15, 16]. This underscores the importance of examining influenza circulation from a global perspective to understand patterns observed at the national level. However, cross-country analyses are complicated by the lack of standardized data. The available data is highly heterogeneous and requires the formulation of ad hoc statistical approaches to answer the epidemiological questions under analysis.

This thesis examines the global circulation of influenza by pursuing two lines of research. On the one hand, it analyzes the reduction in influenza transmission induced by the COVID-19 pandemic. On the other hand, it examines the distribution of influenza (sub)types in different countries, starting from the changes observed during the COVID-19 pandemic and going back to questions about the spatio-temporal patterns of (sub)type co-circulation in the pre-pandemic period. The analyses were mainly based on data from FluNet, a large dataset gathered by the World Health Organization, which collects weekly counts of influenza samples from more than 150 countries, classified by type, subtype, and lineage [17, 18].

The first two studies included in this thesis present analyses driven by concrete epidemiological questions. The first study (Chapter 3) investigates the decline of influenza during the COVID-19 pandemic: To what extent did influenza decline during the first year and a half of the COVID-19 pandemic? What country-level factors were associated with this decline? The second study (Chapter 4) analyzes the coupled dynamics of influenza (sub)types: What regularities emerge in the spatio-temporal patterns of influenza (sub)type co-circulation? How can we predict the relative abundance of (sub)types one year in advance? The novelty of the questions addressed and the complexity of the available data required the application of unconventional statistical methods in the epidemiological literature, such as tree-based methods (Chapter 3) and Compositional Data Analysis combined with multivariate probabilistic forecasting (Chapter 4). In addition, this data stimulated the development of a novel conditional copula model. A third study (chapter 5) is devoted to the definition of this model, the analysis of some of its theoretical properties, and its empirical validation.

Evidence that COVID-19 was altering the circulation of many infectious diseases already emerged in spring 2020. The winter 2019-2020 influenza season had come to an early and abrupt end in many countries [19, 20, 21, 22], coinciding with the start of the implementation of non-pharmaceutical interventions (NPIs) in late February/early March 2020. As COVID-19 spread globally, the world faced an unprecedented scenario characterized by uncertainty about the evolution of the epidemiological context for both COVID-19 and other communicable diseases. In this context, the scientific community began to issue warnings about the potentially harmful effects of a dual epidemic caused by the simultaneous circulation of COVID-19 and influenza [23]. However, the winter influenza epidemic in southern hemisphere countries, which typically occurs between June and September, did not occur [5, 6]. Meanwhile, the US surveillance data for the interseasonal period (May-August 2020) also showed an exceptionally low percentage of positive cases [5]. Later, the seasonal epidemic was also missed in the northern countries between 2020 and 2021, with numbers suggesting an almost complete stop of influenza circulation globally [7, 24].

This PhD started at the end of 2020. As these pieces of the puzzle were added, it became crucial to quantify the phenomenon, i.e. the reduction of influenza, and its global scale, both to shed light on the current situation and to get an idea of possible future scenarios. To address these questions, we developed an ecological analysis covering more than one hundred countries and considering surveillance data over six trimesters from spring 2020 to summer 2021. The results of this study are presented in Chapter 3. The reduction in influenza for each country and trimester is quantified by comparing the rates of positive samples collected during the COVID-19 period with those reported in the five years before the pandemic. The factors associated with influenza decline are then identified, taking into account the ongoing epidemic situation, implemented NPIs, and socio-demographic, geographic, and meteorological factors. These analyses are based on regression trees and random forests to deal with the non-linearity of the problem and the large number of variables involved, some of which were correlated.

During the COVID-19 pandemic, the circulation of influenza viruses was anomalous not only in terms of the number of cases but also in terms of the influenza variants circulating and their geographical distribution. First, the B\Yamagata lineage seems to have become extinct after the pandemic [25, 26, 27]. Second, the (reduced) circulation of the other (sub)types in 2020-21 was characterized by strong spatial segregation [25]. These two observations raise questions about the degree of co-circulation of (sub)types during COVID-19 and how this differs from periods of normal influenza activity.

Chapter 4 proposes a framework for analyzing (sub)type circulation dynamics quantitatively. In addition, some fundamental questions about the coupled dynamics of (sub)types are investigated. We identify regions of the world characterized by a similar alternation of (sub)types and propose statistical methods to predict the relative abundances of (sub)types for a given country one year in advance. The analyses are based on the proportions of cases per (sub)type reported for each country. To treat this data we make use of appropriate transformations derived from Compositional Data Analysis [28].

The relative abundances of influenza (sub)types correspond to multivariate data whose dependence structure can be modeled by *copulas*. Copulas are multivariate cumulative distributions that facilitate the task of describing such complex data through a two-step analysis. First, the univariate marginal distributions are estimated independently. Then, their dependence is modeled by a multivariate function, which is the copula itself [29].

In addition, the dependence between the response variables may sometimes be mediated by external factors. This scenario is even more complex and requires the formulation of conditional copulas. In Chapter 5 we develop a conditional copula model that uses regression trees to incorporate covariates. The proposed method is very flexible as it can include both quantitative and qualitative variables - an advantage for epidemiological research. The asymptotic consistency of the model and the optimality of the model selection procedure are investigated. The model is then tested on both synthetic and real data. In particular, it is used to investigate the variation in the dependence of (sub)type abundances across countries and years.

The manuscript is organized as follows. Chapter 1 provides some notions of global influenza circulation and surveillance and specifies the open questions addressed in the three studies of this thesis. Chapter 2 describes the statistical methods used in the analyses. The three research projects are then presented in chapters 3, 4, and 5. Chapter 3 is based on the article *Global patterns and drivers of influenza decline during the COVID-19 pandemic*, published in the *International Journal of Infectious Diseases* [30]. Chapter 4 examines the coupled dynamics of the influenza (sub)types. It is based on a paper that is under review by the co-authors at the time of writing this thesis and will be submitted in the coming weeks [31]. Chapter 5 proposes a conditional copula model and is based on the paper *Tree-based conditional copula estimation* [32], which has been recently submitted. The three chapters are presented in the form of an academic article, complete with abstract, main manuscript, and supplementary materials. Finally, Chapter 6 concludes the manuscript with some general discussions.

0.1 List of the research projects: publications, codes, and conferences

The research projects related to the thesis are listed below. Additional information is provided about their publication status, the seminars or conferences where the projects were presented, and the links to the code for reproducing the analysis.

1. Francesco Bonacina, Pierre-Yves Boëlle, Vittoria Colizza, Olivier Lopez, Maud Thomas, Chiara Poletto, **Global patterns and drivers of influenza decline during the COVID-19 pandemic**, International Journal of Infectious Diseases. (2023). [Published paper]. [github].

   The project was presented at the following seminars and conferences:

   - Contributed talk at the scientific days of the *Action Cordonnée Modélisation des Maladies Infectieuses* - Bordeaux, France, Nov 2022;
   - Contributed talk at the *France National Conference on Complex Systems (FRCCS2022)* - Paris, France, Jun 2022;
   - Poster at *EPIDEMICS8, the 8th International Conference on Infectious Diseases Dynamics* - Online, Nov 2021
   - Contributed talk at the international *Conference on Complex Systems (CCS2021)* - Lyon, Paris, Oct 2021;
   - Contributed talk at joint meeting of the research group *Statistique et Santé*, between the *Société française de biométrie* and the Société française de statistiques - online, Oct 2021;
   - Invited talk at the *PhD students seminar series* at LPSM - Paris, France, Jun 2021.

2. Francesco Bonacina, Pierre-Yves Boëlle, Vittoria Colizza, Chiara Poletto **Understanding the coupled dynamics of influenza (sub)types: a global analysis leveraging Compositional Data Analysis**. [Manuscript is being finalized, the paper will be submitted soon to a journal of quantitative epidemiology]. [github].

   The project was presented at the following seminars and conferences:

   - Invited talk at the *Modélisation Aléatoire du Vivant research group seminar series* at LPSM - Paris, France, Jan 2024;
   - Poster at *EPIDEMICS9, the 9th International Conference on Infectious Diseases Dynamics* - Bologna, Italy, Nov-Dec 2023;
   - Invited talk at the *SMILE seminar series* at Collège De France - Paris, France, Nov 2023;
   - Poster at the scientific days of the *Action Cordonnée Modélisation des Maladies Infectieuses* - Paris, France, Oct 2023;
   - Invited talk at the *SUMO equipe seminar series* at IPLESP - Paris, France, Jun 2023;
   - Contributed talk at the international *Conference on Complex Systems (CCS2022)* - Palma De Mallorca, Spain, Oct 2022;

3. Francesco Bonacina, Olivier Lopez, Maud Thomas, **Tree-based conditional copula estimation** (2024). [Paper submitted, pre-print available on arXiv]. [github].

Chapter 1

Global circulation and surveillance of human influenza viruses

This chapter presents some epidemiological aspects of human influenza that will serve as the basis for the research work in the continuation of this thesis. First, the main characteristics of influenza viruses are reviewed. Secondly, a summary is provided regarding the current knowledge on the global spatio-temporal patterns of influenza circulation. Thirdly, the risks posed by influenza that motivate this work are highlighted. Fourth, we present the FluNet global surveillance dataset, which is the main source of data for the studies presented in the thesis. Finally, we outline some of the open problems and specify the questions that will be addressed in chapters 3, 4, and 5.

1.1 Characteristics of influenza viruses

Influenza disease. Influenza is a respiratory disease caused by some viruses of the Orthomyxoviridae family - commonly called influenza viruses - which have an endemic circulation globally. The virus can be transmitted via contacts, droplets, or aerosols [33]. It attacks the respiratory tract and causes an illness with a typical course of 1-2 weeks in most cases. However, in some cases, the infection may be accompanied by complications that lead to death, especially in weak individuals, infants, the elderly, or those with chronic diseases.

Structure of influenza viruses. Two types of influenza, namely influenza A and influenza B, are responsible for human epidemics. These viruses share a common structure, comprising a glycoprotein capsule containing genetic material in the form of eight RNA segments. Both influenza A and B carry essential surface proteins, hemagglutinin (H) and neuraminidase (N), pivotal for facilitating viral entry into host cells and subsequent exit after replication, enabling the virus to spread throughout the body. In contrast, influenza A and B differ in other surface proteins, notably the AM2 and BM2 proteins, which are functional for virus replication.

Host of influenza viruses. Influenza B is generally considered a human virus, while influenza A has several animal reservoirs, including wild animals (waterfowl, shorebirds, whales, seals, bears, squirrels, foxes, bats) and farm animals (ducks, chickens, pigs, horses, minks) [34, 35]. This promotes the spread of influenza A viruses around the world, both through long-distance migrations of wild birds [36, 37] and trade of livestock [35, 38]. Not all influenza A variants are transmissible to humans. However, influenza viruses are constantly changing, and there is a constant risk that zoonotic variants will evolve and make the host jump, acquiring the ability to infect humans [39]. In addition, the wide diversity of circulating variants and host species increases

the evolutionary potential of the virus.

Evolution and immune waning. Two main mechanisms contribute to influenza virus evolution: the gradual accumulation of mutations (also called antigenic drift) and the events of genome reassortment, which consist in the exchange of entire segments of genetic material between different strains co-infecting the same cell [40]. The evolution of influenza viruses is particularly fast. As typically occurs for RNA systems, they have a high mutation rate due to limited mutation error correction capabilities. Consequently, they have a great ability to adapt to new hosts [41] and to evolve by escaping the host's immunity. For this reason, the immunity acquired from an infection lasts only for a limited time, which varies with the type of influenza, but does not exceed a few years [42, 43]. The exception is infections that occur at a young age, for which there is a long-lasting immune response, a phenomenon called immune imprinting [13, 44].

Vaccines. Influenza vaccines constitute the most effective way to prevent severe infections. To maximize the protection they contain multiple viruses of both influenza A and B, namely three strains for trivalent vaccines and four strains for the quadrivalent ones. However, due to the fast mutation of influenza viruses, vaccines need to be updated regularly. W.H.O. provides recommendations for vaccine composition twice a year, in February and September, six months before the start of the northern and southern hemisphere winter epidemics, respectively. The selection of the precise strains is made to optimize the vaccine's match with circulating variants, particularly focusing on the one expected to be most prevalent in the upcoming season [45]. However, anticipating the circulating variants months in advance is a real challenge. Consequently, vaccine efficacy largely differs for different influenza strains and from year to year [46].

Nomenclature of influenza viruses. Influenza A subtypes are named according to the type of H and N proteins they contain, with 18 and 9 variants respectively. Most of these subtypes infect mainly non-human species, and nowadays, only A\H1N1 and A\H3N2 strains circulate stably among humans. For influenza B, we refer to lineages rather than subtypes. Since the 1980s, two have been circulating, B\Yamagata and B\Victoria [47], although no cases of B\Yamagata have been reported since April 2020 [26, 27]. Both A subtypes and B lineages are further distinguished into clades and subclades, defined by looking at similarities in genes that encode the H and N proteins. A summary of the influenza virus classification is depicted in figure 1.1, reported from [48]. For further details about virus classification look at [40].

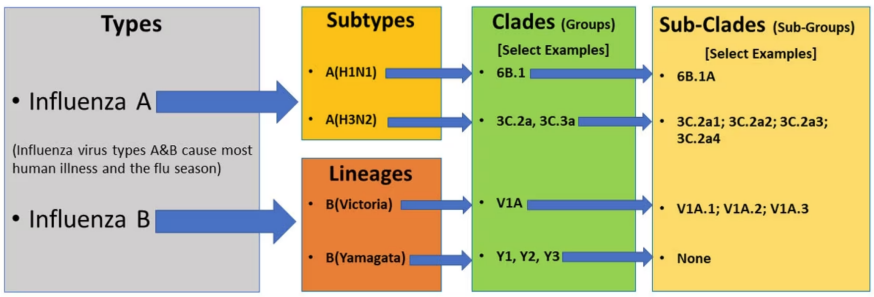


FIGURE 1.1: Classification of seasonal influenza viruses. Influenza virus types A and B are responsible for seasonal influenza epidemics in human populations. Influenza type A splits into subtypes A\H3N2 and A\H1N1, while influenza B separates into lineages B\Yamagata and B\Victoria. All these strains further divide into clades and subclades for finer classification. Source [48]

Influenza surveillance data often provide details on A subtypes but lack information on B lineages. In addition, trivalent vaccines contain one clade for each A subtype (A\H1N1 and A\H3N2) and one clade for type B viruses. Only in the last decade have quadrivalent vaccines been developed that contain two clades for the B virus (one for each lineage) [49]. As a result, in practical terms, influenza viruses are often categorized into three strains—A\H3N2, A\H1N1, and B—often (improperly) labeled as *subtypes* or *(sub)types*. In the remainder of the manuscript, I will also adopt this convention and employ the term *(sub)types* to refer to the A\H3N2, A\H1N1 and B viruses.

Epidemiology of influenza (sub)types. Influenza viruses spread following pronounced seasonal cycles in temperate regions, whereas in tropical and sub-tropical regions they manifest more variable patterns, including bi-annual peaks or year-round epidemics [1, 3]. However, the different (sub)types are characterized by different circulation dynamics on a global scale. Studies found that A\H3N2 does not persist locally between outbreaks but circulates continuously in Southeast Asia and India due to successive and partially overlapping epidemics in the region. From there, the virus is then annually reimported to temperate regions each year. On the other hand, there is evidence for the persistence of A\H1N1 and B between one epidemic and the next [15].

Moreover, Influenza A viruses evolve faster than B viruses, and subtype A\H3N2 faster than A\H1N1 [15, 40, 50]. This means that the rate of substitution of A\H3N2 clades is larger compared to A\H1N1 and B clades, and therefore, overall, a larger portion of the population is susceptible to A\H3N2. Consequently, A\H3N2 usually results in larger and more frequent epidemics [51, 52]. Additionally, A\H3N2 usually infects all ages and is particularly severe for the elderly, while A\H1N1 and B tend to infect younger people. This can be explained by immune imprinting. Older generations weren't exposed to A\H3N2 viruses in their youth as this subtype didn't emerge until the late 1960s [11, 12, 13, 53].

Finally, the (sub)types present shifted epidemic peaks in temperate regions. In particular, B and A\H1N1 may have epidemic activity that extends beyond the winter period, into March in the Northern Hemisphere and into September in the Southern Hemisphere. In contrast, there is considerable overlap in the spread of (sub)types in tropical regions. [10, 54].

Virus interactions. When a new clade of a subtype emerges, it is competitive only if it is antigenically different from those already in circulation, especially concerning H and N proteins, which are the main target of host system antibodies. Antigenic advantage is measured by antigenic maps, that quantify cross-reactivity between tested variants and reference antisera using data from hemagglutinin inhibition assay [45, 46, 55]. In other words, virologists exploit this tool to measure how much the antibodies developed in response to infection of one clade, are also protective against alternative clades. Antigenic maps are used to define the composition of influenza vaccines.

If we backtrack in the classification and shift the focus from clades to subtypes and lineages, understanding the interactions between them becomes more challenging. Evidence for cross-reactive responses to B lineages have been provided by Ferret experiments [56] and in epidemiological studies [13, 57]. In another animal experiment considering (sub)types A\H3N2, A\H1N1, and B, Laurie and colleagues found that not-symmetrical cross-immunity occurred, with A\H1N1 providing temporary immunity for a longer period than B, in turn most effective than A\H3N2 [58]. Similar results were found by Yang and colleagues [59] by estimating the cross-protection matrix for

the A\H3N2, A\H1N1 and B viruses. They adjusted multi-strain compartmental models on the incidence curves of the three strains, considering data from Hong Kong over the period 1998-2018. Furthermore, some epidemiological studies looked at relative abundances of (sub)types [54, 60] and others measured the correlations between the reproductive numbers of the strains estimated over time [61]. These studies corroborate the hypothesis for a competitive interaction between A\H3N2 and A\H1N1 subtypes while disclosing less conclusive results regarding interactions between influenza A and B viruses.

1.2 The spatio-temporal dynamics of influenza viruses

1.2.1 A look at the last century - the punctuated dynamics of influenza viruses

For decades influenza viruses have been showing a circulation characterized by periods of stable circulation, punctuated by events that caused notable shifts [4]. During periods of stability, the same influenza types and subtypes co-circulated, with no major changes, determining epidemics in temperate regions that followed the usual seasonal patterns. When a new variant emerged and achieved a higher transmissibility, it typically caused more severe epidemics, even outside of the period of normal circulation. For example, in Europe, an epidemic wave occurred between the spring and summer of 2009, with a peak in July, caused by the pandemic virus A\H1N109 (data and figures accessible on the FluNet website [62]). The pandemic strain eventually replaced the historical A\H1N1 strain, thanks to its competitive advantage. Similarly, the H2N2 influenza A strain circulating in the 1960s became extinct following the Hong Kong flu pandemic of 1968 when a new influenza A\H3N2 strain appeared [63]. Changement of the circulating variants not always end up in major epidemics. In the 1970s, the ancestral influenza B lineage separated in the B\Victoria and the B\Yamagata lineages [47] and the process was not accompanied by significant epidemiological events. Figure 1.2 reported from [25] summarizes the evolution of influenza strains from 1920 to 2020. The COVID-19 pandemic also determined a major disruption of the circulation of the influenza viruses [5, 7]. In this case, the unprecedented interventions applied worldwide blocked the infectious routes of many pathogens and hindered most communicable diseases [6, 64].

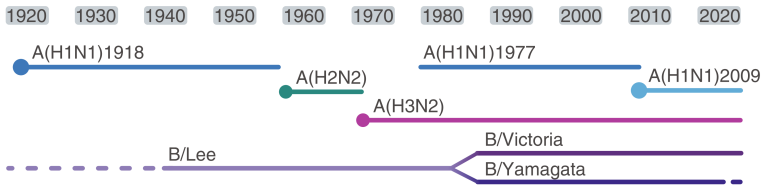


FIGURE 1.2: Timeline of influenza virus circulation from 1920 to 2020. The emergence of pandemic viruses replacing circulating variants is represented by dots. Source [25].

1.2.2 Seasonality and annual cycles

During the period of stable influenza activity, virus circulation has annual regularities that vary from region to region. In temperate regions, influenza epidemics occur during the winter season. In addition, some studies have found a longitudinal gradient in the timing of influenza peaks in these regions: In the Northern Hemisphere, epidemics

begin in Asia and affect Europe and North America later [1, 65]; in the Southern Hemisphere a west-to-east gradient is observed [1]. On the other hand, tropical and subtropical regions are characterized by irregular patterns of influenza circulation, with biannual peaks or even year-round circulation in some countries. These heterogeneities make it difficult to determine the optimal window for vaccine distribution. Indeed, countries typically distribute vaccines in the weeks before winter in their hemisphere, but Alonso and colleagues have shown that some tropical countries may benefit from following the opposite hemisphere's schedule [66].

Climatic factors are important drivers of influenza seasonality [3, 67]. There is evidence for the association of the timing of the epidemics with low temperatures and low relative humidity in temperate regions. These conditions have also been shown to favor droplet transmission in experimental studies in animal models [68]. In tropical regions, however, epidemic peaks often occur during the rainy season, suggesting that transmission is likely to be driven by direct contact rather than droplets in that context [40]. In addition, other important factors contribute to the seasonality of influenza outbreaks. Studies have pointed to the role of seasonal changes in human behavior [69]. In winter and during rainy periods, most activities take place indoors, favoring the transmission of respiratory viruses. In this context, the school calendar is often discussed as a relevant factor, as there is evidence that school-aged children are important drivers of seasonal epidemics [70, 71]. For example, Ewing and colleagues [72] and De Luca and colleagues [73] found that school closures during winter holidays can reduce influenza transmission and delay the peak of epidemics. Similarly, Ajelli and co-authors [74], found that the school closure over the weekends in Italy significantly contributed to reducing the effective reproduction number during the 2009 flu pandemic.

1.2.3 Geographical global patterns

The movement of infected individuals is a major driver of the global spread of influenza viruses [16, 75]. The interaction between the seasonality of virus circulation and human mobility determines the spatial correlations of virus dynamics. In particular, as mentioned above, phylogenetic analyses revealed that A\H3N2 viruses do not persist from one season to the next and are reseeded each year from Southeast Asian countries [15].

Other studies have used the FluNet global surveillance dataset [17, 18] to examine the geographic distribution of (sub)types in a large number of countries. Analyzing data from 19 temperate countries during 1997-2005, Finkelman and colleagues found that A\H3N2 was the predominant (sub)type, dominating or co-dominating in most seasons during this period [54]. They also found that when A\H3N2 was dominant in the Northern Hemisphere, it was also dominant in the Southern Hemisphere, whereas this was not the case for the other (sub)types. More recently, the predominance of A\H3N2 viruses was confirmed by Zanobini and co-authors, who analyzed data from 149 countries from 2010 to 2020 [52]. They found that the predominance was more pronounced in temperate regions and less pronounced in the intertropical belt. He and colleagues studied the period after the 2009 pandemic and found that A\H1N1 showed different dynamics in different macro-regions of the world: in particular, the differences were particularly pronounced for three regions corresponding to North America, Central America, and Europe and Asia [76].

1.3 Burden and risks of human influenza

The impact of an influenza epidemic in a specific country is intricately connected to the circulation of viruses in interconnected regions across the globe. Consequently, to effectively mitigate the impact, it is imperative to monitor virus circulation on a global scale. Furthermore, although until recent years influenza surveillance and academic investigations have neglected the least developed countries, recent studies show that influenza epidemics have a significant impact in all regions of the world [8]. Lastly, a substantial threat arises from the emergence of new influenza variants with pandemic potential, constituting a global menace by definition.

1.3.1 Burden of the seasonal influenza

In a 2018 report, the OECD estimated that respiratory diseases represent the third leading cause of death in Europe (8 percent of the total), and among these, influenza follows in importance chronic obstructive pulmonary disease, pneumonia, and asthma [77]. Recent studies have estimated that deaths directly related to influenza hover around [34000, 58000] per year in Europe [8] and around [4000, 52000] in the US [48]. Influenza has a much higher mortality rate among older people, to the point that among flu deaths, those over 65 are nearly 90% in Europe [77] and US [48] and about 67% on a global scale [8]. Shifting the focus beyond Europe and the US to the rest of the world, we find that the overall mortality rate (considering all causes) is much higher and, as a result, influenza accounts for a significantly smaller portion of mortality. Yet, it would be inaccurate to label influenza as a disease primarily affecting developed countries with aging populations. Examining macro-regions such as Sub-Saharan Africa and Southeast Asia, we find that the mortality rate from influenza per 100K inhabitants is 5.6 and 5.8, respectively — exceeding the estimated 5.3 for Europe [8].

The burden of seasonal influenza is not limited to the number of deaths. A major concern is the pressure on hospitals, particularly in temperate regions where the peak of the epidemic is concentrated in a few weeks. For similar reasons, influenza epidemics in these countries also lead to a spike in workplace absenteeism, which has a significant economic cost [78, 79].

1.3.2 Risk of an influenza pandemic

Since the beginning of the 20th century, five influenza pandemics occurred - Spanish flu in 1918, Asian flu in 1957, Hong Kong flu in 1968, Russian flu in 1977, and Swine flu in 2009 [80]. The Spanish Flu was caused by an A\H1N1 influenza strain which spread worldwide in three waves from 1918 to 1920. It has been estimated to be the most devastating pandemic, causing about 50 million deaths [81]. In recent years, pandemic events have occurred with increasing frequency, and have shown how quickly respiratory viruses can spread in today's hyperconnected world. In 2009, a new H1N1 strain of influenza A caused a pandemic, infecting between 11% and 21% of the total population in just over a year [82], but luckily with no higher risk of severe illness than the seasonal influenza [83]. At the beginning of 2020, SARS-COV-19 disseminated worldwide, at a higher pace than respiratory viruses of previous pandemics [84], despite the unprecedented containment measures put in place by governments, which included social distancing measures, stay-at-home orders and both local and international travel restrictions [85, 86]. Nowadays, the World Health Organization continues to point to influenza viruses as the most likely, or among the most likely, pathogens causing the next pandemic [87]. The danger of influenza lies in the fact that it is a

zoonotic disease for which the probability of a host jump is particularly high. In fact, reassortment events, which often cause spillover, are relatively likely given the large number of host species and variants in circulation. For example, the 2009 pandemic was caused by a virus that jumped from pigs to humans in Mexico after being generated by reassortment of avian, pig, and human viruses [88]. Furthermore, the pandemic risk is exacerbated by the fact that the intensity and frequency of spillover events have increased in recent decades. This is a result of climate change, which pushes animal species to seek new habitats, and the intensification of human activities especially in proximity to biodiversity-rich forest areas, which make contact between humans and potential pathogens increasingly frequent [89, 90, 91, 92]. Recently, an avian H1N5 virus appears to have acquired mutations that allow it to spread among mammals, causing great concern [93, 94].
All this makes clear the need for surveillance for both human pathogens, which are constantly mutating, and animal pathogens that could make the host jump. The Preparedness and Resilience for Emerging Threats (PRET) initiative of W.H.O. elaborates guidelines, tools, and other resources to support countries in their preparedness activities for influenza pandemics and for pandemics caused by other respiratory viruses. Recently, a checklist for respiratory pathogen pandemic preparedness planning [95] and planning for respiratory pathogen pandemics [96] have been released.

1.4 Influenza surveillance

Continuous surveillance of circulating variants is one of the key tools both to prepare for the management of seasonal epidemics and to identify a possible pandemic in its early stages. Many states have a respiratory virus surveillance system, which collects two kinds of data. On the one side, cases of patients with symptoms of influenza-like illness are registered by physicians to monitor the seasonal trend of the epidemic (*clinical surveillance*). Information on age, gender, and co-morbidities are often included for each clinical case. In France, for example, a national network of hundreds of volunteer sentinel physicians participate to clinical surveillance [97]. On the other side, biological data are collected by processing nasopharyngeal or salivary swabs from patients with influenza symptoms (*virological surveillance*). This data provides a picture of circulating viruses, including possible new emerging variants.

Since the 1990s, the W.H.O. has initiated an international coordination and data collection effort to develop a worldwide surveillance system. An important outcome of this project is the global dataset FluNet.

1.4.1 Global surveillance of influenza - the FluNet database

W.H.O.'s Global Influenza Surveillance and Response System (GIRSR) is a global surveillance system that includes National Influenza Centers and other affiliated laboratories [17]. It was established in 1952 to share viruses and information regarding influenza outbreaks. In 1997 it launched FluNet, a public database that reports the weekly number of positive and negative influenza specimens detected in each country. The positive cases are classified by type, subtype/lineage. The data collected are used to produce weekly update bulletins on circulating variants and semiannual reports summarizing the characteristics of seasonal epidemics in the two hemispheres.

The FluNet database is the most comprehensive source for influenza surveillance globally and is a valuable resource for public health agencies and researchers. Nowadays, it collects data from more than 150 countries, with hundreds of thousands of

classified samples each year, and the weekly time detail enables the monitoring of epidemic peaks in different regions of the world. The FluNet data have made it possible to analyze the spatio-temporal dynamics of influenza viruses globally over the past three decades. One of the earliest such studies was by Finkelman and colleagues [54], who examined the seasonality of influenza (sub)type circulation in 19 countries between 1997 and 2005. In recent years, increasing amounts of data have allowed for even larger studies. He et al. in 2015 analyzed the dynamics after the 2009 pandemic in 138 countries [76]. Caini and colleagues [65] considered 47 countries in the WHO European Region between 2010 and 2015, and developed a spatial clustering of countries characterized by similar patterns of influenza activity. Mook and colleagues [98] considered roughly the same countries between 2010 and 2017 and analyzed the onset and duration of the epidemic and the timing of the peak. Zanobini and coauthors [52] examined the (sub)type distribution and temporal characteristics of epidemics in 149 countries between 2010 and 2020. Finally, Zheng and coauthors [51] considered data from 2011 to 2023 and compared influenza activity before and after the COVID-19 pandemic. They observed a global decrease in influenza activity in the early stages of the COVID-19 pandemic, a resumption of circulation in 2022 and 2023, and a change in the duration of influenza epidemics in several countries in the Southern Hemisphere.

It is important to stress that some caution should be exercised when using FluNet data in multi-country analyses. In fact, the absolute number of reported cases is hardly comparable from country to country, and caution should also apply when comparing numbers of infections from year to year for the same country. This arises from the fact that (i) national surveillance procedures vary from one country to another, and (ii) consistency in monitoring efforts over time is not guaranteed even within the same country. Consequently, multi-country analyses should be based on the definition of specific variables and the use of appropriate statistical tools. For instance, it is advisable to prioritize the positivity rates over the absolute number of positive cases, or the proportion of cases per (sub)type rather than the raw counts. These considerations will constitute an essential part of the data treatment in chapters 3 and 4.

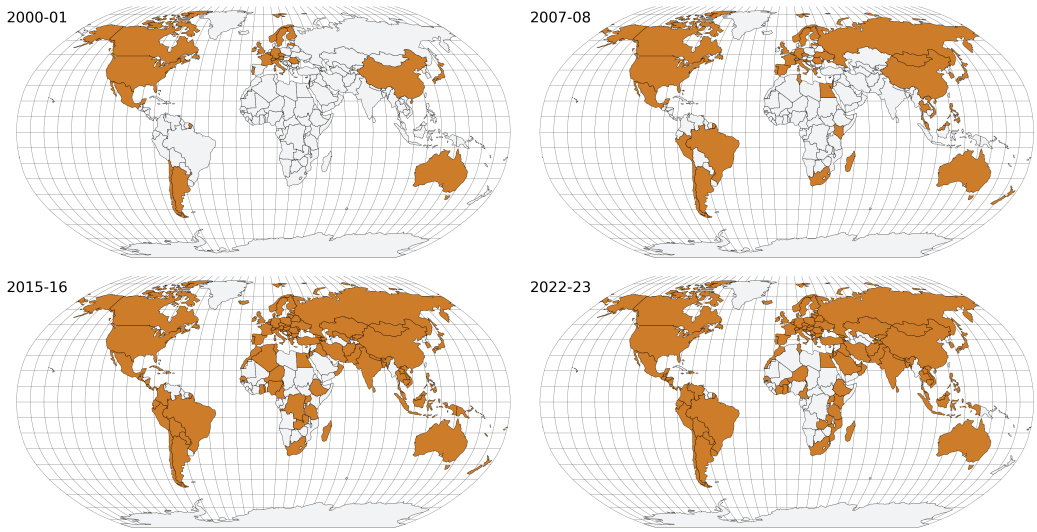


FIGURE 1.3: Spatial coverage of FluNet data over time. Each map refers to data collected in a one-year time window, ranging from April to April. Countries reporting data in the year are shown in orange. The reporting activity of a country is assessed based on two conditions: the country had (i) to have reported data for a minimum of 12 weeks and (ii) to have declared a minimum of 100 infections over the year.

Finally, we point out some limitations of the FluNet dataset that will require a careful step of data cleaning before the analysis of chapters 3 and 4:

- *Limited geographical coverage*. At the beginning of the project, only a few countries contributed to FluNet by reporting data. These countries were largely concentrated in the Global North, and surveillance of underdeveloped countries in tropical and sub-tropical regions was particularly poor (figure 1.3). The number of countries involved in monitoring has since grown over time, with a sudden increase following the 2009 pandemic. Today more than 150 countries report data regularly (figure [1.4). However, some regions of the world are still undermonitored, particularly on the African continent, despite the known importance of surveillance of tropical regions that act as reservoirs of seasonal viruses [99], [100] and, additionally, are considered hot-spots for the emergence of new variants [45].

- *Incomplete strain classification*. Details regarding influenza types (A/B) are often missing for a high proportion of the processed specimens, and the quality declines even further when investigating the classification into subtypes/lineages. However, there has been a notable improvement in this information over time, as evidenced in chapter 4.

- *Data consistency issues*. In addition to missing information, there are cases where reported data lack consistency. For example, the number of influenza A and B infections may exceed the total number of influenza cases.

Despite these limitations, both the number of countries engaged in global surveillance and the quality of the data have largely improved over time, thereby facilitating the exploration of virus circulation across diverse latitudes and over the years.

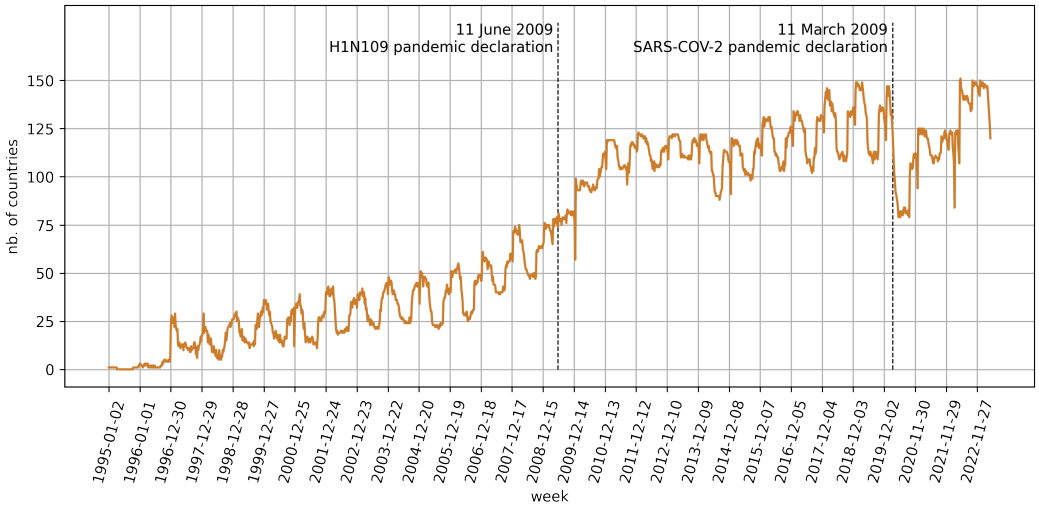


FIGURE 1.4: Number of countries contributing to FluNet over time. The weekly number of countries with no missing information for at least one influenza type (A or B) is reported in orange from Jan 1995 to Apr 2023. Vertical dotted lines indicate the declarations of the 2009 influenza pandemic (June 11, 2009), and the SARS-CoV-2 pandemic (March 11, 2020), by the W.H.O.

1.5 Open questions

We conclude this chapter by outlining some important questions about the global circulation of influenza that remain unanswered today. With the research papers presented in the following chapters, we provide elements to answer some of these questions

Despite the marked seasonality of influenza epidemics in temperate regions, knowing how to predict the onset, peak, and impact of epidemics in these countries remains a challenge. This information is useful for optimizing public health management of the epidemic, including the timing of vaccine distribution and preparation of hospital wards for the peak of the epidemic. In the United States, and more recently in Europe, there are nowcasting projects that attempt to predict the course of the current epidemic [101, 102, 103, 104, 105]. The researchers involved in the project propose a series of statistical and dynamic models to make the predictions, similar to what has been done during recent pandemics [106, 107, 108]. However, forecasting remains a challenge, especially beyond a few weeks. In particular, predicting the onset of an epidemic must take into account influenza activity in associated countries. This is particularly complex because the seeding of an epidemic from a foreign country is an inherently stochastic phenomenon. Metapopulation models that integrate a large amount of data (such as demography and mobility data) are currently used to predict the evolution of a pandemic and evaluate possible scenarios ([109, 110, 111, 112]). These models make it possible to assess the risk of a virus being imported from one country to another ([113, 114]). However, the dynamic of a pandemic virus spreading from an epicenter is very different from the dynamic of seasonal influenza. For seasonal influenza, some authors have tested different statistical approaches to predict peak timing, peak intensity, and type-specific influenza activity in the United States six months in advance using data on influenza activity in the Southern Hemisphere ([115]). In general, however, there are no well-established methods that have been proven to work in this context.

A second aspect regards the prediction of the abundance of the different (sub)types. A few studies proposed national or sub-national analyses to predict the epidemic size of each influenza (sub)type for the ongoing epidemic season. Goldstein and co-authors [60] adopted a statistical approach to estimate the cumulative incidences of the three (sub)types for the entire season in the US, using data from the first weeks, while Kandula and colleagues [116] dealt with a similar problem by using compartmental models. In both studies, the (sub)types were considered as independent viruses and the prediction horizon was limited to the ongoing epidemic. Yang and colleagues adjusted a multi-strain compartmental model on the incidence curves of the A\H1N1, A\H3N2, and B strains from 1998 to 2018 in Hong Kong. This way, they were able to estimate several epidemiological parameters (i.e., the strength of cross-immunity, the duration of immunity, the reproduction number, the infectious period, and more). Other studies have attempted to establish relationships between (sub)type abundances in multi-country analyses [52, 54, 76, 98]. They found evidence of a negative correlation between A\H1N1 and A\H3N2, while interactions with B appeared less clear. In Chapter 4 we will extensively investigate these aspects. We will consider (sub)types as part of a unique coupled ecological system and focus on their relative abundances across countries and years. In a first step, following the objectives of previous studies, we will characterize the spatio-temporal patterns of (sub)type composition and propose a new statistical framework particularly suited for these analyses. In a second step, we will also propose methods to predict (sub)type composition one year in advance, a question not yet addressed in the literature.

An important question, partly related to the previous one, concerns the role of virus

interaction in their spread. Clades of the same influenza subtype indeed interact, and quantifying their degree of interaction is the basis for choosing the optimal vaccine composition. As discussed above, the interaction between (sub)types is less clear but still exists. Another issue is the interaction of influenza with other respiratory viruses. For example, there is conflicting evidence on the interaction between influenza and RSV [117, 118]. Evidence of negative correlations between influenza and other common respiratory viruses was found in the years before the COVID-19 pandemic [117, 119]. More recently, the question has been raised as to whether the interaction between viruses may have been one of the factors explaining the decline in influenza during the COVID-19 pandemic.

Finally, the impact of non-pharmaceutical interventions to limit the spread of communicable diseases has long been studied. During the 2009 pandemic, there was a significant reduction in international travel with Mexico, but this did not slow the spread of the virus as hoped [120]. Other measures related to contact tracing and containment were used during the SARS outbreak in 2003 ([121, 122]). In the context of both seasonal and pandemic influenza, school closure has been often discussed as a possible intervention [71, 73, 74, 123]. With COVID-19, non-pharmaceutical interventions were applied on a global scale and with unprecedented intensity and duration. It has been observed that such non-pharmaceutical interventions hindered most communicable diseases. In Chapter 3, we examine the role that reduced mobility and the measures applied in different countries have played in reducing influenza.

Chapter 2

Statistical tools for analyzing the influenza dissemination

This chapter is dedicated to outlining the statistical methodologies utilized in my research projects. The analyses conducted in Chapters 3 and 4, concerning the decline of influenza during the COVID-19 pandemic and the interconnected dynamics among influenza (sub)types, were prompted by specific epidemiological inquiries. As a biostatistician, my task involved identifying the most suitable statistical approaches to tackle these questions, and customizing them as needed to address the specific challenges at hand. This exploration led me to investigate techniques not commonly used within the field of epidemiology, such as the implementation of tree-based methods (Chapter 3), and the utilization of compositional statistics and multivariate probabilistic forecasting (Chapter 4). Moreover, the nature of the epidemiological inquiries and the available data spurred the development of new methodologies, previously unexplored in the literature. Notably, this led to the formulation of a conditional copula model in Chapter 5, where we delved into both its theoretical formulation and demonstrated its efficacy through analyses of synthetic and real-world data.

The code implemented for each project is publicly accessible. In the appendix of each article, a GitHub directory is specified for easy access to the code. Notably, to demonstrate the functioning of the conditional copula model, we developed an algorithm capable of implementing particularly flexible regression trees, allowing the inclusion of categorical covariates and the definition of custom split rules.

The chapter is organized as follows. In the first section, we present tree-based regression models. In particular, we introduce Classification And Regression Trees (used in Chapters 3 and 5), Random Forests, and Random Forest-based variable selection techniques (employed in Chapter 3). Second, we will discuss compositional data processing in the framework of Compositional Data Analysis (CoDA), introducing the basic principles of CoDA and the log-ratio transformations. These notions will form the basis of data treatment for Chapters 4 and 5. Third, we will present the Bayesian Hierarchical Vector AutoRegressive model used for time series predictions in Chapter 4. Finally, we will introduce some elements of copula theory that we will make extensive use of in Chapter 5.

2.1 Tree-based regression methods

Regression trees and Random Forests are non-parametric supervised learning algorithms introduced by Breiman in 1984 and 2001 respectively [124, 125]. They are extremely flexible, as they can be used for both classification and regression tasks and can include both quantitative and qualitative covariates. Because of these properties, they are used in countless applications in fields ranging from public health [126] to ecology

[127, 128], to economics [129].

In Chapter 3, we use regression trees to identify factors associated with the decline of influenza in different countries and trimesters during the first year and a half of the Covid-19 pandemic. The pandemic had a strong impact on the circulation of influenza viruses but with great heterogeneity across regions and time. Potentially, many factors could be responsible for such heterogeneity. Random forests can handle a large number of covariates without requiring assumptions about their distributions and (lack of) correlations, making them particularly well-suited to our problem. Specifically, we use RFs to identify the set of most significant predictors of influenza reduction through a variable selection procedure. We then consider the selected covariates to feed a regression tree, a simpler and more transparent model that provides interpretable results. The interpretability of the results is a critical aspect of the study because it allows us to advance hypotheses about the underlying epidemiologic mechanisms of influenza reduction. However, we also recognize that the results should be read with caution since regression trees suffer from instability. In our case, we have ensured the robustness of the results by carefully selecting the explanatory variables, optimizing the tree complexity, and by carrying out sensitivity analysis.

In Chapter 5, we combine regression trees with copulas to define a conditional copula model. In such a model, the copula parameters depend on covariates, and we use a regression tree to estimate this dependence in a nonparametric way. Again, the choice to use trees is due to their adaptability to different scenarios, in particular the possibility to include qualitative covariates. In this case, we propose a modified version of the classical regression trees, which requires the implementation of the algorithm from scratch to allow flexible splits, not possible with the standard R and Python packages. We anticipate that in this case, the complexity of the algorithm will make it more difficult to interpret the results since the inference procedure involves several steps.

The CART (Classification And Regression Trees), RF (Random Forest), and VSURF (Variable Selection Using Random Forests) algorithms are briefly described in the Supplementary Material in Chapter 3 and in the main text in Chapter 5. However, we take advantage of the next few paragraphs to provide a more complete presentation of these algorithms, specifying how they are used for our analysis.

2.1.1 Regression Trees - the CART algorithm

Construction of the tree - the growing phase. The CART algorithm is the standard procedure used to implement regression trees and was originally introduced by Breiman in 1984 [124]. Given a regression problem $Y_i = f(\mathbf{X}_i) + \epsilon_i$, with $Y \in \mathbb{R}$ the response variable and $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ the vector of covariates, CART iteratively partitions the covariate space $\mathcal{X}$ with hyperplanes parallel to one of the axes. The procedure aims at minimizing a given cost function and is determined by a binary tree with the following characteristics:

- the root node corresponds to the set of all observations $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$;
- the internal nodes are defined by rules of the type $\{X_i^{(j)} < s\}$, for quantitative covariates, or of the type $\{X_i^{(j)} \in A^{(j)}\}$, for qualitative covariates, where $A^{(j)}$ is a set of modalities of the j-th covariate. The (j, s) pair is chosen by evaluating all eligible covariates and split values to optimize a specific cost function;

- the leaves $(T_\ell)_{\ell=1,\dots,K}$, are the terminal nodes of the tree. They correspond to hyper-rectangles in $\mathcal{X}$ and are such that each observation belongs to only one leaf, i.e. $\sum_{\ell=1}^{K}\sum_{i=1}^{n}\mathbf{1}_{X_i\in T_\ell} = n$.

Given a regression tree, the prediction for an observation is computed by averaging the Y_i values of the data points falling in the same leaf ℓ: $\hat{y}_{n+1} = \frac{1}{n_\ell}\sum_{i=1}^{n}\mathbf{1}_{X_i\in T_\ell}$, with n_ℓ being the number of observations in leaf ℓ. An illustration of the CART algorithm is provided in Figure 2.1.

The growing phase ends when a certain stopping criterion is met. Usually, the maximum number of leaves or the minimum number of observations per leaf is fixed a priori. When one of these criteria is met, the so-called maximal tree is obtained. An example of pseudo-code (1) implementing the construction of a regression tree is provided in Chapter 5.

Selection of the optimal subtree - the pruning phase. The maximal tree often tends

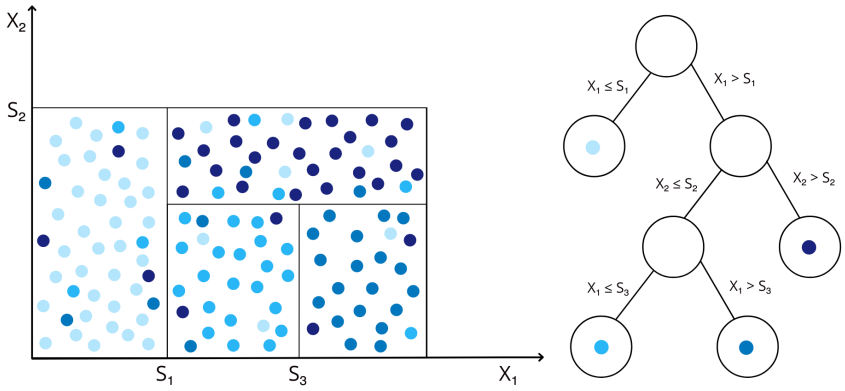


FIGURE 2.1: **Illustration of the CART algorithm**. The observations are divided into different rectangles by perpendicular cuts of the covariate space. The rectangles are defined to be as homogeneous as possible with respect to the response variable, represented here by colors varying from light to dark blue.

to overfit the data and therefore does not offer good predictions. Breiman therefore proposed a pruning procedure to identify the subtree that achieves the optimal compromise between complexity and generalization ability [124]. This is a model selection phase, where the possible models are all the subtrees of the maximal tree. The task may seem extremely costly since the number of subtrees explodes as K increases. However, Breiman also showed that it is not necessary to evaluate all subtrees, as it is sufficient to consider the K trees identified by an iterative procedure in which, starting from the maximal tree, the least advantageous split is eliminated at each step. This procedure results in a sequence of K nested trees (with $K, K-1, \dots, 1$ leaves, respectively) that minimizes the penalized cost function. In practice, pruning is often done by cross-validation, which involves repeatedly splitting the data into training and test sets and minimizing the prediction error on the test set. In this way, an estimate of the prediction error and its standard deviation is calculated for each subtree. Finally, following Breiman's rule, the optimal tree is chosen as the smallest tree with a prediction error less than the minimum error increased by its standard deviation.

Split criteria - adaptation for a likelihood maximization problem. So far, we have

not yet specified the cost function. In the classical CART algorithm, it corresponds to the intra-groups variance, which is the MSE (Mean Squared Error) of the model:

$$Err(T) = \sum_{\ell=1}^{K} \frac{1}{n_\ell} \sum_{i=1}^{n} (Y_i - \bar{Y}_\ell)^2 \mathbf{1}_{X_i \in \mathcal{T}_\ell}, \tag{2.1}$$

with T denoting the tree. This risk is minimized by the maximal tree, while the pruned tree optimizes a penalized risk function of the form $\widetilde{Err}(T) = Err(T) + \lambda|T|$, where $|T|$ is the number of leaves of the tree T and $\lambda > 0$ the penalization constant.

More generally, the CART algorithm can be also used to optimize a likelihood function. In this case, the cost function corresponds to the negative log-likelihood of the model. Specifically, let us assume that the response variable follows a certain parametric distribution, where the value of the parameter varies conditionally upon the covariates, i.e, $Y|\mathbf{X} \sim F_{\theta(\mathbf{X})}$). Then, the goal is to separate the data points into groups as homogeneous as possible regarding the parameter θ of the distribution, such that in each group the distribution is specified by a different parameter. Consequently, the model identified by a tree of K leaves consists of a mixture of K distributions $F_{\hat{\theta}_\ell}$, where the $\hat{\theta}_{\ell=1,\dots,K}$ are the parameters estimated for each leaf. In particular, denoting with $\mathcal{L}$ the log-likelihood function of the model, the vector $(\hat{\theta}_1, \dots, \hat{\theta}_1)$ is the maximum likelihood estimator of the optimal parameters

$$(\theta_1^*, \dots, \theta_K^*) = \arg \max_{(\theta_1,\dots,\theta_K)} \sum_{\ell=1}^{K} \mathbb{E}\left[\mathcal{L}(\theta_\ell; \mathbf{Y}) \mathbf{1}_{\mathbf{X} \in \mathcal{T}_\ell}\right].$$

In practice, to optimize the likelihood of the mixture model, each split of the tree is chosen to maximize the gain in log-likelihood. That is, given a P (parent) node, the left (L) and right (R) child nodes are identified so as to maximize the quantity

$$\mathcal{L}(\theta_L; \mathbf{Y}_i) \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_L} + \mathcal{L}(\theta_R; \mathbf{Y}_i) \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_R} - \mathcal{L}(\theta_P; \mathbf{Y}_i) \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_P}.$$

In Chapter 5, we use this framework to define a conditional copula model - see Section 2.4 at the end of the chapter for a presentation of copulas. In that case, the $F_\theta(\mathbf{X})$ distribution is a parametric copula, and the response variable is a multivariate vector of uniform margins. We prove the asymptotic consistency of the model, first analyzing the behavior of the maximal tree, and then focusing on the subtree selected by the pruning procedure. In particular, we exploit tools from the theory of empirical processes [130, 131] to control the stochastic term of the error.

This custom split poses technical difficulties in terms of implementation. First, Rpart is designed to deal with 'classical' regression problems involving a univariate response variable, whereas in our model we have multivariate response variables (the uniform margins $\mathbf{U}$ of the copula). We make it work by using a workaround: as the response variable, we simply specify a list of indexes, which are then used by the split function to select the observations to be included from an external matrix. This allows us to run the simulations in Chapter 5 using Rpart. However, a second difficulty arises when we include categorical variables in the model. Indeed, the treatment of such variables involves an additional step of sorting the modes before evaluating the optimal split. In principle, in the presence of a categorical variable with M classes, the splits to be evaluated are all those that divide the classes into 2 groups ($M(M-1)/2$ possibilities). However, [132] explains that an efficient way to proceed is rather to sort the categories and then follow this sorting to evaluate only the $M-1$ splits. Specifically,

the classes are sorted by increasing values of $\bar{Y}_{m=1,\dots,M}$, for the classical version of the splits implemented in Rpart, and by increasing values of the parameter $\theta_{m=1,\dots,M}$ in our case.

Therefore, to apply the conditional copula model to influenza surveillance data with qualitative covariates in chapter 5, we implemented a custom procedure of the CART algorithm that is flexible in several ways. *(i)* It performs user-defined splits that optimize a likelihood function passed as an argument. *(ii)* It accepts multivariate vectors for both response variables and covariates. *(iii)* An argument specifies which covariates are qualitative, triggering the necessary sorting step. *(iv)* A pruning function allows the identification of the optimal K-1 subtrees of a K-leaf tree and returns their performance in terms of log-likelihood.

2.1.2 Random Forest and Variable Selection Using Random Forests

Random Forests. We have mentioned that regression trees have the disadvantage of being unstable models, i.e., even a small change in the data may cause the estimated tree to vary significantly. Random forests (RFs) [125] overcome this problem by aggregating the estimates of many random trees. The algorithm relies on two main ingredients. First, the aggregation of a large number of trees leads to a stable estimator. On the other hand, stochasticity is introduced so that the trees are maximally diverse and maximally explore the correlations between covariates and the response variable. Stochasticity is introduced in two ways. First, trees are constructed from bootstrap samples of the data. Second, in constructing the trees, a random subset of the covariates is selected at each split, and the optimal covariate for partitioning the data is chosen only among them. The final prediction of a forest is given by the average of the predictions provided by each tree.

RFs are powerful predictive models that work well without careful tuning of the parameters (number of trees, number of covariates to consider for each split, different criteria for stopping tree construction, and more), but they lose the interpretability of the trees. However, an understanding of how the covariates relate to the response variable can be gained by examining the predictive power of each covariate, which provides a proxy for its importance. This is especially useful in a high-dimensional model. We use RFs for this purpose, adopting the VSURF procedure proposed by Genuer and colleagues.

Variable Selection Using Random Forests. The VSURF [133] procedure estimates the importance of variables using permutation measures, a technique commonly used in machine learning [134, 135, 136], and identifies a set of significant covariates. Consider p covariates $(X^{(1)}, \dots, X^{(p)})_{i=1,\dots,n}$ that are used to predict the values $Y_{i=1,\dots,n}$ of a response variable. Then, the VSURF algorithm works as follows:

1. *Estimation of variable importance (VI).* The importance of a variable is defined as the increase in prediction error when the variable of interest is randomly mixed among observations.

    - First, an RF is trained on the data, and the out-of-bag prediction error (i.e., the mean squared error) is estimated for each tree in the forest. Recall that each tree is built on a bootstrap sample of the data and that the observations not included in the bootstrap constitute the out-of-bag sample;
    - Then the variable importance of the j-th covariate, with $j \in \{1, \dots, p\}$, is calculated. The $X^{(j)}$ values for the observations of the out-of-bag sample are

shuffled and the prediction error is evaluated on the perturbed out-of-bag sample ($\widetilde{errOOB}^j$). The variable importance for the j-th covariate writes

$$VI^j = \frac{1}{ntree} \sum_{t=1}^{ntree} \left(\widetilde{errOOB}_t^j - errOOB_t \right),$$

where t is the index that runs over the $ntree$ trees in the forest.

- The procedure is repeated for numerous forests to obtain a statistic for the VI of each covariate. This gives an estimate of the mean $\overline{VI}$ and standard deviation σ_{VI}.

2. *Removal of the variables with near-zero importance.* If the importance of a covariate is estimated to be close to zero, it means that the variable does not contribute to the predictions of the model. Consequently, all variables with an importance below a certain threshold are discarded. The idea is to use the smallest standard deviation ($\min\{\sigma_{VI}^j\}$) as the threshold, although this definition can be refined with a fitting procedure (see [133]).

3. *Selection of covariates that form the best predictive model*. In the final step, the variables are sorted by decreasing values of VI, and a forward selection procedure is used to identify the smallest set of covariates that constitute an 'optimal' model. The forward procedure includes in the first model only the covariate with the largest VI, in the second model the two covariates with the largest VI, and so on until the model includes all covariates not excluded in step 2. For each model, its prediction error is calculated with its standard deviation. Finally, following Breiman's rule, the best model is defined as the smallest model with a prediction error less than the minimum prediction error increased by its standard deviation.

2.2 Treatment of compositional data - Compositional Data Analysis

In Chapters 4 and 5, we will carry on a cross-country analysis of the spatio-temporal distributions of three influenza strains, namely influenza A\H1N1, A\H3N2, and B. Our investigation will be based on FluNet data, which consists of virological samples reported by different countries and classified by influenza (sub)type. However, the absolute number of samples is hardly comparable from country to country, due to differences in surveillance systems. Percentages of infections by (sub)type are more robust to biases and were already used to perform cross-country comparisons in previous studies [51, 52, 54, 98].

We are interested in understanding when a (sub)type was dominant and when, on the contrary, there was a high mixing of the three strains, which are the spatial patterns of (sub)type co-circulation and which is the (sub)type that will probably dominate the next season. To answer these questions, the relevant information is the relative abundance of the three (sub)types. Thus, in our analyses, we will forget about the absolute information (i.e. the counts of infections) and only focus on the proportions of (sub)types. However, the treatment of this data requires some caution, as widely inspected by Compositional Data Analysis (CoDA), a branch of statistics that studies this kind of data [28].

In the following, we will present what compositional data is, and the main issues typically encountered in their investigation. Then, we will list the CoDA principles

and introduce the Aitchison geometry which defines a proper metric to compare compositions. Finally, we will present the log-ratio transformations which provide a convenient way to treat compositional data and which we will be employed in the data pre-processing stages of Chapters 4 and 5.

The rest of the section is mainly based on the textbook of Filzmoser, Hron, and Templ [137]. Other sources will be referenced when used.

2.2.1 The compositional data

Compositional data, or, simply, *compositions*, are vectors that sum up to a constant. Following the definition of Filzmoser and co-authors [137], they constitute '*multivariate observations where relative rather than absolute information is relevant for the analysis*' and the data units are typically proportions or percentages. The element of a compositions are called *parts*, and D-dimensional compositions live in the D-part simplex:

$$S^D = \{\mathbf{x} = (x_1, \dots, x_D) \in \mathbb{R}^D | x_i > 0, \sum_{i=1}^{D} x_i = k\},$$

where k is a constant. It follows that $S^D \subset \mathbb{R}^{D-1}$.

The operation that projects D-dimensional vectors into S^D is called *closure*, and it is defined as:

$$\mathcal{C}(\mathbf{z}) = k * \left(\frac{z_1}{\sum_{i=1}^{D} z_i}, \dots, \frac{z_D}{\sum_{i=1}^{D} z_i} \right), \text{ for } \mathbf{z} \in \mathbb{R}^D, k \in \mathbb{R}.$$

Then, a composition $\mathbf{x} \in S^D$ constitutes an equivalence class of the vectors $\mathbf{z} \in \mathbb{R}^D$ such that $\mathcal{C}(\mathbf{z}) = \mathbf{x}$.

Furthermore, given a composition $\mathbf{x} = (x_1, \dots, x_D)$, a subcomposition $\mathbf{x}_s$ with s parts is obtained from the closure of a subvector of $\mathbf{x}$ of length s.

2.2.2 Challenges posed by compositional data

Compositional data are difficult to analyze because they can not be studied using operations from the standard Euclidean geometry. In fact, calculations as the sum of compositions, or the multiplication of a composition by a constant, might result in vectors outside the simplex.

Moreover, Euclidean differences are often misleading when comparing compositions as, since compositions live in a bounded space, a variation of one unit is all the more important the closer one gets to the edges. For example, suppose we monitor changes in the incidence of some diseases. Then, an increase in incidence by 1% can be interpreted as a normal fluctuation for diseases such as influenza, which infects up to 15% of the population each year, while the same increase would raise alarm for rare diseases that usually infect a minimal percentage of the population. This points to the fact that, when working with compositions, the relevant information lies in the ratios between parts, rather than in their differences.

The importance of focusing on ratios of parts is also supported by the fact that compositions might contain spurious correlations between the parts, determined by the common denominator. This problem had been introduced by Pearson already in 1897 [138], and then investigated by Chayes [139] and Aitchison [28, 140], and it is often referred to as the negative bias problem or closure problem. Consider a situation where $\mathbf{x}$ is given by the closure of $\mathbf{z}$. Then a change in one component of $\mathbf{z}$, while the other components remain constant, induces a negative correlation between the parts

of $\mathbf{x}$. However, this correlation is spurious, since it is simply determined by the constraint on the sum, and not by the fact that the components vary in a synchronized manner. In addition, correlations between parts of a composition may be incoherent with correlations calculated on the parts of a subcomposition. To illustrate this aspect, we propose a simple example. Suppose we have soil samples and measure the abundances of some minerals and water. In all samples, the amounts of minerals show little variation around small values, while the amount of water, the most abundant element, varies greatly depending on the climatic conditions in the field at the time of sampling. We then calculate the relative mineral abundances (i) over the entire dataset and (ii) after excluding the water measurements. When we calculate the correlations between the mineral proportions in the two cases, they can take very different values, even with opposite signs, suggesting conflicting evidence about the composition of the soil under analysis. For other examples of spurious correlations, the interested reader can refer to [141, 142, 143]. The problem of spurious correlations can be overcome by looking at the ratios of the parts.

2.2.3 Principles of Compositional Data Analysis

We can now introduce some principles that any analysis of compositional data should meet. As already mentioned, the main motivation for working with compositional data is that we believe the relevant information lies in the ratio of the parts and not in the parts themselves. This leads to the formulation of the principle of scale invariance, according to which the choice of the normalization constant is irrelevant, or, equivalently, the multiplication of a composition by a positive number does not change the results. Second, the results should be invariant under permutations of the parts, since permutations do not change the information embodied in the composition, similar to what happens in standard multivariate statistics. Third, to overcome the problem of relative scales illustrated earlier, dissimilarities are expressed by ratios of parts rather than by differences between them. Finally, the principle of subcompositional coherence states that the information carried by a composition should not contrast with the information carried by its subcompositions. This is assured by the subcompositional dominance, according to which the distance between two compositions cannot be less than the distance between their subcompositions, and by the fact that subcompositions preserve the ratios of parts.

2.2.4 The Aitchison geometry

The Aitchison geometry aims to equip the simplex with an Euclidean vector space structure so that all calculations involving compositions are well-defined. The first step consists of defining two operations to replace the standard vector addition and scalar multiplication. In the Aitchison geometry such operations are called *perturbation* and *powering*, denoted by $\oplus$ and $\odot$, respectively.

- **Perturbation:** given two compositions $\mathbf{x}, \mathbf{y} \in S^D$, then

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}\left((x_1 * y_1, \ldots, x_D * y_D)'\right).$$

- **Powering:** given a composition $\mathbf{x} \in S^D$ and a scalar $\alpha \in \mathbb{R}$, then

$$\alpha \odot \mathbf{x} = \mathcal{C}\left((x_1^\alpha, \ldots, x_D^\alpha)'\right).$$

The set $(S^D, \oplus, \odot)$ constitutes a vector space since the following eighth axioms are satisfied:

1. perturbation is commutative: $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$;
2. perturbation is associative: $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$;
3. there exists the neutral element of perturbation: $\mathbf{e} = \mathcal{C}((1, 1, \ldots, 1)')$, that is the barycentre of $\mathcal{S}^D$;
4. for any compositions $\mathbf{x}$, there exists its inverse element with respect to the perturbation: $\mathbf{x}^{-1} = \mathcal{C}\left((x_1^{-1}, x_2^{-1}, \ldots, x_D^{-1})'\right)$, such that $\mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{e}$ and $\mathbf{x} \oplus \mathbf{y}^{-1} = \mathbf{x} \ominus \mathbf{y}$;
5. powering is associative: $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \cdot \beta) \odot \mathbf{x}$;
6. powering is distributive with respect to the perturbation: $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$;
7. powering is distributive with respect to the scalar addition: $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$;
8. there exists the neutral element of the powering: $1 \odot \mathbf{x} = \mathbf{x}$.

Next, the simplex can be equipped with an inner product, such that norms and distances are defined in the Aitchison sense.

- **Aitchison inner product**: given two compositions $\mathbf{x}, \mathbf{y} \in S^D$, their inner product is defined as
$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}.$$
- **Norm of a composition**: the norm of $\mathbf{x} \in S^D$ is defined via the inner product
$$\|\mathbf{x}\|_A = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_A}.$$
- **Distance between compositions**: the distance between two compositions $\mathbf{x}, \mathbf{y} \in S^D$ is
$$d_A(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}.$$

Therefore, the simplex $(S^D, \oplus, \odot, \langle \cdot, \cdot \rangle_A)$ is a $(D-1)$-dimensional Euclidean vector space, denoted as the Aitchison geometry.

Once a proper geometry is defined in the simplex, any operation between compositions is well defined. For example, we can now evaluate dissimilarities between compositions by computing the Aitchison distance. Of course, the problem of evaluating dissimilarities between vectors of proportions is widespread in the literature, and alternative measures such as the Shannon entropy or the Kullback-Leibler divergence are commonly used in ecology or information theory. Thus, the Aitchison distance might seem to be just an alternative choice, not necessarily simpler or better than other more common measures. However, the point here is that the Aitchison geometry defines a general framework for performing any statistical analysis or computation on compositions. It is a way to operate on compositions as if we were in standard Euclidean space. For example, given a sample of compositions, we can compute the center of

the distribution and its dispersion around the center. But we can also fit some probabilistic models, estimate their parameters, evaluate the presence of outliers, and more. These analyses require the standard Euclidean geometry calculations to be translated into Aitchison geometry. For example, given a sample of n compositions of D-parts, the average composition is computed as:

$$\overline{\mathbf{x}}_n = \frac{1}{n} \odot \bigoplus_{i=1}^{n} \mathbf{x}_i = \mathcal{C}\left(\left(\prod_{i=1}^{n} x_{i1} \right)^{1/n}, \left(\prod_{i=1}^{n} x_{i2} \right)^{1/n}, \ldots, \left(\prod_{i=1}^{n} x_{iD} \right)^{1/n} \right)'.$$

However, this translation work can be laborious in practice. A more convenient approach is to map the compositions into Euclidean space, where standard operations can be directly applied, with the added benefit that computational tools for statistical analysis are also already available. The results can then be transformed back to the simplex. This is usually done with log-ratio transformations, where the logarithm of the ratios of the parts is calculated. Three commonly used log-ratio transformations are presented in the next section.

2.2.5 The log-ratio transformations

Aitchison, in his seminal studies of Compositional Data Analysis in the 1980s [28, 140], proposed two transformations for mapping compositions into the Euclidean space, namely the additive log-ratio (*alr*) and the centered log-ratio (*clr*) transformations. In 2003, Egozcue and co-authors [144] formulated a third map, the isometric log-ratio (*ilr*) transformation, to overcome some limitations of the previous two.

All three transformations provide a one-to-one mapping of compositions into the real space. They are based on log-ratios, which are easier to handle mathematically than ratios. Moreover, the use of the logarithm allows to map the neutral element of the perturbation $\mathcal{C}\left((1,\ldots,1)'\right)$ into the neutral element of the addition in the standard Euclidean space. The two transformations proposed by Aitchison have a long historical usage, but they do not fully fulfill all the principles listed above. The isometric log-ratio transformation, on the other hand, satisfies all the principles but sacrifices some interpretability. The three transformations are defined hereafter.

The additive log-ratio transformation. Given a compositions $\mathbf{x} \in S^D$, the *alr* transformation is the map $alr : S^D \to \mathbb{R}^{D-1}$ such that

$$\mathbf{z} = alr(\mathbf{x}) = \left(\frac{\ln x_1}{\ln x_D}, \ldots, \frac{\ln x_{D-1}}{\ln x_D} \right)', \mathbf{z} \in \mathbb{R}^{D-1}. \tag{2.2}$$

The *alr* transformation is invariant under multiplication by a constant, i.e. $alr(\mathbf{x}) = alr(k * \mathbf{x}), k \in \mathbb{R}$. Moreover, the operations of perturbation and powering translate into the standard addition and multiplication under the *alr* transformation: $alr(\mathbf{x} \oplus \mathbf{y}) = alr(\mathbf{x}) + alr(\mathbf{y})$ and $alr(\mathbf{x} \odot \mathbf{y}) = alr(\mathbf{x}) * alr(\mathbf{y})$, with $\mathbf{x}, \mathbf{y} \in S^D$. However, this is not true for the Aitchison inner product and the Aitchison distances, meaning that distances and angles are not preserved by the *alr* transformation. In addition, from the formula 2.2 it is clear the D-th part of the composition plays a predominant role, and consequently the *alr* transformation is not invariant under permutation.

The centered log-ratio transformation. Given a compositions $\mathbf{x} \in S^D$, the *clr* transformation is the map $clr : S^D \to \mathbb{R}^D$ such that

$$\mathbf{z} = clr(\mathbf{x}) = \left(\frac{\ln x_1}{\sqrt[D]{\prod_{k=1}^{D} x_k}}, \dots, \frac{\ln x_D}{\sqrt[D]{\prod_{k=1}^{D} x_k}} \right)', \mathbf{z} \in \mathbb{R}^D. \tag{2.3}$$

The denominator of the components is the *geometric mean* of the compositions $\mathbf{x}$:

$$g_m(\mathbf{x}) = \sqrt[D]{\prod_{k=1}^{D} x_k},$$

which provides a notion of the 'average part'. It means that, under the *clr* map, the parts of $\mathbf{x}$ are rescaled with respect to their 'center', and this overcomes the problem of permutation invariance. Moreover, it can be proved that the *clr* is an isometric transformation (i.e. that preserves distances and angles). However, an important drawback of *clr* coordinates is that they are redundant. In fact, they always sum to 0, meaning that the vectors $clr(\mathbf{x})$ live on a hyper-plane of $\mathbb{R}^D$, and are not expressed in terms of a basis. This also implies that some problems of spurious correlations might persist.

Moreover, in our analyses, it will be important to minimize the dimensionality of the data. Consequently, we won't consider the *clr* transformation, and we will only use transformations that map 3-part compositions to 2-dimensional vectors.

The isometric log-ratio transformation. Given a compositions $\mathbf{x} \in S^D$, the *ilr* transformation is the map $ilr : S^D \to \mathbb{R}^{D-1}$ such that

$$\mathbf{z} = ilr(\mathbf{x}) = (z_1, \dots, z_{D-1})', \tag{2.4}$$

where

$$z_j = \sqrt{\frac{D-j}{D-j+1}} \ln \frac{x_j}{\sqrt[D-j]{\prod_{k=j+1}^{D} x_k}} \text{ for } j = 1, \dots, D-1.$$

The *ilr* transformation is scale-invariant and permutation-invariant. Furthermore, it is an isometry, meaning that: $\langle \mathbf{x}, \mathbf{y} \rangle_A = \langle ilr(\mathbf{x}), ilr(\mathbf{y}) \rangle$ and $d_A(\mathbf{x}, \mathbf{y}) = d(ilr(\mathbf{x}), ilr(\mathbf{y}))$. Furthermore, the *ilr* coordinates constitute an orthonormal basis in the hyper-plane formed by the *ilr* coordinates. The drawback of these coordinates is that they can be difficult to interpret, especially in high dimension. However, in our studies, we will always work with 3-part compositions, for which an interpretation is pretty straightforward. Given a composition $\mathbf{x} = (x_1, x_2, x_3)' \in S^3$, the *ilr* coordinates are:

$$\begin{cases} z_1 &= \sqrt{\frac{2}{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}} \\ z_2 &= \sqrt{\frac{1}{2}} \ln \frac{x_2}{x_3} \end{cases}$$

In this case, z_1 measures the relative amount of the first part (x_1) with respect to the average abundance of the other two (x_2 and x_3), while z_2 measures the relative abundance of x_2 vs. x_3. Note that any analysis will be invariant under the permutation of the parts, but the interpretation of the results will be more or less straightforward depending on their order. This will be particularly relevant in our analysis, where we consider three influenza strains (A\H1N1, A\H3N2, and B) whose mutual roles are not equivalent. In fact, we expect the subtypes of influenza A to interact more with

each other than with B. Therefore, we will always consider the proportion of B as the first part, so that z_1 will express the relative abundance between B and the average abundances of A, while z_2 will measure the proportion of A\H1N1 vs. A\H3N2.

2.3 Time series probabilistic forecasting

In Chapter 4, we predict the composition of influenza (sub)types by country one year in advance. This constitutes a time series prediction problem, which is tackled by using several models, from naive estimators to hierarchical autoregressive models.

Data consists of country-year proportions of influenza infections by (sub)type, i.e. vectors of the form (B%, A\H1N1%, A\H3N2%)$_{t,c}$, where t denotes the year and c the country. Before any analysis, these vectors are mapped into the Euclidean space $\mathbb{R}^2$ through the isometric log-ratio transformation (*ilr*) (Cf. section 2.2.5). Let's name (u, v) the *ilr* coordinates in the Euclidean space, and $y_{t,c} = ilr\left((\text{B\%}, \text{A\textbackslash H1N1\%}, \text{A\textbackslash H3N2\%})'\right) = (u, v)'_{t,c}$ the correspondent country-year vectors. Then, the time series for country c is a bivariate trajectory of the form $(y_1, \ldots, y_T)_c$.

In our setting, forecasting is particularly challenging due to the very short time series we have, consisting of 10 points as the years under analysis range from 2010 to 2019. In addition, we have to keep the last part of the series for model validation, and then we can only consider the first 7-9 points (depending on the scenario) to train the prediction algorithms. The shortness of the time series, together with their multivariate nature, imposes the use of very simple models to keep the number of parameters to be estimated as low as possible. Among the simplest and most common methods for time series forecasting are autoregressive models, which express predictions as linear functions of past observations. In Chapter 4, we make predictions using the multivariate version of the autoregressive models, namely the vector autoregressive (VAR) models.

Although the time series are short, we have several of them, as our analysis covers 81 countries. Thus, we can compensate for the lack of data over time by using information from similar countries to improve the predictions. To do this, we consider a hierarchical vector autoregressive (HVAR) model [145], a model designed to provide estimates for an ensemble of similar trajectories. This model assumes that all trajectories in the ensemble are generated by similar VAR processes, implying that individual country trajectories have similar VAR coefficients sampled from some latent common distributions. By using appropriate priors for the latent distributions, it is possible to write the likelihood of each coefficient, conditional on the other parameters. This allows the optimal coefficients to be estimated via efficient Monte Carlo Gibb sampling.

Both the VAR and HVAR models provide probabilistic forecasts, i.e. point forecasts with confidence intervals. On the one hand, for the VAR process, the Least Squares estimator of the coefficients can be derived analytically, and its dispersion is evaluated asymptotically. On the other hand, the Monte Carlo procedure directly provides forecast distributions for the HVAR coefficients. Probabilistic forecasts are particularly important to assess the goodness of the predictions and to allow model comparison, and they are commonly used to evaluate epidemiological models [9, 146, 147].

In the following, we present the vector autoregressive (VAR) model and its extension to the hierarchical version (i.e., the HVAR model), both of which are used in chapter 4. The main goal of this section is to illustrate the structure of the HVAR model without going into the details of the marginal likelihood calculations, which can be found in the original paper of [145].

2.3.1 The Vector AutoRegressive model (VAR)

Hereafter, we present the vector autoregressive processes following the textbook of Lütkepohl [148].

Let's consider a multivariate time series of the form $(y_1, \dots, y_T)$, where each vector of the series has size K. Let's assume that the time series has been generated by a stable VAR process of order p, denoted as VAR(p). Then, the vector at time t can be expressed as a linear function of the previous p vectors:

$$y_t = \nu + A^{(1)} y_{t-1} + \cdots + A^{(p)} y_{t-p} + \epsilon_t,$$

where ν is a vector of K intercept terms, $A^{(i)}$ are $K \times K$ coefficient matrices, and ϵ is Gaussian noise. The goal is to estimate the model's coefficients from the available data.

Since we have a sample of T vectors y_t, with $t = 1, \dots, T$, then the same VAR(p) process can be written for each one of the last $T - p$ vectors. Thus, the model's coefficients can be estimated from the following equations:

$$\begin{cases} y_{p+1} & = \nu + A^{(1)} y_p + \cdots + A^{(p)} y_1 + \epsilon_{p+1} \\ \vdots & \\ y_T & = \nu + A^{(1)} y_{T-1} + \cdots + A^{(p)} y_{T-p} + \epsilon_T \end{cases}$$

However, following Chapter 3 of [148], the model can be written in a more compact form. Let's define the following quantities:

$$\begin{aligned} Y &= (y_{p+1}, \dots, y_T), \\ B &= (\nu, A^{(1)}, \dots, A^{(p)}), \\ Z_t &= \begin{pmatrix} 1 \\ y_p \\ \vdots \\ y_1 \end{pmatrix}, \\ Z &= (Z_p, \dots, Z_{T-1}), \\ U &= (\epsilon_{p+1}, \dots, \epsilon_T). \end{aligned}$$

Then, the VAR(p) process can be written as $Y = BZ + U$.

The least squares estimator $\hat{B}$ of the process can be derived, and results in $\hat{B} = YZ'(ZZ')^{-1}$. Then, the prediction for the year $T + 1$ is computed as $\hat{y}_{T+1} = \hat{B} Z_T$.

In addition, under the hypothesis of Gaussian white noise, the consistency and asymptotic normality of the least squares estimator are guaranteed, and a plausible estimator of the asymptotic dispersion is provided. In our analysis, the confidence intervals of the prediction are computed via an unbiased estimator of the white noise covariance matrix, corrected for short time series. For bivariate time series, it writes

$$\hat{\Sigma}_\epsilon = \frac{T - p + 1}{(T - p)(T - 3p - 1)} (YY' - YZ'(ZZ')ZY').$$

2.3.2 The Bayesian Hierarchical Vector AutoRegressive (HVAR) model

Now let's consider an ensemble of similar time series. We use the Bayesian Hierarchical Vector AutoRegressive (HVAR) algorithm proposed by Lu and colleagues [145] to

model all the series in the ensemble simultaneously[1]. Specifically, all the time series are assumed to be generated by similar VAR(p) processes, meaning that the coefficients of each VAR(p) process are sampled from common distributions defined by some parameters to be estimated.

In our analysis of Chapter 4, the ensemble of time series corresponds to an ensemble of country-trajectories of (sub)type compositions over time, with trajectories defined in a 2-dimensional space. Thus, for each country c of the ensemble, we denote with $(y_1, \dots, y_T)_c$ the corresponding trajectory, and with $y_{t,c} = (u, v)'_{t,c} \in \mathbb{R}^2$ the country-year vectors.

The VAR(p) process for country c can be explicitly written as:

$$\begin{pmatrix} u \\ v \end{pmatrix}_{t,c} = \begin{pmatrix} \nu_u \\ \nu_v \end{pmatrix}_c + \begin{bmatrix} A^{(1)}_{uu} & A^{(1)}_{uv} \\ A^{(1)}_{vu} & A^{(1)}_{vv} \end{bmatrix}_c \begin{pmatrix} u \\ v \end{pmatrix}_{t-1,c} + \dots + \begin{bmatrix} A^{(p)}_{uu} & A^{(p)}_{uv} \\ A^{(p)}_{vu} & A^{(p)}_{vv} \end{bmatrix}_c \begin{pmatrix} u \\ v \end{pmatrix}_{t-p,c} + \begin{pmatrix} \epsilon_u \\ \epsilon_v \end{pmatrix}_{t,c},$$

where $(\epsilon_u, \epsilon_v)'_{t,c}$ is Gaussian noise specified by a precision matrix $\Lambda \in \mathbb{R}^{2,2}$. Following the notation defined in the previous section, the same VAR(p) process is written compactly as $Y_c = B_c Z_c + U_c$.

Lu and co-authors propose a model where a hierarchical structure is assumed by assuming that the coefficients of each country are the sum of two contributions: $B_c = W + V_c$. W is a matrix encoding the average behavior of the group and is the same for all the trajectories, while V_c is the matrix for the single trajectory adjustment. Assumptions on the distribution of the model coefficients (W, V_c) and of the Gaussian noise (U_c) are summarized hereafter. We adopt a notation very close to [145], the interested reader can refer to the original paper for additional details on the model.

- Elements of W are sampled from a multivariate normal distribution centered in 0 and with precision matrix D:

  $$vec(W)|D \sim MVN(0, D^{-1}).$$

  $vec(W)$ is a column vector containing the columns of W stacked one after the other. D is a diagonal matrix of size $K(Kp+1)$, with entries determined by the parameters $\lambda_{2,j}$ and τ_j^2, for $j = 1, \dots, K(Kp+1)$, such that

  $$diag(D) = (\lambda_{2,1} + \frac{1}{2\tau_1^2}, \dots, \lambda_{2,K(Kp+1)} + \frac{1}{2\tau^2_{K(Kp+1)}}).$$

  Furthermore, coefficients $2\tau_j^2$ follow independent exponential distributions, each of rate $\lambda^2_{1,j}/2\xi_j^2$.

  It remains to precise the priors for $\xi_j^2, \lambda_{1,j}$ and $\lambda_{2,j}$. The ξ_j^2 parameters derive from calculations that involve the precision matrix Λ^{-1} (refer to [145] for the precise formula). The $\lambda_{1,j}$ and the $\lambda_{2,j}$ coefficients have $\Gamma(\mu_1, \nu_1)$ and $\Gamma(\mu_2, \nu_2)$ priors, respectively.

- The country-level coefficients V_c have multivariate normal priors as well, such that

  $$vec(V_c)|\Theta \sim MVN(0, \Theta^{-1}),$$

  with $diag(\Theta) = (\theta_1, \dots, \theta_{K(Kp+1)})$. All the θ_j coefficients, with $j = 1, \dots, K(Kp+1)$, have a unique Gamma prior $\Gamma(\alpha, \beta)$.

[1] Lu and coauthors' model includes VAR processes that are defined without the intercept term. We have slightly modified their model to include such a term.

- For all countries, the Gaussian noise is defined by the same time-invariant precision matrix $\Lambda \in \mathbb{R}^{K,K}$. That is,

$$vec(U_c) \sim MVN(0, \mathbb{I}_T \otimes \Lambda^{-1}).$$

  $\mathbb{I}_T \otimes \Lambda^{-1}$ is a block diagonal matrix of size (KT, KT), with blocks Λ on the diagonal. Furthermore, Λ is assumed to be a positive definite matrix, sampled from a Wishart distribution of parameters S, the scale matrix of size $K \times K$, and d, the degrees of freedom.

- The hyper-parameters $(\mu_1, \nu_1, \mu_2, \nu_2, \alpha, \beta, S, \nu)$ are assumed to be known and in practice are sampled from uniform distributions at the initialization step of the Monte Carlo sampling.

2.4 Copulas - modeling the dependence of multivariate variables

A copula is a multivariate cumulative function with uniform margins. The name *copula* was introduced by Sklar in 1959 by adopting a Latin term for *bound*. In fact, Sklar's theorem states that any multivariate joint cumulative distribution can be expressed in terms of uniform marginal distributions that are *coupled*, *bound* through a function called copula. Hence, this tool allows the investigation of multivariate joint distributions in two steps. Given multivariate random variables, at first, the marginal cumulative distributions are estimated independently. Next, these are combined by choosing an appropriate copula to describe their dependence structure.

In the following sections, we will introduce key concepts of copula theory along with Archimedean copulas, an important parametric class that will be applied in chapter 5. We will then outline the most common methods for copula inference. Finally, we will look at conditional copula models, an active research topic in the field, and detail our contributions. Specifically, we'll introduce the model examined in detail in Chapter 5. Our primary reference is Nelsen's textbook [29], with additional sources referenced when used. To simplify the discussion, we often present results and formulas that apply to bivariate copulas. Extensions to multivariate cases are straightforward in most cases.

2.4.1 Copulas, Sklar's theorem, Fréchet-Hoeffding bounds and measures of dependence

Definition 1 (Copula) *A bidimensional copula is a function* $C : [0,1]^2 \to [0,1]$*, with the following properties.*

1. *C is grounded:* $\forall u, v \in [0,1], \quad C(u,0) = C(0,v) = 0.$
2. *C has uniform margins:* $\forall u, v \in [0,1], \quad C(u,1) = u$ *and* $C(1,v) = v.$
3. *C is 2-increasing:*

$$\forall u_1, u_2, v_1, v_2 \in [0,1], \text{ with } u_1 \leq u_2 \text{ and } v_1 \leq v_2,$$
$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$$

Sklar's theorem provides the connection between copula functions and probability distributions and constitutes a central element of the theory of copulas.

Theorem 1 (Sklar's theorem) *Consider H a joint cumulative distribution function with margins F and G. Then, H admits a copula representation:*

$$H(x, y) = C(F(x), G(y)). \tag{2.5}$$

If F, G are continuous, then the C copula is unique. Conversely, if C is a copula and F, G are cumulative distribution functions, then the function H defined by 2.5 is a joint cumulative distribution function with margins F and G.

Notice that, under the assumption of continuity for F and G, the inverse formula holds, and the C copula can be expressed as:

$$C(u, v) = H(F^{-1}(u), G^{-1}(v)). \tag{2.6}$$

This is an equivalent formulation of Sklar's theorem which offers a practical way to build copulas from joint distribution functions, often useful in applications.

Given that a copula is a joint cumulative distribution function, then the properties enumerated in definition 1 can be also written in terms of probability. Considering the random variables U, V uniformely distributed in $[0, 1]$, and $C(u, v) = \mathbb{P}(U \leq u, V \leq v)$, then property 1) implies that $\mathbb{P}(U \leq u, 0) = \mathbb{P}(0, V \leq v) = 0$. Property 2) corresponds to $\mathbb{P}(U \leq u, V \leq 1) = \mathbb{P}(U \leq u) = u$ and $\mathbb{P}(U \leq 1, V \leq v) = \mathbb{P}(V \leq v) = v$, and thus finds that the margins are uniforms. Property 3) is equivalent to $\mathbb{P}(u_1 \leq U \leq u_2, v_1 \leq V \leq v_2) \geq 0$, which is a basic requirement of any cumulative distribution function.

From the definition, we know that bivariate copulas are particular functions confined within the unit cube $[0, 1]^3$. However, this understanding can be enriched. Specifically, there exist two peculiar copulas, known as the Fréchet-Hoeffding bounds, which delineate the extremities encompassing all other copulas.

- The Fréchet-Hoeffding lower bound is the copula W such that: $W(u, v) = max(u + v - 1; 0)$.
- The Fréchet-Hoeffding upper bound is the copula M such that: $M(u, v) = min(u; v)$.

For each copula C holds: $W(u, v) \leq C(u, v) \leq M(u, v), \forall (u, v) \in [0, 1]^2$. Moreover, another important copula that is often used for reference is the product copula: $\Pi(u, v) = uv$. It gives the joint probability distribution of independent random variables. In fact, for two uniform and independent random variables U, V, it holds that $\mathbb{P}(U, V) = \mathbb{P}(U)\mathbb{P}(V) = uv$.

We have said that copulas allow the description of dependencies between random variables. However, when speaking of dependency measures, reference is often made to simpler indicators, such as Pearson's linear correlation coefficient or Kendall and Spearman's rank correlation coefficients. Kendall's τ and Spearman's ρ coefficient, in particular, along with copulas enjoy the property of scale invariance, i.e. their value does not change under strictly increasing transformations of the random variables. These coefficients are related to copulas by the following expressions:

$$\begin{aligned} \tau(X, Y) &= \tau_C = 4 \int_0^1 \int_0^1 C(u, v)\, dC(u, v) - 1; \\ \rho(X, Y) &= \rho_C = 12 \int_0^1 \int_0^1 uv\, dC(u, v) - 3; \end{aligned} \tag{2.7}$$

with X, Y independent random variables, $u = F_X(x)$ and $v = G_Y(y)$. These coefficients take values in $[-1, 1]$ and it can be verified that $\tau_W = \rho_W = -1$, $\tau_\Pi = \rho_\Pi = 0$ and $\tau_M = \rho_M = 1$.

2.4.2 Archimedean copulas

Among the parametric copulas, the Archimedean copulas stand out as one of the most significant classes. Their ease of construction and ability to model a wide range of dependency structures make them particularly important. Examples of their applications can be found in [149, 150, 151, 152, 153, 154], the interested reader can find more reference of applied studies in [155]. The model proposed in Chapter 5 applies to any parametric copulas, and its operation will be demonstrated using three commonly encountered Archimedean copulas.

A parametric copula is generally denoted as $C_\theta(u, v)$, whit $\theta \in \Theta \subseteq \mathbb{R}^m$. Archimedean copulas constitute a specific class of parametric copulas which are uniquely defined by a function known as the *generator*. They are defined as follows:

Definition 2 (Archimedean copula) *A bidimensional Archimedean copula is a copula that can be expressed as:*

$$C(u, v) = \varphi^{[-1]}\left(\varphi(u) + \varphi(v)\right),$$

where φ is the copula generator and $\varphi^{[-1]}$ is its pseudo-inverse. φ and $\varphi^{[-1]}$ satisfy the following properties:

1. $\varphi : [0, 1] \to [0, \infty]$ *is continous, strictly decreasing, convex and such that* $\varphi(1) = 0$;
2. $\varphi^{[-1]} : [0, \infty] \to [0, 1]$, *such that*

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t), & 0 \leq t \leq \varphi(0), \\ 0, & \varphi(0) \leq t \leq \infty. \end{cases}$$

   Consequently, $\varphi^{[-1]}$ is continuous and non-decreasing on $[0, \infty]$;

3. $\varphi^{[-1]}(\varphi(u)) = u$ *and* $\varphi\left(\varphi^{[-1]}(t)\right) = \min(t, \varphi(0))$;
4. *Finally, if $\varphi(0) = \infty$, then $\varphi^{[-1]} = \varphi^{-1}$ and the copula is said to be a strict Archimedean copula.*

Genest and Rivest [150] provided a simplified formulation of Kendall's τ coefficients for Archimedean copulas in terms of the generator function:

$$\tau_\theta = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi\prime(t)} dt. \tag{2.8}$$

Here, we anticipate that this relationship offers a direct means of estimating copulas using the method-of-moments, which we will discuss in the following section. Additionally, it's worth noting that for one-parameter copulas, θ and τ are linked through a bijective function. Consequently, one can opt to define the problem using either the θ or τ values. Often, the τ scale is preferred for comparing goodness-of-fit measures across different copula families [151, 153].

The Clayton, Frank, and Gumbel families are three Archimedean copulas that are often considered in the literature and will be in use in Chapter 5. They are defined by a single parameter and are therefore particularly convenient for applications. Their characteristics are summarised in Table 2.1.

Copula	$\varphi_\theta(t)$	$C_\theta(u,v)$	$\theta \in$	τ
Clayton	$\frac{t^{-\theta}-1}{\theta}$	$\left(\max\left(^{-\theta}+v^{-\theta}-1,0\right)\right)^{-1/\theta}$	$[-1,\infty)\backslash\{0\}$	$\frac{\theta}{\theta+2}$
Frank	$\ln\frac{e^{-\theta}-1}{e^{-\theta t}-1}$	$-\frac{1}{\theta}\ln\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$	$(-\infty,\infty)\backslash\{0\}$	$1+\frac{4}{\theta}(D_1(\theta)-1)$
Gumbel	$(-\ln t)^\theta$	$\exp\left(-\left((-\ln u)^\theta+(-\ln v)^\theta\right)^{1/\theta}\right)$	$[1,\infty)$	$\frac{\theta}{\theta+1}$

TABLE 2.1: Clayton, Frank, and Gumbel copula families. D_1 is the Debye function of order 1: $D_1(\theta) = \int_0^\theta \frac{t}{\theta}(e^t-1)dt$.

2.4.3 Copula inference

In applications with multivariate observations, usually, these are considered realizations of random variables whose dependence can be modeled using copulas. As already mentioned, copula estimation usually involves two independent steps: margin estimation, which returns the so-called pseudo-observations ($\tilde{U}_i = \widehat{F}_X(X_i), \tilde{V}_i = \widehat{G}_Y(Y_i)$), and estimation of the copula itself. In general, both steps of inference can be approached with either parametric or nonparametric estimators. It is not mandatory to adopt the same framework for the two steps. A copula estimator is said to be parametric if both the margins and the copula are modeled with parametric distributions, while it is said to be semi-parametric when the pseudo-observations are empirically estimated and injected into the parametric copula estimator. In the following, we will present some of the most common inference methods.

Empirical estimation. A simple empirical estimator that is often considered a reference model is the empirical distribution function:

$$\widehat{C}_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i\le x, Y_i\le y}$$

This estimator can be refined by kernel smoothing techniques. The summation then becomes a weighted average as follows:

$$\widehat{C}_{n,\mathbf{H}}(x,y) = \frac{\sum_{i=1}^n K_{\mathbf{H}}((x-X_i,y-Y_i))\mathbf{1}_{X_i\le x,Y_i\le y}}{\sum_{i=1}^n K_{\mathbf{H}}((x-X_i,y-Y_i))},$$

where weights are given by the kernel smoother K which depends on the bandwidth $\mathbf{H}$. $\mathbf{H}$ is a symmetric and positive definite matrix which controls the extent of the smoothing along each direction. For instance, for a Gaussian smoother, it corresponds to the covariance matrix of the kernel distribution. Furthermore, $|\mathbf{H}|$ goes to zero as n increases. There exist several eligible distributions for kernel smothers, for more details the interested reader can refer for example to [156, 157]. Let's precise that the $\mathbf{H}$ parameter does not introduce any assumption about the shape of the *true* unknown copula approximated by $\widehat{C}_{n,\mathbf{H}}$.

Maximum Likelihood based methods. For parametric copulas, the model's likelihood is determined by the copula density evaluated at the observations. In two dimensions,

the copula density is equivalent to

$$c(u,v;\theta) = \frac{\partial^2 C(u,v;\theta)}{\partial u \partial v}.$$

Consequently, the log-likelihood of a full parametric model-with both parametric distributions for margins and the copula- is written as:

$$\mathcal{L}_n(\theta,\theta_X,\theta_Y) = \sum_{i=1}^{n} \log c(F_X(X_i;\theta_X), G_Y(Y_i;\theta_Y);\theta) + \sum_{i=1}^{n} \log f_X(X_i;\theta_X) + \sum_{i=1}^{n} \log g_Y(Y_i;\theta_Y),$$

whit (θ_X,θ_Y) the parameters for the marginal distributions and θ the one defining the copula. Thus, the maximum likelihood parameter estimators are:

$$(\hat{\theta},\hat{\theta}_X,\hat{\theta}_Y) = \arg\max_{(\theta,\theta_X,\theta_Y)} \mathcal{L}_n((X_i,Y_i)_{i=1,\dots,n};\theta,\theta_X,\theta_Y).$$

The computational expense of this estimator can escalate quickly, particularly for models with dimensions exceeding two. As a result, in practical scenarios, it's common to optimize parameters sequentially rather than all at once. Following the approach outlined by Shih and Louis [158], parameters for the margins are initially fine-tuned and then injected into the likelihood function of the copula for estimating θ. Another drawback is that the estimator might be biased if the margins are misspecified. To overcome this problem, Genest and co-authors and Shih and Louis proposed a semi-parametric approach, namely the maximum pseudo-likelihood estimator also known as omnibus estimator [150, 158, 159]. It corresponds to a maximum-likelihood estimator where the likelihood of the copula is expressed as a function of empirical pseudo-observations, thus without any assumption about the marginal distributions. The maximum pseudo-likelihood estimator is defined as:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log c(\widehat{F}_n(X_i), \widehat{G}_n(Y_i);\theta). \tag{2.9}$$

The empirical estimation of the margins can be done using the same methods defined in equations 2.4.3 and 2.4.3.

Method of moments. A last, widely used, estimator is the method-of-moment estimator, based on the inversion of dependence measures such as Kendall's τ. As seen previously, Kendall's τ coefficient can be expressed in terms of double integrals of the copula distribution (2.7), and, in the case of Archimedean copulas, the expression turns into a simpler formula depending on the generator φ. A close related quantity is the univariate distribution $K(t) = \mathbb{P}(C(u,v) \leq t)$, that for Archimedean copulas equals to $K_\theta(t) = t - \frac{\varphi_\theta(t)}{\varphi_\theta{}'(t)}$. According to the method-of-moments, the first few moments of $K(t)$ can be expressed in terms of θ and equaled to their empirical estimations. Then, the copula parameters are derived either analytically or numerically. In practice, this method is mostly used when θ is unidimensional and it can be immediately calculated from the empirical Kendall's τ:

$$\widehat{\theta}_n = f(\tau_n), \text{ with } \tau_n = \frac{2}{n(n-1)} \sum_{i<j} sgn(X_i - X_j) sgn(Y_i - Y_j). \tag{2.10}$$

This method has the advantage of being extremely direct for one-parameter copulas; however, it usually performs less well than the maximum pseudo-likelihood estimator

[160].

In the simulations presented in Chapter 5, we will use the method of moments to estimate the copula parameters for the different subpopulations of the conditional model. However, the choice of such subpopulations will be carried out by maximizing the pseudo-likelihood of the model.

2.4.4 Conditional copulas

So far we have discussed models for studying the dependence between two (or more) random variables. However, there are cases where external factors are also present, which could mediate the dependence between the variables in question and further complicate the problem. For example, we can imagine a medical context where it is useful to monitor certain physiological parameters of a patient and study their dependence, which however could vary conditionally upon several factors (age, gender, presence of chronic or prior illnesses, etc.). Hence the interest in formulating copula models that also include covariates.

A first approach is to include covariate information in the margin modeling but not in the copula modeling, adopting what is called the *simplyfing assumption*. On the one hand, several authors adopt this assumption, which substantially simplifies statistical analyses and in some cases seems to give good results even when not fully satisfied (Hobaek 2010). On the other hand, nowadays several conditional copula models have been proposed that avoid the adoption of such a stringent assumption. (For a discussion of this, the reader may refer to [161]).

Sklar's theorem in the case of conditional copulas is analogous to the original formulation [162] and it is expressed as

$$F(\mathbf{y}|\mathbf{X}) = C(F^{(1)}(y^{(1)}|\mathbf{X}), \dots, F^{(k)}(y^{(k)}|\mathbf{X})). \tag{2.11}$$

Here $\mathbf{Y} \in \mathbb{R}^k$ is the vector of multivariate random variables with distribution F and realizations $\mathbf{y}$, $\mathbf{U} = (F^{(1)}(\mathbf{y_1}), \dots, F^{(k)}(\mathbf{y_k}))$ is the vector of uniforms and $\mathbf{X} \in \mathbb{R}^d$ is the vector of covariates. So, we slightly modified the notation compared to the previous part of the chapter (where X was used for one of the copula margins), to be consistent with the notation of Chapter 5 and with what is typically used in the literature of conditional copula models.

Some current conditional copula models. Similar to what was seen earlier, for conditional models the estimation of the copula usually takes place independently of the estimation of the marginal distributions, and different procedures may be adopted for the two steps. In the following, we present some of the models often cited in the literature, to give the reader an idea of current models involving different techniques and degrees of parameterization.

Gijbels and co-authors propose a model in which both the margins and the copula are estimated empirically [157]. They use an estimator similar to equation 2.4.3, but conditioned on covariates. In practice, the conditioning is carried out by the kernel smoother, which weights the relative contribution of the observations by looking at their distance in terms of the covariates. This approach has the advantage of not making any assumptions about the dependence relationships.

Other models instead require the copula family to be chosen a priori and typically consider Archimedean copulas with a one-dimensional parameter $\theta \in \Theta$ as well as univariate covariates. Then, θ is expressed as a function of the covariate in a way such that

$g(\theta(x)) = \eta(x)$, or equivalently $\theta(x) = g^{-1}(\eta(x))$, assuming that g^{-1} exists and that $g^{-1} : \mathbb{R} \to \Theta$. In this framework, g^{-1} is the inverse link function, a monotonic function that serves to rescale $\eta(x)$ values in the interval Θ. g^{-1} is chosen a priori, together with the copula family. For example, for Frank's copulas the parameter θ lives in the interval $\Theta = [1, \infty)$, so a possible choice for the inverse link function is $g^{-1}(t) = \exp(t) + 1$ [151]. In contrast, the focus is on the η function, which is unknown and needs to be estimated. It is a real-valued function, called calibration function, which allows the level of dependence to be adjusted according to the covariate. Different models adopt different forms for η. Acar and co-authors propose a non-parametric estimator based on the maximization of a local log-likelihood, defined as the sum of local contributions [151]. In their model, the contribution of each observation $(U^{(1)}, U^{(2)}, X)_i$ in the surroundings of x is given by $\log c(U_i^{(1)}, U_i^{(2)} | g^{-1}(\beta_0 + \beta_1(X_i - x))$, and these terms are further weighted by a kernel function. The local log-likelihood to be maximized then turns out to be a function of the parameters (β_0, β_1) that holds:

$$\mathcal{L}(\beta_0, \beta_1 | x, h) = \sum_{i=1}^{n} \log c(U_i^{(1)}, U_i^{(2)} | g^{-1}(\beta_0 + \beta_1(X_i - x)) K_h(X_i - x), \tag{2.12}$$

where the kernel bandwidth h represents a hyper-parameter to be tuned a priori. Abegaz et al. formulate a similar model, in which $\eta(x)$ is estimated locally by polynomial functions of order p [152], while Sabeti and colleagues assume that the θ is a function of the covariate via an additive model with terms consisting of cubic spline functions [153].

In all these models, there are decisions to be made regarding the hyperparameters (the bandwidths of the smoothing kernels) and/or the family of copulas to be considered, necessitating a model selection phase. Gijbels and colleagues present formulas for the asymptotic bias and variance, which are dependent on the kernel bandwidths [157, 163]. Consequently, they suggest choosing these hyperparameters to optimize the model's asymptotic performance. Acar and co-authors determine the copula family and kernel bandwidths by assessing cross-validated prediction errors, while Sabeti and colleagues employ a method based on cross-validated pseudo-marginal likelihoods.

Our contribution – copulas meeting regression trees. We introduce a conditional model that essentially is a mixture of copula models, with the subpopulations within the mixture identified non-parametrically through a regression tree. Initially, we designate a copula family and then postulate that the copula parameter can vary across sets of observations sharing similar covariate values. Additionally, the procedure involves a model selection step, that corresponds to the optimization of the tree depth to prevent overfitting. This step is based on cross-validated pseudo-likelihood calculation. The precise formulation of the model and the presentation of the asymptotic convergence results are deferred to Chapter 5. Nonetheless, in the subsequent discussion, we elucidate the main distinctions of our approach from those presented earlier.

The models previously presented go in the direction of proposing estimators that are simultaneously smooth and highly flexible to capture all patterns in the data. For our part, we propose a stepwise estimator that may seem coarser. However, it has important advantages and can be in some ways complementary to the models already in the literature. First, our model allows for the inclusion of categorical covariates. Such variables are ubiquitous in practical applications in fields such as medicine, public health, sociology, economics, risk and insurance, and it is essential to have models that can deal with them. Second, the inclusion of multidimensional covariates does not present

particular technical difficulties in our case, as may be the case with estimators involving smoother kernels [152]. Furthermore, even from a computational perspective, our model appears to be more scalable than other approaches as the size of the covariates grows. In fact, in our model, the computation time is proportional to the number of covariates and their number of unique values - and this also implies that the tree construction is very fast when we include categorical covariates with a number of modalities $M \ll n$, where n is the number of observations. In contrast, for other methods based on likelihood maximization, the addition of covariates implies a quadratic increase in the parameters to be optimized. Among the studies presented above, only Sabeti and colleagues present applications with more than one covariate. In Chapter 5, we will illustrate applications with two covariates. Thirdly, the other models do not discuss the possibility of estimating copulas with multidimensional parameters, whereas our model does not require any special modification to apply to this scenario as well. However, this is not entirely true, since technical difficulties arise in evaluating the optimal split when we have both categorical covariates and multidimensional theta.

Chapter 3

Global patterns and drivers of influenza decline during the COVID-19 pandemic

This Chapter is based on the study entitled *Global patterns and drivers of influenza decline during the COVID-19 pandemic* and published in the *International Journal of Infectious Diseases* [30]. It is a joint work with Pierre-Yves Böelle (Sorbonne Université), Vittoria Colizza (Sorbonne Université, Tokyo Institute of Technology), Olivier Lopez (Sorbonne Université), Maud Thomas (Sorbonne Université), and Chiara Poletto (Sorbonne Université, University of Padova).

The code for the reproducibility of the analysis is available at `https://github.com/FrancescoBonacina/flu-reduction-during-covid-19`.

3.1 Abstract

Objectives: The influenza circulation reportedly declined during the COVID-19 pandemic in many countries. The occurrence of this change has not been studied worldwide nor its potential drivers.
Methods: The change in the proportion of positive influenza samples reported by country and trimester was computed relative to the 2014-2019 period using the FluNet database. Random forests were used to determine predictors of change from demographical, weather, pandemic preparedness, COVID-19 incidence, and pandemic response characteristics. Regression trees were used to classify observations according to these predictors.
Results: During the COVID-19 pandemic, the influenza decline relative to prepandemic levels was global but heterogeneous across space and time. It was more than 50% for 311 of 376 trimesters-countries and even more than 99% for 135. COVID-19 incidence and pandemic preparedness were the two most important predictors of the decline. Europe and North America initially showed limited decline despite high COVID-19 restrictions; however, there was a strong decline afterward in most temperate countries, where pandemic preparedness, COVID-19 incidence, and social restrictions were high; the decline was limited in countries where these factors were low. The "zero-COVID" countries experienced the greatest decline.
Conclusion: Our findings set the stage for interpreting the resurgence of influenza worldwide.

3.2 Introduction

Starting with the global spread of SARS-CoV-2, observations of a sharp decline in influenza circulation were reported. In the first months of 2020, the flu season was shortened in some northern-hemisphere and tropical countries [19, 164]. During the following 18 months, influenza incidence showed an all-time low in New Zealand [6], Australia [5], the United States [24, 165, 166] and the WHO European Region [7]. The circulation was still low in 2021.

The measures adopted in response to the COVID-19 pandemic are likely to have hindered influenza transmission at the same time, since the routes of transmission are identical. Indeed, influenza decline, as well as that of other transmissible diseases, coincided with non-pharmaceutical interventions against COVID-19 [19, 64, 165, 166, 167, 168].

Understanding how this decline occurred may help interpret the current influenza trends and anticipate future viral circulation. While the issue has been described for specific countries or regions [5, 6, 7, 19, 24, 166, 169, 170, 171], little work has been done at the global scale [25, 26, 172].

Here we provide a global quantitative analysis of the influenza reduction based on the Global Influenza Surveillance and Response System FluNet database [18, 62]. We considered the period between March 2020 and September 2021 and estimated influenza reduction by country and trimester relative to a pre-pandemic period (2014-2019). We identified geographical, demographical, health preparedness and COVID-19 status characteristics predictive of influenza decline using random forests and clustered observations with similar decline in time and space using a regression tree.

3.3 Materials and methods

3.3.1 Overview of the methods

We used data from the FluNet influenza repository [18, 62] to quantify the global influenza change during the COVID-19 pandemic (March 2020 to September 2021) compared to the pre-pandemic period (December 2014 to December 2019). We mapped influenza decline by trimester and country. We then used random forests to identify the most significant predictors of decline and a regression tree to classify countries-trimesters based on these predictors. Potential predictors included a wide range of covariates, among them country factors (geographical, meteorological, demographic and health preparedness factors) and variables associated with the COVID-19 pandemic assembled from sources detailed below.

3.3.2 Influenza data and definition of influenza reduction

The FluNet influenza repository [18, 62] provides weekly counts of influenza specimens by country. For our analysis we considered records from 2014 to 2021. To account for influenza seasonality, we defined 13 weeks-long "influenza trimesters" beginning on the first Monday following December 11, March 12, June 11 and September 11. These dates were chosen so that the middle of the December 11 trimester coincided with the peak of a typical influenza circulation in the northern hemisphere. We refer to these trimesters as Dec-Mar, Mar-Jun, Jun-Sep and Sep-Dec, respectively. Data from FluNet was aggregated by trimester-country. The 20 trimesters from Dec-Mar 2014-15 to Sep-Dec 2019 defined the reference "pre-pandemic" period, the six trimesters from Mar-Jun 2020 to Jun-Sep 2021 the "pandemic" period. The trimester from Dec 2019 to

Mar 2020 was excluded as it overlapped the period of COVID-19 emergence. We also discarded trimesters having less than 10 processed influenza specimens per week on average and those typically unaffected by influenza epidemics (i.e. having less than 5% of the annual positive cases on average during the pre-pandemic period, e.g. the summers in temperate regions). We computed the percentage of influenza positive cases as the ratio of positive to positive plus negative samples during the trimester (adding 0.5 to avoid division by zero issues). We computed the "log relative influenza level" as the base-10 logarithm of the ratio between the percentage of positive cases during a trimester and the average percentage of positive cases in the corresponding pre-pandemic trimesters [166, 169]. Under the assumption that influenza surveillance was not substantially altered during the pandemic, this quantifies the reduction in influenza circulation. We also tested for secular trends that could potentially bias this indicator (Supplementary Materials).

3.3.3 Variables for prediction of influenza reduction

We collected the covariates described in Table 3.1 from public sources and IATA. Additional details on computation are provided in the Supplementary Materials.

TABLE 3.1: **Definition, computation and source of the variables used as predictors of influenza change.**

Variable	Description	Source	Min, max
Age	Median age of the country population	[173, 174]	15.1, 48.2
Longitude	Population-weighted average of longitude for cities with more than 300 K inhabitants by country or country capital longitude, from -180 (W) to 180 (E)	[175]	$-100.7, 174.4$
Latitude	Population-weighted average of latitude for cities with more than 300 K inhabitants by country or country capital latitude, from $-90(\text{ S})$ to $90(\text{ N})$	[175]	$-38.7, 60.4$
T	Average temperature (in Celsius degrees) over the trimester-country	[176]	$-8.8, 37.8$
RH	Average relative humidity over the trimester-country	[176]	17.3, 93.5
IDVI	Infectious disease vulnerability index, country level indicator of the vulnerability to health emergencies from 0 (most vulnerable) to 1 (less vulnerable)	[177]	0.15, 1
COVID-19 daily cases	Average daily reported cases of COVID-19 per million inhabitants over the trimester-country	[178]	0, 553.5
Workplace presence reduction	Median percentage of reduction of daily presence in workplaces over the trimester-country. Reduction from the first 5 weeks in 2020 in the same location	[179]	$-22.5\%, 69.0\%$
Reduction of international flights	Average percentage of reduction in the inbound and outbound air passengers over the trimester-country with respect to the same trimester-country of 2019	[180]	$-16.8\%, 100\%$

nb days of school closure	For each country, number of days over the trimester where policies related to schools and universities closure were implemented	[181]	0,91
nb days of workplace closure	For each country, number of days over the trimester where policies related to workplaces closure were implemented	[181]	0,91
nb days of public event restrictions	For each country, number of days over the trimester where policies related to event restrictions were implemented	[181]	0,91
nb days of gathering restrictions	For each country, number of days over the trimester where policies related to social gathering restrictions were implemented	[181]	0,91
nb days of public transport restrictions	For each country, number of days over the trimester where policies related to public transport restrictions were implemented	[181]	0,91
nb days of stay at home requirements	For each country, number of days over the trimester with "shelter-in-place" and otherwise confine to the home orders	[181]	0,91
nb days of international travel restrictions	For each country, number of days over the trimester with airport screening, quarantine of arrival passengers or restrictions of international travels	[181]	0,91
nb days of facial covering requirements	For each country, number of days over the trimester with policies on the use of facial coverings outside the home	[181]	0,91
nb days of testing implementation	For each country, number of days over the trimester with government policy on who has access to testing for current infection (polymerase chain reaction tests)	[181]	0,91
nb days of contact tracing implementation	For each country, number of days over the trimester with government policy on contact tracing after a positive diagnosis	[181]	0,91
nb days of elderly shielding	For each country, number of days over the trimester with policies to protect older adults (as defined locally) in long-term care facilities and/or community and home-based settings	[181]	0,91

3.3.4 Clustering and regression tree analysis

We used the VSURF algorithm based on random forests to select the covariates that were highly predictive of influenza reduction [133]. Importance is defined as the increase in prediction-error when the variable of interest is randomly reshuffled across

observations. We discarded variables with close to zero importance in a univariable analysis. Then, we carried out a forward selection of predictors, including variables in their order of importance one at a time. Following Breiman's rule [124], we retained the model with the least variables having a prediction error less than the minimum prediction error plus one standard deviation. Using the variables selected above, we fit a regression tree in order to obtain an interpretable model [124]. The details of the approach are provided in the Supplementary Materials. Analyses were performed with R version 4.2.1 [182] and packages *vsurf* [29] and *rpart* [132].

3.3.5 Robustness and sensitivity analyses

The details of the robustness checks and the sensitivity analysis are reported in the Supplementary Materials. In summary, we checked the robustness of the regression analysis to stochastic fluctuations in the dataset and to criteria for including the FluNet records in the analysis; we explored alternative definitions for covariates: COVID-19 daily deaths instead of COVID-19 daily cases; Oxford COVID-19 Government Response Tracker stringency index instead of governmental response [181]; alternative Google mobility reports instead of presence in workplaces. We also explored separate inclusion of age and IDVI as these were highly correlated ($\rho_{Spearman} = 0.87$, $p_{val} < 0.01$).

3.4 Results

3.4.1 Decline of influenza in space and time

One hundred sixty-six (166) countries contributed data to FluNet between December 2014 and September 2021. Figure 3.1A shows the time course of the reports. In the pre-pandemic period, the percentage of positive tests varied seasonally between 4% and 33%, with major peaks during seasonal epidemics in northern countries and lower peaks for southern countries. The global number of tests for influenza remained within the range of historical levels throughout the whole COVID-19 pandemic period, but the percentage of influenza positive tests dropped sharply, to a minimum level of 0.04% during the months of July and August 2020.

One hundred twelve countries remained for analysis, contributing 376 trimester-country observations (see Table 3.2 in the supplementary materials). During the pandemic the percentage of influenza positive tests varied across countries and trimesters over five orders of magnitude (from less than 0.002% to a maximum of 49%, as reported in 3.1B), compared to only two orders of magnitude over the pre-pandemic period (between 1% and 95%). For 135 out of the 376 observations, the percentage of positive influenza tests was more than 100 times smaller than expected. The reduction of influenza positivity could be dramatic, as shown by the 0 positive tests out of 26114 processed tests reported in Japan during Mar-Jun 2021, compared to the average 75% expected in the pre-pandemic period. An increase in the percentage of positive tests was seen in 22 observations: This was for example the case for Haiti during Dec-Mar 2020-21, where the percentage of positive tests was 15% compared to an expected 2.2% before the pandemic.

The spatial variation of the influenza decline is mapped in 3.2 over the 6 pandemic trimesters. For the majority of countries, the decline remained limited during Mar-Jun 2020, with 46 out of 65 countries reporting less than 90% reduction from the pre-pandemic period (i.e. log relative influenza level > -1). The decline became more pronounced in the subsequent trimesters, especially in North America, Europe, Mexico and Japan during Dec-Mar 2020-21 and Mar-Jun2021. The decline was also strong in

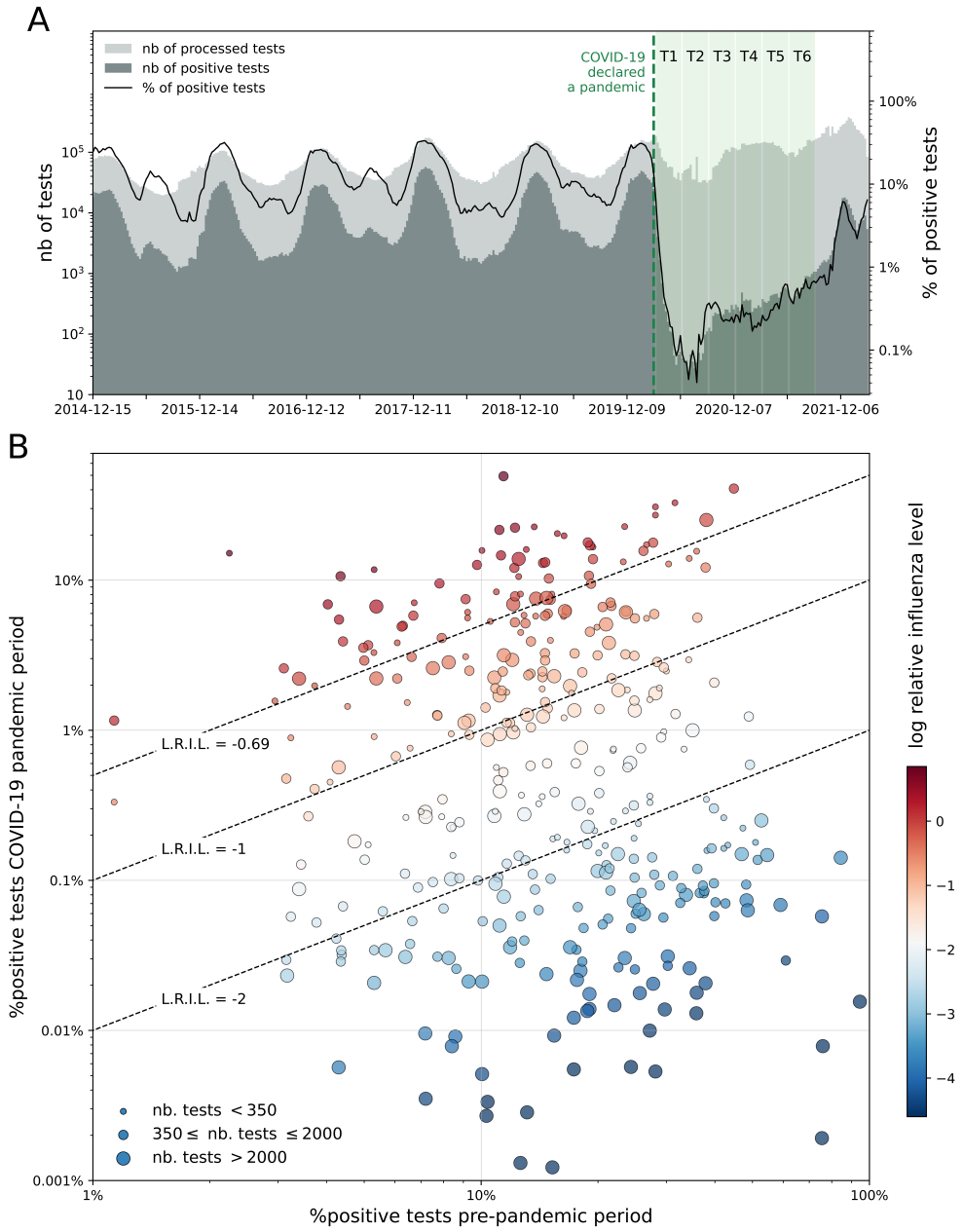


FIGURE 3.1: **Change in influenza circulation during the COVID-19 pandemic relative to the pre-pandemic period. A.** Weekly count of processed and positive tests of influenza reported to FluNet for all 166 countries included in the database from Dec 2014 to Jan 2022. The green shaded area indicates the COVID-19 pandemic period considered in the study. The six blocks indicate the trimesters. The week in which COVID-19 was declared a pandemic by WHO is reported as reference. **B.** Percentage of positive tests for the pre-pandemic and COVID-19 pandemic periods (Dec 2014 - Dec 2019 and Mar 2020 - Sep 2021, respectively), for all 376 countries and trimesters satisfying the filtering criteria on the FluNet data. For each trimester-country, the x coordinate is the average percentage of positive tests of the five years included in the pre-pandemic period, while the y coordinate is the percentage of positive tests during the COVID-19 pandemic period. The size of the dots is proportional to the number of samples found in FluNet for the pandemic period. Dots' colour indicates the log relative influenza level. As guides to the eyes, the three dashed lines indicate the level curves of the log relative influenza level (L.R.I.L.) equal to -2, -1 and -0.69, which correspond to flu reductions of 99%, 90% and 50%, respectively.

the majority of Southern-hemisphere countries during both Jun-Sep 2020 and Jun-Sep 2021. Conversely, a number of countries in South Asia (e.g. Bangladesh, Afghanistan), Africa (e.g. Mali, Senegal, Nigeria, Kenya, Zambia) and Central America (e.g. Honduras, Haiti) showed limited influenza reduction throughout the whole COVID-19 pandemic period (log relative influenza levels > -1). The levels of reduction changed over the period. Interestingly, the log relative influenza level was as low as -2.4 during Jun-Sep 2020 in China but increased again starting Sep-Dec 2020. A similar increasing trend was observed also in a few other countries, e.g. in Kenya and Nigeria.

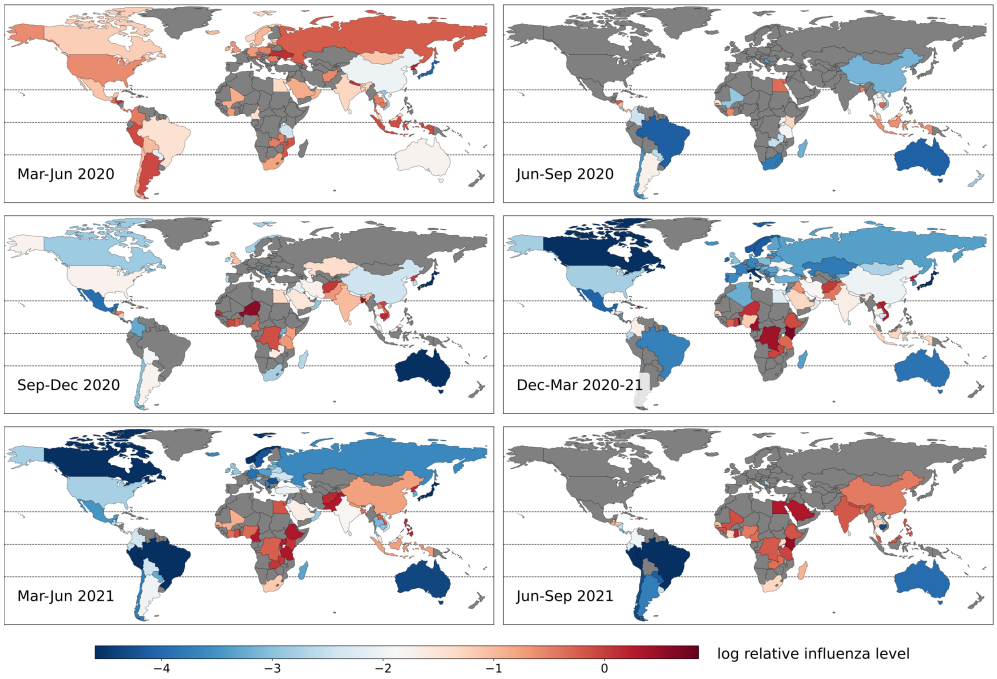


FIGURE 3.2: **Influenza decline during the first 18 months of COVID-19 pandemic by trimester-countries.** Maps of the log relative influenza level for the 6 trimesters considered in the analysis. The grey color indicates trimesters-countries not included in the analysis.

3.4.2 Clustering and regression tree analysis

The analysis was carried out on 93 countries for which covariates were available, totaling 330 trimester-country observations (see Table 3.2 in the supplementary materials). Among the 20 covariates tested, 11 were selected as predictors of the log relative influenza level (Figure 3.3). Sociodemographic, preparedness, geographical, weather and COVID-19 management aspects contributed all to explaining the changes, though COVID-19 daily cases and IDVI were the most important.

The full regression tree built from the data accounted for 69% of the variance of the log relative influenza level (R^2=0.69) (see Figure 3.7 and Table 3.3 in the supplementary materials). To interpret the relationships between the selected variables and the trimesters-countries, we focus here on the first four splits based on IDVI, COVID-19 daily cases, longitude, and workplace mobility reduction (3.4A). The five groups identified by these splits (labeled 1 to 5, Figure 3.4A) showed a gradient in average log relative influenza level ranging from -3.03 (reduction by 99.9%) to -0.71 (reduction by 80%). How the observations in each group rank with respect to the whole dataset is shown in Figure 3.4B. Group 1 included 109 countries-trimesters with high influenza decline, corresponding to the lower quartile of the whole dataset distribution. This

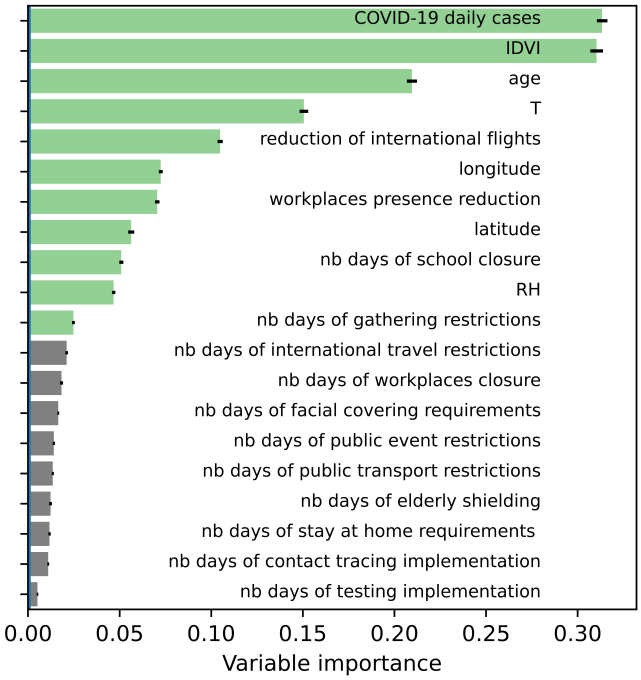


FIGURE 3.3: **Importance of covariates predicting influenza decline in random forest analysis.** Importance of covariates as predictors of the log relative influenza level. In green the 11 covariates selected as significant to build the model with the minimum prediction error following the Breiman's rule. Black segments show the standard deviations of the importance.

group was characterised by high IDVI (median value corresponding to the 71th pc. of the whole dataset), high COVID-19 daily cases (83rd pc.), old population (70th pc.), low temperatures (25th pc.). Median reduction of workplace presence and median number of days with school closure were close to the whole-population median but were higher than other groups, except for group 4 discussed below. Population gathering restrictions were especially high (82nd pc.). The corresponding countries-trimesters included countries in Europe and North America during the 2020-21 influenza season, countries in temperate South America, and high-IDVI countries in Central America and Tropical Asia (see Table 3.3 in the supplementary materials).

Group 2 was the smallest and clustered observations with the largest influenza decline (median log relative influenza level corresponding to the least 8% of all data points). It gathered all observations from Australia, Japan, New Zealand and South Korea. These trimesters-countries showed low COVID-19 daily cases (29th pc.), high IDVI (91st pc) and high reduction of international flights (88th pc.). Reduction of workplace presence, and number of days of school closure and gathering restrictions were comparatively low (23rd, 23rd, 13rd, pcs., respectively).

Group 3 corresponded to 45 observations with intermediary log relative influenza level. Covariates were also close to the median of all data points. Singapore from Sep-Dec 2020 to Jun-Sep 2021 is part of this group (larger tree in Figure 3.8). Covariates of these observations are close to the second group - e.g. high influenza reduction, low COVID-19 daily cases, high reduction of international flights. Other observations of group 3 (e.g. Southeast Asia countries, such as Malaysia, Vietnam, Indonesia and Thailand) were similar to Singapore, but had lower population's age and IDVI. They showed, however, a more limited influenza decline.

Group 4 had 39 observations corresponding to Europe and North America during the Mar-Jun 2020 trimester. At this period, influenza decline was limited (median log

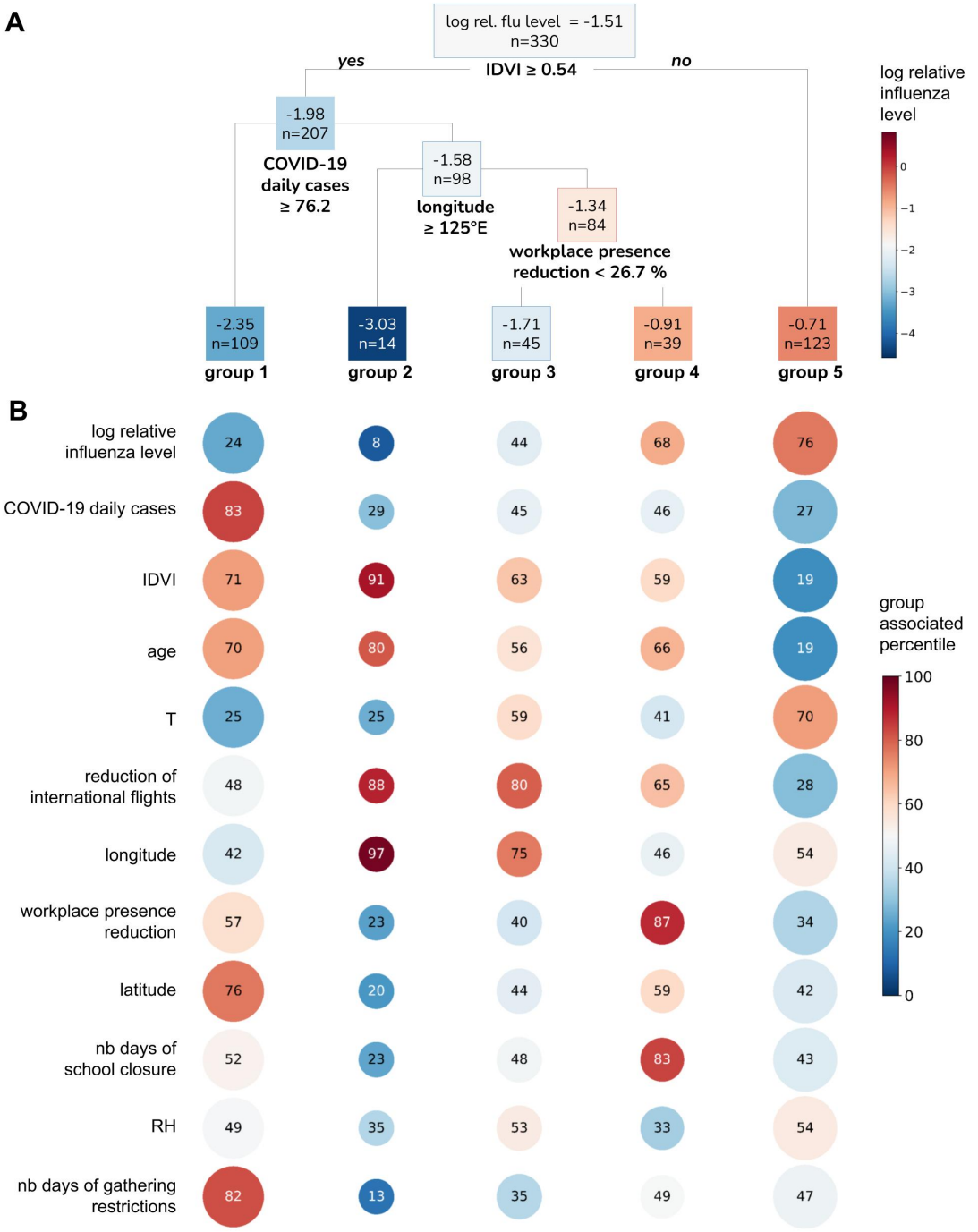


FIGURE 3.4: **Regression tree analysis of influenza decline and characteristics of the identified subgroups. A.** Regression tree obtained with the variables selected in Figure 3.3. We report here the first four splits, which partition the observations in five groups. The full tree is reported in Figure 3.8 of the supplementary materials. For each node the average log relative influenza level and the number of observations are reported (the former is also indicated with a colour scale). **B.** Characteristics of each group. For each variable the colour of the circle indicates the percentile of the whole dataset distribution the median of the group corresponds to. The percentile value is also indicated inside the circle. The size of the circle increases with the number of observations of the group.

relative influenza level corresponding to the 68th pc. of all data points), but there were already a strong response to the COVID-19 pandemic as quantified e.g. by the reduction in the workplace presence (87th pc.) and number of days of school closure (83rd pc.).

Finally, group 5 included 123 trimesters-countries with the lowest decrease in influenza relative to the pre-pandemic period (log relative influenza level 76th pc.). In this group, there was a low number of COVID-19 cases (27th pc.), young population (19th pc.), low IDVI (19th pc.) and high temperatures (70th pc.). The response to the COVID-19 pandemic was mild, with limited reduction of international flights (28th pc.), as well as workplace presence reduction (34th pc.) and number of days of school closure (43rd pc.) small compared to the whole population. This group was largely formed by tropical countries, e.g. in Africa, South and Southeast Asia, Central America and the Caribbean (see Table 3.3 of the supplementary materials).

3.4.3 Robustness and sensitivity analyses

Variable selection and tree structure were robust to stochastic fluctuations. The five-group classification was robust to small perturbations in the data set, as was the selection of predictive variables. In some cases, for example with different inclusion criteria for FluNet data, observations in Group 1 and Group 2 tended to cluster together. More details are reported in the supplementary materials.

3.5 Discussion

The systematic analysis of influenza circulation across all continents and climatic regions shows that the influenza decline was global during the spread of the COVID-19 pandemic. This decrease was heterogeneous across countries and trimesters between March 2020 and September 2021. Demographic, socio-economic, weather and COVID-19 characteristics explained a large part of this heterogeneity.

Influenza circulation is characterised by marked seasonal epidemics in temperate countries but a more complex annual pattern in the tropics [67]. Surveillance may be reinforced in epidemic times. Using the log relative influenza level allowed adjusting for such changes. We found that influenza declined nearly everywhere and remained low compared to the pre-pandemic period during the 18 first months of the COVID-19 pandemic. Importantly, the global number of influenza tests remained roughly the same in the pre-pandemic and pandemic period, ruling out change in surveillance as the likely explanation. The largest reduction was between July and August 2020, and a progressive increase was seen again till September 2021. Temperate countries had the largest reduction, while it was limited in the tropics [25, 171, 172, 183].

Influenza circulation could a priori change during the COVID-19 pandemic because of governmental measures, self-adopted behavioural changes and direct interaction with SARS-CoV-2. We indeed found that reduction of international flights, presence at workplaces, school attendance and mass gatherings explained part of the reduction, although the impact was non-linear. Initial strong restrictions against COVID-19 had to be relaxed in some low-resource countries [171, 184, 185] allowing renewed influenza circulation. Conversely, countries where a strong response against the COVID-19 pandemic could be maintained saw exceptionally low influenza circulation, except in Mar-Jun 2020 where strong local restrictions in Europe and the USA likely occurred at the end of the influenza season, in the majority of cases [5]. For the rest of the time, temperate countries in Europe, North America and South America that

adopted a COVID-19 response centred over local restrictions by reducing workplace presence, school attendance and gatherings had large reduction in influenza circulation, irrespective of the reduction of international flights. This was very different in four "zero-Covid" nations (Australia, New Zealand, Japan and South Korea) where influenza dropped though local restrictions were limited [37], suggesting a key role for border controls in preventing seeding from abroad. Reducing international flights by 94-97% however did not prevent influenza introduction in Vietnam from neighbouring Cambodia [14] likely due to the difficulty of controlling land borders.

Limitation of gatherings or public events, imposed international travel restrictions and school closure were previously found to be the main drivers in suppressing influenza [169, 170]. Actual behaviour, i.e. volume of flights rather than imposed international travel restrictions; or percentage presence at the workplace rather than mandatory reduction was however more predictive of influenza reduction than governmental restrictions. Behavioural proxies may indeed capture adhesion to restrictions that depended on place and stage of the pandemic [186, 187, 188].

Reduction of influenza could also stem from direct viral interference with SARS-CoV-2, for example through competition for cellular resources or interferon production [189, 190]. Infection rates with influenza reportedly changed according to SARS-CoV2 status and vice versa [190]. In this respect, we found high influenza decline with high COVID-19 incidence in group 1, and low influenza decline with low COVID-19 incidence in group 5, but also low levels for both in zero-COVID countries. Underreporting of COVID-19 cases may be an alternative explanation to low COVID-19 reporting in the low-income countries of group 5 [184, 185, 191].

The characterisation of influenza decline in space and time may come of use to analyse its resurgence over time. Loss of exposure to the influenza virus may lead to more severe waves or out of season waves [165, 166] and may increase the susceptible pool, especially in children [192, 193]. In the first half of 2022, influenza circulation was remarkably late in Europe [194] and early in the Southern hemisphere, with a peak above average in Australia [20]. Other epidemiological changes could occur regarding the exposed population and the seeding from the tropics [67] as global air transportation resumes. Deciding on the composition of the vaccine may also prove more difficult due to the change in the evolutionary dynamics of circulating strains [25]. Between March 2020 and August 2022 no B\Yamagata-lineage circulation was confirmed globally [25, 195].

Our study is affected by limitations. We assumed that influenza surveillance was not substantially altered during the pandemic period. The number of samples in the FluNet databases indeed did not change substantially over time, as many countries maintained influenza surveillance or quickly resumed it after initial disruption [5, 6]. Influenza positivity rate may have been affected by changes in surveillance protocols due to the COVID-19 pandemic. We did not account for influenza vaccination, due to limited information at the global scale. Vaccination rates are highly heterogeneous among countries [49]. While targeted recommendations increased coverage in the elderly during the 2020-21 season in 9 northern hemisphere countries and Australia [49], the efficacy of the influenza vaccine during the study period remains unknown. Lineages circulating in Southeast Asia during autumn 2020 were not included in the recommendations for the 2020-21 Northern Hemisphere season [14]. Last, we relied on the FluNet database, which integrates worldwide influenza records aggregating countries with highly diverse influenza surveillance quality and coverage. Results from the sensitivity analysis showed that the reported results were similar in varying exclusion criteria.

3.6 Supplementary Materials

3.6.1 Additional Methods

Definition of trimesters. The trimesters considered in the analyses are periods of 13 weeks, defined in such a way that the Dec-Mar trimester best covers the typical period of flu epidemics in northern countries. Northern countries are defined as countries with latitude above the Tropic of Cancer. The first and last day of the trimester (a window of 91 days) are identified based on FluNet data from 1995 to 2019, as follows: For each of the 365 possible starting dates, we computed the annual proportion of positive cases falling in the period, averaged over all northern countries and years. We then define the Dec-Mar trimester as the one that contains the highest annual proportion of positive cases. The Mar-Jun, Jun-Sep and Sep-Dec trimesters are identified accordingly. Also, we verified that the Jun-Sep trimester according to this definition roughly contains the highest proportion of positive cases for southern countries. The Dec-Mar, Mar-Jun, Jun-Sep and Sep-Dec trimesters obtained begin the first Monday following 12 December, 12 March, 11 June and 11 September, respectively. Certain years have 53 weeks instead of 52, thus trimesters may occasionally have 14 weeks.

Details on the computation of the log relative influenza level. Data reported on FluNet were partial or not consistent in some cases. Number of processed tests was sometimes different from the sum of positive and negative tests. In this case, the sum of positive and negative tests was used as the number of processed tests. When one of the three records (processed, positive and negative tests) was missing, this could be computed from the other two. The number of processed tests, when missing, was computed from the sum of positive and negative tests, the number of positive tests, when missing, was computed from the difference between processed and negative tests (provided the former was larger or equal to the later), and so on. We discarded weeks in which only processed and either positive or negative tests were present and the number of processed tests was smaller than the number of positive/negative tests. We also discarded weeks with only one record. Russia showed some irregularities with certain weeks having the number of processed tests nearly equal to the number of positive tests differently from the preceding or following weeks, signalling sudden changes in the data collection and sharing protocol. These weeks were removed from the analysis.

Before calculating the percentage of positive influenza tests, 0.5 positive cases are added to each trimester-country so that the positivity rate always results greater than zero. This allows distinguishing countries without influenza and with a massive surveillance system from countries without influenza but processing only a few tests.

When working with percentages - e.g. the percentage of influenza positive samples or the percentage of annual influenza samples falling in a certain trimester - the centre of the distribution was computed from the closure of the geometric mean, that was proved to be a BLU (best linear unbiased) estimator, unlike the standard arithmetic mean [164].

The log relative influenza level quantified the influenza positivity rate during the COVID-19 pandemic period relative to the average value of the pre-pandemic trimesters satisfying the inclusion criteria. The latter represents the expected value for the positivity rate if no secular trend was present before the pandemic. We tested for the existence of secular trends for the trimesters-countries included in the regression tree by fitting a linear regression to the positivity rates of pre-pandemic trimesters according to time. No significant trend was detected - smallest adjusted p-value 0.09 after Bonferroni correction for 215 tests.

Definition of the covariates included in the main analyses

- **age**: median age of population, UN projection for 2020.
- **longitude**: longitude of the centre of population of the country in degrees, from -180 (W) to 180 (E). Longitude of the country is computed as the average longitude of all the cities of the country with more than 300K inhabitants. The average is weighted for the population size of each city. If there are no cities in the country with at least 300K inhabitants, the longitude of the capital is considered.
- **latitude**: latitude of the centre of population of the country in degrees, from -90 (S) to 90 (N). The latitude is calculated analogously to the *longitude*.
- **T**: average temperature (in Celsius degrees) of the trimester-country. For each country, the temperature is computed as the average temperature of all the cities within the country with more than 300K inhabitants, weighted by the population size. If there are no cities in the country with at least 300K inhabitants, the capital is considered. Temperature data are taken from the ERA5 dataset, which provides hourly estimates of weather variables for all locations identified by a regular lat-lon grid of 0.25 degrees. The temperature of a city is calculated by looking at the closest grid point to the city and averaging the temperatures for the hours 0h00, 6h00, 12h00 and 18h00 of each day of the trimester.
- **RH**: average relative humidity, computed analogously to the *temperature*.
- **IDVI**: Score for the preparedness of a country in facing infectious diseases, from 0 (most vulnerable) to 1 (less vulnerable).
- **COVID-19 daily cases**: number of reported daily cases of COVID-19 per million of inhabitants averaged over the trimester.
- **workplace presence reduction**: median over the trimester of the daily percentage reduction of presence at workplaces.
- **reduction of international flights**: average percentage of reduction in the inbound and outbound air passengers of the country for each trimester with respect to the same trimester of 2019. The reduction for a trimester is calculated as the weighted average of the monthly reduction for the 4 months covering the trimester, with first and last months of the trimester, partially covered by the trimester, weighted 0.5, while the other months, fully covered by the trimester, weighted 1. The reduction for the month m and the year $y = \{2020, 2021\}$ is defined as $1 - w_{m,y}/w_{m,2019}$, with w being the number of passengers flying to or from the country.
- **nb days of school closure**: number of days over the trimester when policies related to schools and universities closure were implemented. The OxGRT dataset provides 2 daily variables: (i) the level of severity of the policy as measured on an ordinal scale (0=no measure, 1=altered openings for schools, 2=closing certain levels/categories of schools, 3=complete closure), and (ii) the geographical scope, i.e. whether that policy is enforced locally or nationally. Based on the values of these variables different definitions of school closure are possible - severity equal or above 1, 2, or 3, and each of these severity levels being implemented either locally or nationally. To choose the most convenient definition we used an

unsupervised approach. We first computed the number of days with school closure for each data point (trimester-country) for all possible definitions. We then computed the distribution of the number of days with school closure over all data points and picked the definition with maximum resolution power, i.e. that maximises the number of observations with values not falling in the extremes. We obtained that schools are considered to be closed if certain levels/categories of schools were closed on a national scale or if there was at least one complete closure on a local scale.

- **nb days of workplace closure**: The severity levels defined in the OxGRT dataset were: 0=no measures, 1=recommend closings, 2=require closing for some sectors/categories of workers, 3=require closing for all-but-essential workplaces. With the unsupervised procedure described for school closure we obtained that workplaces were defined as closed for stringency level at least 2 nationwide, or for stringency level 3 locally.

- **nb days of public event restrictions**: The severity levels defined in the OxGRT dataset were: 0=no measures, 1=recommend cancelling, 2=require cancelling. With the unsupervised procedure we obtained that public events were defined as closed when there was a countrywide enforcement.

- **nb days of gathering restrictions**: The severity levels defined in the OxGRT dataset were: 0=no restrictions, 1=restrictions above 1000 people, 2=restrictions between 101-1000 people, 3=restrictions between 11-100 people, 4=restrictions on gatherings of 10 people or less. With the unsupervised procedure we defined as gatherings restriction a nationwide ban of gatherings of more than 100 people.

- **nb days of public transport restrictions**: The severity levels defined in the OxGRT dataset were: 0=no measures, 1=recommend closing, 2=require closing. With the unsupervised procedure we obtained that public transports were defined as closed when a recommendation (level 1) was issued at local or national level.

- **nb days of stay at home requirements**: The severity levels defined in the OxGRT dataset were: 0=no measures, 1=recommend not leaving house, 2= require not leaving house with exceptions for 'essential' trips, 3=require not leaving house with minimal exceptions. With the unsupervised procedure described for school closure we obtained that staying at home was implemented for severity level 1 or more, locally or nationally.

- **nb days of international travel restrictions**: The severity levels defined in the OxGRT dataset were: 0=no restrictions, 1=screening arrivals, 2=quarantine arrivals from some or all regions, 3=ban arrivals from some regions, 4=ban on all regions or total border closure. With the unsupervised procedure we obtained that international travels were defined as enacted for severity level 3 or 4.

- **nb days of facial covering requirements**: The severity levels defined in the OxGRT dataset were: 0=no policy, 1=recommended, 2=required in some specified shared/public spaces with other people present, 3=required in all shared/public spaces with other people present, 4=required at all times regardless of location or presence of other people. With the unsupervised procedure we obtained that mask use was implemented for severity level 3 or 4.

- **nb days of testing implementation**: The severity levels defined in the OxGRT dataset were: 0=no testing policy, 1=only those who both (a) have symptoms AND (b) meet specific criteria, 2=testing of anyone showing COVID-19 symptoms, 3=open public testing. With the unsupervised procedure we obtained that testing policies were defined as implemented for stringency level 3.
- **nb days of contact tracing implementation**: The severity levels defined in the OxGRT dataset were: 0=no contact tracing, 1=not for all cases, 2=contact tracing for all identified cases. With the unsupervised procedure we obtained that contact tracing was defined as implemented for severity level 2.
- **nb days of elderly shielding**: The severity levels defined in the OxGRT dataset were: 0=no measures, 1=recommended isolation, hygiene, and visitor restriction measures in LTCFs and/or elderly people to stay at home, 2=narrow restrictions for isolation, hygiene in LTCFs, some limitations on external visitors and/or restrictions protecting elderly people at home, 3=extensive restrictions for isolation and hygiene in LTCFs, all non-essential external visitors prohibited, and/or all elderly people required to stay at home and not leave the home with minimal exceptions, and receive no external visitors. With the unsupervised procedure described for school closure we obtained that protection of elderly people was defined as implemented when it was enforced at least at level 2 locally or nationally.

Definition of the covariates included in the sensitivity analyses:

- **COVID-19 daily deaths**: number of reported daily deaths of COVID-19 per million of inhabitants averaged over the trimester.
- **station presence reduction**: median over the trimester of the daily percentage reduction of presence in public transport stations and transportation hubs.
- **recreation place presence reduction**: median over the trimester of the daily percentage reduction of presence at restaurants, bars, shopping malls and other recreation places.
- **home presence rise**: median over the trimester of the daily percentage rise of presence in residential places.
- **stringency index**: average of the daily stringency index provided by OxCGRT. This index combines eight indicators of containment and closure policies and an indicator regarding the presence of public information campaigns related to the pandemic. The daily index ranges from 0 for countries with no measures, to 100 for countries adopting maximally stringent policies regarding all nine indicators. Seven of the eleven indicators considered in the previous covariates are included in this index.

Covariate distributions. We provide in Figure 3.5 the distributions of the log relative influenza level and the covariates across trimesters-countries.

Algorithms for the regression analysis

Clustering and regression trees. We relied on the CART algorithm [19] to classify trimesters-countries based on a target variable, here the log relative influenza level. In a nutshell, observations are iteratively splitted into two groups through a binary partition over a

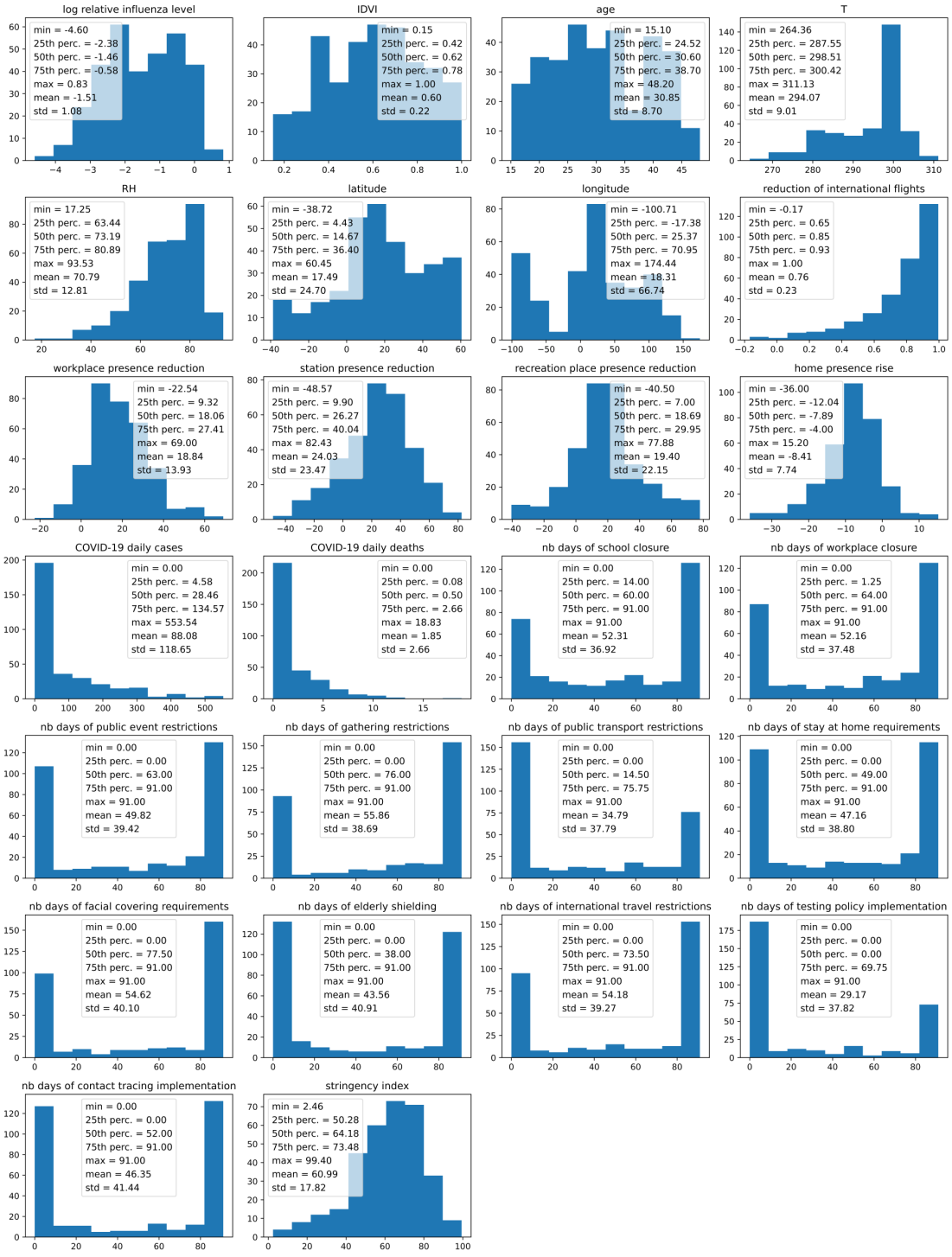


FIGURE 3.5: **Distributions of the 26 variables considered in the regression analysis, for all the 330 observations included in the study.** For each plot, summary values of the distributions are shown in the legend. We included here both the covariates considered in the main analysis and the covariates considered in the sensitivity analysis.

covariate that is selected at each iteration in such a way that the intra-group variance of the target variable is minimised.

The procedure leads to building the maximal tree. Then, it is pruned in order to avoid overfitting. In particular, we controlled the structure of the tree through two parameters tunable on the R package [196]: *minbucket* is the minimum number of observations for each terminal node, *cp* is a regularisation parameter that penalises the increase in the number of terminal nodes - through the addition of a linear term in the error function. The two parameters are optimised by cross-validation.

Cross-Validation for the hyperparameter tuning of the regression tree. The search for the optimal values of *minbucket* and *cp* is run over the following grid of parameters:

$$\begin{cases} minbucket \in \{4 \leq m \leq 18; m \in \mathbb{N}\} \\ cp \in \{0.001 * c; 0 \leq c \leq 20; c \in \mathbb{N}\} \end{cases}$$

For each point of the parameter space, 2000 trees are generated. Each tree is created on a random sample of 70% of the data and its prediction error (1-coefficient of determination) is calculated on the remaining 30% of the observations. Following the Breiman's rule, the optimal tree is identified as the simplest tree that has mean prediction error less than the minimum error increased by its standard deviation. The simplest tree is identified by looking at the smallest number of splits on average, the largest *minbucket*, and the largest *cp*, in order. The optimal parameters identified were $cp = 0.011, minbucket = 4$.

Variable selection through permutation risk measures for covariate importance. The Variable Selection Using Random Forests (VSURF) algorithm [133] has been exploited to identify the predictors associated with the reduction of influenza. This method evaluates the importance of each variable by measuring the prediction error increase when values of one variable at a time are permuted. This is a classical method used in the framework of Random Forests and, more in general, in machine learning algorithms. Also, some studies pointed out that permutation risk measures of variable importance are often more effective than alternative methods based on Sobol's indices or Shapley values [5]. The VSURF algorithm was run using the following parameters: *ntrees=8000, nfor.thres=100, nfor.interpr=100, nfor.pred=100* (and *mtry=6* by default).

3.6.2 Additional results

Countries included in the analysis

There were 166 countries that contributed to FluNet during the period from 15 Dec 2014 to 12 Sep 2021. All FluNet records for these 166 countries were included in Figure 3.1A. Upon filtering based on the quality and extent of the FluNet records, 112 countries were included in the descriptive study (Figure 3.1B and Figure 3.2). For those countries, only the trimesters satisfying the inclusion criteria were included. Among the 112 countries, covariates were available only for 93 countries. This last group of countries was included in the regression analysis. The list of countries discarded at each step and included among the 93 countries is reported in Table 3.5.

Regression tree

Additional details of the 5-group classification. We provide in the following additional details on the 5-group repartition presented in Figure 4 of the main paper: the box

	countries in Fig. 1A (166)	countries in Fig. 1B and Fig. 2 (112)	countries for the regression analyses (93)
Central and South America	AIA, ATG, ABW, BHS, BRB, BLZ, VGB, CYM, CUB, DMA, DOM, GUF, GRD, GLP, MTQ, KNA, LCA, VCT, SUR, TTO, TCA, VEN		ARG, BOL, BRA, CHL, COL, CRI, ECU, SLV, GTM, GUY, HTI, HND, JAM, MEX, NIC, PAN, PRY, PER, URY
North America and Europe	BEL, BMU, CZE, GRC, MLT, MNE, NLD, SVK, CHE	ALB, ISL, OWID_KOS, MKD	AUT, BLR, BIH, BGR, CAN, HRV, DNK, EST, FIN, FRA, DEU, HUN, IRL, ITA, LVA, LTU, LUX, NOR, POL, PRT, MDA, ROU, RUS, SRB, SVN, ESP, SWE, UKR, GBR, USA
Africa	BFA, CAF, TCD, MRT, MAR, RWA, SYC, SLE, SSD, TUN, ZWE	DZA, COD, ETH, GIN, MDG	AGO, CPV, CMR, COG, CIV, EGY, GMB, GHA, GNB, KEN, MLI, MUS, MOZ, NAM, NER, NGA, SEN, ZAF, SDN, TGO, UGA, TZA, ZMB
Western, Central and South Asia	BHR, CYP, KWT, MMR, SYR, TJK, TKM, ARE, UZB, YEM	ARM, AZE, BTN, IRN, MDV, TLS, PSE	AFG, BGD, KHM, GEO, IND, IDN, IRQ, ISR, JOR, KAZ, KGZ, LAO, LBN, MYS, NPL, OMN, PAK, PHL, QAT, SAU, SGP, LKA, THA, TUR, VNM
Easthern Asia and Oceania	FJI, PNG	CHN, PRK, NCL	AUS, JPN, MNG, NZL, KOR

TABLE 3.2: **Scheme of countries included in the different steps of the study.** Countries are indicated with their 3-letter code, OWID_KOS is for Kosovo. Countries are grouped into five regions, aggregating different influenza transmission zones [172]: Central and South America (Temperate South America, Tropical South America and Central America and Caribbean), North America and Europe (North America, Northern Europe, South West Europe and Eastern Europe), Africa (Northern Africa, Western Africa, Middle Africa, Eastern Africa, Southern Africa), Western, Southern and Central Asia (Western Asia, Southern Asia, South-East Asia, Central Asia), Eastern Asia and Oceania (Eastern Asia, Oceania Melanesia Polynesia).

plot of covariate values for observations in each group (Figure 3.6), and the list of trimesters-countries belonging to each group (Table 3.3).

Full Regression Tree. The regression tree selected using the the algorithm had 14 terminal leaves and a coefficient of determination $R^2 = 0.69$. The leaves identified by the model were well-defined (Figure 3.7), i.e. distinct from each other and characterised by homogeneous values of log relative influenza level - only for two of them the interquartile width of the observed log relative influenza level was greater than unity.

The tree is shown in Figure 3.8. The first four splits are done according to IDVI, COVID-19 daily cases, longitude and workplace presence reduction as discussed in the main paper. The other variables are used for a finer partition in smaller groups. The classification of trimesters-countries in leaves is reported in the supplementary data [197].

Group 1 (109 observations) is split into leaves 1 and 2. In leaf 1, lower temperatures relative to leaf 2 are associated with a greater reduction in influenza. Leaf 1 consists largely of temperate countries in Europe, North and South America, during the 2020-2021 influenza season, which had a greater influenza reduction. Leaf 2 includes a more limited number of observations (26, compared with the 83 in leaf 1) from countries of tropical and subtropical areas with IDVI>0.54, e.g. Panama, Costa Rica, Colombia, Malaysia. Influenza reduction for these countries was less strong compared with leaf 1, but still substantial if compared with low-IDVI tropical countries, classified in group 5.

Group 2 is formed by a single leaf with well defined properties detailed in the main paper.

The 45 observations of the group 3 are distributed in four leaves including a few trimesters-countries each. Similarly to the split within group 1, a first split based on temperature separates countries with higher temperature and higher log relative influenza level (Saudi Arabia and Qatar) from countries with lower temperature and

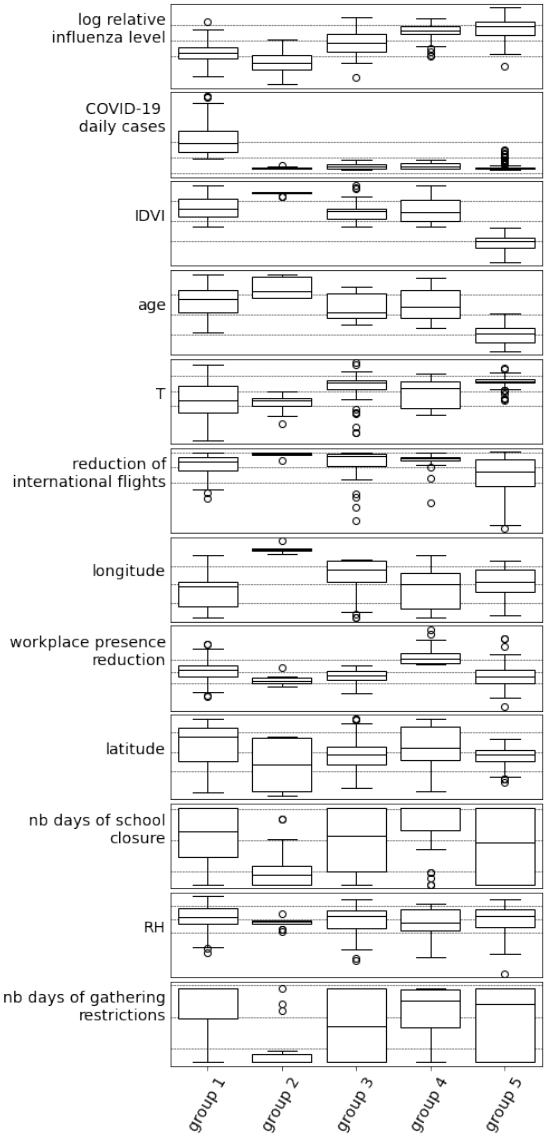


FIGURE 3.6: **Variable distributions for the five-group partitioning by means of the regression tree.** For each variable the boxplot shows the distribution in the group. Horizontal lines show median and quartiles of the whole dataset for comparison.

	Spring 2020	Summer 2020	Autumn 2020	Winter 2020-21	Spring 2021	Summer 2021
Group 1: leaves 1,2						
Central and South America	CHL	ARG, BRA, CHL, COL, CRI, PAN	ARG, CHL, COL, CRI	BRA, COL, CRI, MEX	ARG, BRA, CHL, COL, CRI, ECU, PAN, PER, PRY	ARG, BRA, CHL, COL, CRI, PAN, PRY, URY
North America and Europe			CAN, GBR, IRL, PRT, USA	AUT, BGR, BLR, CAN, DEU, DNK, ESP, EST, FRA, GBR, HRV, HUN, IRL, ITA, LUX, LVA, MDA, NOR, POL, PRT, ROU, RUS, SRB, SVN SWE, UKR, USA	AUT, BLR, CAN, DEU, DNK, EST, HRV, IRL, LTU, LUX, LVA, MDA, NOR, POL, ROU, SVN, SWE, UKR, USA	
Africa		ZAF			GEO, JOR, LBN, MYS, OMN, QAT, TUR	ZAF
Western, Southern and Central Asia	QAT, SGP	QAT	ISR, JOR, LBN, OMN	GEO, ISR, JOR, LBN, QAT, TUR		LKA, MYS, PHL, THA
Eastern Asia and Oceania						
Group 2: leaf 3						
Central and South America						
North America and Europe						
Africa						
Western, Southern and Central Asia						
Eastern Asia and Oceania	AUS, JPN	AUS, NZL	AUS, JPN, KOR	AUS, JPN, KOR	AUS, JPN, KOR	AUS
Group 3: leaves 4,5,6,7						
Central and South America	BRA, PRY	PRY	MEX		MEX	ECU, PER, SLV
North America and Europe	SWE		NOR	FIN	RUS	
Africa		MUS	ZAF		ZAF	
Western, Southern and Central Asia	THA, VNM	IDN, LKA, MYS, THA, VNM	KAZ, MYS, QAT, SAU, SGP, THA, VNM	KAZ, LKA, SAU, SGP, THA, VNM	IDN, SAU, SGP, THA, VNM	QAT, SAU, SGP
Eastern Asia and Oceania	MNG			MNG		
Group 4: leaves 8,9						
Central and South America	ARG, COL, CRI, MEX, PAN, PER, SLV	SLV		ECU		
North America and Europe	AUT, CAN, DEU, DNK, EST, GBR, IRL, LVA, NOR, POL, ROU, RUS, SVN, UKR, USA				GBR	
Africa	MUS, ZAF					
Western, Southern and Central Asia	IDN, LKA, MYS, OMN, SAU	SGP	LKA	IDN, KGZ, OMN	LKA, PHL	
Eastern Asia and Oceania						
Group 5: leaves 10,11,12,13,14						
Central and South America	BOL, GTM, HND, HTI, JAM	HND, HTI, NIC	BOL, HND, HTI, NIC	GTM, HND, HTI	BOL, GTM, HND, HTI, JAM	GTM, HND, HTI, JAM, NIC
North America and Europe						
Africa	CIV, CMR, EGY, MLI, MOZ, TZA, ZMB	CIV, EGY, KEN, MLI, SEN, TZA, UGA, ZMB	CIV, CMR, EGY, GHA, KEN, NER, SEN, TGO, TZA, UGA, ZMB	CIV, CMR, EGY, GHA, KEN, NER, NGA, SEN, TGO, TZA, ZMB	CIV, CMR, EGY, GHA, KEN, MLI, NGA, SEN, TGO, TZA, UGA, ZMB	CIV, CMR, EGY, GHA, KEN, MLI, NGA, SEN, TGO, TZA, UGA, ZMB
Western, Southern and Central Asia	AFG, BGD, IND, KHM, LAO, NPL	BGD, KHM, LAO, NPL	AFG, BGD, IND, IRQ, KHM, LAO, NPL, PAK	AFG, IND, IRQ, KHM, LAO, NPL, PAK	AFG, BGD, IND, KHM, LAO, NPL, PAK	BGD, IND, KHM, LAO, NPL
Eastern Asia and Oceania						

TABLE 3.3: **Classification of trimesters-countries according to the high-level partitioning in five groups by means of the regression tree.** Countries are grouped into five regions, aggregating different influenza transmission zones [172]: Central and South America (Temperate South America, Tropical South America and Central America and Caribbean), North America and Europe (North America, Northern Europe, South West Europe and Eastern Europe), Africa (Northern Africa, Western Africa, Middle Africa, Eastern Africa, Southern Africa), Western, Southern and Central Asia (Western Asia, Southern Asia, South-East Asia, Central Asia), Eastern Asia and Oceania (Eastern Asia, Oceania Melanesia Polynesia).

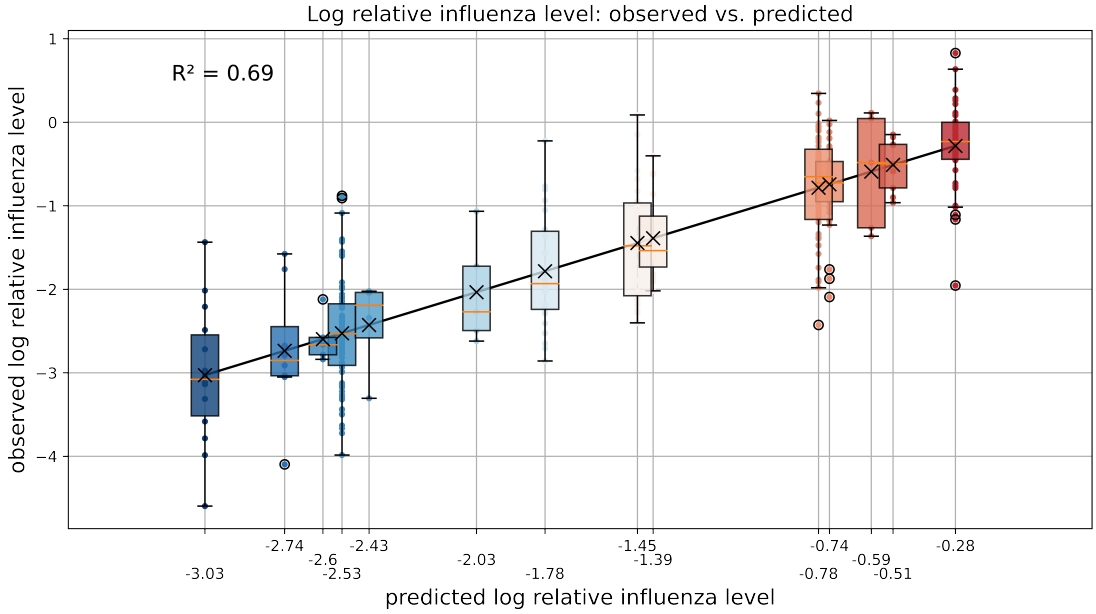


FIGURE 3.7: **Goodness of fit of the regression tree.** The 14 boxplots display the distributions of the log relative influenza level for trimesters-countries of the 14 leaves. The black crosses identify the mean values of the distributions (also shown through the color scale) that correspond to the predicted log relative influenza level. R^2 is the coefficient of determination.

lower log relative influenza level. This second branch splits based on the reduction of international flights. On the left side of the split there is a group of 8 trimesters-countries where low values of log relative influenza level were associated with lower-than-average reduction of international flights, reduction of workplace presence and number of days with gathering restrictions. This group was characterised by a number of days of school closure higher than average. The right side of the split has two leaves that are discussed in the main paper, i.e. leaf 5 including mainly Singapore and leaf 6 including mainly other Southeast Asia countries.

Group 4 consists of 39 observations. This includes mainly countries during Mar-Jun 2020 that are grouped in leaf 9 (34 out of 39 observations). Five observations are separated by the other because they have a more limited number of days with school closure and a lower log relative influenza level. This is a heterogeneous set of countries, mainly between Dec-Mar 2021 and Mar-Jun 2021.

Group 5 contains a significant proportion of all observations (123 out of 330) that are separated into five leaves (leaves 10, 11, 12, 13 and 14). Interestingly, the five leaves show a clear trend with increasing log relative influenza level that is, in general, associated with a decrease in four of the five COVID-19 response variables - COVID-19 daily cases, reduction of international flights, reduction of workplace presence, and number of days with school closure. Limitations on gatherings remain moderate for all five leaves. Two splits are based on the reduction of international flights, between leaf 10 and leaves 11 and 12, and between leaf 13 and 14. For these two splits a greater reduction of international flights is associated with a lower influenza log ratio. Finally, leaves 11 and 12 differ in relative humidity. The two leaves contain almost the same set of countries for different trimesters - e.g. Guatemala, Honduras, India, Nepal, and Zambia. These are tropical countries characterised by a dry and a rainy season throughout the year, where influenza usually peaks twice a year with the main peak during the rainy season [7, 64, 166, 167]. Our analysis shows that for rainy seasons the reduction of influenza was smaller.

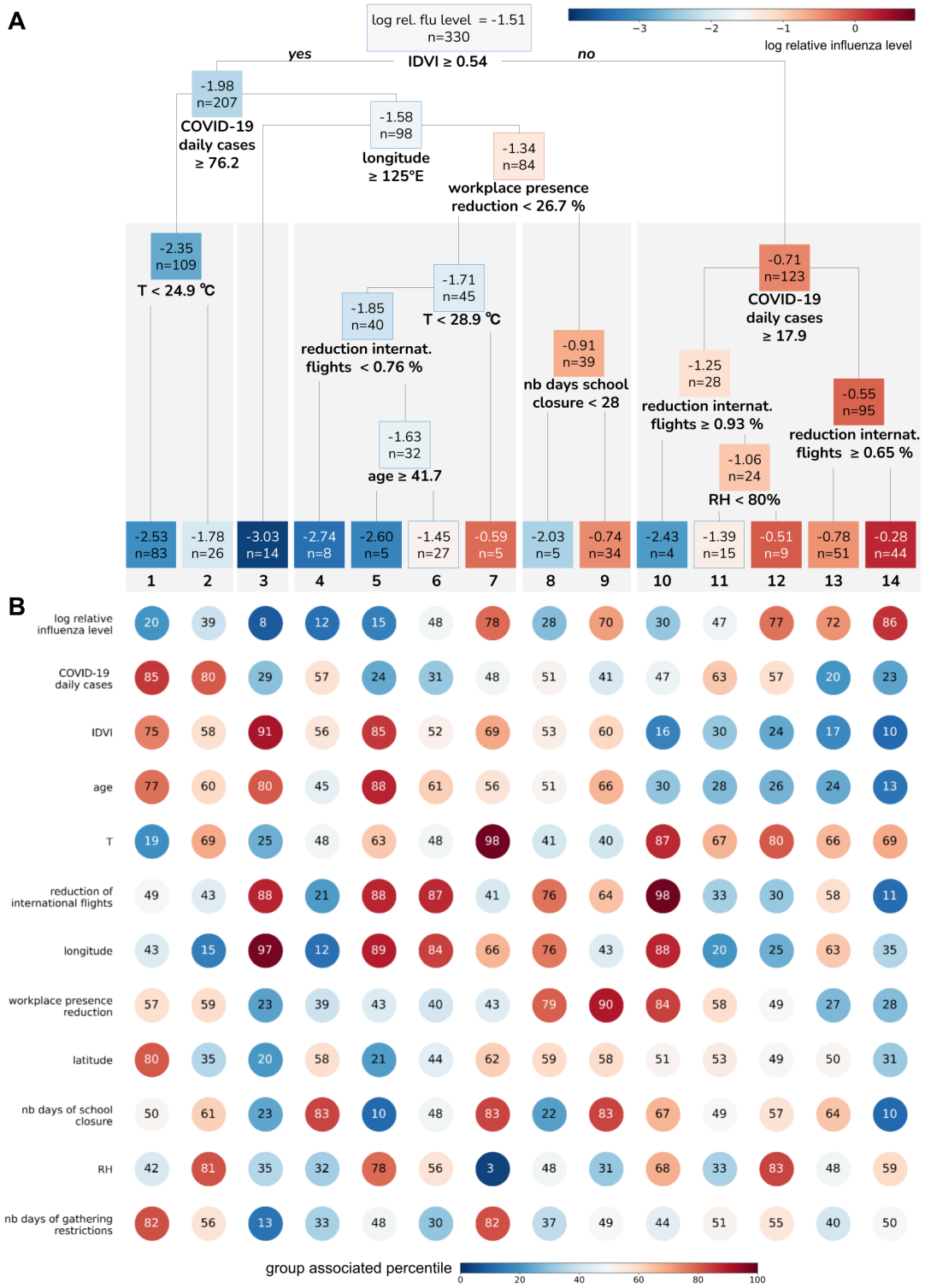


FIGURE 3.8: **Regression tree. A** Regression tree obtained with the variables selected in Figure 3. For each node the average log relative influenza level and the number of observations are reported (the former is also indicated with a colour scale). **B** Properties of each leaf. For each covariate the percentile of the whole dataset distribution the median of the group corresponds to is indicated with the colour scale and reported in the bubble.

3.6.3 Robustness checks and sensitivity analyses

Robustness of the variable selection procedure. The variable selection procedure is a stochastic algorithm that may lead to different results when repeated. Therefore, variable importance was estimated by averaging over 100 stochastic realisations. To be sure results were stable we repeated the procedure 20 times. Each time the same 11 predictors are selected as the important covariates for predicting the log relative influenza ratios.

Robustness of the tree structure optimization. The regression tree was regularised by two parameters (*minbucket* and *cp*) optimised with cross-validation through the stochastic procedure discussed in the Additional methods section of the supplementary material. To be sure that the procedure converges to a stable result we repeated it 10 times. The regularisation parameters selected each time were very similar and led to the construction of trees almost identical to the tree described in the Results section and in the supplementary material. In particular, the classification of trimesters-countries in the five high-level groups was robust.

Robustness of the tree under small perturbations of the dataset. We assessed the robustness of the tree under small perturbations in the dataset. We built ten regression trees on a random subsample of 314 observations (~95% of the total) keeping the same 11 predictors and hyperparameters. The ten resulting trees (i.e. the perturbed trees) were compared with the tree built from the entire dataset (i.e. the reference tree) using the Adjusted Rand Index (ARI) [168]. Specifically, we compared the classification in 5 groups to assess the robustness of the 5-group repartition discussed in the main paper. The average score value for the ten comparisons is 0.86, indicating good agreement between the perturbed and reference trees.

Sensitivity of the variable selection under changes on the assumption made

We tested whether the predictors of the log relative influenza level changed with the choices made throughout the analysis. The creation of the dataset of observations is based on two main assumptions: (i) *k=0.5* positive cases had been added to each trimester-country in order to remove zero counts of influenza cases, and (ii) a threshold *s=130* for the minimum number of tests processed per quarter was set to discard trimesters-countries with poor data. In addition, some choices were made when defining the covariates. We tested the robustness of our results to all these choices by analysing the following alternative models (the baseline model is here referred as *base 0* model):

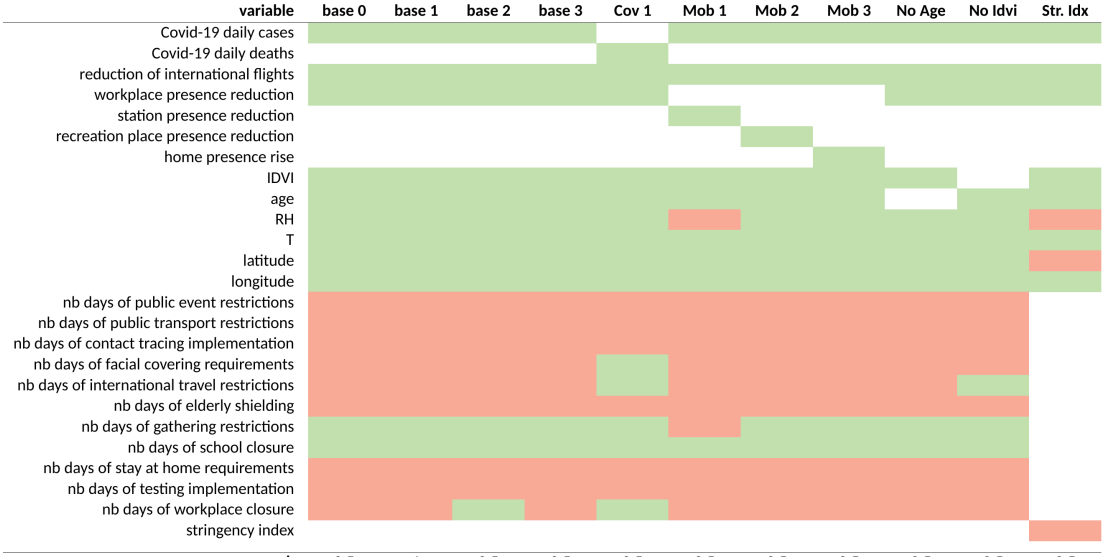
- *base 1*: k=1 (instead of 0.5), s=130, same covariates of the *base 0* model;
- *base 2*: k=0.5, s=26 (instead of 130), same covariates of the *base 0* model;
- *base 3*: k=0.5, s=260 (instead of 130), same covariates of the *base 0* model;
- *Cov 1*: k=0.5, s=130, COVID-19 daily deaths is used in alternative to COVID-19 daily cases to quantify the intensity of the COVID-19 epidemic;
- *Mob 1:* k=0.5, s=130, the public transport station presence reduction is tested in alternative to workplace presence reduction to capture changes of social activity;
- *Mob 2:* k=0.5, s=130, the recreation place presence reduction is tested in alternative to workplace presence reduction to capture changes of social activity;

- *Mob 3:* k=0.5, s=130, the increase in home presence is tested in alternative to workplace presence reduction to capture changes of social activity;
- *No Age:* k=0.5, s=130, the variables age is removed among the set of covariates to be included in the regression;
- *No IDVI:* k=0.5, s=130, the variables IDVI is removed among the set of covariates to be included in the regression;
- *Str. Idx:* k=0.5, s=130, all variables associated with NPIs are replaced by the stringency index.

For all the alternative models the selected sets of important factors are highly similar (Table 3.4): impact of COVID-19, international mobility, workplace presence reduction (or the alternative proxy of social activity considered), IDVI, age, temperature and longitude always result significant, while latitude and RH are discarded only once and twice respectively. All proxies of social activity tested were classified as important. The stringency index when included was not selected, indicating that the aggregate information it carries is not important in explaining the influenza reduction. This is consistent with the fact that only 2 out of the 11 governamental response variables were selected as important.

We used the ARI similarity index to compare the sensitivity trees and the baseline tree up to the five-group repartitions. Models Mob 1, Mob 2, Mob 3, No Age, Str. Idx. led to an excellent recovery of the baseline tree (ARI > 0.9). The 5-group repartition showed the same behaviour as in the baseline model. The tree obtained with Base 1 was in good agreement with the baseline tree. Interestingly, removing IDVI from the set of covariates led to only a moderate recovery of the baseline tree (ARI= 0.69), despite IDVI and Age being highly correlated. Age is not able to fully compensate for IDVI in creating the 5 group repartition discussed in the main paper. This is consistent with the fact that both covariates were found to be important by the VSURF procedure.

The comparison of trees obtained with base 2, base 3 and Cov 1 and the baseline tree up to the five-groups repartition led to lower similarity values (ARI between 0.53 and 0.66). Still, we found that the four trees shared in large part the same behaviour. In particular, the first split was the same (i.e. based on IDVI with the same threshold value), meaning that all partitions had the distinction between high IDVI, low influenza countries and low IDVI, high influenza ones. Also, the 2020 Mar-Jun trimesters of temperate countries (comprising the bulk of group 4) remained in great part grouped together and separated from the rest. In all three sensitivity trees, countries of group 2 of the baseline tree (zero-covid countries) were grouped together and included in a larger group together with countries of group 1 of the baseline tree.



variable	base 0	base 1	base 2	base 3	Cov 1	Mob 1	Mob 2	Mob 3	No Age	No Idvi	Str. Idx
Covid-19 daily cases											
Covid-19 daily deaths											
reduction of international flights											
workplace presence reduction											
station presence reduction											
recreation place presence reduction											
home presence rise											
IDVI											
age											
RH											
T											
latitude											
longitude											
nb days of public event restrictions											
nb days of public transport restrictions											
nb days of contact tracing implementation											
nb days of facial covering requirements											
nb days of international travel restrictions											
nb days of elderly shielding											
nb days of gathering restrictions											
nb days of school closure											
nb days of stay at home requirements											
nb days of testing implementation											
nb days of workplace closure											
stringency index											
k	0.5	1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
s	130	130	26	260	130	130	130	130	130	130	130
nb observations	330	335	390	279	321	330	330	330	330	330	330
nb tested predictors	20	20	20	20	20	20	20	20	19	19	10
nb selected predictors	11	11	12	11	14	9	11	11	10	11	7
Similarity index (ARI)	–	0.82	0.61	0.53	0.66	0.94	0.95	0.92	1.00	0.69	1.00

TABLE 3.4: **Scheme of predictors selected for 11 alternative models.** Each model includes only variables associated with a coloured cell, green is for the selected variables, red for the rejected ones. Additional information about the parameters *k,s* used for the definition of the observation set is provided. Also, the similarity index is reported: it measures the similarity of the five-group classifications made by each alternative model compared with the reference model *(base 0)*.

Chapter 4

Understanding the coupled dynamics of influenza (sub)types: a global analysis leveraging Compositional Data Analysis

This Chapter is based on the study entitled *Understanding the coupled dynamics of influenza (sub)types: a global analysis leveraging Compositional Data Analyis* [31]. It has been conducted under the supervision of my PhD advisor Chiara Poletto (Sorbonne Université, University of Padova), with the collaboration of Pierre-Yves Böelle (Sorbonne Université) and Vittoria Colizza (Sorbonne Université, Georgetown University). It is currently under review by the co-authors and will be submitted in the coming weeks.

The code for the reproducibility of the analysis is available at https://github.com/FrancescoBonacina/coupled-dynamics-flu-subtypes.

4.1 Abstract

Background & aims of study. (Sub)type composition of seasonal influenza waves varies in space and time. Different (sub)types tend to have different impacts on different population groups, therefore understanding the drivers of (sub)types' co-circulation and anticipating (sub)type composition is important for epidemic preparedness and response. In this study, we propose the application of Compositional Data Analysis (CoDA) to quantitatively analyze the proportions of flu (sub)types.
Methods & results. Influenza (sub)type compositions – i.e. vectors of frequencies of A\H1N1, A\H3N2, and B infections – sum to one and therefore are not defined in a metric space. CoDa is widely used in geology and ecology to treat this kind of data. In accordance with CoDA's precepts, we can apply an isometric log-ratio transformation to map compositions into points of a metric space, thus opening the path to statistical analyses of compositions' trajectories. From FluNet, we reconstructed (sub)type compositions by country and year from 2000 to 2022 and analyzed them in the CoDA framework. First, we looked at global ecological trends by comparing annual statistics of (sub)type mixing by country. Distributions were similar across years except for atypical years of extraordinarily low mixing occurring in correspondence with a new clade emergence (2003/2004 season), A1N109pdm emergence, and the COVID-19 pandemic. Second, we addressed the geographical structuring by clustering the trajectories of annual countries' compositions. We identified two macroregions with synchronous (sub)type alternation and constantly strong (sub)type mixing, respectively. Finally, we probed the potential of the CoDA framework for forecasting the annual time series of

countries' compositions. Among the tested models, the Bayesian Hierarchical Vector AutoRegressive model was the best performing, improving the predictions obtained with naïve approaches.
Implications. CoDA allowed identifying meaningful patterns in the spatiotemporal dynamics of influenza (sub)types and showed its potential for the forecast of (sub)types compositions. The statistical and visualization tools presented here provide a synthesis of surveillance data that could enable novel hypotheses on influenza drivers. A similar technique could be applied to different spatiotemporal scales or any epidemiological data in the form of percentages.

4.2 Introduction

Since 2009 the H1N1pdm and H3N2 subtypes of influenza type A and influenza type B co-circulate in the human population [40, 198]. The influenza viral diversity profoundly impacts the epidemiological characteristics of influenza epidemics [8, 9, 199]. Due to the mechanism of immune imprinting [44, 192, 200], the H3 hits more severely the elderly [11, 13, 53]. As a consequence of this, combined with the higher transmissibility of H3 [59], influenza waves dominated by this subtype are often more severe. On the other hand, influenza A/H1N1 and B cause a higher burden among the younger population [13, 53, 201]. In the northern temperate regions, the peak of B infections typically occurs a few weeks after the peak of A infections [54, 65]. These examples show that anticipating the (sub)type composition of approaching seasons is key to improving our preparedness for seasonal waves, e.g. allocating hospital capacities and optimizing vaccine distribution among age groups.

Still anticipating the (sub)type co-circulation is complicated by the fact that flu viruses interact with each other. Evidence of viral interference has emerged from experiments with ferret models [56, 58] and from population-level epidemiological analyses [54, 59, 60, 76, 199, 202]. In particular, past studies have shown that cross-immunity is an important ingredient of models aiming at reproducing plausible influenza dynamics [203, 204]. In other words, Influenza (sub)types form a coupled ecological system that needs to be studied as a whole. A second source of complication is represented by the fact that influenza rapidly spreads globally and viral compositions in different countries are interdependent [15]. This makes the study of worldwide influenza circulation essential for interpreting the viral patterns observed at the country scale.

In response to these needs, the Global Influenza Surveillance and Response System (GISRS) [17] gathers and makes available through the FluNet portal a weekly number of samples by (sub)types and country. The quality and quantity of the data is constantly increasing. Yet, surveillance systems are not standardized across countries, therefore counts of infections cannot be compared from one country to another. Percentages of infections by (sub)type are more robust to biases and amenable to cross-country comparison. They also more directly describe the patterns of dominance/codominance among (sub)types. Previous works analyzed these data with descriptive statistics, e.g. computing the minimum/maximum percentages of infections by (sub)type or the number of seasons in which each (sub)type was predominant [51, 52, 54, 65, 98]. This has led to important findings. It has provided evidence for the alternation of A/H1N1pdm and A/H3N2 in temperate regions [98], has pointed to the predominance of A/H3N2 among all (sub)types [54], and has addressed the altered (sub)type circulation following the emergence of the A/H1N1pdm (sub)type [76] and the COVID-19 pandemic [51]. However, further developments in this direction are complicated by

the fact that treating percentage data is not easy. More sophisticated quantitative analyses cannot be used without defining a proper metric space.

Here, we investigated the coupled dynamics of (sub)types by analyzing their relative abundances across countries and years through the Compositional Data Analysis (CoDA) framework [137, 140]. This approach is used in ecology and geology to map percentage data into a metric space thus easing quantitative analyses. Through CODA we defined for each country a trajectory in the (sub)type composition space. We quantified how these trajectories evolved in time and their spatial structure. We then proposed an approach able to leverage this structure to forecast (sub)type relative abundance in each country one year ahead with improved accuracy compared with naive estimators.

4.3 Results

4.3.1 A Compositional Data Analysis framework for studying the relative abundances of flu (sub)types

We study the relative abundance of flu (sub)types for different countries-years, defined by vectors of percentages of the form (B%, H1%, H3%) - for brevity, we will now on use H1 for the A/H1N1 strains (both the historical and the pandemic ones), and H3 for A/H3N2. We consider weekly surveillance data reported in FluNet [17, 18] from 2000 to 2023, for up to 151 countries. Data are aggregated annually - year beginning at the end of April - to define country-year vectors, as depicted in Figure 4.1A (for details see the Methods). In mathematical terms, these vectors are multivariate observations that live in the 3-part simplex S^3 (Figure 4.1B). In the jargon of Compositional Data Analysis [140], they are called *compositions*, i.e. vectors of positive components that sum to a constant. The study of this type of data involves several problems well-known in compositional statistics. For example, the fact that components are inherently non-independent makes it difficult to assess correlations between compositions [138, 139, 140]. Also, since compositions live in a bounded space, a variation of one unit is all the more important the closer one gets to the edges [137]. Therefore, the statistical tools developed for unconstrained data cannot be used for studying compositions [137]. In the 1980s, Aitchison developed a formal study of compositions [28, 140]. In particular, he proposed log-ratio transformations to map constrained data into new unconstrained data, thus making them suitable for analysis with standard statistics. This method is commonly applied in fields such as ecology, geology, and environmental science [142, 205, 206, 207, 208], where it is common to work with data in the form of proportions.

Here, we apply the isometric log-ratio (*ilr*) transformation proposed by [144]:

$$\begin{cases} u &= \sqrt{\frac{2}{3}} \ln \frac{B\%}{\sqrt{H1\% * H3\%}} \\ v &= \sqrt{\frac{1}{2}} \ln \frac{H1\%}{H3\%} \end{cases} \tag{4.1}$$

The vectors (B%, H1%, H3%)$\in S^3$ were then mapped into vectors $(u, v) \in \mathbb{R}^2$, where u denotes the relative abundance between influenza B and the average proportion of influenza A subtypes, while v denotes the relative amount of H1 vs. H3. The relative abundance of (sub)types for 132 countries in 2017 is depicted in the (u, v) coordinates in Figure 4.1C. The metric space thus defined allows us to compute statistics on (u, v) points, but also to define trajectories of points to follow the trend of (sub)types in a given country over time (Figure 4.1D). We can then identify countries with similar trajectories and compute statistics on the ensemble of trajectories. Together with the

continuous representation of the relative abundances of (sub)types, we can introduce discrete states corresponding to (sub)type dominance or co-dominance. We considered one (sub)type to be dominant if it accounted for at least 50% of the samples, otherwise, the three (sub)types were co-dominant. These states correspond to the four regions of the simplex, or equivalently, of the Euclidean space identified by the dotted lines in Figures 4.1B and 4.1C.

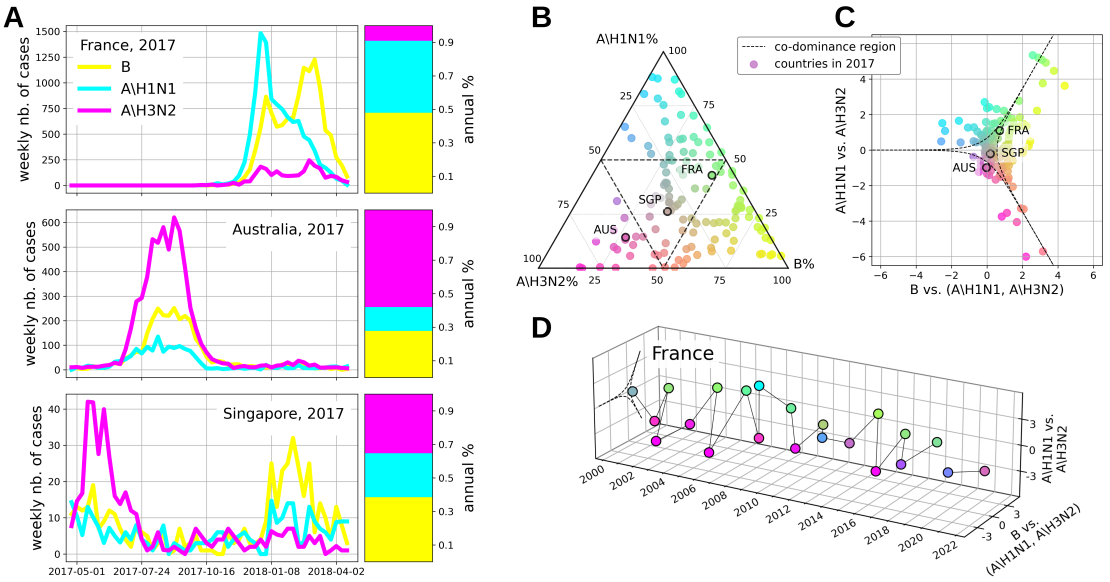


FIGURE 4.1: **Relative abundances of influenza (sub)types H1, H3, and B. A) Incidence by (sub)types for France, Australia, and Singapore.** Incidence curves from April 2017 to April 2018 are used to estimate the relative abundances of the (sub)types in 2017. **B) Proportions of H1, H3, and B represented in the simplex.** Observations for 132 countries in 2017 are shown in the 3-part simplex, each point corresponds to one of the countries. The dotted line identifies the co-dominance region at the center, while the other three portions of the simplex correspond to the dominance of a (sub)type. The colors of the points are associated with their position in the simplex and so with the relative abundance of the (sub)types: gray indicates a perfect co-circulation, while yellow, cyan, or magenta designate the total dominance of B, H1, or H3, respectively. **C) Relative abundances of (sub)types represented in $\mathbb{R}^2$.** The same 132 points are represented in the 2D Euclidean space after an isometric log-ratio transformation. **D) Trajectory of relative abundances of (sub)types in $\mathbb{R}^2$ for France, from 2000 to 2022.** The point for 2020 is missing because <50 classified cases were reported in France in the period Apr 2020 to 2021. Points follow the same triangular color code in Figures B) and C) and D).

4.3.2 Degree of (sub)type mixing over time

The ensemble of (sub)type composition trajectories for the 151 countries is provided in Figure 4.2A. The number of countries contributing to FluNet was initially limited (in 2000, only 30 countries contributed more than 50 samples, our threshold for inclusion), but steadily increased throughout the following years (e.g. 62 countries contributed in 2008) and then steeply rose after the 2009 influenza pandemic (109 in 2010). Concurrently to the global spread of SARS-COV-2, influenza incidence reduced massively [30, 172], to the point that only 26 countries reported at least 50 cases in 2020. Circulation of influenza viruses gradually resumed in 2021 and 2022 [51, 209, 210].

Figure 4.2A shows strong country-to-country variations in the (sub)type compositions. We first analyze global trends focusing on the variation in time of (sub)type mixing distribution across countries. To this end, we introduce the *mixing score*. This

indicator ranges from positive values for compositions within the co-dominance region - i.e. when (sub)types co-circulate in similar proportions - to negative values when one strain is dominant, and equals zero when one strain is responsible for exactly 50% of the infections. The mixing score quantifies (sub)type mixing in each country thus enabling defining distributions as synthetic indicators of the global (sub)type co-circulation in a given year. The comparison among distributions of different years (Figure 4.2B) clearly shows anomalous events. Results did not change when computing the mixing score for points represented with alternative log-ratio coordinates (Supplementary Material).

Boxplots of the mixing scores in Figure 4.2B have interquartile intervals ranging between -1.82 and 0.13. Co-dominance does not happen frequently and for the majority of years one (sub)type is dominant - negative values of mixing score corresponding to a (sub)type accounting for >50% infections. At the same time, however, strong dominance of a (sub)type (e.g. the (sub)type accounting for >75% of the infections) is also rare. For example, in 2017, one strain accounted for more than 75% of the infections in only 22 out of 132 countries. Still, Figure 4.2B highlights anomalous years, i.e. 2003, 2009, 2020, and 2021, when the mixing score reached extraordinarily low values - interquartile values ranging from -3.66 and -0.10. H3 was strongly dominant in 35 out of 45 countries in 2003, and H1 in 80 out of 96 countries in 2009 (Figure 4.2C). In 2020 only 4 out of 26 countries experienced (sub)type co-dominance, while in 13 countries one (sub)type was strongly dominant. In 2021, H3 strongly dominated in 70 countries out of 100, H1 in 3, and B in 3.

Atypical distributions reveal relevant ecological events that occurred at the global scale in those years. A new clade of influenza H3N2 - the A/Fujian/411/2002-like (H3N2) - was first isolated in 2002 and quickly disseminated in 2003 and 2004, becoming the dominant strain worldwide thanks to immunity evasion [211, 212]. Similarly, in 2009 a new strain of H1 hosted in pigs entered the human population in Mexico and disseminated globally, giving rise to the 2009 swine flu pandemic declared in June 2009 by W.H.O. [88, 213, 214]. Recently, Dhanasekaran and co-authors described the change in the (sub)type distribution in the aftermath of the SARS-COV-2 pandemic [25]. They reported counts and genetic classification of flu specimens tested in the few countries that experienced flu outbreaks from April 2020 to July 2021. Their findings unveiled spatial segregation (see also Figure 4.5), likely attributable to the unprecedented reduction in international mobility [215], which acted as a barrier preventing localized epidemics from spreading beyond their origins [25, 30]. From April 2021 to April 2022, some spatial segregation of (sub)types persisted, accompanied by the return of a global influenza circulation - though not yet at pre-pandemic levels - dominated by H3 [216].

4.3.3 Countries with similar trajectories of flu (sub)type dynamics

Besides interannual variability, the co-circulation dynamics of (sub)types present spatial structures. We analyze trajectories of 81 countries with complete data in the longest period of regular circulation of influenza, namely from 2010 to 2019, to exclude the effects of the flu swine and the SARS-COV-2 pandemics declared in June 2009 and March 2020, respectively, and we looked at countries with similar patterns of flu (sub)type alternation.

By performing Ward's linkage hierarchical clustering, we identified two groups of countries with similar trajectories of (sub)type relative abundances: Group I consists of 39 spiked trajectories oscillating synchronously, and Group II of 42 flatter trajectories (Figure 4.3A). The two groups interestingly correspond to distinct spatial regions, with Group I including countries from Europe (32), West Asia (4), and North Africa (2) as

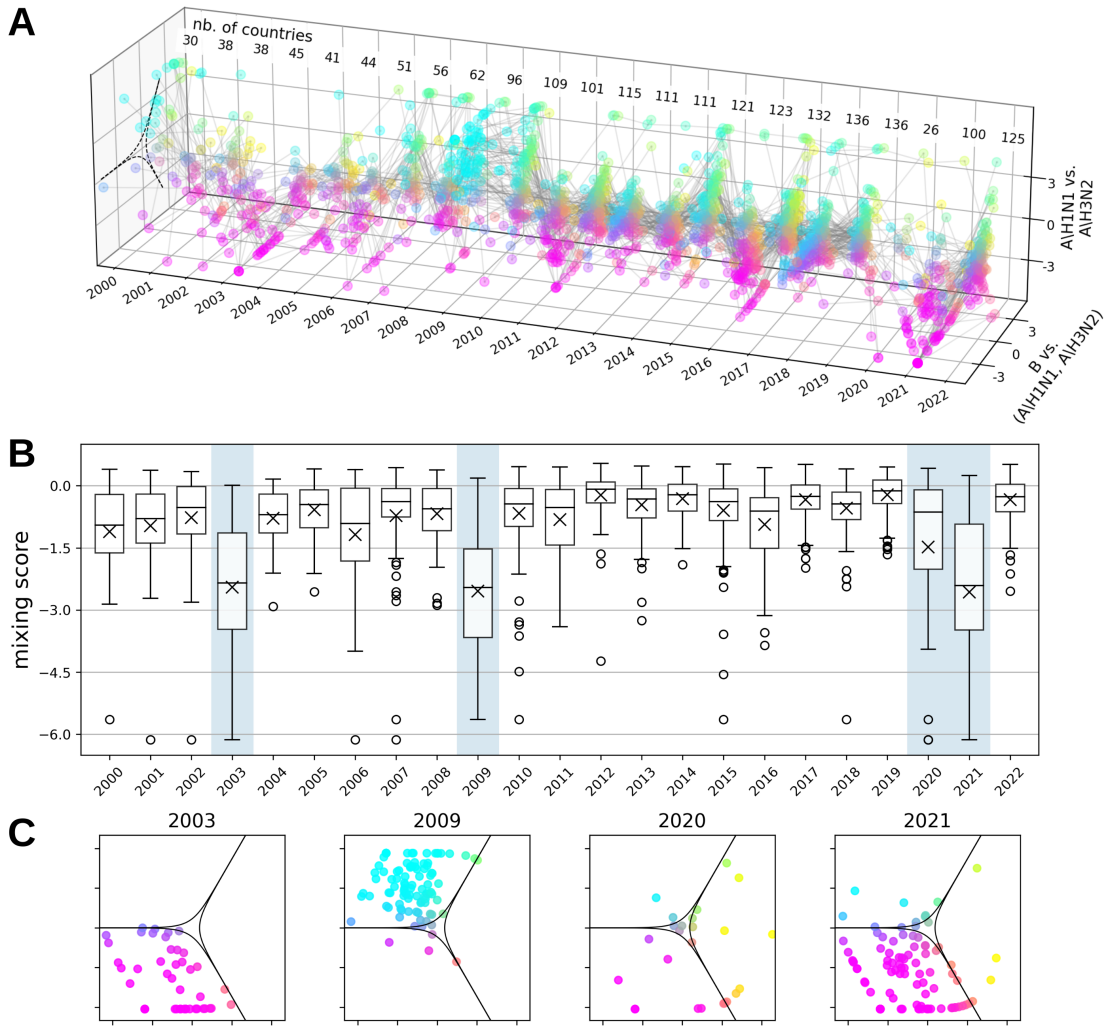


FIGURE 4.2: **A) Trajectories of relative abundances of influenza (sub)types H1, H3, and B.** We present trajectories for 151 countries, including years from 2000 to 2022 for which the number of flu cases classified by type was ≥ 50. The number of countries included by year is shown at the top. **B) Degree of mixing of flu (sub)types over time.** For the years 2000 to 2022, the mixing score of flu (sub)types was computed for each country, and their distributions are depicted through the boxplots. Positive scores represent countries where each (sub)type is responsible for <50% cases, negative scores denote the dominance of one (sub)type. **C) Flu (sub)type abundances for atypical years.** In 2003, 2009, 2020, and 2021, (sub)type mixing was unusually low. In 2003 and 2009, almost only one (sub)type - respectively H3 and H1 - circulated in most countries. Spatial segregation of flu (sub)types occurred in 2020 and 2021 and only a few countries experienced co-circulation of multiple (sub)types.

well as South Korea, and Group II including countries from all over the world except Europe (Figure 4.3B). Groups are robust when testing for other clustering techniques, other data inclusion criteria and alternative log-ratio coordinates for representing trajectories (supplementary material).

Our grouping matches well with the Influenza Transmission Zones (ITZs) defined by W.H.O. [217], in the sense that all Group I countries belong to Europe, North Africa, and West Asia, except for Iran from the Southern Asian ITZ and the Republic of Korea from the East Asian ITZ, while Group II countries belong to the other ITZs, except for Oman and Qatar from the West Asian ITZ (Figure 4.6).

The most evident result of the clustering is that the Group I countries have a strong alternation of (sub)types, well characterized and different from the rest of the world. Group II, on the other hand, includes countries that overall tend to have less alternation. In the tropics, in particular, there is a continuous circulation of all the (sub)types resulting in a higher mixing than in temperate regions overall. This is probably due to the different seasonality of influenza epidemics in the different climatic regions [2]. Notably, the combined effect of partial extinction of influenza viruses during the summer season, importation of new variants from abroad, and cross-immunity of (sub)types promotes more pronounced alternation of (sub)types in temperate regions. Yet, in Group II, alongside tropical nations, there are also several countries located in temperate regions across both hemispheres. Six Central and South American countries (Mexico, Nicaragua, El Salvador, Costa Rica, Argentina, Chile) and five Northern Hemisphere countries (United States, Canada, Japan, Mongolia, Kazakhstan) are classified within Group II but are then separated into two distinct subgroups by successive iterations of the hierarchical clustering (Supplementary Materials). Both of these subgroups show a marked alternation of (sub)types which is not synchronized with Group I. The five Northern Hemisphere countries, specifically, differentiated from Group I for a different pattern of H1 and H3 dominance in the early years following the 2009 pandemic (Supplementary Materials) [76, 98].

4.3.4 One-year forecasting of flu (sub)type relative abundances

The representation of (sub)type dynamics in terms of trajectories enables forecasting. We considered the same 81 country-specific trajectories of the previous section and tested five methods for predicting the relative abundance of flu (sub)types one year ahead. For each trajectory, we calculated the forecasts for the years 2017, 2018, and 2019, each time training the five algorithms on all the previous years since 2010.

First, we defined two forecasting methods that do not require CoDA. The simplest approach is to assert that the probability of encountering a specific dominance state (H1, H3, B, or co-dominance) in the coming year within a country corresponds to the frequency of observation of the state during the previous years. We considered this method as the null model, *M0 frequency-of-past-states*. A second simple model, the *M1 H1-H3-alternation*, is based on the knowledge acquired from past analyses of (sub)type dominance at large spatio-temporal scales, which revealed that H1 and H3 viruses tend to alternate, while influenza B often co-dominates [54, 76, 98]. Based on this, we predicted the composition at year y by taking the composition at year $y1$ with reversed percentages of H1 and H3. Going a step further, thanks to CoDA we were able to define more sophisticated methods based on statistical tools otherwise difficult to apply to percentage data. The *M2 average* model calculates the average of the time series. The *M3 VAR* model is a Vector AutoRegressive (VAR) model with lag=1, i.e. a linear model in which each composition is computed from the previous year's composition [148]. Finally, we took advantage of the spatial patterns identified before, to define a Bayesian

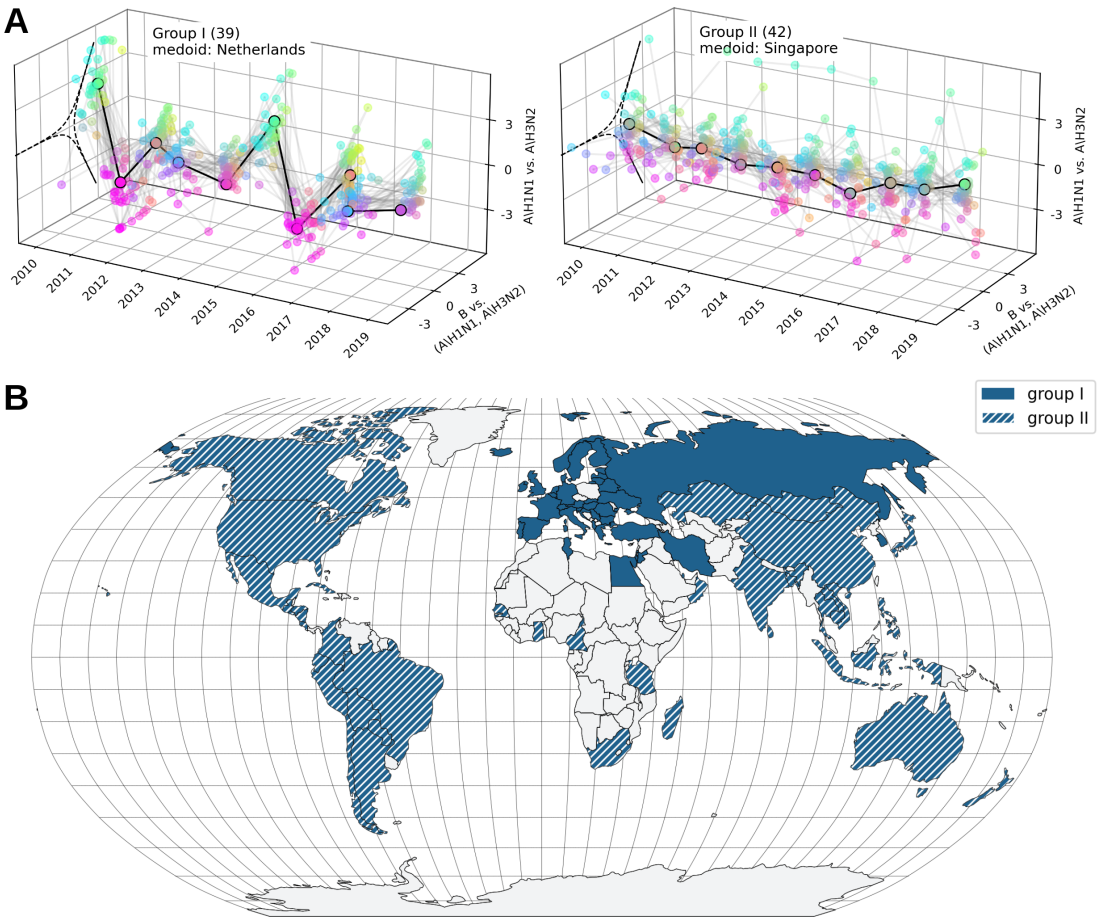


FIGURE 4.3: **Countries with similar trajectories of flu strain alternation. A) Hierarchical clustering of trajectories of H1, H3, and B relative abundances.** Trajectories for 81 countries from 2010 to 2019 are clustered in two groups - Groups I and II - with 39 and 42 countries, respectively. The medoid trajectories - i.e. the most central trajectories - of the two groups are highlighted in black and are reported in the legends. **B) Geographic positioning of Group I and Group II countries.** (Source of shape files for the map: Natural Earth [218].)

Hierarchical Vector AutoRegressive model (*M4 HVAR*), as applied in [145]. Through this algorithm, countries were no longer considered independent, and the forecast now presupposed similarity in VAR processes for the trajectories of countries within the same group. In practice, we applied the *M4 HVAR* model separately on Group I and Group II countries, each time simultaneously estimating the VAR coefficients for each trajectory within the group. Predictions for France (Group I) and Australia (Group II) are illustrated in Figure 4.4 as examples.

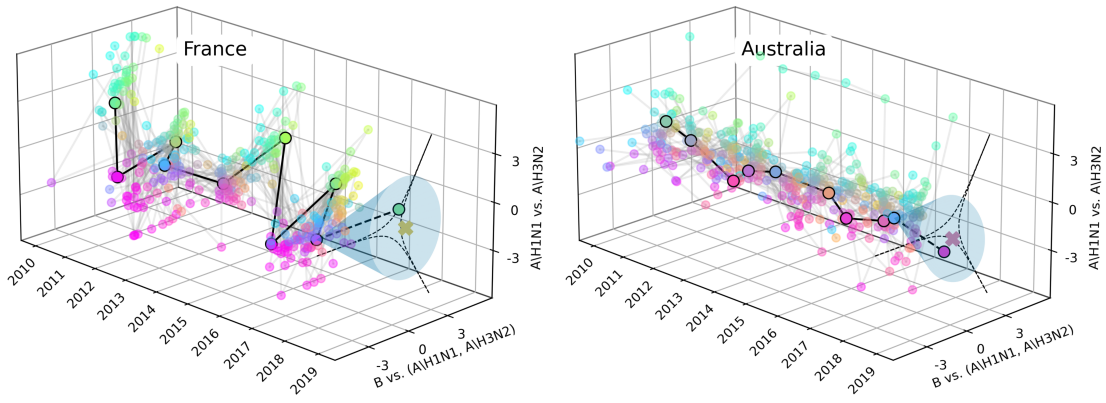


FIGURE 4.4: **Prediction of relative abundances of flu (sub)types for France in 2019 and for Australia in 2019.** Predictions are computed using the *M4 HVAR* model of order 2 and 1, respectively. The observed trajectories from 2010 to 2018 are represented with black solid lines, while the dashed segments link the points of 2019 to predict. The thinner gray lines correspond to the trajectories of the other countries within the respective group - Group I for France, and Group II for Australia - that are considered to train the model. The crosses depict the predictions and the shadow areas the ellipses associated with the 95% confidence intervals. Dot colors follow the triangular color code such that yellow, cyan, and magenta indicate a predominance of B, H1, and H3, respectively. The co-dominance regions are also shown with dashed lines as a reference.

We looked at the percentages of correctly predicted dominance states - the *Dominance State Accuracy* - to assess the forecast performances of the models (Table 4.1, left panel). By using the *M0* frequency of past states model we were able to correctly predict the dominance state 19% of the time for Group I countries, 36% for Group II countries, and 28% for all the countries. In this case, the accuracy was approximately comparable to a random guess, where one of the four possible dominance states is sampled with probability $1/4$. Estimates substantially improved with the *M4 HVAR* model, with which we obtained 32%, 40%, and 36% correct predictions, respectively. The other models overall did not provide more accurate predictions than the *M0* model.

The dominance state is the most interesting datum from an epidemiological point of view, but at the same time, it only provides quantized information, as the result of coarse discretization of (sub)type proportions. Models *M1* to *M4* have the advantage of operating on continuous values and allow us to estimate (sub)type compositions, including confidence intervals for models *M2*, *M3*, and *M4*. Thus, we also wanted to assess the performances of models *M1*,*M2*, *M3*, and *M4* by looking at the predicted vs. observed compositions. To this end, we relied on the *Energy Score* [219] - a metric for evaluating probabilistic forecasting often used in epidemiology [146] - (Table 4.1, right panel), as well as the *Dawid-Sebastiani Score* [219] and *Variogram Score* [220] as alternative scores for additional robustness checks (Supplementary materials). Those metrics are all designed to account for both calibration and sharpness of predictions, and all agreed on the supremacy of the *M4 HVAR* model. It is worth noting that, as expected, discrepancies in model performances were less accentuated for Group II countries that

had flat trajectories, while the *M4 HVAR* model revealed its potential to capture fluctuations in Group I trajectories. We also tested predictions on trajectories defined with alternative log-ratio coordinates and obtained very similar findings (Supplementary materials).

score	Dominance State Accuracy					Energy Score				
method	M0 frequency past states	M1 AH1/AH3 alternation	M2 average	M3 VAR(1)	M4 HVAR (p_I=2, p_{II}=1)	M0 frequency past states	M1 AH1/AH3 alternation	M2 VAR(1)	M3 average	M4 HVAR (p_I=2, p_{II}=1)
group I (39)	0.19	0.20	0.20	0.21	**0.32**	/	3.85	1.82	2.20	**1.46**
group II (42)	0.36	0.33	0.37	0.29	**0.40**	/	2.04	1.35	1.63	**1.20**
all countries (81)	0.28	0.27	0.28	0.26	**0.36**	/	2.91	1.58	1.90	**1.32**

TABLE 4.1: **Evaluating methods for influenza (sub)type forecasting.** In the columns: two scores are considered to compare five prediction methods. On the rows: average prediction scores are computed for Group I countries, Group II countries, and all the countries. Methods that perform best by country groupings are highlighted for both the Dominance State Accuracy and the Energy Score. We precise that Accuracy is positively oriented, whereas the Energy Score is negatively oriented. This implies that as the model's performance improves, Accuracy increases, while the Energy Score decreases.

A close look at the estimated coefficients for the *M4 HVAR* model reveals interesting perspectives on the coupled dynamics of the (sub)types. First, we find that within the same group, the coefficients estimated are overall robust from country to country (Figures 4.8, supplementary material). Second, the offset of the model estimated for Group I countries identifies a co-dominance of the (sub)types as the most likely situation, with H3 abundance close to 50 percent. This is in line with the fact that overall H3 is the most circulating subtype [51, 52] due to its fast mutations that make it highly transmissible [15, 40]. For Group II countries, we found the same tendency, but with a higher mixing of (sub)types on average and a higher variability from country to country. Finally, the other coefficients in the models regulate the viruses' alternation and are independent of the specific country. They are almost all negative, indicating that the (sub)types tend to alternate from one year to the next, as illustrated by the reversal of colors in Figure 4.9 (Supplementary Materials). This is consistent with other studies [98] and with the cross-immunity effect described in the literature [58, 59, 60, 61, 202].

4.3.5 One-year forecasting of the (i) dominance/non-dominance and of the (ii) circulation/non-circulation of each (sub)type

We predicted the relative abundances of (sub)types one year ahead in terms of compositions - i.e. (B%, H1%, H3%) - and in terms of dominance states - i.e. dominance of B, H1, H3 or co-dominance. This is a complex task because it involves multilabel classification and so, even proposing methods that substantially improve naive estimations, we still end up with results that are not informative enough for public health response. Here, we simplify the problem by asking binary questions. In particular, we focus on one (sub)type at the time and we ask whether this specific (sub)type:

(i) will be dominant (>50% of cases) in the next year (true or false?);

(ii) or it will have a negligible impact (<10% of cases) in the next year (true or false?).

Hereafter, for each (sub)type we answer questions (i) and (ii), by using the same five methods applied before. We imagine a situation where public health authorities have some standard procedures that are applied to face the coming influenza season and those procedures can be modified and optimized only in the presence of additional reliable information regarding the (sub)type dominance. In such a scenario, above

all, it is important to avoid false positives and that's why we evaluate predictions by computing the *precision*. Other metrics are considered for robustness checks and show consistent results (see Supplementary Materials).

Results in Table 4.2 show that the *M4 HVAR* model is the best-performing model for predictions on B and H3 (sub)types, while the average is the best for H1. In particular, the *M4 HVAR* provides the largest improvement when applied to trajectories of Group I countries. For example, for those countries, the *M4 HVAR* model correctly identified 31 of the 41 cases in which B accounted for less than 10% of infections, with only five false positives (not shown). This corresponds to a precision in predicting B negligibility of 0.86 for *M4 HVAR*, compared to 0.17÷0.53 for the other models. We find a similar improvement when predicting the negligibility (dominance/non-dominance) of H3, where precision for *M4 HVAR* was 0.49 (0.58) compared to 0.13÷0.36 (0.0÷0.38) for the other models. Furthermore, results were robust when predictions were run on trajectories expressed in alternative coordinates (Supplementary Materials).

	B will be dominant (>50%) in the next season: T/F ?					B will be negligible (<10%) in the next season: T/F ?				
score	Precision					Precision				
method	M0 frequency past states	M1 AH1/AH3 alternation	M2 average	M3 VAR(1)	M4 HVAR (p_I=2, p_{II}=1)	M0 frequency past states	M1 AH1/AH3 alternation	M2 average	M3 VAR(1)	M4 HVAR (p_I=2, p_{II}=1)
group I (39)	0.13	0.00	0.00	0.14	**0.27**	0.28	0.17	0.25	0.53	**0.86**
group II (42)	**0.13**	0.00	0.00	0.07	0.00	0.11	0.12	0.13	**0.15**	**0.15**
all countries (81)	0.13	0.00	0.00	0.11	**0.26**	0.17	0.14	0.18	0.29	**0.67**

	A\H1N1 will be dominant (>50%) in the next season: T/F ?					A\H1N1 will be neglegible (<10%) in the next season: T/F ?				
score	Precision					Precision				
method	M0 frequency past states	M1 AH1/AH3 alternation	M2 average	M3 VAR(1)	M4 HVAR (p_I=2, p_{II}=1)	M0 frequency past states	M1 AH1/AH3 alternation	M2 average	M3 VAR(1)	M4 HVAR (p_I=2, p_{II}=1)
group I (39)	0.29	0.25	**0.33**	0.11	0.32	0.00	0.03	**0.15**	0.00	0.00
group II (42)	0.37	**0.47**	0.43	0.34	0.31	0.08	0.16	0.11	0.06	**0.40**
all countries (81)	0.32	0.34	**0.41**	0.23	0.31	0.07	0.11	**0.13**	0.04	0.12

	A\H3N2 will be dominant (>50%) in the next season: T/F ?					A\H3N2 will be neglegible (<10%) in the next season: T/F ?				
score	Precision					Precision				
method	M0 frequency past states	M1 AH1/AH3 alternation	M2 average	M3 VAR(1)	M4 HVAR (p_I=2, p_{II}=1)	M0 frequency past states	M1 AH1/AH3 alternation	M2 average	M3 VAR(1)	M4 HVAR (p_I=2, p_{II}=1)
group I (39)	0.16	0.36	0.13	0.33	**0.49**	0.00	0.29	0.25	0.38	**0.58**
group II (42)	0.32	0.32	0.28	0.30	**0.41**	0.00	**0.17**	0.05	0.12	**0.17**
all countries (81)	0.23	0.33	0.22	0.32	**0.43**	0.00	0.24	0.07	0.24	**0.44**

TABLE 4.2: **Evaluating methods for predicting the (i) dominance/non-dominance and the (ii) circulation/non-circulation of each (sub)type one year ahead.** Results are summarized in six panels: for each one of the three (sub)types (on the rows) we answered two questions (in the columns). In each panel, the average precision is computed to compare predictions for Group I countries, for Group II countries, and for all countries, by using five different prediction methods. We highlight the method performing best by country groupings.

4.4 Discussion

We provided a fully quantitative representation of spatio-temporal influenza (sub)type dynamics in terms of country-year trajectories. We treated percentages of flu cases by (sub)type through the framework of CoDA thanks to which we developed visualization tools, synthetic ecological indicators, and quantitative analyses.

First, we focused on events of disruption of (sub)type mixing that typically occur when a particularly transmissible flu variant emerges. Previous studies often investigated single events, analyzing the virus sequences and the antigenic maps to quantify

the advantage of the new variant [212, 221]. Here, we focused on the global scale and addressed the detection of those events over time and the concise quantification of their intensity worldwide. This framework will enable us to follow the changes in flu circulation after the COVID-19 pandemic.

Then, we considered the intra-pandemic period from 2010 to 2019 and investigated the spatial patterns of the (sub)type alternation. We found that Europe and neighboring countries were characterized by a strong and well-synchronized alternation of (sub)types, clearly distinguishable from patterns observed for the rest of the world. This geographical structure proved to be valuable information for making predictions about the relative abundances of (sub)types for the following year.

This forecasting problem is essentially new in the literature. Previous studies looked at the ongoing influenza epidemic and predicted influenza incidence weeks in advance [222, 223, 224, 225, 226]. Other studies focused on the evolution of influenza viruses, considering genetic mutations and antigenic characteristics to predict the growth, decline, and replacement of circulating clades for an individual influenza subtype/lineage [46, 227, 228, 229, 230]. In this study, we investigated the conventional subject of influenza (sub)type distribution, while also posing a novel question—predicting their relative abundances a year ahead. We showed that within the CoDA framework, it was possible to define sophisticated statistical methods that substantially improve naive estimations. The goodness of predictions was highly variable, with the *M4 HVAR* algorithm generally performing well for Group I countries (39 European and neighboring countries) and for B and H3 (sub)types. In some cases, we got surprisingly accurate results: for Group I countries, in 86% of the cases where we predicted negligible influenza B circulation for 2017-2019, influenza B was actually responsible for less than 10% of the cases. Predicting the relative abundances of (sub)types might help tailor interventions, by identifying the cohorts most at risk, distributing vaccines accordingly, informing public health practitioners, and allocating beds and intensive care in hospital wards.

We point out some limitations that can be enhanced by future research. First, we did not consider vaccine coverage. However, it is clear that this element could improve predictive models, as the amount of circulating (sub)types may to some extent be the result of vaccine coverage. Unfortunately, global data on flu vaccine coverage is currently not publicly available. Additionally, the inclusion of this type of data is challenging, both because vaccine efficacy largely varies over time and by (sub)type, and also because vaccine efficacy is typically estimated against severe symptoms, making it complicated to evaluate the vaccine's impact on virus transmission. Second, we included spatial dependencies only implicitly, by distinguishing countries in two groups before the application of the *M4 HVAR* model. However, it would be of interest to explicitly incorporate spatial correlations, similar to the approach of Paul et al. in their analysis of flu epidemics in the South of Germany [231]. Lastly, further research might include demography, climate, and air travel.

We also specify that our forecast framework is applicable for periods of stable strain circulation, as was the case for years 2010-2019. Following the COVID-19 pandemic, the global circulation of influenza viruses was highly perturbed, making any prediction more difficult.

Similar analyses can be applied to surveillance data of influenza at different spatio-temporal scales or to any epidemiological data in the form of percentages. For example, now that PCR swabs to simultaneously test for SARS-COV-2, Influenza, and RSV infections are available, the possibility arises to comprehensively investigate the interconnected dynamics of these respiratory diseases (for previous studies in this direction see [117, 119]). More broadly, there is a variety of circulating viruses with distinct strains

capable of causing different diseases. For instance, enteroviruses encompass more than 100 strains, most of which result in asymptomatic infections, while some can lead to serious conditions like hand-foot-and-mouth disease or polio [232, 233, 234]. Another context where data in percentage form are often used concerns the monitoring of antibiotic usage. As an illustration, the World Health Organization reported the proportional consumption of antibiotics based on AWaRe categorization for 65 countries in its latest published report [217].

4.5 Methods

4.5.1 FluNet data

FluNet collects influenza surveillance data from different countries and provides the number of weekly cases classified by type/subtype [17, 18]. We determine influenza B infections by summing B, B, and unspecified cases, and H1 infections by summing pre-2009 pandemic H1 and post-pandemic H1 cases. It may happen that for some influenza A samples the subtype is not specified. If so, we redistributed these counts between H1 and H3 in accordance with the proportions of the classified cases reported in the same week. In case no influenza A case was subtyped for a specific week and country, we looked at the proportions of H1 and H3 in the five weeks centered around the week or, alternatively, in the year. Later, we aggregated influenza B, H1, and H3 cases over a one-year time frame and calculated the respective percentages to define the relative abundance of the three influenza strains. The time frame of a year considered in the analyses goes from April to April of the following year. More precisely, the beginning of the year coincides with the first Monday following April 23. This date was chosen so as to minimize the risk of splitting the influenza epidemic of a country in the temperate areas into two consecutive years. Specifically, we looked at FluNet data from 1995 to 2019 for all countries and calculated the week that on average had the lowest proportion of annual positive cases. In all analyses, we discarded countries-years with fewer than 50 classified influenza cases. We also tested an alternative threshold of 500 cases for robustness check.

4.5.2 Log-ratio transformations

We used the isometric log-ratio (*ilr*) transformation to map points from the Simplex to the Euclidean space (defined in section Results) and the additive log-ratio (*alr*) transformation as an alternative map for robustness check [28]. This latter is given by the formula 4.2:

$$\begin{cases} u & = \ln \frac{B\%}{H3\%} \\ v & = \ln \frac{H1\%}{H3\%} \end{cases}. \tag{4.2}$$

Nevertheless, these transformations are not defined when any component equals zero. Therefore, we first replaced zero components with small percentages, by using a Bayesian-multiplicative treatment [235, 236]. Such treatment assumes that the zero counts are the result of insufficient sample sizes, rather than a real absence of the virus.

4.5.3 Definition of the mixing score

The mixing score is defined as the distance between the point in the Euclidean plane (u, v), representing the (sub)type composition, and the boundary of the co-dominance

region, taken with a positive sign when the point is within that region and with a negative sign otherwise. Let's recall that the boundary of the co-dominance region in the simplex is identified by the points (B%, H1%, H3%) for which one component corresponds to exactly 50%. The *ilr* (or *alr*) transformation defines the co-dominance region in the Euclidean space. For example, the points such that H1%=50%, under the *ilr* transformation are mapped into the coordinates $(u, v(u))$ such that

$$v(u) = \sqrt{2}\left(\ln\left(e^{u\sqrt{3/2}} + \sqrt{e^{u\sqrt{6}} + 4}\right) - \ln 2\right).$$

4.5.4 Clustering of trajectories

Clustering techniques require the definition of (i) a distance between objects and (ii) a procedure for grouping elements given their relative distances. In our analysis, we defined the distance between two trajectories – represented in (u, v) coordinates – as the average Euclidean distance of the corresponding points. Then we applied Ward's linkage hierarchical clustering (Ward Jr., 1963). As alternative clustering algorithms, we tested the weighted linkage and the k-medoid clusterings, other than Ward's linkage method applied to the 29 countries which had at least 500 classified cases per year. See [237, 238] for the k-medoid method, the web page of the *scipy.cluster.hierarchy.linkage* python module for the other algorithms. Hierarchical clusterings provide a hierarchy of nested partitions, each one made of $1, 2, \ldots$ to N groups. Here, we relied on the widely used Silhouette coefficients [239] to retrieve the best partition. The optimal partition consistently grouped countries into two groups, highly similar across the sensitivity analyses (Supplementary Materials). The only exception occurred with Ward's linkage clustering, where the first split separated Qatar from all other countries. However, the latter were then further divided into two groups. This is in line with the classification discussed in the results.

4.5.5 Forecasting of trajectories

We studied bivariate trajectories of compositions of the form $(y_1, \ldots, y_T)$, such that $y_t = (u, v)'_t = ilr\left((\text{B\%}, \text{H1\%}, \text{H3\%})'_t\right)$, with $t = 1, \ldots, T$. Each composition corresponded to one of four possible dominance states - dominance of B, H1 or H3, or co-dominance. We used five methods for predicting the alternation of (sub)types one year in advance. In particular, we considered a naive method by which we predicted only the dominance state (*M0 frequency-of-past-states*), and four alternative methods by which we predicted both the dominance state and the composition. For three of these (*M2*, *M3*, and *M4*), we could also compute the confidence intervals of the predictions. A summary of the observable that can be predicted by the approach (i.e. (sub)type composition and/or dominance state) and of the evaluation scores that can be used for each method is presented in the supplementary material.

We detail hereafter the forecasting methods we used:

- *M0 frequency-of-past-states*: the probability of observing a given dominance state in year $T + 1$ is given by the percentage of times the state has been observed in years 1 to T. Accordingly, the dominant state in year $T + 1$ is defined as the most observed state in the past if that state is unique, or it is defined by uniformly randomly choosing one of the most observed states.

- *M1 H1-H3-alternation*: the estimated composition is $\hat{y}_{T+1} = (u, -v)'_T$, or, equivalently in the simplex, $(\text{B\%}_{T+1} = \text{B\%}_T, \text{H1\%}_{T+1} = \text{H3\%}_T, \text{H3\%}_{T+1} = \text{H1\%}_T)'$.

- *M2 average*: the estimated composition is $\hat{y}_{T+1} = \frac{1}{T}\sum_{t=1}^{T} y_t$, and the prediction's confidence interval is given by the empirical covariance matrix $\hat{\Sigma} = \frac{1}{T(T-1)}\sum_{t=1}^{T}(y_t - \hat{y}_{T+1})'(y_t - \hat{y}_{T+1})$.
- *M3 VAR*: assuming that the trajectory has been generated by a VAR process of lag p, then the composition of year t can be written as a linear function of the previous p compositions:

$$y_t = \nu + A^{(1)}y_{t-1} + \cdots + A^{(p)}y_{t-p} + \epsilon_t,$$

where ν is a vector of two intercept terms, $A^{(I)}$ are 2×2 coefficient matrices, and ϵ_t is Gaussian noise. Following Chapter 3 of [148], we define:

$$\begin{aligned} Y &= (y_{p+1}, \ldots, y_T), \\ B &= (\nu, A^{(1)}, \ldots, A^{(p)}), \\ Z_t &= \begin{pmatrix} 1 \\ y_p \\ \vdots \\ y_1 \end{pmatrix}, \\ Z &= (Z_p, \ldots, Z_{T-1}), \\ U &= (\epsilon_{p+1}, \ldots, \epsilon_T). \end{aligned}$$

Then, the VAR(p) process can be written as $Y = BZ + U$. The corresponding least squares estimator is $\hat{B} = YZ'(ZZ')^{-1}$. The prediction for the year $T+1$ is computed as $\hat{y}_{T+1} = \hat{B}Z_T$, and the prediction's confidence interval is estimated via the empirical covariance matrix corrected for short time series:

$$\hat{\Sigma}_\epsilon = \frac{T-p+1}{(T-p)(T-3p-1)}(YY' - YZ'(ZZ')ZY').$$

- *M4 HVAR*: we consider groups of similar trajectories and we assume that each trajectory followed a VAR process, such that the process for country c written in the compact form is $Y_c = B_c Z_c + U_c$. Then, the hierarchical structure is imposed by assuming that the VAR processes for the trajectories in the group are similar. Specifically, we define $B_C = W + V_C$, with W being the matrix of coefficients encoding the average behavior of the group, that is the same for all the trajectories in the group, and V_C being the coefficient matrix for the single trajectory adjustment. Moreover, we assume that elements in (W, V_C, U_C) are independent random variables, sampled from distributions parametrized by some latent variables. A detailed representation of the model is shown in Section 2.3.2. For each coefficient, it is possible to write the likelihood of the model conditional on the other parameters. Hence, coefficients can be estimated via a Gibb sampling. For the conditional distributions and the code for implementing the Gibb sampler, we followed [145] slightly modifying their model to introduce the intercept terms. From the Monte Carlo chains, we obtained several predictions $\hat{y}_{T+1}$, that we used to estimate the final prediction and the confidence intervals.

We compared model performances across several metrics. We defined the *Dominance State Accuracy* as the percentage of dominance states correctly predicted, in order to compare dominance state predictions, while we relied on probabilistic forecasting evaluation scores to compare predictions of compositions. Specifically, for each

country-year prediction, we used the *Energy Score* to compare the composition $y \in \mathbb{R}^2$ corresponding to the observation with the forecast distribution F defined by N samples $X_1, \ldots, X_N$, with $X_i \in \mathbb{R}^2$, from the posterior distribution. Then, all the energy scores were averaged to obtain a single value for the model's performance. The formula for the energy score is

$$ES(F, y) = \frac{1}{N}\sum_{i=1}^{N}\|X_i - y\| - \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\|X_i - X_j\|,$$

where the two terms take into account the calibration and the sharpness of the prediction, respectively. It is worth noting that this is a multivariate generalization of the more common *Continuous Rank Probability Score*, often used in epidemiology [146]. Moreover, in the case of point forecast, it coincides with the *Mean Absolute Error*, and it is therefore also suitable for evaluating the *M2 H1-H3-alternation* method for which we don't have confidence intervals.

For robustness check, we computed alternative evaluation scores. The *Dawid-Sebastiani Score* is the multivariate generalization of the *Logarithmic Score* [219], also commonly used in epidemiology [146], and is defined as a function of the mean μ_X and covariance matrix Σ_X of the samples X_i:

$$DS(F, y) = -\log(|\Sigma_X|) - (y - \mu_X)'\Sigma_X^{-1}(y - \mu_X).$$

Finally, the *Variogram Score* is more suitable for evaluating the correct or incorrect estimation of the correlations between components of the multivariate quantity [220] and it is defined as

$$VS^p(F, y) = \sum_{i=1}^{d}\sum_{j=1}^{d} w_{i,j}(|y^{(i)} - y^{(j)}|^p - \frac{1}{N}\sum_{k=1}^{N}|X_k^{(i)} - X_k^{(j)}|^p)^2.$$

We considered $p = 0.5$ and $p = 1$ with constant weights $w_{i,j} = 1$ as standard choices. All these probabilistic forecast evaluation scores are negatively-oriented, such that they decrease when the forecast improves. Furthermore, they are *proper scores*, i.e. designed to be optimized when the forecast distribution coincides with the true distribution of the observations [219].

4.6 Code and data availability

Analysis was implemented in R (version 4.3.2) and Python (version 3.8.5). Other than standard packages for data treatment, plots and calculations (mainly the Python packages *pandas, matplotlib, numpy, os*), we relied on the following packages for specific tasks:

- *zComposition 1.4.0-1* (R) for zero imputation in the pre-processing of compositions [236];
- *robCompositions 2.3.1* (R) for mapping compositions from the Simplex to the Euclidean space and back [240];
- *ternary* (python) for drawing ternary plots [241];
- *scipy 1.6.2* (python) for clustering analysis;

- *sklearn 1.3.2* (python) for clustering analysis and computation of some of the prediction evaluation scores [242];
- R code developed by Lu and colleagues [145], based on which we performed the VAR and HVAR predictions;
- *scoringRules 1.0.2* (R) for calculation of proper scores for probabilistic forecast evaluation [243].

Code and data for reproducible analyses are available at https://github.com/FrancescoBonacina/coupled-dynamics-flu-subtypes.

4.7 Supplementary Materials

4.7.1 Table of computable quantities for each forecasting method

We used five different methods to forecast the (sub)type compositions one year in advance. In Table 4.3 we provide a summary of the observables that can be predicted by each approach (i.e., (sub)type compositions and/or dominance states) and of the evaluation scores that can be computed to assess the goodness of the predictions.

Quantity		Is the quantity computable for the specific forecasting method ?				
Notation	Description	M0	M1	M2	M3	M4
Predicted quantities						
$(\widehat{u}, \widehat{v})$	Predicted composition one year ahead. It corresponds to the center of the forecast distribution.	X	✓	✓	✓	✓
$\widehat{\Sigma}_{(u,v)}$	Covariance matrix defining the width of the forecast distribution.	X	X	✓	✓	✓
$\hat{s}$	Predicted dominance state among {B, A\H1N1, A\H3N2, co-dominance}. It is defined by looking at the dominance region where $(\widehat{u}, \widehat{v})$ falls.	✓	✓	✓	✓	✓
$\widehat{\pi}_s$	Vector containing the estimated probabilities of observing the four possible dominance states. It is computed by sampling 1000 points (u,v), from the forecast distribution $N((\widehat{u}, \widehat{v}), \widehat{\Sigma}_{(u,v)})$ and looking at the percentage of points falling in each region of dominance.	✓	X	✓	✓	✓
Scores for evaluating the forecasts of the composition						
ES	The Energy Score (Gneiting and Raftery 2007). It coincides with the Mean Absolute Error for point forecasts.	X	✓	✓	✓	✓
DSS	The Dawid-Sebastiani Score (Gneiting and Raftery 2007). It is defined only when the mean and the covariance matrix of the forecast distribution exist.	X	X	✓	✓	✓
VS05	The Variogram Score of order 0.5 (Scheuerer and Hamill 2015). It is meaningful only for probability forecasts.	X	X	✓	✓	✓
VS1	The Variogram Score of order 1 (Scheuerer and Hamill 2015). It is meaningful only for probability forecasts.	X	X	✓	✓	✓
Scores for evaluating the forecasts of the dominant state						
Accuracy	The percentage of correct predictions, used for evaluating classification tasks.	✓	✓	✓	✓	✓
Precision	The ratio TP/(TP+FP), used for evaluating classification tasks. With TP=true positives and FP=false positives.	✓	✓	✓	✓	✓
Average Precision (from PR-curve)	Score used to summarize a precision-recall curve for binary classification, built by comparing observed labels with probabilities estimated for the two classes. For multilabel classification, several ways to trace the problem back to a binary problem exist. Here we averaged scores from four definitions implemented in *sklearn.metrics* (Pedregosa et al. 2011).	✓	✓	✓	✓	✓
AUC ROC	The Area Under the Receiver Operating Characteristic Curve. Score used to summarize a ROC curve computed for binary classification, built by comparing observed labels with probabilities estimated for the two classes. For multilabel classification, several ways to trace the problem back to a binary problem exist. Here we averaged scores from four definitions implemented in *sklearn.metrics* (Pedregosa et al. 2011).	✓	✓	✓	✓	✓

TABLE 4.3: **Computable quantities for each forecasting method.**

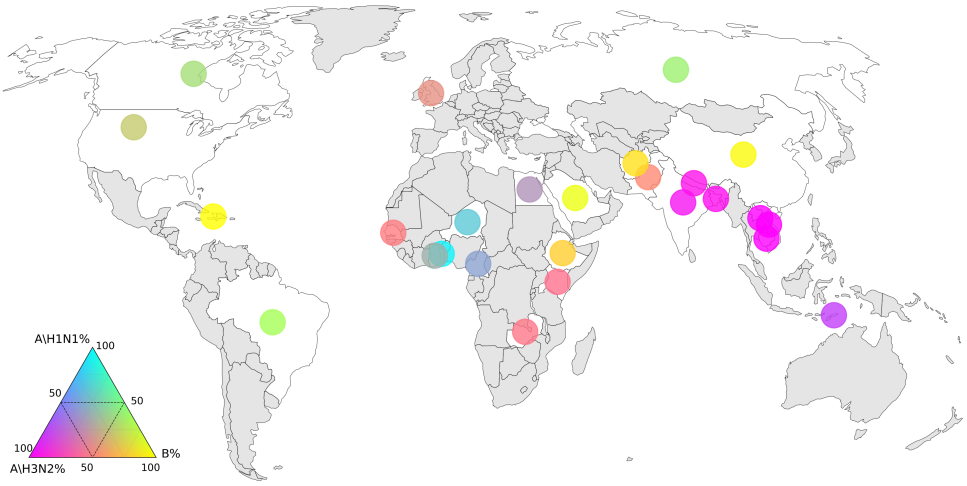


FIGURE 4.5: **Relative abundances of influenza (sub)types from April 2020 to April 2021.** Countries in gray didn't report a minimum of 50 classified cases of influenza in the period. For the other countries, the circle's color represents the relative abundance of the (sub)types according to the scale defined by the ternary diagram at the bottom left.

4.7.2 Additional Results

Geographical segregation of flu (sub)types in 2020-2021

From April 2020 to April 2021 only 26 countries reported a minimum of 50 classified cases of influenza. An unusual spatial segregation of (sub)types occurred during that period (Figure 4.5). Only 4 countries presented a situation of (sub)type co-dominance, while in as many as 13 countries one (sub)type was responsible for more than 75% of the cases. H3 was almost the only circulating influenza strain in seven countries of South-East Asia (India, Nepal, Bangladesh, Cambodia, Vietnam, Laos, and Timor-Leste). Influenza B accounted for more than 80% of flu cases in five countries (Saudi Arabia, Haiti, China, Etiopia, Afghanistan) and it was the dominant strain in the other American countries and in the Russian Federation. H1 circulated mainly in Togo, Niger, Cameroon, Ghana and Egypt.

Detailed hierarchical clustering of trajectories in 2010-2019

Grouping of countries up to the six-group partition. In Figure 4.6 we report a diagram of the grouping of countries identified by Ward's hierarchical clustering algorithm, developed up to the six-group partition. Our country groups are also compared with the Influenza Transmission Zones defined by the W.H.O.

Differences between groups in terms of trajectories. We compare average trajectories for the different country groups identified by the hierarchical clustering to highlight the main differences in patterns of (sub)type alternation (Figure 4.7). In particular, we compare Group I vs. Group II (excluding Qatar, since it has a specific stand-alone behavior), and Group I vs. the three individual subgroups that compose Group II - namely *South and Central America (II)*, *Temperate North (II)* and *Tropics (II)*. The results indicate that tropical countries (Tropics II) experienced very limited alternation of (sub)types during 2010-2019, compared to the other groups. On the other hand, alternation of (sub)types in *South and Central America (II)* and *Temperate North (II)* countries, although pronounced, was not synchronized with Group I countries and that's why those countries were

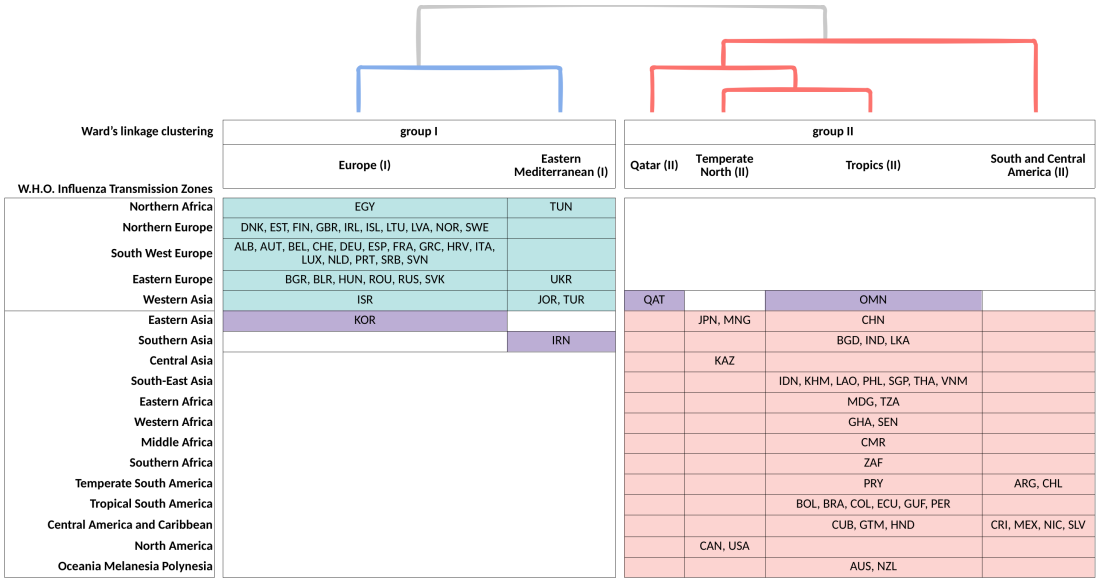


FIGURE 4.6: **Hierarchical clustering of country trajectories compared with the W.H.O. Influenza Transmission Zones.** The columns correspond to the country groups identified by Ward's linkage hierarchical clustering up to the six-group partition. The nested structure of the clustering is specified by the dendrogram at the top. The rows designate the Influenza Transmission Zones of W.H.O.. The 81 countries considered in the analysis are identified by their three-letter iso-codes. We used blue and red to show the correspondence between our clustering and the W.H.O. ITZs. The four countries in purple are the only ones for which the two groupings do not match.

not grouped with Europe and neighboring regions. Furthermore, on average, Group I countries had a greater circulation of influenza H1 compared to H3 (right column in Figure 4.7), while the abundance of influenza B vs. influenza A (left column) was similar in all the three groups. We further investigate the trend for the United States of America and Australia as specific countries of the temperate regions included in Group II.

Insights into the estimated HVAR models

In Figure 4.8 we report distributions of the VAR coefficients estimated through the *M4 HVAR* model for Group I and Group II trajectories (left and right panel, respectively). The model's lag was fine-tuned for each group by testing lags 1 and 2 and comparing the average Energy Scores for predictions across all countries within the group and all the predicted years (2017, 2018, and 2019). Lag=2 was chosen for Group I and lag=1 for Group II. Let us make explicit that the VAR process for countries of Group I is defined as

$$\begin{pmatrix} u \\ v \end{pmatrix}_t \simeq \begin{pmatrix} \nu_u \\ \nu_v \end{pmatrix} + \begin{bmatrix} A^{(1)}_{uu} & A^{(1)}_{uv} \\ A^{(1)}_{vu} & A^{(1)}_{vv} \end{bmatrix} \begin{pmatrix} u \\ v \end{pmatrix}_{t-1} + \begin{bmatrix} A^{(2)}_{uu} & A^{(2)}_{uv} \\ A^{(2)}_{vu} & A^{(2)}_{vv} \end{bmatrix} \begin{pmatrix} u \\ v \end{pmatrix}_{t-2}.$$

The same process without the last term was used to model Group II trajectories.

From Figure 4.8 we see that the VAR processes adjusted for different countries through the *M4 HVAR* algorithm are similar to each other. We can then focus on the coefficients of a specific country to get an intuition of the functioning of the VAR process that is valid for all countries. A VAR process of order p consists of a map that determines the coordinates of point $t+1$, starting from points $(t-p+1,\dots,t)$. To make graphical representation possible, we consider a VAR process of order 1. Specifically,

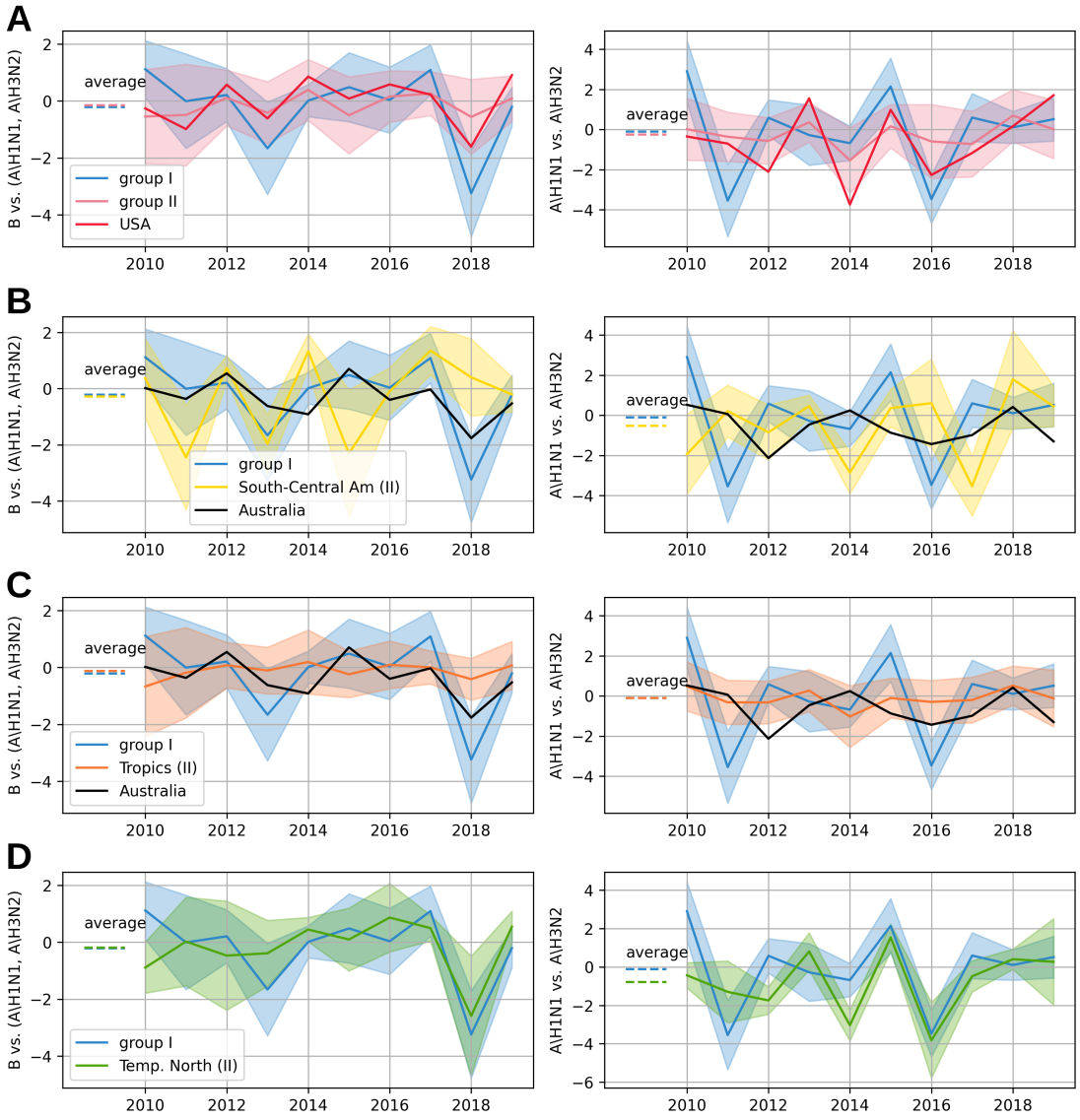


FIGURE 4.7: **Typical trajectories of (sub)type alternation by country groups.** Bivariate trajectories by group are compared by looking at average trends (solid line) with standard deviation confidence intervals (shaded area). Trends for the individual coordinates are shown in the two columns. Average values over the entire period are depicted with the dotted lines. **A)** Comparison of Group I, Group II, and the United States of America. **B), C), D)** Comparison of Group I with the different subgroups of Group II. In **B)** and **C)** the trajectory for Australia is also shown as an additional comparison.

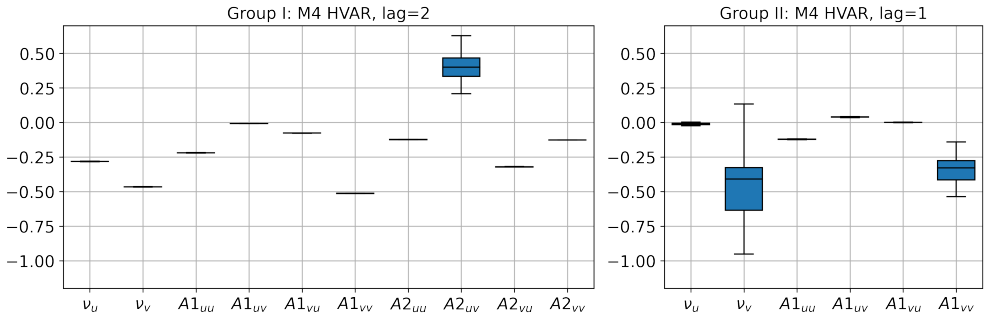


FIGURE 4.8: **Distributions of coefficients of the HVAR models estimated for Group I (lag=2) and Group II (lag=1) trajectories.**

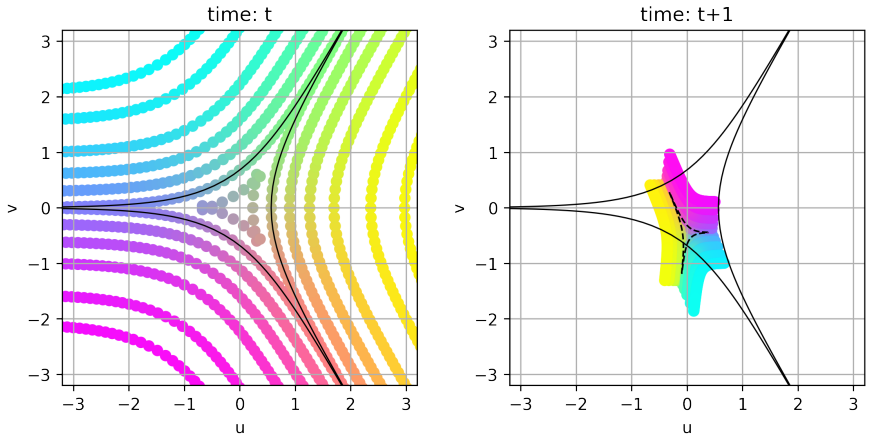


FIGURE 4.9: **Illustration of a VAR transformation.** Here, we consider the VAR coefficients estimated through the *M4 HVAR* model with lag=1 for the trajectory of Australia from 2010 to 2018. Specifically, points of the left panel are mapped into points of the right panel via the transformation 4.3.

we consider the process estimated for Australia, which is defined as

$$\begin{pmatrix} u \\ v \end{pmatrix}_t = \begin{pmatrix} 0.013 \\ -0.448 \end{pmatrix} + \begin{bmatrix} -0.121 & 0.040 \\ -0.001 & -0.265 \end{bmatrix} \begin{pmatrix} u \\ v \end{pmatrix}_{t-1} \tag{4.3}$$

.

In Figure 4.9 we illustrate the effect of the transformation on a grid of points defined in the *ilr* coordinates. We notice that:

1. the offset of the transformation defines a point within the co-dominance region, close to the level line H3%=50% (i.e. the bottom left boundary of the co-dominance region);
2. points of the same color have reversed positions moving from the left to the right plot. This means that the (sub)type abundances, in general, tend to reverse;
3. The transformation results in a contraction of the points toward the co-dominance region, indicating that the strong dominance of a (sub)type is unlikely to be predicted.

4.7.3 Robustness checks and sensitivity analyses

Alternative computations of the *mixing score* over time

We computed (sub)type *mixing scores* considering alternative Euclidean coordinates for the country-year relative abundances of (sub)types. In particular, we choose the additive log-ratio (*alr*) transformation as an alternative to the isometric log-ratio (*ilr*) transformation [28] for computing the Euclidean coordinates. The distributions of the *mixing scores* over time are highly similar in the two cases (Figure 4.10) and the same four atypical years stand out.

Alternative clusterings of trajectories

We identified countries with similar alternation of (sub)types in the period 2010-2019. The clustering analysis discussed in the manuscript was performed via Ward's linkage algorithm applied on 81 country trajectories in *ilr* coordinates, and the optimal

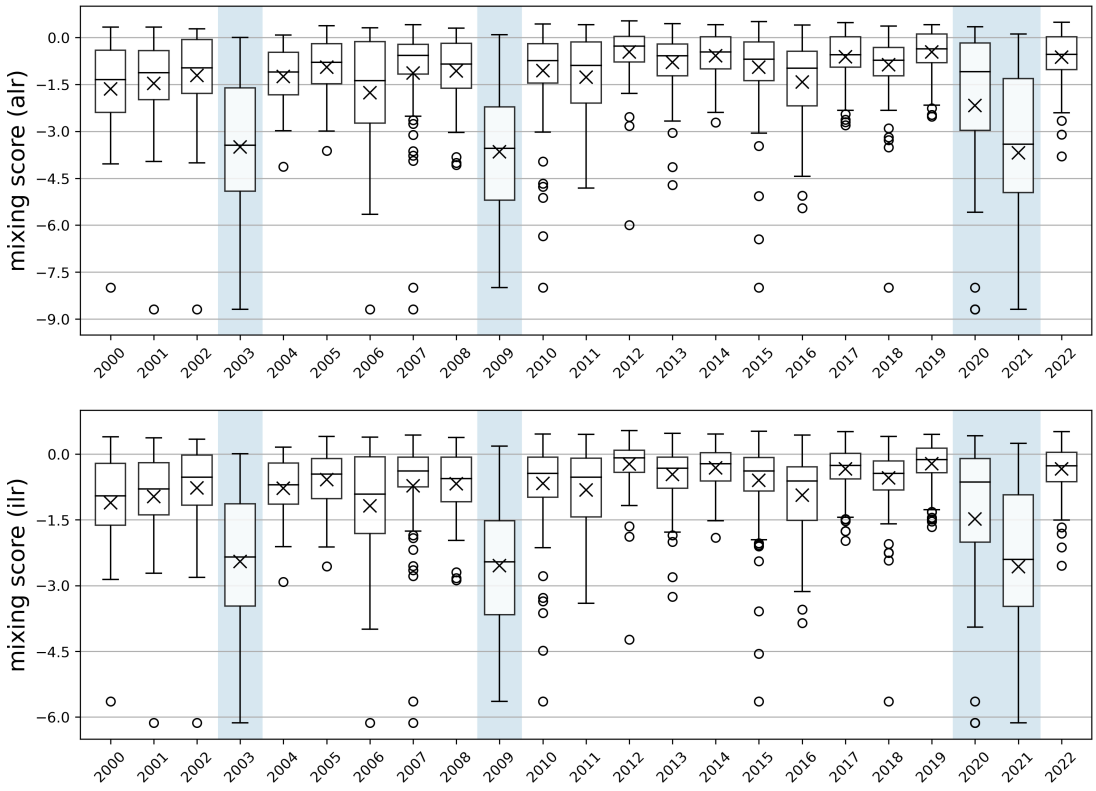


FIGURE 4.10: **Comparison of the distributions of the degree of (sub)type mixing over time, calculated under the *alr* transformation (upper graph) and the *ilr* transformation (lower graph)**

partition resulted in two groups with 39 and 42 countries each (Figure 4.3 and Figure 4.6). The 81 countries were selected because they reported at least 50 cases of flu classified by (sub)type for the whole period. To test the robustness of this classification (hereafter named as the *reference classification*), we repeated the analysis considering different clustering methods, different criteria for data inclusion, and different log-ratio transformations.

Alternative methods

- We applied the weighted linkage hierarchical clustering (see [244]) on the 81 trajectories in the *ilr* coordinates. The algorithm first separated Qatar from all the other 80 countries, which were further split into two clusters of 39 and 41 countries. These latter exactly match the groups of the *reference classification*, except for Mongolia and the Republic of Korea. However, we precise that, according to the Silhouette coefficient, the 2-group partition was better than the 3-group partition, which in turn was better than the subsequent partitions.

- We tested the *k-medoid clustering* on the same trajectories. This algorithm corresponds to the better-known k-mean methods, where barycenters are substituted with medoids - see chp. Partitioning Around Medoids in ([237]). In this case, the 2-group partition was the optimal one, with 39 and 42 countries in each group. Seven countries changed groups with respect to the reference classification: Croatia, Jordan, Mongolia, the Republic of Korea, the Russian Federation, Tunisia, and Turkey.

Alternative criteria for data inclusion

- To understand whether the less reliable estimates of the (sub)type relative abundances had a significant impact on the definition of the clusters, we adopted a more stringent country selection criterion. We considered only the 29 countries that reported at least 500 cases of flu classified by (sub)types for each year from 2010 to 2019 and we applied Ward's linkage clustering. The optimal partition identified two groups. None of the 29 countries changed group with respect to the *reference classification*.

Alternative log-ratio transformation

- The same 81 trajectories were considered in the *alr* coordinates and the Ward's linkage hierarchical clustering was applied. The optimal partition consisted of three groups: the first two groups (39 and 41 countries, respectively) were identical to the groups of the *reference classification*, except Qatar which separated from all the other countries and constituted the third group on its own.

Robustness of model performances for predicting the trajectories of (sub)type abundances

Hereafter we compute alternative metrics to evaluate the model performances in predicting the trajectories of (sub)type abundances. In addition, we repeat predictions for the same 81 trajectories expressed in *alr* coordinates. The results consistently demonstrate the superior performance of the *M4 HVAR* model overall. For Group II countries, it is not the best model, but still one of the top-performing (Table 4.4).

Robustness of model performances for predicting the (i) dominance/non-dominance and the (ii) circulation/non-circulation of one specific (sub)type

Hereafter we compute alternative metrics to evaluate the model performances in predicting the (i) dominance/non-dominance and the (ii) circulation/non-circulation of one specific (sub)type one year ahead. In addition, we repeat predictions for the same 81 trajectories expressed in *alr* coordinates. Results in table 4.5 show that H3 and B are more predictable than H1 and for those (sub)types *M4 HVAR* overtakes the other models (although its supremacy is less clear concerning predictions about B's circulation in Group II countries). For H1 predictions the *M2 average* model seems to perform better, but *M4 HVAR* is still a competitive choice. Results are consistent for predictions on trajectories in *alr* coordinates (Table 4.6).

	group	method	Scores for evaluating predictions of **compositions** ES	DSS	VS05	VS1	Scores for evaluating predictions of the **dominance states** Error Rate	Average Precision (from PR-curve)	AUC ROC
trajectories in the **ILR coordinates**	group I (39)	**M0 frequency past states**	/	/	/	/	0.81	0.32	0.46
		M1 AH1/AH3 alternation	3.85	/	/	/	0.80	0.29	0.45
		M2 average	1.82	31.03	0.99	7.28	0.80	0.31	0.47
		M3 VAR(1)	2.20	5.69	1.41	14.20	0.79	0.33	0.48
		M4 HVAR (p_I=2, p_{II}=1)	**1.46**	**4.42**	**0.84**	**6.43**	**0.68**	**0.38**	**0.60**
	group II (42)	**M0 frequency past states**	/	/	/	/	0.64	0.39	0.58
		M1 AH1/AH3 alternation	2.04	/	/	/	0.67	0.34	0.52
		M2 average	1.35	30.10	0.72	5.25	0.63	**0.42**	**0.59**
		M3 VAR(1)	1.63	**2.42**	1.13	12.10	0.71	0.37	0.53
		M4 HVAR (p_I=2, p_{II}=1)	**1.20**	3.27	**0.66**	**5.22**	**0.60**	0.39	0.57
	all countries (81)	**M0 frequency past states**	/	/	/	/	0.72	0.35	0.53
		M1 AH1/AH3 alternation	2.91	/	/	/	0.73	0.31	0.49
		M2 average	1.58	30.55	0.85	6.22	0.72	0.37	0.56
		M3 VAR(1)	1.90	3.99	1.27	13.11	0.74	0.35	0.53
		M4 HVAR (p_I=2, p_{II}=1)	**1.32**	**3.82**	**0.75**	**5.80**	**0.64**	**0.39**	**0.62**
trajectories in the **ALR coordinates**	group I (39)	**M0 frequency past states**	/	/	/	/	0.81	0.32	0.46
		M1 AH1/AH3 alternation	5.59	/	/	/	0.79	0.29	0.45
		M2 average	2.39	32.13	1.37	11.09	0.80	0.31	0.48
		M3 VAR(1)	2.81	6.79	1.39	13.64	0.79	0.36	0.54
		M4 HVAR (p_I=2, p_{II}=1)	**1.82**	**5.49**	**0.96**	**8.25**	**0.68**	**0.44**	**0.64**
	group II (42)	**M0 frequency past states**	/	/	/	/	0.64	0.39	**0.58**
		M1 AH1/AH3 alternation	2.92	/	/	/	0.69	0.34	0.52
		M2 average	1.91	31.20	**0.63**	**4.26**	**0.63**	**0.40**	0.53
		M3 VAR(1)	2.20	**3.52**	1.17	11.24	0.71	0.36	0.51
		M4 HVAR (p_I=2, p_{II}=1)	**1.63**	4.33	0.70	4.84	**0.63**	0.37	0.57
	all countries (81)	**M0 frequency past states**	/	/	/	/	0.72	0.35	0.53
		M1 AH1/AH3 alternation	4.21	/	/	/	0.74	0.31	0.48
		M2 average	2.14	31.65	0.99	7.55	0.72	0.36	0.56
		M3 VAR(1)	2.49	5.09	1.27	12.39	0.74	0.36	0.57
		M4 HVAR (p_I=2, p_{II}=1)	**1.72**	**4.89**	**0.83**	**6.48**	**0.65**	**0.41**	**0.66**

TABLE 4.4: **Robustness of model performances for predicting the trajectories of (sub)type abundances.** Seven scores (on the columns) are used to evaluate predictions of relative abundances of (sub)types one year ahead. Average scores are computed to compare predictions for Group I countries, Group II countries, and for all countries, by using five different prediction methods. Predictions are calculated for trajectories in both *ilr* and *alr* coordinates (blue and red panel, respectively). Methods performing best by country grouping and score are highlighted. *ES, DSS, VS05, VS1,* and *Error Rate* scores are negatively oriented, meaning that superior performance is indicated by smaller values. In contrast, increasing values of *Average Precision (from the PR curve)* and *AUC ROC* designate an improvement of the model.

		trajectories in the **ILR coordinates**					
		B will be dominant (>50%) in the next season: T/F ?			B will be neglegible (<10%) in the next season: T/F ?		
group	**method**	**Precision**	**Average Precision (from PR-curve)**	**AUC ROC**	**Precision**	**Average Precision (from PR-curve)**	**AUC ROC**
group I (39)	**M0 frequency past states**	0.13	0.28	0.43	0.28	0.35	0.43
	M1 AH1/AH3 alternation	0.00	0.25	0.34	0.17	0.28	0.20
	M2 average	0.00	0.25	0.50	0.25	0.30	0.39
	M3 VAR(1)	0.14	0.22	0.43	0.53	0.51	0.73
	M4 HVAR (p_I=2, p_{II}=1)	**0.27**	**0.32**	**0.66**	**0.86**	**0.84**	**0.89**
group II (42)	**M0 frequency past states**	**0.13**	**0.09**	**0.68**	0.11	0.12	0.50
	M1 AH1/AH3 alternation	0.00	0.05	0.46	0.12	0.12	0.51
	M2 average	0.00	**0.09**	0.65	0.13	0.13	0.52
	M3 VAR(1)	0.07	0.07	0.48	**0.15**	0.14	0.57
	M4 HVAR (p_I=2, p_{II}=1)	0.00	0.05	0.36	0.08	**0.18**	**0.67**
all countries (81)	**M0 frequency past states**	0.13	0.18	0.54	0.17	0.21	0.42
	M1 AH1/AH3 alternation	0.00	0.14	0.41	0.14	0.18	0.29
	M2 average	0.00	0.19	**0.65**	0.18	0.18	0.36
	M3 VAR(1)	0.11	0.15	0.52	0.29	0.26	0.60
	M4 HVAR (p_I=2, p_{II}=1)	**0.26**	**0.23**	0.63	**0.67**	**0.58**	**0.82**
		A\H1N1 will be dominant (>50%) in the next season: T/F ?			A\H1N1 will be neglegible (<10%) in the next season: T/F ?		
group	**method**	**Precision**	**Average Precision (from PR-curve)**	**AUC ROC**	**Precision**	**Average Precision (from PR-curve)**	**AUC ROC**
group I (39)	**M0 frequency past states**	0.29	**0.40**	**0.51**	0.00	0.09	0.36
	M1 AH1/AH3 alternation	0.25	0.31	0.40	0.03	0.08	0.31
	M2 average	**0.33**	0.36	0.42	**0.15**	**0.17**	**0.61**
	M3 VAR(1)	0.11	0.27	0.32	0.00	0.06	0.24
	M4 HVAR (p_I=2, p_{II}=1)	**0.32**	0.32	0.44	0.00	0.06	0.23
group II (42)	**M0 frequency past states**	0.37	0.39	0.48	0.08	0.13	0.45
	M1 AH1/AH3 alternation	**0.47**	0.40	0.58	0.16	0.15	0.57
	M2 average	0.43	0.40	0.54	0.11	0.11	0.39
	M3 VAR(1)	0.34	0.39	0.56	0.06	0.14	0.47
	M4 HVAR (p_I=2, p_{II}=1)	0.31	**0.46**	**0.63**	**0.40**	**0.23**	**0.63**
all countries (81)	**M0 frequency past states**	0.32	**0.38**	0.50	0.07	0.11	0.45
	M1 AH1/AH3 alternation	0.34	0.35	0.49	0.11	0.11	0.48
	M2 average	**0.41**	0.36	0.49	**0.13**	0.11	**0.50**
	M3 VAR(1)	0.23	0.32	0.45	0.04	0.10	0.39
	M4 HVAR (p_I=2, p_{II}=1)	0.31	0.36	**0.52**	0.12	**0.12**	0.45
		A\H3N2 will be dominant (>50%) in the next season: T/F ?			A\H3N2 will be neglegible (<10%) in the next season: T/F ?		
group	**method**	**Precision**	**Average Precision (from PR-curve)**	**AUC ROC**	**Precision**	**Average Precision (from PR-curve)**	**AUC ROC**
group I (39)	**M0 frequency past states**	0.16	0.19	0.38	0.00	0.24	0.43
	M1 AH1/AH3 alternation	0.36	0.26	0.60	0.29	0.29	0.64
	M2 average	0.13	0.19	0.43	0.25	0.19	0.37
	M3 VAR(1)	0.33	0.37	0.70	0.38	0.37	0.74
	M4 HVAR (p_I=2, p_{II}=1)	**0.45**	**0.41**	**0.81**	**0.58**	**0.55**	**0.81**
group II (42)	**M0 frequency past states**	0.32	0.29	0.59	0.00	0.11	0.37
	M1 AH1/AH3 alternation	0.32	0.28	0.58	**0.17**	0.17	**0.63**
	M2 average	0.28	0.28	0.61	0.05	0.10	0.38
	M3 VAR(1)	0.30	0.26	0.54	0.12	0.16	0.53
	M4 HVAR (p_I=2, p_{II}=1)	**0.41**	**0.33**	**0.64**	**0.17**	**0.19**	0.59
all countries (81)	**M0 frequency past states**	0.23	0.22	0.50	0.00	0.15	0.39
	M1 AH1/AH3 alternation	0.33	0.27	0.59	0.24	0.23	0.64
	M2 average	0.22	0.24	0.54	0.07	0.14	0.39
	M3 VAR(1)	0.32	0.29	0.61	0.24	0.24	0.66
	M4 HVAR (p_I=2, p_{II}=1)	**0.43**	**0.37**	**0.74**	**0.44**	**0.38**	**0.75**

TABLE 4.5: Robustness of model performances for predicting the (i) dominance/non-dominance and the (ii) circulation/non-circulation of one specific sub)type. Trajectories are expressed in *ilr* coordinates.

		trajectories in the **ALR coordinates**					
		B will be dominant (>50%) in the next season: T/F ?			B will be neglegible (<10%) in the next season: T/F ?		
group	**method**	**Precision**	**Average Precision (from PR-curve)**	**AUC ROC**	**Precision**	**Average Precision (from PR-curve)**	**AUC ROC**
group I (39)	**M0 frequency past states**	0.13	0.28	0.43	0.28	0.35	0.43
	M1 AH1/AH3 alternation	0.00	0.25	0.34	0.18	0.28	0.21
	M2 average	0.00	0.27	0.54	0.19	0.29	0.35
	M3 VAR(1)	0.14	0.34	0.66	0.62	0.68	0.81
	M4 HVAR (p_I=2, p_{II}=1)	**0.23**	**0.51**	**0.77**	**0.93**	**0.87**	**0.92**
group II (42)	**M0 frequency past states**	**0.13**	**0.09**	**0.68**	0.11	0.12	0.50
	M1 AH1/AH3 alternation	0.03	0.05	0.46	0.13	0.12	0.52
	M2 average	0.00	0.06	0.41	0.12	0.13	0.51
	M3 VAR(1)	0.07	0.05	0.42	**0.15**	0.15	0.57
	M4 HVAR (p_I=2, p_{II}=1)	0.00	0.06	0.45	0.00	**0.16**	**0.62**
all countries (81)	**M0 frequency past states**	0.13	0.18	0.54	0.17	0.21	0.42
	M1 AH1/AH3 alternation	0.02	0.14	0.38	0.15	0.19	0.33
	M2 average	0.00	0.21	0.65	0.14	0.17	0.34
	M3 VAR(1)	0.11	0.23	0.67	0.28	0.32	0.64
	M4 HVAR (p_I=2, p_{II}=1)	**0.22**	**0.40**	**0.78**	**0.80**	**0.61**	**0.83**
		A\H1N1 will be dominant (>50%) in the next season: T/F ?			A\H1N1 will be neglegible (<10%) in the next season: T/F ?		
group	**method**	**Precision**	**Average Precision (from PR-curve)**	**AUC ROC**	**Precision**	**Average Precision (from PR-curve)**	**AUC ROC**
group I (39)	**M0 frequency past states**	0.29	**0.40**	0.51	0.00	0.09	0.36
	M1 AH1/AH3 alternation	0.28	0.32	0.43	0.04	0.08	0.34
	M2 average	0.33	0.35	0.44	**0.11**	**0.17**	**0.60**
	M3 VAR(1)	0.11	0.28	0.38	0.00	0.06	0.22
	M4 HVAR (p_I=2, p_{II}=1)	**0.36**	0.38	**0.53**	0.00	0.07	0.26
group II (42)	**M0 frequency past states**	0.37	0.39	0.48	0.08	0.13	0.45
	M1 AH1/AH3 alternation	**0.50**	0.43	0.61	0.18	0.17	0.61
	M2 average	0.43	0.41	0.55	0.11	0.11	0.39
	M3 VAR(1)	0.34	0.40	0.56	0.09	0.14	0.48
	M4 HVAR (p_I=2, p_{II}=1)	0.43	**0.45**	**0.63**	**0.50**	**0.28**	**0.64**
all countries (81)	**M0 frequency past states**	0.32	0.38	0.50	0.07	0.11	0.45
	M1 AH1/AH3 alternation	0.37	0.36	0.52	**0.12**	**0.12**	**0.51**
	M2 average	**0.41**	0.37	0.50	0.11	0.11	0.47
	M3 VAR(1)	0.23	0.34	0.48	0.06	0.10	0.39
	M4 HVAR (p_I=2, p_{II}=1)	0.38	**0.40**	**0.57**	0.11	**0.12**	0.47
		A\H3N2 will be dominant (>50%) in the next season: T/F ?			A\H3N2 will be neglegible (<10%) in the next season: T/F ?		
group	**method**	**Precision**	**Average Precision (from PR-curve)**	**AUC ROC**	**Precision**	**Average Precision (from PR-curve)**	**AUC ROC**
group I (39)	**M0 frequency past states**	0.16	0.19	0.38	0.00	0.24	0.43
	M1 AH1/AH3 alternation	0.32	0.24	0.57	0.28	0.28	0.62
	M2 average	0.13	0.18	0.42	0.00	0.21	0.40
	M3 VAR(1)	0.33	0.34	0.68	0.33	0.38	0.72
	M4 HVAR (p_I=2, p_{II}=1)	**0.48**	**0.41**	**0.80**	**0.67**	**0.55**	**0.82**
group II (42)	**M0 frequency past states**	**0.32**	0.29	0.59	0.00	0.11	0.37
	M1 AH1/AH3 alternation	0.31	0.25	0.52	0.13	0.13	0.52
	M2 average	0.28	0.26	0.56	0.00	0.10	0.38
	M3 VAR(1)	0.30	0.25	0.52	0.13	0.15	0.55
	M4 HVAR (p_I=2, p_{II}=1)	0.29	**0.30**	**0.61**	**0.33**	**0.18**	**0.58**
all countries (81)	**M0 frequency past states**	0.23	0.22	0.50	0.00	0.15	0.39
	M1 AH1/AH3 alternation	0.31	0.24	0.54	0.20	0.20	0.57
	M2 average	0.22	0.23	0.51	0.00	0.14	0.41
	M3 VAR(1)	0.32	0.27	0.60	0.24	0.26	0.66
	M4 HVAR (p_I=2, p_{II}=1)	**0.42**	**0.35**	**0.71**	**0.56**	**0.40**	**0.75**

TABLE 4.6: Robustness of model performances for predicting the (i) dominance/non-dominance and the (ii) circulation/non-circulation of one specific sub)type. Trajectories are expressed in *alr* coordinates.

Chapter 5

Tree-based conditional copula estimation

This Chapter is based on the paper *Tree-based conditional copula estimation* [32] which has been recently submitted. This work was carried out under the supervision of my PhD advisors Olivier Lopez (Ensae IP Paris) and Maud Thomas (Sorbonne Université).

The code used for the simulations and the real-data application is publicly available at https://github.com/FrancescoBonacina/tree-based-conditional-copula-estimation.

5.1 Abstract

This paper proposes a regression tree procedure to estimate conditional copulas. The associated algorithm determines classes of observations based on covariate values and fits a simple parametric copula model on each class. The association parameter changes from one class to another, allowing for non-linearity in the dependence structure modeling. It also allows the definition of classes of observations on which the so-called "simplifying assumption" [161] holds reasonably well. When considering observations belonging to a given class separately, the association parameter no longer depends on the covariates according to our model. In this paper, we derive asymptotic consistency results for the regression tree procedure and show that the proposed pruning methodology, that is the model selection techniques selecting the appropriate number of classes, is optimal in some sense. Simulations provide finite sample results and an analysis of data of cases of human influenza presents the practical behavior of the procedure.

5.2 Introduction

Since Sklar's seminal result, copula theory has emerged as a practical means of describing the dependence between random variables. Allowing one to distinguish between the marginal behavior of each component of a random vector and the dependence structure (represented by a copula function), Sklar's theorem opens the way to flexible modeling of various forms of dependence (see [29]). In this paper, we propose a new method to perform conditional copula analysis based on regression trees and derive consistency results for this procedure.

Various estimation procedures and analyses of copulas have been studied in the statistical literature [245, 246, 247, 248, 249]. In the presence of covariates, conditional copula analysis consists of fitting a copula function to the conditional distribution of a random vector. From an application point of view, Dupuis and colleagues [250] have shown their importance in modeling certain natural disasters such as hurricanes, or

the dependence between different expense lines in actuarial problems. Lopez and co-authors [251] and Farkas and co-authors [252] have used this type of model for insurance claim management. Another important application, for example in finance, can be found in [253]. More generally, the study of conditional copulas also appears particularly important in Vine copulas [254]. Previous studies [152, 157, 255] have studied both semi-parametric and non-parametric procedures for performing this analysis. Finally, Fermian and Lopez [256] have examined the case of high-dimensional covariates and relied on a dimension reduction approach to perform the analysis.

We propose here to use regression trees to perform this conditional copula analysis. Regression trees, along with the *Classification And Regression Tree* (CART) algorithm, were originally introduced by [124] and are now classic tools used for several applications (e.g. see [257, 258, 259]. Apart from the computational efficiency of the CART algorithm, an interesting feature of this approach is the ability to construct classes of individuals (based on their characteristics) with similar behavior. In the context of copula analysis, this corresponds to classes of individuals with the same copula (i.e. dependence) structure. This model can be seen as a means to easily generalize the "simplifying assumption" considered by many authors (see for example [161, 260]). According to this hypothesis, only the marginal distributions of each component depend on the covariates, while the dependence structure does not vary with them. In contrast, in our model, the copulas are different for each cluster determined by the regression tree, and thus the simplifying assumption holds separately for each cluster.

The rest of the paper is organized as follows. In Section 5.3, we describe the general framework of the regression trees and the algorithm used to fit them to the data. Section 5.4 is devoted to proving the theoretical results on the consistency of this procedure. Particular attention is paid to the part of model selection, known as the "pruning step", which consists of selecting an appropriate sub-tree from the maximal tree obtained by iterative partitioning of the data set. In Section 5.5, the practical behavior of the model is investigated through a simulation study and a real data analysis. The proofs of the theoretical results are gathered in the Appendix.

5.3 Regression trees for conditional copula analysis

5.3.1 Model and notations

We consider a set of observations $(\mathbf{Y}_i, \mathbf{X}_i)_{1\leq i\leq n}$ consisting of independent identically distributed copies of the random vector $(\mathbf{Y}, \mathbf{X})$, where $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ are covariates, and $\mathbf{Y} = (Y^{(1)}, \ldots, Y^{(k)}) \in \mathbb{R}^k$ is a random vector of response variables $Y^{(j)}, j = 1, \ldots, k$. The marginal conditional cumulative distribution functions (c.d.f.) of the random vector $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ are defined as

$$F^{(j)}(t^{(j)}|\mathbf{x}) = \mathbb{P}\left(Y^{(j)} \leq t^{(j)}|\mathbf{X} = \mathbf{x}\right), \quad t^{(j)} \in \mathbb{R},\ j = 1, \ldots, k.$$

From Sklar's Theorem [261], the joint conditional c.d.f. $F(\mathbf{t}|\mathbf{x}) = \mathbb{P}(\mathbf{Y} \leq \mathbf{t}|\mathbf{X} = \mathbf{x})$ can be expressed as

$$F(\mathbf{t}|\mathbf{x}) = \mathfrak{C}_{\mathbf{x}}(F^{(1)}(t^{(1)}|\mathbf{x}), \ldots, F^{(k)}(t^{(k)}|\mathbf{x})), \text{ for all } \mathbf{t} = (t^{(1)}, \ldots, t^{(k)}) \in \mathbb{R}^k. \tag{5.1}$$

Where, for all $\mathbf{x}$, $\mathfrak{C}_{\mathbf{x}}$ is a copula function, that is a c.d.f. on $[0,1]^k$ with margins uniformly distributed over $[0,1]$. The copula function $\mathfrak{C}_{\mathbf{x}}$ in (5.1) is unique if the distribution of $\mathbf{Y}$ is continuous, which is the assumption that we will make throughout the paper. In

general, the analyses of the marginal distributions and the dependence structure are therefore made separately.

In the following, we will consider a semi-parametric assumption as in [152] or [251] by introducing a parametric family of copula functions $\mathcal{C} = \{C_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ with $\Theta \subset \mathbb{R}^m$. We denote $c_{\boldsymbol{\theta}}$ the copula density associated with $C_{\boldsymbol{\theta}}$, that is

$$c_{\boldsymbol{\theta}}(\mathbf{u}) = \frac{\partial^k C_{\boldsymbol{\theta}}(\mathbf{u})}{\partial u^{(1)} \dots \partial u^{(k)}}, \quad \boldsymbol{u} = (u^{(1)}, \dots, u^{(k)}) \in [0,1]^k.$$

In the sequel, we assume that, for all $\mathbf{x} \in \mathbb{R}^d$, $\mathfrak{C}_{\mathbf{x}} \in \mathcal{C}$, meaning that there exists a unique $\boldsymbol{\theta}^0(\mathbf{x}) \in \Theta$ such that

$$\mathfrak{C}_{\mathbf{x}} = C_{\boldsymbol{\theta}^0(\mathbf{x})}. \tag{5.2}$$

Our aim is then to retrieve the function $\boldsymbol{\theta}^0(\mathbf{x})$ from the data $(\mathbf{Y}_i, \mathbf{X}_i)_{1 \leq i \leq n}$.

Our estimation strategy is based on regression trees. A tree $\mathbb{T}$ of size K is a partition of $\mathcal{X}$, that is, $\mathbb{T} = (\mathcal{T}_\ell)_{\ell=1,\dots,K}$ where $\mathcal{T}_\ell \cap \mathcal{T}_{\ell'} = \emptyset$ for $\ell \neq \ell'$ and $\cup_{\ell=1}^K \mathcal{T}_\ell = \mathcal{X}$. The sets $\mathcal{T}_{\ell=1,\dots,K}$ are called leaves, and each leaf $\mathcal{T}_\ell$ is obtained as the intersection of conditions of the type $x^{(j)}_{-,\ell} \leq x^{(j)} \leq x^{(j)}_{+,\ell}$ if $X^{(j)}$ is continuous, and of the type $x^{(j)} \in \mathcal{A}^{(j)}_\ell$ where $\mathcal{A}^{(j)}_\ell$ is a set of potential modalities for a discrete covariate. This particular structure of the partition is associated with a binary tree structure, where the nodes of the tree correspond to conditions on a given covariate and the leaves of the tree to the final classification. The CART algorithm described in Section 5.3.2 will make this tree structure more obvious.

Given a tree $\mathbb{T}$ with K leaves, we thus consider the estimator of $\boldsymbol{\theta}^0(\mathbf{x})$ to be constant on each leaf of $\mathbb{T}$, that is, of the type $\sum_{\ell=1}^K \theta_\ell \mathbf{1}_{\mathcal{T}_\ell}(\mathbf{x})$, with $\theta_\ell \in \mathbb{R}^m$. In other words, individuals are divided into K classes, for each of which the dependence structure is described by a different copula (from the same parametric family, but with a specific parameter θ_ℓ). In the ideal case, the target function $\boldsymbol{\theta}^0(\mathbf{x})$ is constant on each leaf of the tree $\mathbb{T}$, meaning that $\boldsymbol{\theta}^0(\mathbf{x}) = \theta^0_\ell$ for $\mathbf{x} \in \mathcal{T}_\ell$, where

$$\theta^0_\ell = \arg\max_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\left[\log c_{\boldsymbol{\theta}}(\mathbf{U}_i) \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell}\right].$$

$c_{\boldsymbol{\theta}}$ is the copula density associated with the copula function $C_{\boldsymbol{\theta}}$, and $\mathbf{U}_i$ is the random variable defined by

$$\mathbf{U}_i = (F^{(1)}(Y_i^{(1)}|\mathbf{X}_i), \dots, F^{(k)}(Y_i^{(k)}|\mathbf{X}_i)),$$

which has uniform margins over $[0,1]$, and is jointly distributed according to the c.d.f. $\mathfrak{C}_{\mathbf{X}_i} = C_{\boldsymbol{\theta}^0(\mathbf{X}_i)}$.

However, in practice, a misspecification bias is expected, since the target function $\boldsymbol{\theta}^0(\mathbf{x})$ is not a piecewise constant function while the estimator function is. For a given tree $\mathbb{T}$, the corresponding estimator $\widehat{\boldsymbol{\theta}}(\mathbf{x}|\mathbb{T})$ is defined as

$$\widehat{\boldsymbol{\theta}}(\mathbf{x}|\mathbb{T}) = \sum_{\ell=1}^K \widehat{\theta}_\ell \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell},$$

where

$$\widehat{\theta}_\ell = \arg\max_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^n \log c_{\boldsymbol{\theta}}(\widehat{\mathbf{U}}_i) \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell},$$

and $(\widehat{\mathbf{U}}_i)_{1 \leq i \leq n}$ are pseudo-observations, that is the estimated versions of $(\mathbf{U}_i)_{1 \leq i \leq n}$.

Typically, these pseudo-observations are the result of a preliminary estimation of the marginal distribution, namely $\widehat{U}_i^{(j)} = \widehat{F}^{(j)}(Y_i^{(j)}|\mathbf{X}_i)$, but alternative procedures are

possible. For example, a parametric model can be used to handle the margins. In Section 5.3.3, we also discuss the possibility of relying on tree-based methods to estimate the margins as well, although there is no obligation to use the same type of technique for the dependence structure as for the margins. Therefore, in the following, we will try to keep our results as general as possible, expressing convergence conditions that this step should verify, but without imposing a specific method. However, let us point out that an interesting feature of regression trees is their ability to deal with both quantitative and qualitative covariates, which requires relying on estimation techniques for the margins that satisfy the same requirements.

The rest of the section is devoted to presenting our estimation procedure based on regression trees. We describe the CART procedure consisting of two steps. First, the construction of the maximal tree (Section 5.3.2), which determines the proper decomposition of the covariate space $\mathcal{X}$ to obtain the regression tree $\mathbb{T}$ and deduce an estimator $\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T})$. Second, the pruning step (Section 5.3.2), which corresponds to a selection model step. However, fitting the dependence structure requires a preliminary estimation of the margins, which is done once and for all before starting the algorithm. Various methods may be used to deal with this preliminary step, the only requirement being that they satisfy the conditions under which our theoretical results hold. Examples of possible methods to estimate the margins are presented in Section 5.3.3.

5.3.2 Regression tree estimation of the dependence structure

Regression trees provide an easy and transparent way to group observations that have similar behavior in terms of the response variable Y. They constitute a nonparametric regression model capable of reproducing highly nonlinear trends in the data and are thus able to approximate a wide class of functions. In addition, they can include both quantitative and categorical (non-ordinal) covariates.

Originally proposed by [124], regression trees are implemented through the CART algorithm, which involves a two-step process. Initially, a maximal tree is constructed, forming a binary structure that assigns observations to numerous classes (leaves), often leading to overfitting. Subsequently, the maximal tree is pruned to identify the subtree that offers the best compromise between complexity and generalization ability.

Section 5.3.2 describes how the construction of the optimal tree takes place, making explicit our split criterion based on the maximization of the log-likelihood of the copula mixture model. In Section 5.3.2, we define the penalization criterion and discuss the pruning phase.

Construction of the maximal tree

Recall that, as mentioned before, the computation of the pseudo-observations $\widehat{\mathbf{U}}_i$ is done once and for all before starting the algorithm. Then, the CART procedure is applied to $(\widehat{\mathbf{U}}_i, \mathbf{X}_i)_{1\leq i\leq n}$, with the aim of maximize the log-likelihood of the copula mixture model. Such log-likelihood function can be written as the sum of the log-likelihoods of the parametric copulas estimated for the individual leaves of the tree:

$$\mathcal{L}_n(\theta_1, \ldots, \theta_K) = \sum_{\ell=1}^{K} \Big(\frac{1}{n} \sum_{i=1}^{n} \log c_{\theta_\ell}(\widehat{\mathbf{U}}_i) \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_\ell}\Big).$$

More precisely, the log-likelihood of the model is maximized conditionally on the covariates as a consequence of the recursive partitioning of the observations. In fact, at

each split the observations are separated by looking at their values for one of the covariates. Formally, if we denote by $D_P = (\widehat{\mathbf{U}}_i, \mathbf{X}_i)_{i \in P}$ the observations that belong to a certain node P (parent) and by $R_P(\mathbf{X})$ the condition of the covariates that identifies those observations—such that $R_P(\mathbf{X}_i)$ is 1 if $(\widehat{\mathbf{U}}_i, \mathbf{X}_i) \in D_P$, 0 otherwise - then the left and right child nodes are determined by conditions of the type $\{X_i^{(j)} \leq s, (\widehat{\mathbf{U}}_i, \mathbf{X}_i) \in D_P\}$ and $\{X_i^{(j)} > s, (\widehat{\mathbf{U}}_i, \mathbf{X}_i) \in D_P\}$. Such split is uniquely determined by the pair (j, s), with $j = 1, \ldots, p$ and $s \in \mathbb{R}$. (This is true for quantitative covariates, while in the presence of qualitative covariates, the split is performed following Remark 1.)

Initially, all the observations are in the root of the tree, implying that the dependence among the variables $\widehat{\mathbf{U}}_i$ is modeled by a single copula with parameters $\boldsymbol{\theta}^0_{root}$, which are optimized using maximum likelihood estimation (MLE). Subsequently, each split is carried out to maximize the increase in the model log-likelihood. This gain simply corresponds to the sum of the log-likelihoods estimated for the child nodes—once more, evaluated in correspondence with the parameters optimized via MLE—from which the log-likelihood of the parent node is subtracted. In practical terms, the optimal gain, and thus the optimal split, is determined by testing all the possible splits. The splitting process ends when further splits fail to improve the log-likelihood, or more commonly, upon meeting specific stopping criteria. For instance, a common criterion is setting a minimum number of observations per leaf.

The pseudocode summarizing the construction of the maximal tree is presented in Algorithm 1.

Remark 1 *The implementation of the CART algorithm illustrated in example 1 requires quantitative (or binary) covariates, for which an ordering of values is straightforward. For qualitative variables with $M > 2$ modalities, the algorithm should include an ordering step preliminary to the split research, as suggested in [132]. Specifically, first, the modalities are sorted by increasing values of the $\widehat{\theta}$ parameter estimated by considering the observations associated with each modality. Then, the $M - 1$ possible splits are evaluated and the optimal one is identified. An example of this procedure is available in the code we implemented for the application on the human influenza data (5.5.2).*

Pruning step

Obtaining the maximal tree from the CART algorithm is not sufficient to have a proper estimation of the objective function $\boldsymbol{\theta}^0$, since this decomposition leads to overfitting. A subtree must be extracted from this maximal tree. This subtree will achieve a proper compromise between goodness of fit and complexity.

The complexity is here measured in terms of the number of leaves of a given tree $\mathbb{T}$. The selected subtree is thus obtained through the maximization of the following penalized criterion,

$$\bar{\boldsymbol{\theta}}(\mathbf{x}) = \arg\max_{\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T})} \frac{1}{n} \sum_{i=1}^{n} \log c_{\widehat{\boldsymbol{\theta}}(\mathbf{X}_i|\mathbb{T})}(\widehat{\mathbf{U}}_i) - \lambda dim(\mathbb{T}), \tag{5.3}$$

where the arg max is taken over all subtrees $\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T})$ of $\mathbb{T}$, and $dim(\mathbb{T})$ is the number of leaves of $\mathbb{T}$. This criterion could give the impression that one needs to consider all the possible subtrees within the maximal tree, and then select the optimal one. Fortunately, the particular shape of the penalty in (5.3) ensures that the best tree with K leaves (according to this criterion) is a subtree of the best tree with $K + 1$ leaves [124]. This selection is then performed through validation on a test sample or cross-validation.

Algorithm 1: Construction of the maximal tree

Data: $\mathsf{D} \leftarrow (\mathbf{X}_i, \widehat{\mathbf{U}}_i)_{i=1,\dots,n}$

function `StoppingCriteria(D)`:
- `#Define conditions to stop tree growth`
- `#For example limit the minimum observations per leaf`
- **return** true *if stopping criteria met, otherwise* false

function `FindOptimalSplit(`D_P`)`:
- `#Initialization`
- best_gain, best_j, best_s $\leftarrow (0, -999, -999)$
- `#Grid search over all the possible features and split values`
- **for** *each possible* (j, s) **do**
  - $D_\ell \leftarrow D_P[X^j \leq s]$
  - $D_R \leftarrow D_P[X^j > s]$
  - gain $\leftarrow$ `LogL(`D_ℓ`)` $+$ `LogL(`D_R`)` $-$ `LogL(`D_P`)`
  - **if** gain > best_gain **then**
    - best_gain, best_j, best_s $\leftarrow$ (gain, j, s)
  - **end**
- **end**
- **return** *(*best_j, best_s*)*

function `BuildTree(D)`:
- `#Initialization`
- $R_{root} \leftarrow \{\forall \mathbf{X}\}$
- ListRulesInternalNodes $\leftarrow [R_{root}]$
- ListRulesLeaves $\leftarrow [\ \ ]$
- `#Tree construction`
- **while** `size(`ListRulesInternalNodes`)`>0 **do**
  - `#Retrieve the rule and the observations of the parent node to be split`
  - $R_P \leftarrow$ ListRulesInternalNodes [0]
  - $D_P \leftarrow$ D [R_P(D)]
  - **if** `StoppingCriteria(`D_P`)` **then**
    - `#Move the rule defining this node in the list of the leaves`
    - ListRulesInternalNodes $\leftarrow$ `RemoveItem(`ListRulesInternalNodes, R_P`)`
    - ListRulesLeaves $\leftarrow$ `AddItem(`ListRulesLeaves, R_P`)`
  - **else**
    - `#Find the optimal split and compute the rules defining the left/right children`
    - $(j^*, s^*) \leftarrow$ `FindOptimalSplit(`D_P`)`
    - $R_\ell \leftarrow R_P \wedge \{X^{j^*} \leq s^*\}$
    - $R_R \leftarrow R_P \wedge \{X^{j^*} > s^*\}$
    - `#Replace the rule of the parent node with the ones of its children`
    - ListRulesInternalNodes $\leftarrow$ `RemoveItem(`ListRulesInternalNodes, R_P`)`
    - ListRulesInternalNodes $\leftarrow$ `AddItem(`ListRulesInternalNodes, R_ℓ`)`
    - ListRulesInternalNodes $\leftarrow$ `AddItem(`ListRulesInternalNodes, R_R`)`
  - **end**
- **end**
- **return** ListRulesLeaves

5.3.3 Estimation of the margins

Let us consider a given margin $Y^{(j)}$. If the components of $\mathbf{X}$ are all continuous covariates, a simple non-parametric estimator can be obtained via, for example, kernel smoothing. Following [255], it writes

$$\widehat{F}^{(j)}(t^{(j)}|\mathbf{x}) = \frac{\sum_{i=1}^{n} K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right) \mathbf{1}_{Y_i^{(j)} \leq t^{(j)}}}{\sum_{i=1}^{n} K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right)}, \tag{5.4}$$

where $h > 0$ is the bandwidth and K the kernel smoother. Several choices are possible for K (e.g. see [157]). In general, $K(\mathbf{x}) = \prod_{j=1}^{d} k(x^{(j)})$, with k a positive function such that $\int k(u)\mathrm{d}u = 1$. As it is classical for kernel estimators, the rate of uniform convergence is $O(h^2 + [\log n]^{1/2} n^{-1/2} h^{-d/2})$ [262], where h^2 corresponds to the bias term.

On the other hand, this estimator is not valid if $\mathbf{X}$ contains some qualitative components. In this perspective, consider the case where $\mathbf{X}$ has M modalities, then, a possible non-parametric estimator is

$$\widehat{F}^{(j)}(t^{(j)}|\mathbf{x}) = \frac{1}{n_{\mathbf{x}}} \sum_{i=1}^{n} \mathbf{1}_{Y_i^{(j)} \leq t^{(j)}} \mathbf{1}_{\mathbf{X}_i = \mathbf{x}},$$

where $n_{\mathbf{x}} = \sum_{i=1}^{n} \mathbf{1}_{\mathbf{X}_i = \mathbf{x}}$. Note that since the covariates are assumed to be random, so is $n_{\mathbf{x}}$. If $n_{\mathbf{x}}$ were not random, the rate of convergence would be typically the same as for an empirical c.d.f., that is $n_{\mathbf{x}}^{-1/2}$.

However, this approach quickly reaches its limits since, when M is large, the number of observations such that $\mathbf{X}_i = \mathbf{x}$ becomes quite small, diminishing considerably the rate of convergence. An alternative is to build classes of modalities, that is decomposing the set of covariates into $m < M$ modalities, as is the case if regression trees are also applied to the margins. Consider this decomposition into m modalities, and let $\mathcal{M}(\mathbf{x})$ denote the class to which $\mathbf{x}$ belongs to, and

$$\widehat{F}^{(j)}(t^{(j)}|\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbf{1}_{Y_i^{(j)} \leq t^{(j)}} \mathbf{1}_{\mathbf{X}_i \in \mathcal{M}(\mathbf{x})}}{\sum_{i=1}^{n} \mathbf{1}_{\mathbf{X}_i \in \mathcal{M}(\mathbf{x})}}.$$

In this case, a bias term appears, since $\widehat{F}^{(j)}(t^{(j)}|\mathbf{x})$ converges towards $\mathbb{P}(Y^{(j)} \leq t^{(j)} | \mathcal{M}(\mathbf{X}) = \mathcal{M}(\mathbf{x}))$.

An alternative way to proceed is also to consider a parametric model for the margins, like, for example, Generalized Linear Models [263]. In this case, and under proper assumptions, the convergence rate in the estimation of the margins can become $n^{-1/2}$, up to a strong assumption on the distributions.

Due to the variety of possible approaches in estimating the margins, we will keep the rest of the paper as general as possible regarding this point, only requiring generic convergence assumptions on this preliminary step.

5.4 Consistency results

This section is dedicated to presenting the main theoretical results that validate the asymptotic behavior of the copula tree estimation procedure. We gather and discuss the list of assumptions required to obtain these results in Section 5.4.1. Moving on to Section 5.4.2, we explore the consistency of a single tree (with a given number of leaves K that may tend to infinity). We chose to focus on the stochastic part of the

error, while the approximation error is expected to decrease with K. The rate of the decrease depends on the specific shape of the target function $\theta^0(\mathbf{x})$, which remains an open problem in the regression tree literature. In Section 5.4.3, we investigate the ability of the penalized criterion to achieve a similar performance as if the optimal number of leaves were known.

5.4.1 Conditions and assumptions

First, Assumption 3 controls the rate of consistency of the pseudo-observations, which represents the c.d.f. of the margins.

Assumption 3 *Assume that*

$$\sup_{\substack{i=1,\dots,n \\ j=1,\dots,k}} \left| \frac{U_i^{(j)}}{\widehat{U}_i^{(j)}} + \frac{1-U_i^{(j)}}{1-\widehat{U}_i^{(j)}} \right| = O_P(1)\,. \tag{5.5}$$

For some $0 < \alpha < 1/2$,

$$\sup_{\substack{i=1,\dots,n \\ j=1,\dots,k}} \left| \frac{\widehat{U}_i^{(j)} - U_i^{(j)}}{\left[U_i^{(j)}(1-U_i^{(j)})\right]^{\alpha}} \right| = O_P(\varepsilon_n), \tag{5.6}$$

for some sequence ε_n *that tends to 0 as* n *tends to infinity.*

Assumption 5.5 is here to control the behavior of the pseudo-observations near the border of $[0,1]^k$. If the margins are estimated via empirical distribution functions, this assumption easily holds from Remark *ii* in [264]. In case this assumption would not hold for more complex estimators, it can be avoided through the introduction of trimming, i.e., by removing those points too close to the boundaries of the unit square. This would introduce a bias that can be controlled through (5.6) (see remark 2).

As it will appear in the theoretical results of Sections 5.4.2 and 5.4.3, this rate is expected to go faster to zero than the part related to the estimation of the tree itself, otherwise it will be predominant. It is important to note that (5.6) is similar to the slightly stronger condition

$$\sup_{\substack{t\in\mathbb{R} \\ j=1,\dots,d}} \left| \frac{\widehat{F}^{(j)}(t|\mathbf{x}) - F^{(j)}(t|\mathbf{x})}{\left[F^{(j)}(t|\mathbf{x})(1-F^{(j)}(t|\mathbf{x}))\right]^{\alpha}} \right| = O_P(\varepsilon_n).$$

If we consider the estimation of the (unconditional) c.d.f. $F^{(j)}(t) = \mathbb{P}(Y^{(j)} \leq t)$ by the empirical distribution function, this condition is easily satisfied with $\varepsilon_n = n^{-1/2}$ (see Example 19.12 in [131]). In the case of the kernel-based estimator, Section 5.7.5 shows that the rate is slower, namely $\varepsilon_n = (h^2 + [\log n]^{1/2} n^{-1/2} h^{-d/2})$, and, in the case of discrete covariates, Section 5.7.6 shows that $\varepsilon_n = n^{-1/2}$.

Before presenting the rest of the assumptions, we introduce two conditions on classes of functions which will be necessary in the following.

Condition 4 *A class of functions* $\mathcal{F} = \{\mathbf{u} \mapsto \varphi_{\boldsymbol{\theta}}(\mathbf{u}) : \boldsymbol{\theta} \in \Theta\} \subset L^2(\mathbb{R}^k)$ *(for some* $k > 0$*) is said to satisfy Condition 4 if*

$$|\varphi_{\boldsymbol{\theta}}(\mathbf{u}) - \varphi_{\boldsymbol{\theta}'}(\mathbf{u})| \leq B(\mathbf{u}) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_1, \quad \mathbf{u} \in [0,1]^k$$

where B is a function in $\mathbb{R}^k$ such that $\mathbb{E}[B(\mathbf{U})^2] < \infty$.

For such a class, there exists an envelope function, that is a function Φ such that, for all $\boldsymbol{\theta} \in \Theta$, $|\phi_{\boldsymbol{\theta}}(\mathbf{u})| \leq \Phi(\mathbf{u})$ and $\mathbb{E}[\Phi(\mathbf{U})^2] < \infty$. Taking any point $\widetilde{\boldsymbol{\theta}} \in \Theta$, Φ can be chosen as $\Phi(\mathbf{u}) = \varphi_{\widetilde{\boldsymbol{\theta}}}(\mathbf{u}) + diam(\Theta)B(\mathbf{u})$, where $diam(\Theta)$ denotes the diameter of the compact set Θ.

Condition 5 *A class of functions $\mathcal{F} = \{\mathbf{u} \mapsto \varphi_{\boldsymbol{\theta}}(\mathbf{u}) : \boldsymbol{\theta} \in \Theta\} \subset L^2(\mathbb{R}^k)$ (for some $k > 0$) is said to satisfy Condition 5 if*

1. *there exist an envelope Φ and a universal constant A_1 such that, for all $\varphi \in \mathcal{F}$,*

$$|\varphi(\mathbf{u})| \leq \Phi(\mathbf{u}) \leq A_1 \sum_{r=1}^{k} \frac{1}{\{u^{(r)}[1-u^{(r)}]\}^{\beta_1}}, \quad \mathbf{u} \in [0,1]^k,$$

   with $0 \leq \beta_1 < 1/2$.

2. *there exists a universal constant A_2 such that for all $\varphi \in \mathcal{F}$,*

$$|\partial_j \varphi(\mathbf{u})| \leq \frac{A_2}{\{u^{(j)}[1-u^{(j)}]\}^{\beta_2}} \sum_{r=1}^{d} \frac{1}{\{u^{(r)}[1-u^{(r)}]\}^{\beta_3}},$$

   with $0 \leq \beta_2 \leq 1$, $\beta_3 < 1/2$, and where ∂_j denotes the partial derivative with respect to the $j-th$ component of $\mathbf{u}$.

These conditions allow controlling the complexity of the class of functions and are related to classical assumptions used for the consistency of classical maximum likelihood estimators (see [130]). The second condition is required to control the behavior of the copula log-likelihood and of its derivatives close to the boundaries of $[0,1]^d$. These conditions are similar to the one used in [246, 265]. They hold for many classical classes of copula functions, like the Gaussian, Clayton, Frank, and Gumbel families.

We then consider the two following assumptions.

Assumption 6 *Let*

$$\mathcal{F}_1 = \{\mathbf{u} \mapsto \log c_{\boldsymbol{\theta}}(\mathbf{u}), \boldsymbol{\theta} \in \Theta\}.$$

Assume that $\mathcal{F}_1$ satisfies Conditions 4 and 5.

Assumption 7 *Let*

$$\mathcal{F}_2 = \{\mathbf{u} \mapsto \nabla_{\boldsymbol{\theta}} \log c_{\boldsymbol{\theta}}(\mathbf{u}), \boldsymbol{\theta} \in \Theta\}.$$

Assume that $\mathcal{F}_2$ satisfies Conditions 4 and 5.

5.4.2 Asymptotic theory for a single tree

In this section, we consider a tree $\mathbb{T} = (\mathcal{T}_\ell)_{\ell=1,\dots,K}$ with K leaves.

Let

$$\boldsymbol{\theta}^0 = (\theta_1^0, \dots, \theta_K^0) = \arg \max_{(\theta_1,\dots,\theta_K)} \sum_{\ell=1}^{K} \mathbb{E}\left[\log c_{\theta_\ell}(\mathbf{U}) \mathbf{1}_{\mathbf{X} \in \mathcal{T}_\ell}\right],$$

where the maximum is supposed to be achieved at a unique point $(\theta_1^0, \dots, \theta_K^0)$, and we denote

$$\boldsymbol{\theta}^0(\mathbf{x}|\mathbb{T}) = \sum_{\ell=1}^{K} \theta_\ell^0 \mathbf{1}_{x \in \mathcal{T}_\ell}.$$

Proposition 2 presented below is a consistency result. To that purpose, we consider the L^1-norm to compare our maximum likelihood estimator $\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T})$ and $\boldsymbol{\theta}^0(\cdot|\mathbb{T})$:

$$\|\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T}) - \boldsymbol{\theta}^0(\cdot|\mathbb{T})\|_1 = \int \|\widehat{\boldsymbol{\theta}}(\mathbf{x}|\mathbb{T}) - \boldsymbol{\theta}^0(\mathbf{x}|\mathbb{T})\|_1 \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}) = \sum_{\ell=1}^{K} |\widehat{\theta}_\ell - \theta_\ell^0| \mathbb{P}(\mathbf{X} \in \mathcal{T}_\ell),$$

where $\mathbb{P}_{\mathbf{X}}$ is the distribution of the covariates $\mathbf{X}$.

Proposition 2 *Under Assumptions 3 to 6, and if* $n[K \log K]^{-1} \to \infty$,

$$\|\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T}) - \boldsymbol{\theta}^0(\cdot|\mathbb{T})\|_1 = o_P(1).$$

By considering an additional assumption on the copula family, namely Assumption 7, and conditions on the Hessian matrix, we obtain the convergence rate.

Theorem 3 *Under Assumptions 3 to 7, and with the additional condition that for* $\ell = 1, \ldots, K$, *the Hessian matrix* $\nabla^2_\theta \log c_{\boldsymbol{\theta}}(\theta^0_\ell)$ *is invertible, then,*

$$\|\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T}) - \boldsymbol{\theta}^0(\cdot|\mathbb{T})\|_1 = O_P\left(\frac{[K \log K]^{1/2}}{n^{1/2}} + \varepsilon_n\right).$$

It is not surprising to notice that the stochastic part of the error deteriorates with K, due to the increase in the complexity of the model. On the other hand, although this part is harder to track, the approximation error is supposed to decrease with K, which means that $\boldsymbol{\theta}^0(\cdot|\mathbb{T})$ is supposed to be closer to the "true" target function $\boldsymbol{\theta}^0(\cdot)$ when the number of leaves of $\mathbb{T}$ increases.

5.4.3 Oracle property for the pruning step

Let us define the optimal subtree extracted from the maximal tree $\mathbb{T}_{\max}$ (which has $K_{\max}$ leaves) as

$$\boldsymbol{\theta}^0(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}^0(\cdot|\mathbb{T})} \mathbb{E}\left[\log c_{\boldsymbol{\theta}^0(\mathbf{X}|\mathbb{T})}(\mathbf{U})\right]. \tag{5.7}$$

Let K^0 denote the number of leaves of $\boldsymbol{\theta}^0$. If K^0 were known, Theorem 3 shows that one may expect a convergence rate of $\sqrt{K^0 \log K^0 / n}$ for the stochastic part. The next result shows that the penalized procedure has the ability to asymptotically achieve this optimal rate even though the number K^0 is unknown. This, of course, requires conditions on the penalizing constant λ.

Theorem 4 *Assume that the assumptions of Theorem 3 hold for all the subtrees of the maximal tree with* $K_{\max}$ *leaves. Then, if* $\lambda \to 0$, *and if* $\lambda n^{1/2}[K_{\max} \log K_{\max}]^{-1/2} \to \infty$, *it holds*

$$\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_1 = O_P\left(\frac{[K^0 \log K^0]^{1/2}}{n^{1/2}} + \varepsilon_n\right).$$

All the proofs are postponed to the Appendix section.

5.5 Empirical evidence

5.5.1 Simulation study

In this section, we present the functioning of the conditional copula analysis on simulated data.

The regression framework

We consider a bivariate random variable $\mathbf{U} = (U^{(1)}, U^{(2)})$, with uniform margins over $[0,1]$ and distributed according to an Archimedean copula $C_{\theta(\mathbf{X})}$. Archimedean copulas is a standard family of copulas often used in modeling applications, see [266, 267]. They are determined by a single parameter $\theta \in \mathbb{R}^1$ which is associated with Kendall's τ coefficient through a bijective relationship [245]. In our framework, θ, and thus also τ, depends on two covariates $\mathbf{X} = (X^{(1)}, X^{(2)})$, which are random variables uniformly distributed in $[0,1]$. Moreover, we assume that $\mathbf{U}_i$ are samples of the true cumulative marginal distributions of some bivariate response variables $\mathbf{Y} = (Y^{(1)}, Y^{(2)})$ conditionally on the covariates $\mathbf{X}$. Specifically, we assume normal distributions for these margins, with the mean parameters being a linear function of $\mathbf{X}$.

The first step of the simulations consists of generating synthetic data for $(\mathbf{X}_i, \theta_i, \tau_i, \mathbf{U}_i, \mathbf{Y}_i)$. Hence, our goal is to estimate the parameters θ_i (or τ_i) from $(\mathbf{X}_i, \mathbf{Y}_i)$, pretending not to know the true observations $\mathbf{U}_i$, as it is usually the case in a real data scenario. Therefore, as a preliminary step to the conditional copula analysis, we first compute the pseudo-observations. We do that by considering two different approaches, a parametric and a non-parametric one, which will result in two vectors of pseudo-observations, namely $\mathbf{V}$ and $\mathbf{W}$. Eventually, we fit the conditional copula model to both the $\mathbf{V}$ and $\mathbf{W}$ pseudo-observations, other than to the true margins $\mathbf{U}$ for additional comparison. The goodness of the three fits is evaluated against a benchmark model.

Definition of different scenarios

To investigate different scenarios, we consider three Archimedean copulas - the Clayton, Frank, and Gumbel copulas. We also consider three types of dependence between τ and the covariates $(X^{(1)}, X^{(2)})$, which we report below:

(i) a step-wise function:

$$\tau_i = \begin{cases} 0.3 & \text{if } X_i^{(1)} < 0.4, X_i^{(2)} < 0.75 \\ 0.5 & \text{if } X_i^{(1)} \geq 0.4, X_i^{(2)} < 0.75 \\ 0.7 & \text{if } X_i^{(1)} < 0.4, X_i^{(2)} \geq 0.75 \\ 0.9 & \text{if } X_i^{(1)} \geq 0.4, X_i^{(2)} \geq 0.75 \end{cases}$$

(ii) a steep sigmoid:

$$\tau_i = 0.3 - \frac{0.2}{1 + \exp(-40(X_i^{(1)} - 0.4))} - \frac{0.4}{1 + \exp(-40(X_i^{(2)} - 0.75))}$$

(ii) a gentle sigmoid:

$$\tau_i = 0.3 - \frac{0.2}{1 + \exp(-15(X_i^{(1)} - 0.4))} - \frac{0.4}{1 + \exp(-15(X_i^{(2)} - 0.75))}$$

With these constraints, having fixed the covariates $(X^{(1)}, X^{(2)})$, we obtain nine different conditional copulas, from which we sample $\mathbf{U}$ observations. Let us specify that these conditional copulas are defined such that Kendall's τ coefficients always range in the interval $[0.3, 0.9]$, to ensure comparability.

Finally, in all scenarios the response variables $\mathbf{Y}$ is defined as follow:

$$\begin{cases} Y_i^{(1)} &= \Psi^{-1}(U_i^{(1)} - 1 - 0.2X_i^{(1)} - 0.05X_i^{(2)}) \\ Y_i^{(2)} &= \Psi^{-1}(U_i^{(2)} - 1 + 0.1X_i^{(1)} - 0.2X_i^{(2)}) \end{cases}$$

where Ψ is the c.d.f. of the distribution $\mathcal{N}(0,1)$.

Pseudo-observation computation

We consider two alternative methods to compute the pseudo-observations.

First, in a parametric approach, we assume that the marginal distributions of $Y^{(j)}$ conditionally on $\mathbf{X}$ can be approximated by normal distributions with variance fixed at 1. Thus, we estimate the mean parameter through a linear model, i.e. $\hat{\mu}_i^{(j)} = LM(\mathbf{X}_i)$, and we compute the pseudo-observations $\mathbf{V}_i^{(j)} = \Psi^{-1}(Y_i^{(j)} - \hat{\mu}_i^{(j)})$.

Second, to avoid assumptions on the form of the margins, we perform a kernel estimation depending on the covariates as defined in (5.4). We consider a simple Gaussian kernel, with the bandwidth h optimized depending on the scenario, specifically, we used $h = 0.4$ for Clayton and Frank copulas, and $h = 0.3$ for the Gumbel copula. This way, the pseudo-observations $\mathbf{W}_i$ are computed as empirical percentiles, where in the calculation of the empirical cumulative distribution function the different observations are weighted differently according to their distance in terms of covariates.

Model's evaluation

As a reference model, we simply fit the Archimedean copula to $\mathbf{U}$ (and to $\mathbf{V}$ and $\mathbf{W}$), ignoring the additional information carried by the covariates. It means that we estimate a unique value for τ, which corresponds to the estimation provided by the root of the regression tree of the conditional copula model. Hence, the prediction errors of the conditional copula model and of the benchmark model are compared. For comparison, we consider the Mean Squared Errors for both the estimates of the τ coefficients and the values of the cumulative copula and the log-likelihood values of the models toward the observations/pseudo-observations.

Simulation results

For each one of the nine settings presented above, we build 500 triples of datasets, containing 1000 observations $\mathbf{U}_i$, 1000 pseudo-observations $\mathbf{V}_i$, and 1000 pseudo-observations $\mathbf{W}_i$, respectively. Results are presented in Figure 5.1. In all scenarios, the conditional copula model outperforms the benchmark model, both in terms of log-likelihood values and estimates for the τ coefficients and for the cumulative copula values. As expected, the predictions worsen when the dependence on the covariates changes from a step function, which can be perfectly captured by a regression tree, to smoother functions. Finally, no significant changes are noticed when models are fitted to observations or pseudo-observations. We notice that the conditional copula model most of the time identifies five or six groups of observations. That corresponds to a slightly overfitting of the model, as four groups are expected.

5.5.2 Real data example

In this section, we present an application of the conditional copula model on epidemiological data of cases of human influenza.

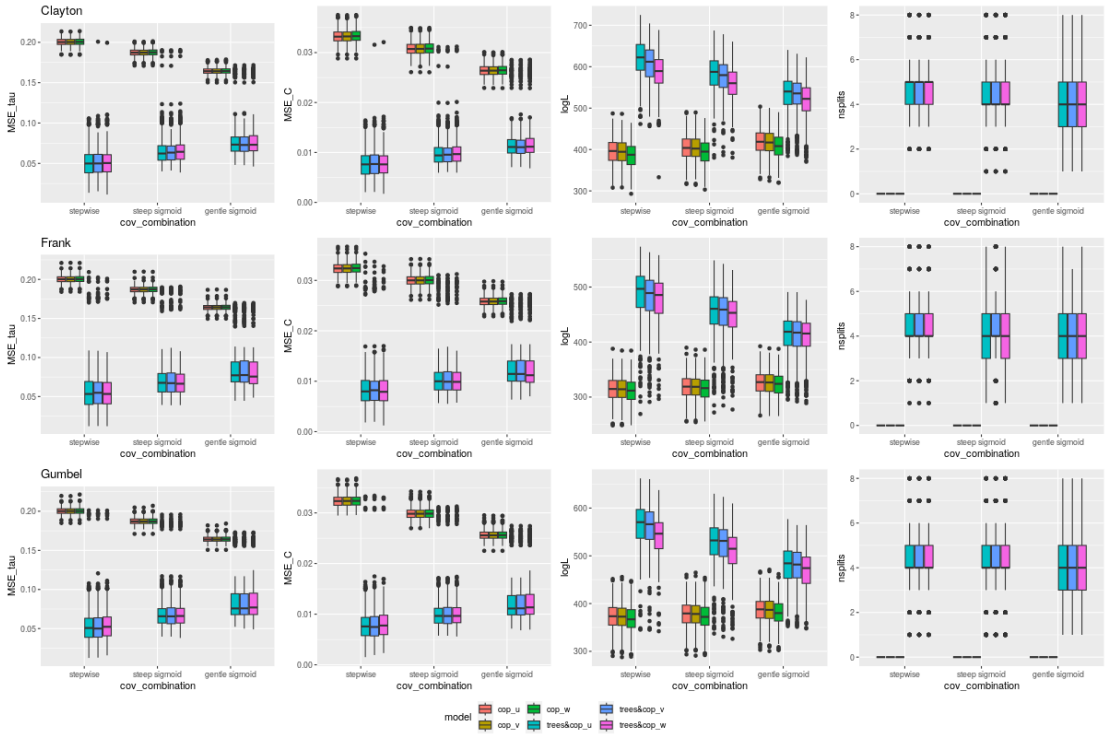


FIGURE 5.1: **Results of simulations.** Results for the Clayton, Frank, and Gumbel copulas are depicted on the different rows. For each copula, the results for the three types of covariate dependence are reported on the x-axis. The six colors identify different models: red, orange, and green are for the conditional copula model fitted on the observations **U** and on the pseudo-observations **V** and **W**, respectively, while cyan, blue and magenta are for the benchmark model. In the first two columns, we show results in terms of MSE for the τ estimates and the cumulative copula estimates, in the third column in terms of log-likelihood. In the fourth column, we report the distributions of the number of splits, i.e. the number of leaves minus 1, identified by the regression trees of the conditional copula models. Each boxplot represents the results for 500 datasets of 1000 points each.

The human influenza: context and data

Three main influenza strains co-circulate worldwide and infect humans: influenza A/H1N1pdm, influenza A/H3N2, and influenza B. The relative proportions of the three viruses are highly variable in time and space and the unpredictability of the strains' (co-)dominance patterns poses a major limitation to the mitigation of the upcoming epidemic wave in terms of intervention design and vaccination. Here, we use the conditional copula model to capture some trends of the coupled dynamic of influenza subtypes. In particular, first, we assume that we can use Archimedean copulas to describe the dependence structure of the relative abundances of influenza subtypes across regions and years. Second, we implement the conditional copula model to identify spatio-temporal patterns of such dependence structure.

The World Health Organisation provides data on influenza surveillance for several countries, consisting of weekly counts of cases classified by subtype [17, 18]. We consider data from 80 countries that reported a minimum of 50 classified cases per year in the period from April 2010 to April 2019. Then, we aggregate counts annually (from April to April) and for each country-year (800 observations in all) we compute the proportion of cases of A/H1N1pdm, A/H3N2, and B. We consider the relative abundances of subtypes as the response variables to be modeled with an Archimedean copula, testing Clayton, Frank, and Gumbel families, and the year and the Influenza Transmission Zone (ITZ) as the relevant covariates. The ITZs are 18 groups of countries with similar influenza transmission patterns identified by the W.H.O. (see [217] for the precise definition of the groups). Before fitting the conditional copula model, we perform a preprocessing step by applying an additive log-ratio transformation to the relative proportion of subtypes [140]. This is a common procedure when working with percentage data [142, 205, 207], and it allows us to map bounded vectors (A/H1N1pdm %, A/H3N2 %, B %)$\in S^3$ into unbounded vectors $(Y^{(1)}, Y^{(2)}) \in \mathbb{R}^2$, where S^3 is the so-called 3-part simplex. In particular, we use the isometric log-ratio transformation proposed by [144]:

$$\begin{cases} Y^{(1)} &= \sqrt{\frac{2}{3}} \ln \frac{B\%}{\sqrt{A\backslash H1N1pdm\% * A\backslash H3N2\%}} \\ Y^{(2)} &= \sqrt{\frac{1}{2}} \ln \frac{A\backslash H1N1pdm\%}{A\backslash H3N2\%} \end{cases}$$

Thus, the actual response variable is $\mathbf{Y} = (Y^{(1)}, Y^{(2)})$, with $Y^{(1)}$ describing the relative abundance between influenza B and the average proportion of influenza A subtypes, while $Y^{(2)}$ denotes the relative amount of A/H1N1pdm and A/H3N2.

Model implementation

Estimation of the margins. We consider regression trees to model the relationship between the response variables $Y^{(j)}$ and the covariates year and ITZ. Both the covariates are treated as categorical variables, meaning that a priori values have no precise sorting and an ordering step is needed preliminary to the split search, as explained in Remark 1. This allows maximum flexibility to the splitting procedure so that the tree can effectively capture the trends in the data. Once the trees are optimized by cross-validation, the pseudo-observations $\widehat{\mathbf{U}}^{(j)}$ are computed from a mixture of empirical cumulative distribution functions defined over the groups of points identified by the

optimal tree. That is

$$\widehat{\mathbf{U}}^{(j)}(t^{(j)}|\mathbf{X}) = \sum_{l=1}^{K} \left(\frac{1}{n_l} \sum_{i=1}^{n_l} \mathbf{1}_{Y_i^{(j)} \leq t^{(j)}} \mathbf{1}_{\mathbf{X}_i \in \mathcal{T}_l^{(j)}} \right),$$

with $\mathcal{T}_{l=1,\dots,K}^{(j)}$ being the terminal nodes of the tree fitted on the $Y^{(j)}$ response variable.

Fit of the conditional copula model. We test two-dimensional Clayton, Frank, and Gumbel copulas to model $\widehat{\mathbf{U}}$ conditionally on the covariates year and ITZ, again treated as categorical covariates. We implement a 3-fold cross-validation repeated 50 times to optimize the pruning of the trees with Breiman's rule used to identify the optimal tree. The best conditional model is the one with the Frank copula, which leads to the highest value of log-likelihood. It results in a tree with five leaves (Figure 5.2) which will be discussed in the next paragraph.

Results and discussions

Thanks to the conditional copula model we were able to ameliorate the adjustment to the data; the log-likelihood of the Frank simple copula on all the 800 $\widehat{\mathbf{U}}$ pseudo-observations was 1.2, and it increased to 13.8 for the Frank copula mixture model identified by the optimal tree.

In the estimation of the response variables $Y^{(1)}$ and $Y^{(2)}$, the years are used more often than the regions to perform the splits, indicating that the relative abundances of B vs. A and A/H1N1pdm vs. A/H3N2, taken independently, varied more in time than in space (see Figures 5.3 in the Appendix). In other words, to a first approximation, we find consistent temporal dynamics worldwide, going a step further we also identify significant spatial patterns. It is interesting to note that each time the spatial information is used to perform the split, European regions are grouped together, sometimes with other neighboring regions (mainly North Africa and Western and Central Asia), and always separated from countries of the southern hemisphere. These spatial country groupings overall match well the geographical clustering found in other studies with different methods (Cf. Chapter 4). Previous studies found evidence for an annual reseeding of influenza viruses from tropical and subtropical countries to temperate regions, especially for A/H3N2 viruses [15, 75, 99, 100]. These dynamics could also contribute to determining the patterns in subtype compositions that emerged from our analysis. However, our purely descriptive analysis does not allow us to speculate on any underlying mechanism.

Once $\widehat{U}^{(1)}$ and $\widehat{U}^{(2)}$ are computed, the conditional copula model identifies significant changes in the pseudo-observations dependence across space and time. It results in a tree with five leaves, characterized by different degrees of correlation (Kendall's τ among the $\widehat{U}^{(1)}$ and $\widehat{U}^{(1)}$ pseudo-observations range from -0.09 to 0.3). However, we note that a single leaf contains most of the data points (630 out of 800), meaning that the simplifying assumption would probably provide a reasonable approximation for the majority of the countries-years in our analysis. However, the other leaves allow us to refine the fit of the data and further separate a few country years that are mainly characterized by a proportion of B infections higher than the average.

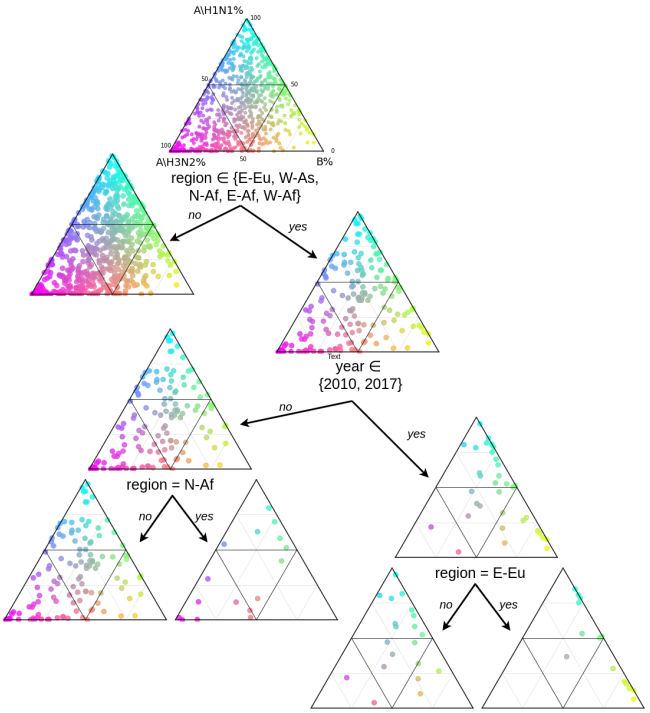


FIGURE 5.2: **Optimal tree identified by the Frank conditional copula model applied to data of relative abundances of influenza subtypes across countries and regions.** Data on the relative abundances of influenza subtypes are considered for 800 countries-years (corresponding to the 800 points in the top ternary plot). Similarly, for each node of the tree, a simplex represents the subtype relative abundances of the countries-years clustered in the node. We use a ternary color code to distinguish countries-years with dominance of A/H1N1pdm (cyan), A/H3N2 (pink), and B (yellow). For each split, the condition used to partition the observations is indicated. From top-left to bottom-right, the number of observations in each leaf is 630, 120, 16, 20 and 14. In the same order, Kendall's τ coefficients equal -0.06, -0.09, 0.3, -0.03, and 0.25.

5.6 Conclusion

In this paper, we proposed a new methodology to model conditional copulas, which is based on regression trees. The technique is applicable under the assumption that the conditional copulas all belong to the same family of parametric copulas, with the association parameter changing with the value of the covariates. The procedure presents many advantages. First, the tree structure theoretically allows capturing any form of the conditional association parameter. Second, the simplicity of the final model, if restricted to a single leaf of the tree, allows one to obtain a tractable output. We note that our approach allows a relaxation of the simplifying assumption [161], but this remains valid for each of the subsets of data identified by the tree. Another interesting feature is the ability to deal with quantitative and/or qualitative covariates.

In addition, let us point out that this method can be easily extended to the case where several families of copulas are tested at each node. This would give a more complex final structure, since not only the association parameter but also the copula family could vary from one leaf to another. However, it would increase the complexity of the implementation of the algorithm. Finally, let us note that the potential weakness of the procedure is its instability. Like every regression tree procedure, our method can be very sensitive to new incoming data, as new information may considerably change the structure of the tree and the classes that are made. Careful attention should be given to this aspect. On the other hand, a direct extension that could reduce this instability would be to consider the corresponding random forest algorithm, i.e., computing many small copula trees on separate bootstrap samples and then aggregating them. The aggregation of these trees would be a way to stabilize the result, but of course, would reduce the interpretability of the model.

R codes: The R codes are publicly available at https://github.com/FrancescoBonacina/tree-based-conditional-copula-estimation.

5.7 Appendix

The Appendix section is organized as follows. We first start with preliminary results that are needed to prove our results in Section 5.7.1, including some results on the complexity of the class of functions defined by the model in Section 5.7.1, and section 5.7.1 provides a general result that will be used repeatedly to handle deviations of the score function. We then prove Proposition 2 in Section 5.7.2, Theorem 3 in Section 5.7.3, and Theorem 4 in Section 5.7.4. Results on the convergence rates of the margins are then shown in Sections 5.7.5 and 5.7.6.

5.7.1 Preliminary results

In all this section, let us denote

$$\mathfrak{F} = \left\{ (\mathbf{u}, \mathbf{x}) \mapsto \phi(\mathbf{u}; \mathbf{x}) = \sum_{\ell=1}^{K} \varphi_\ell(\mathbf{u}) \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell} \text{ with for } \ell = 1, \dots, K, \, \varphi_\ell \in \mathcal{F} \text{ satisfying Condition 4} \right\},$$

and, for $\phi \in \mathfrak{F}$, , and

$$\mathcal{Z}(\phi) = \mathbb{E}\left[\phi(\mathbf{U}; \mathbf{X})\right], \quad \mathcal{Z}_n^*(\phi) = \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{U}_i; \mathbf{X}_i) \quad \text{and} \quad \widehat{\mathcal{Z}}_n(\phi) = \frac{1}{n} \sum_{i=1}^{n} \phi(\widehat{\mathbf{U}}_i; \mathbf{X}_i).$$

Bracketing numbers

We first introduce the concept of bracketing numbers to measure the complexity of a class of functions $\mathcal{F}$. For $\varepsilon > 0$, a $\varepsilon-$bracket $[\mathfrak{a}, \mathfrak{b}]$ is the set of functions f such that for all $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{u} \in \mathbb{R}^k$, $\mathfrak{a}(\mathbf{u}, \mathbf{x}) \leq f(\mathbf{u}, \mathbf{x}) \leq \mathfrak{b}(\mathbf{u}, \mathbf{x})$, with the condition that

$$\int (\mathfrak{a}(\mathbf{u}, \mathbf{x}) - \mathfrak{b}(\mathbf{u}, \mathbf{x}))^2 \mathrm{d}\mathbb{P}(\mathbf{u}, \mathbf{x}) \leq \varepsilon^2 .$$

We then define $\mathcal{N}(\varepsilon, \mathcal{F})$ as the minimal number of ε-brackets required to cover the class of functions $\mathcal{F}$. More details on bracketing numbers can be found in Chapter 19 of [131], and in Chapter 2.2. of [130].

Lemma 5 *For $\varepsilon > 0$,*

$$\mathcal{N}(\varepsilon, \mathfrak{F}) \leq \left(\frac{K^{m/2} C_1 \|\Phi\|_2^m}{\varepsilon^m} \right)^K ,$$

for some constant C_1 depending only on Θ and m.

Consider an element $\phi \in \mathfrak{F}$. It can be written as $\phi = \sum_{\ell=1}^K \varphi_\ell \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell}$, where each φ_ℓ is in $\mathcal{F}$, and satisfies Condition 4. Then, from Example 19.7 of [131], for each $\ell = 1, \ldots, K$, for all $\varepsilon > 0$

$$\mathcal{N}(\varepsilon, \mathcal{F}) \leq \frac{C_1(m, \Theta) \|\Phi\|_2^m}{\varepsilon^m},$$

where C_1 is a constant depending on $\mathrm{diam}(\Theta)$ and m.

Therefore, for each φ_ℓ, the set of $\varepsilon K^{-1/2}$-brackets $[a_{i(\ell)}, b_{i(\ell)}]$ for $i = 1, \ldots, K^{m/2} C_1(m, \Theta) \|\Phi\|_2^m \varepsilon^{-m}$ with $a_{i(\ell)} \leq b_{i(\ell)}$ covers $\mathcal{F}$.

Now, for $i = 1, \ldots, K^{m/2} C_1(m, \Theta) \|\Phi\|_2^m \varepsilon^{-m}$, define

$$\mathfrak{a}_i(\mathbf{u}, \mathbf{x}) = \sum_{\ell=1}^K a_{i(\ell)}(\mathbf{u}) \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell} \quad ; \quad \mathfrak{b}_i(\mathbf{u}, \mathbf{x}) = \sum_{\ell=1}^K b_{i(\ell)}(\mathbf{u}) \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell}. \tag{5.8}$$

Clearly, $\mathfrak{a}_i \leq \mathfrak{b}_i$ and $\phi \in [\mathfrak{a}_i, \mathfrak{b}_i]$. Moreover,

$$\int (\mathfrak{b}_i(\mathbf{u}, \mathbf{x}) - \mathfrak{a}_i(\mathbf{u}, \mathbf{x}))^2 \mathrm{d}\mathbb{P}(\mathbf{u}, \mathbf{x}) \leq \sum_{\ell=1}^K \int (b_{i(\ell)}(\mathbf{u}) - a_{i(\ell)}(\mathbf{u}))^2 \mathrm{d}\mathbb{P}(\mathbf{u}) \leq \varepsilon^2.$$

Thus, the set of brackets $[\mathfrak{a}_i, \mathfrak{b}_i]$, for $i = 1, \ldots, K^{m/2} C_1(m, \Theta) \|\Phi\|_2^m \varepsilon^{-m}$ defined in (5.8) and deduced from the brackets $[a_{i(\ell)}, b_{i(\ell)}]$ are ε-brackets covering $\mathfrak{F}$, and their number is less that

$$\left(\frac{K^{m/2} C_1(m, \Theta) \|\Phi\|_2^m}{\varepsilon^m} \right)^K ,$$

leading to the result.

General results on sums involving pseudo-observations

The first result of this section shows how to replace pseudo-observations $\mathbf{U}_i$ by their estimated version $\widehat{\mathbf{U}}_i$ in studying the asymptotic behavior of sums involving these quantities. Going back to $\mathbf{U}_i$ then simplifies considerably the study of such quantities, since one goes back to classical i.i.d. quantities.

Lemma 6 *Assume furthermore that there exist* $0 \leq \beta_1, \beta_3 < 1/2,$, $0 \leq 1\beta_2 < 1$ *and two universal constants* A_1 *and* A_2 *such that for all* $\varphi \in \mathcal{F}$ *satisfies Condition 5.*

Then, under Assumption 3,

$$\sup_{\phi \in \mathfrak{F}} \left| \widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}_n^*(\phi) \right| = O_P(\varepsilon_n),$$

with ε_n *tends to 0 when n tends to* ∞.

First, recall that

$$\sup_{\phi \in \mathfrak{F}} \left| \widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}_n^*(\phi) \right| = \sup_{\phi \in \mathfrak{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \phi(\widehat{\mathbf{U}}_i; \mathbf{X}_i) - \phi(\mathbf{U}_i; \mathbf{X}_i) \right\} \right|.$$

Then, from a Taylor expansion,

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \phi(\widehat{\mathbf{U}}_i; \mathbf{X}_i) - \phi(\mathbf{U}_i; \mathbf{X}_i) \right\} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} \partial_j \phi(U_i^{(1)}, \ldots, \widetilde{U}_i^{(j)}, \ldots, U_i^{(k)}; \mathbf{X}_i) \left[\widehat{U}_i^{(j)} - U_i^{(j)} \right],$$

where $\widetilde{U}_i^{(j)}$ is between $U_i^{(j)}$ and $\widehat{U}_i^{(j)}$. From Condition 5,

$$|\partial_j \varphi(\mathbf{u})| \leq \frac{A_2}{[u^{(j)}(1-u^{(j)})]^{\beta_2}} \sum_{r=1}^{k} \frac{1}{[u^{(r)}(1-u^{(r)})]^{\beta_3}}, \quad j \in \{1, \ldots, k\},$$

Thus,

$$|\partial_j \phi(U_i^{(1)}, \ldots, \widetilde{U}_i^{(j)}, \ldots, U_i^{(k)}; \mathbf{X}_i)| \leq \frac{A_2}{[\widetilde{U}_i^{(j)}(1-\widetilde{U}_i^{(j)})]^{\beta_2}} \times \left\{ \sum_{r \neq j} \frac{1}{[U_i^{(r)}(1-U_i^{(r)})]^{\beta_3}} + \frac{1}{[\widetilde{U}_i^{(j)}(1-\widetilde{U}_i^{(j)})]^{\beta_3}} \right\}.$$

Note that

$$\frac{1}{\widetilde{U}_i^{(j)}(1-\widetilde{U}_i^{(j)})} \leq \sup_{i=1,\ldots,n} \left(\frac{U_i^{(j)}}{\widehat{U}_i^{(j)}} + \frac{1-U_i^{(j)}}{1-\widehat{U}_i^{(j)}} \right) \frac{1}{U_i^{(j)}(1-U_i^{(j)})},$$

leading to

$$|\partial_j \phi(U_i^{(1)}, \ldots, \widetilde{U}_i^{(j)}, \ldots, U_i^{(k)}; \mathbf{X}_i)| \leq \frac{A_2 [\max(1, \left(\frac{U_i^{(j)}}{\widehat{U}_i^{(j)}} + \frac{1-U_i^{(j)}}{1-\widehat{U}_i^{(j)}} \right)]^{\beta_3}}{[U_i^{(j)}(1-U_i^{(j)})]^{\beta_2}} \left\{ \sum_{r=1}^{k} \frac{1}{[U_i^{(r)}(1-U_i^{(r)})]^{\beta_3}} \right\}.$$

Let

$$Z_i = \sum_{r=1}^{k} \frac{1}{[U_i^{(r)}(1-U_i^{(r)})]^{\beta_3}}.$$

Since $\beta_3 < 1/2$, $E[Z_i^2] < \infty$. Hence,

$$\begin{aligned} \left| \widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}_n^*(\phi) \right| \quad \leq \quad & \frac{A_2}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{Z_i}{[U_i^{(j)}(1-U_i^{(j)})]^{\beta_2 - \beta'}} \sup_{\substack{i=1,\ldots,n \\ j=1,\ldots,k}} \left| \frac{\widehat{U}_i^{(j)} - U_i^{(j)}}{\left[U_i^{(j)}(1-U_i^{(j)}) \right]^{\beta'}} \right| \\ & \times \sup_{\substack{i=1,\ldots,n \\ j=1,\ldots,k}} \left(\frac{U_i^{(j)}}{\widehat{U}_i^{(j)}} + \frac{1-U_i^{(j)}}{1-\widehat{U}_i^{(j)}} \right)^{\beta_3} \end{aligned}$$

with $\beta' = \min(\beta_3, \alpha)$. Then, first, from Cauchy-Schwarz inequality,

$$\mathbb{E}\left[\left\{\frac{Z_i}{[U_i^{(j)}(1-U_i^{(j)})]^{\beta_2-\beta'}}\right\}^2\right] \leq \mathbb{E}[Z_i^2]^{1/2}\mathbb{E}\left[\frac{1}{[U_i^{(j)}(1-U_i^{(j)})]^{2[\beta_2-\beta']}}\right]^{1/2} < \infty\,.$$

Second, from Assumption 3,

$$\sup_{\substack{i=1,\dots,n\\ j=1,\dots,k}}\left|\frac{\widehat{U}_i^{(j)} - U_i^{(j)}}{\left[U_i^{(j)}(1-U_i^{(j)})\right]^{\beta'}}\right| = O_p(\varepsilon_n)\,,$$

and

$$\sup_{\substack{i=1,\dots,n\\ j=1,\dots,k}}\left|\frac{U_i^{(j)}}{\widehat{U}_i^{(j)}} + \frac{1-U_i^{(j)}}{1-\widehat{U}_i^{(j)}}\right|^{\beta_3} = O_P(1)\,.$$

Remark 2 *If (5.5) does not hold, the estimation procedure can be modified by introducing some trimming, that is multiplying each term of the log-likelihood by* $\mathbf{1}_{\min(1-\hat{U}_i^{(j)},\hat{U}_i^{(j)})\geq\eta_n}$. *If* η_n *tends to zero slower than* ε_n, $\hat{U}_i^{(j)} \geq U_i^{(j)}/2$ *for* n *large enough due to (5.6) for the indexes* i *where this indicator function is not zero. However, the introduction of trimming induces some bias for the estimator, which can be controlled thanks to Assumption 6.*

With at hand Lemma 6 and the complexity bound of Lemma 5, one can derive the main result of this section, which will be used several times in the proof of our main theorems.

Proposition 7 *Assume furthermore that there exist* $0 \leq \beta_1, \beta_3 < 1/2$, $0 \leq /\beta_2 < 1$ *and two universal constants* A_1 *and* A_2 *such that for all* $\varphi \in \mathcal{F}$ *satisfies Condition 5. Then,*

$$\sup_{\phi\in\mathfrak{F}}\left|\widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}(\phi)\right| = O_P\left(\sqrt{\frac{K\log K}{n}} + \varepsilon_n\right).$$

Writing

$$\sup_{\phi\in\mathfrak{F}}\left|\widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}(\phi)\right| \leq \sup_{\phi\in\mathfrak{F}}\left|\widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}_n^*(\phi)\right| + \sup_{\phi\in\mathfrak{F}}|\mathcal{Z}_n^*(\phi) - \mathcal{Z}(\phi)|$$

For the first term, from Lemma 6,

$$\sup_{\phi\in\mathfrak{F}}\left|\widehat{\mathcal{Z}}_n(\phi) - \mathcal{Z}_n^*(\phi)\right| = O_P\left(\varepsilon_n\right).$$

For the second term, introduce for $\delta > 0$, $J(\delta, \mathfrak{F}) = \int_0^\delta \sqrt{\log \mathcal{N}(\varepsilon, \mathfrak{F})}\mathrm{d}\varepsilon$. From Corollary 19.35 of [131],

$$\sqrt{n}\mathbb{E}\left[\sup_{\phi\in\mathfrak{F}}|\mathcal{Z}_n^*(\phi) - \mathcal{Z}(\phi)|\right] \leq A_3 J(\|\Phi\|_2, \mathfrak{F}),$$

for some universal constant $A_3 \geq 0$. Then, from Lemma 5,

$$J(\|\Phi\|_2, \mathfrak{F}) \leq \int_0^{\|\Phi\|_2} K^{1/2}\left\{\frac{m}{2}\log K + \log(C_1\|\Phi\|_2^m) + m\log\left(\frac{1}{\varepsilon}\right)\right\}^{1/2} d\varepsilon.$$

Hence,

$$\sqrt{n}\mathbb{E}\left[\sup_{\phi\in\mathfrak{F}}|\mathcal{Z}_n^*(\phi)-\mathcal{Z}(\phi)|\right]\leq C_2(m,\Theta)\sqrt{K\log K}.$$

5.7.2 Proof of Proposition 2

We are now ready to prove Proposition 2.

Recall that

$$\|\widehat{\boldsymbol{\theta}}(\cdot|\mathbb{T})-\boldsymbol{\theta}^0(\cdot|\mathbb{T})\|_1=\sum_{\ell=1}^K|\widehat{\boldsymbol{\theta}}_\ell-\boldsymbol{\theta}_\ell^*|\mathbb{P}(\mathbf{X}\in\mathcal{T}_\ell),$$

so that it suffices to show that

$$\sup_{\ell=1,\ldots,K}|\widehat{\boldsymbol{\theta}}_\ell-\boldsymbol{\theta}_\ell^0|=o_P(1). \tag{5.9}$$

Let

$$\begin{aligned}
\widehat{\mathcal{L}}_n(\theta_1,\ldots,\theta_K) &= \frac{1}{n}\sum_{i=1}^n\sum_{\ell=1}^K\log c_{\theta_\ell}(\widehat{\mathbf{U}}_i)\mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell},\\
\mathcal{L}_n^*(\theta_1,\ldots,\theta_K) &= \frac{1}{n}\sum_{i=1}^n\sum_{\ell=1}^K\log c_{\theta_\ell}(\mathbf{U}_i)\mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell},\\
\mathcal{L}(\theta_1,\ldots,\theta_K) &= \mathbb{E}\left[\sum_{\ell=1}^K\log c_{\theta_\ell}(\mathbf{U}_i)\mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell}\right].
\end{aligned}$$

From Corollary 3.2.3 of [130], (5.9) holds if

$$\sup_{\theta_1,\ldots,\theta_\ell}|\widehat{\mathcal{L}}_n(\theta_1,\ldots,\theta_\ell)-\mathcal{L}(\theta_1,\ldots,\theta_\ell)|=o_P(1).$$

Let us introduce

$$\mathfrak{F}_1=\left\{(\mathbf{u},\mathbf{x})\mapsto\sum_{\ell=1}^K\log c_{\boldsymbol{\theta}}(\mathbf{u})\mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell}\text{ with }\boldsymbol{\theta}=(\theta_\ell)_{\ell=1,\ldots,K}\in\Theta^K\right\}. \tag{5.10}$$

From Assumption 6, Proposition 7 applies to $\mathfrak{F}_1$, leading to

$$\sup_{\theta_1,\ldots,\theta_\ell}|\widehat{\mathcal{L}}_n(\theta_1,\ldots,\theta_\ell)-\mathcal{L}(\theta_1,\ldots,\theta_\ell)|=\sup_{\phi\in\mathfrak{F}_1}|\mathcal{Z}_n^*(\phi)-\mathcal{Z}(\phi)|=O_P\left(\sqrt{\frac{K\log K}{n}}+\varepsilon_n\right)$$

which tends to zero under the condition on K and the result follows.

5.7.3 Proof of Theorem 3

Introduce

$$\begin{aligned}
\dot{\mathcal{L}}_n(\theta_1,\ldots,\theta_K) &= \frac{1}{n}\sum_{\ell=1}^K\sum_{i=1}^n\nabla_{\boldsymbol{\theta}}\log c_{\theta_\ell}(\widehat{\mathbf{U}}_i)\mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell},\\
\dot{\mathcal{L}}(\theta_1,\ldots,\theta_K) &= \sum_{\ell=1}^K\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\log c_{\theta_\ell}(\mathbf{U})\mathbf{1}_{\mathbf{X}\in\mathcal{T}_\ell}\right],
\end{aligned}$$

and

$$\mathfrak{F}_2 = \left\{ (\mathbf{u}, \mathbf{x}) \to \sum_{\ell=1}^{K} \nabla_{\boldsymbol{\theta}} \log c_{\theta_\ell}(\mathbf{u}) \mathbf{1}_{\mathbf{x} \in \mathcal{T}_\ell} : (\theta_\ell)_{l=1,\dots,K} \in \Theta^K \right\}.$$

From Proposition 7,

$$\sup_{\theta_1,\dots,\theta_\ell} \left| \dot{\mathcal{L}}_n(\theta_1, \dots, \theta_K) - \dot{\mathcal{L}}(\theta_1, \dots, \theta_K) \right| = O_P\left(\frac{[K \log K]^{1/2}}{n^{1/2}} + \varepsilon_n \right). \tag{5.11}$$

Then, write

$$\begin{aligned} \dot{\mathcal{L}}(\theta_1^0, \dots, \theta_K^0) - \dot{\mathcal{L}}(\widehat{\theta}_1, \dots, \widehat{\theta}_K) &= \left\{ \dot{\mathcal{L}}(\theta_1^0, \dots, \theta_K^0) - \dot{\mathcal{L}}_n(\theta_1^0, \dots, \theta_K^0) \right\} \\ &+ \left\{ \dot{\mathcal{L}}_n(\theta_1^0, \dots, \theta_K^0) - \dot{\mathcal{L}}_n(\widehat{\theta}_1, \dots, \widehat{\theta}_K) \right\} \\ &+ \left\{ \dot{\mathcal{L}}_n(\widehat{\theta}_1, \dots, \widehat{\theta}_K) - \dot{\mathcal{L}}(\widehat{\theta}_1, \dots, \widehat{\theta}_K) \right\}. \end{aligned}$$

The rates of the first and last brackets in this decomposition are given by (5.11), while the middle one is

$$\left\{ \dot{\mathcal{L}}_n(\theta_1^0, \dots, \theta_K^0) - \dot{\mathcal{L}}_n(\widehat{\theta}_1, \dots, \widehat{\theta}_K) \right\} = \dot{\mathcal{L}}_n(\theta_1^0, \dots, \theta_K^0),$$

has also the same rate. This shows that

$$\left| \dot{\mathcal{L}}(\theta_1^0, \dots, \theta_K^0) - \dot{\mathcal{L}}(\widehat{\theta}_1, \dots, \widehat{\theta}_K) \right| = O_P\left(\frac{[K \log K]^{1/2}}{n^{1/2}} + \varepsilon_n \right).$$

From the assumption on the Hessian matrix, and since Proposition 2 applies (which guarantees that each $\widehat{\theta}_\ell$ is in an arbitrary small neighborhood of θ_ℓ^* for n large enough), we get

$$\left| \dot{\mathcal{L}}(\theta_1^0, \dots, \theta_K^0) - \dot{\mathcal{L}}(\widehat{\theta}_1, \dots, \widehat{\theta}_K) \right| \geq \mathfrak{a} \|\widehat{\theta} - \theta^0\|_1,$$

for some $\mathfrak{a} > 0$, from a Taylor expansion, and the result follows.

5.7.4 Proof of Theorem 4

Let $\widehat{\boldsymbol{\theta}}^K$ denote the best tree with K leaves, with respect to the log-likelihood (and $\boldsymbol{\theta}^{0,K}$ its corresponding limit), and $\widehat{K}$ denote the number of leaves of $\bar{\boldsymbol{\theta}}$.

Write

$$\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^0 = [\widehat{\boldsymbol{\theta}}^{K^0} - \boldsymbol{\theta}^0] \mathbf{1}_{\widehat{K}=K^0} + \sum_{K \neq K^0} \left[\widehat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^0 \right] \mathbf{1}_{\widehat{K}=K}.$$

Let $R = \sum_{K \neq K^0} \left[\widehat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^{0,K} \right] \mathbf{1}_{\widehat{K}=K}$, and note that

$$\mathbb{P}(R \geq t) \leq \mathbb{P}(\widehat{K} > K^0) + \mathbb{P}(\widehat{K} < K^0).$$

The result is then shown if we prove that $\mathbb{P}(\widehat{K} > K^0)$ and $\mathbb{P}(\widehat{K} < K^0)$ tend to zero when n tends to infinity, which is done below studying each probability separately.

First case: $\mathbb{P}(\widehat{K} > K^0)$. We will use the notation $\mathcal{L}_n^K$ to denote the log-likelihood associated with $\widehat{\boldsymbol{\theta}}^K$. If $\widehat{K} > K^0$, this means that there exists some $K^0 < K < K_{\max}$ such that

$$\mathcal{L}_n^K - \mathcal{L}_n^{K^0} \geq \lambda(K - K^0),$$

that is

$$\mathcal{L}_n^K(\widehat{\boldsymbol{\theta}}^K) - \mathcal{L}_n^K(\boldsymbol{\theta}^{0,K}) \geq \lambda(K - K^0),$$

since $\mathcal{L}_n^K(\boldsymbol{\theta}^0) = \mathcal{L}_n^{K^0}(\boldsymbol{\theta}^{0,K})$ for $K \geq K^0$. Whence,

$$\mathbb{P}(\widehat{K} > K^0) \leq \mathbb{P}(\exists K > K^0 : \mathcal{L}_n^K(\widehat{\boldsymbol{\theta}}^K) - \mathcal{L}_n^K(\boldsymbol{\theta}^{0,K}) \geq \lambda(K - K^0)).$$

Since $\lambda(K - K^0) \geq \lambda$, and since $\mathcal{L}_n^K(\widehat{\boldsymbol{\theta}}^K) - \mathcal{L}_n^K(\boldsymbol{\theta}^{0,K}) \leq \mathcal{L}_n^{K_{\max}}(\widehat{\boldsymbol{\theta}}^{K_{\max}}) - \mathcal{L}_n^{K_{\max}}(\boldsymbol{\theta}^{0,K_{\max}})$,

$$\mathbb{P}(\widehat{K} > K^0) \leq \mathbb{P}\left(\mathcal{L}_n^{K_{\max}}(\widehat{\boldsymbol{\theta}}^{K_{\max}}) - \mathcal{L}_n^{K_{\max}}(\boldsymbol{\theta}^{0,K_{\max}}) \geq \lambda\right). \tag{5.12}$$

In the proof of Proposition 2, we showed that

$$\mathcal{L}_n^{K_{\max}}(\widehat{\boldsymbol{\theta}}^{K_{\max}}) - \mathcal{L}_n^{K_{\max}}(\boldsymbol{\theta}^{0,K}) = O_P([K_{\max} \log K_{\max}]^{1/2} n^{-1/2} + \varepsilon_n).$$

Hence, the right-hand side of (5.12) tends to zero provided that $\lambda n^{1/2}[K_{\max} \log K_{\max}]^{-1/2} \to \infty$.

Second case: $\mathbb{P}(\widehat{K} < K^0)$.
In this case $\mathcal{L}_n^K - \mathcal{L}_n^{K^0} \leq \mathcal{L}_n^{(K^0-1)} - \mathcal{L}_n^{K^0}$. From the proof of Proposition 2, $\mathcal{L}_n^{(K^0-1)} - \mathcal{L}^{*(K^0-1)} = O_P([K^* \log K^0]^{1/2} n^{-1/2})$, and $\mathcal{L}_n^{(K^0)} - \mathcal{L}^{*(K^0)} = O_P([K^* \log K^0]^{1/2} n^{-1/2})$. Then, similarly to the first case,

$$\mathbb{P}(\widehat{K} < K^0) \leq \mathbb{P}\left(\mathcal{L}_n^{(K^0-1)} - \mathcal{L}^{*(K^0-1)} - \mathcal{L}_n^{K^0} + \mathcal{L}^{*K^0} \geq \frac{\lambda}{2}\right) + \mathbb{P}\left(\mathcal{L}^{*(K^0-1)} - \mathcal{L}^{*K^0} \geq \frac{\lambda}{2}\right).$$

The first probability tends to zero under the same conditions as in the first case, while the second is equal to $\mathbf{1}_{\mathcal{L}^{*(K^0-1)} - \mathcal{L}^{*(K^0)} \geq \lambda/2}$, since the quantity $\mathcal{L}^{*(K^0-1)} - \mathcal{L}^{*K^0}$ is deterministic. This indicator function tends to zero when n tends to infinity if λ tends to zero.

5.7.5 Convergence rate for the margins for kernel estimators

In this section, we show that Assumption 3 holds for the kernel estimator (5.4). This is a consequence of Theorem 4 in [262]. We show the result under three additional assumptions on the model:

1. the density of $\mathbf{X}$ is bounded away from zero on $\mathcal{X}$, that is $\inf_{\mathbf{x} \in \mathcal{X}} f_{\mathbf{X}}(\mathbf{x}) > 0$;
2. we have
$$\sup_{\mathbf{x} \in \mathcal{X}, y} \left| \frac{F^{(j)}(y)}{F^{(j)}(y|\mathbf{x})} + \frac{1 - F^{(j)}(y)}{1 - F^{(j)}(y|\mathbf{x})} \right| \leq \mathfrak{a},$$
for some finite constant $\mathfrak{A}$;
3. the kernel function is a continuous and bounded function, symmetric around 0, such that $\int u^2 K(u) du < \infty$, the density $\mathbf{x} \mapsto f_{\mathbf{X}}(\mathbf{x})$ and $\mathbf{x} \mapsto F^{(j)}(t|\mathbf{x})$ are twice continuously differentiable with respect to $\mathbf{x}$, with uniformly bounded derivatives up to order 2.

The first assumption is required to avoid the denominator, in the kernel weights, going too close to zero. The second one is a way to consider that there is some kind of uniform domination of the behavior of the conditional distributions when $\mathbf{x}$ changes. Finally,

the third assumption is classical in kernel regression and will help to control the bias term involved in smoothing techniques.

Introducing the kernel estimator of the density of $\mathbf{X}$,

$$\widehat{f}_X(\mathbf{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right),$$

we can write, for $t \leq 1/2$,

$$\widehat{f}_X(\mathbf{x})\frac{\widehat{F}^{(j)}(t|\mathbf{x})}{\left[F^{(j)}(t|\mathbf{x})(1-F^{(j)}(t|\mathbf{x}))\right]^{\alpha}} = \frac{1}{nh^p}\sum_{i=1}^{n} K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right) f_t(Y_i^{(j)}),$$

where

$$f_t(y) = \frac{\mathbf{1}_{y\leq t}}{\left[F^{(j)}(t|\mathbf{x})(1-F^{(j)}(t|\mathbf{x}))\right]^{\alpha}} \leq \frac{1}{[F^{(j)}(y|\mathbf{x})]^{\alpha}[1-F^{(j)}(1/2|\mathbf{x})]^{\alpha}} \leq \frac{\mathfrak{A}^{\alpha}}{[F^{(j)}(y)]^{\alpha}[1-F^{(j)}(1/2|\mathbf{x})]^{\alpha}}.$$

Since

$$\mathbb{E}\left[\left(\frac{1}{[F^{(j)}(Y_i^{(j)})]^{\alpha}}\right)^{p}\right] < \infty,$$

for some $p > 2$ for $\alpha < 1/2$, and since the covering number of the class of functions f_t is controlled (see Example 19.12 of [131]), then Theorem 4 of [262] applies, showing that

$$\sup_{t\leq 1/2,x}\left|\frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right) f_t(Y_i^{(j)}) - \mathbb{E}\left[f_t(Y_i^{(j)})K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right)\right]\right| = O_P\left([\log n]^{1/2}n^{-1/2}h^{-d/2}\right).$$

Then, from a Taylor expansion and the third assumption of this section, we get

$$\mathbb{E}\left[f_t(Y_i^{(j)})K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h}\right)\right] = \mathbb{E}\left[f_t(Y_i^{(j)})|\mathbf{X}_i = \mathbf{x}\right] f_{\mathbf{X}}(\mathbf{x}) + O(h^2).$$

Let us note that this h^2 rate can be improved if one uses a degenerate kernel with a sufficiently high number of moments equal to zero. Then, from the rate of uniform convergence of $\widehat{f}_X(\mathbf{x})$ from Theorem 1 of [262], we get

$$\sup_{t\leq 1/2,\mathbf{x}}\left|\frac{\widehat{F}^{(j)}(t|\mathbf{x}) - F^{(j)}(t|\mathbf{x})}{\left[F^{(j)}(t|\mathbf{x})(1-F^{(j)}(t|\mathbf{x}))\right]^{\alpha}}\right| = O_P(h^2 + [\log n]^{1/2}n^{-1/2}h^{-d/2}).$$

Studying the supremum for $t > 1/2$ can be done in the same way, by studying $1 - F^{(j)}$ instead of $F^{(j)}$.

5.7.6 Convergence rate for the margins for discrete covariates

For discrete covariates, recall that

$$\widehat{F}^{(j)}(t|\mathbf{x}) = \frac{\sum_{i=1}^{n}\mathbf{1}_{Y_i^{(j)}\leq t}\mathbf{1}_{\mathbf{X}_i\in\mathcal{C}(\mathbf{x})}}{\sum_{i=1}^{n}\mathbf{1}_{\mathbf{X}_i\in\mathcal{C}(\mathbf{x})}}.$$

From the central limit theorem,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\mathbf{X}_i\in\mathcal{C}(\mathbf{x})} = \mathbb{P}\left(\mathbf{X}\in\mathcal{C}(\mathbf{x})\right) + O_P(n^{-1/2}).$$

The upper part can be studied using similar arguments as Example 19.12 of [131], noticing that the class of functions $f_t(Y_i^{(j)})\mathbf{1}_{\mathbf{X}_i\in\mathcal{C}(\mathbf{x})}$ (where f_t is defined in section 5.7.5) has a similar covering number as the class of functions f_t. This leads to

$$\sup_{t,\mathbf{x}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{Y_i^{(j)}\leq t}\mathbf{1}_{\mathbf{X}_i\in\mathcal{C}(\mathbf{x})} - F^{(j)}(t|\mathbf{x})\mathbb{P}\left(\mathbf{X}\in\mathcal{C}(\mathbf{x})\right)\right| = O_P(n^{-1/2}).$$

Then, we get

$$\sup_{t,\mathbf{x}}\left|\frac{\widehat{F}^{(j)}(t|\mathbf{x}) - F^{(j)}(t|\mathbf{x})}{\left[F^{(j)}(t|\mathbf{x})(1-F^{(j)}(t|\mathbf{x}))\right]^{\alpha}}\right| = O_P(n^{-1/2}).$$

5.7.7 Regression trees for margin estimation in the real data example

We report here the regression trees resulting from the fits of the variables $Y^{(1)}$ and $Y^{(1)}$ as functions of the covariates year and Influenza Transmission Zone. (see Section 5.5.2).

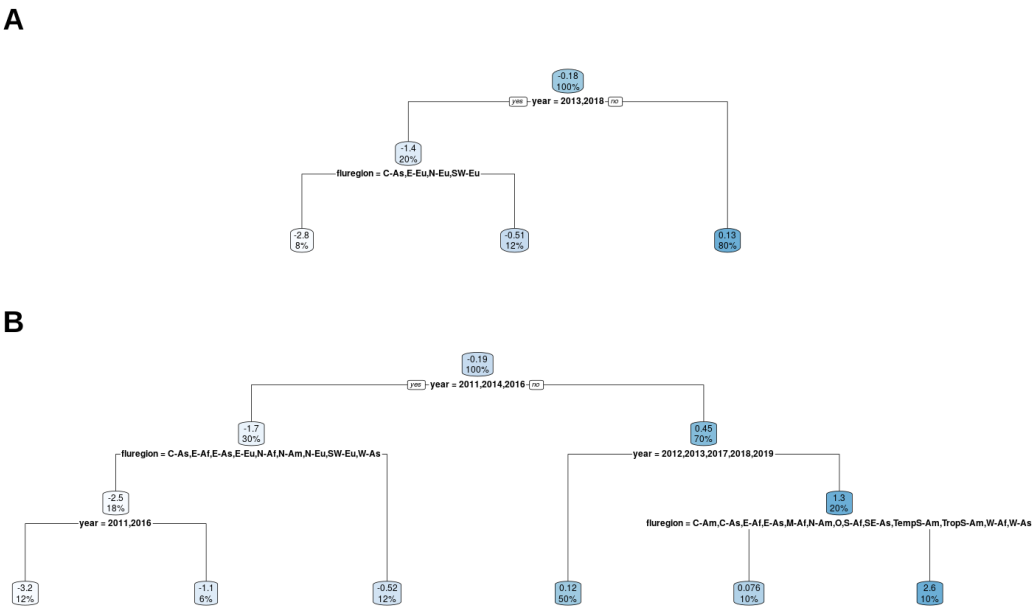


FIGURE 5.3: **Optimal trees for margins estimation.** Years and Influenza Transmission Zones are classified by regression trees to approximate the response variables $Y^{(1)}$ (plot A) and $Y^{(2)}$ (plot B). The coefficients of determination of the two fits are 0.29 and 0.5, respectively. For each node, the average value of the response variable and the percentage of the observations included are indicated.

Chapter 6

Conclusions and discussions

The research projects presented in this thesis were motivated by the need to understand and anticipate the patterns of influenza circulation on a global scale. Many types and subtypes of influenza viruses co-circulate worldwide contributing to annual epidemics. Their dissemination follows heterogeneous trends in space and time, shaped by the interplay of multiple factors: human behavior, population susceptibility shaped by previous infections and vaccination, and seasonal effects. Moreover, the annual cycles of influenza circulation might be altered by the occurrence of major epidemiological events. Since the 1950s, four new influenza variants emerged causing pandemics and, more recently, the measures applied during the COVID-19 pandemic disrupted the influenza circulation. In this thesis, we proposed novel statistical approaches to investigate the spatio-temporal patterns of influenza circulation. We tackled the problem from a global perspective, analyzing data from FluNet, a public repository for the virological surveillance of human influenza worldwide.

During the COVID-19 pandemic, influenza circulation was strongly altered, making it critical to assess the situation. This was especially needed to anticipate possible post-pandemic scenarios. In Chapter 3, we quantified the impact of the COVID-19 pandemic on influenza circulation from April 2020 to September 2021. We found that influenza circulation was significantly lower (by about two orders of magnitude) than the pre-pandemic levels, worldwide and throughout the study period.

At the time of our analyses, the evidence of the massive drop in influenza circulation raised concerns about the particularly severe influenza outbreaks that could have possibly occurred following the relaxation of the anti-COVID-19 interventions. Indeed, the stop of influenza circulation for around two years determined an increase in population susceptibility, especially among children. In 2022, some authors investigated this aspect and used models to predict the epidemic size of the upcoming season (i.e. the winter season of 2022-2023) in several countries. The results of Lei and colleagues [268] and Ali and colleagues [269] were particularly alarming, with epidemic sizes estimated up to 4-5 times the typical size of pre-pandemic seasons, as a consequence of an increase in population susceptibility estimated between 10% and 60% [269]. Other authors estimated more conservative increases in susceptibility ([-1.6%,8.2%]) for five European countries [270]. Nowadays, predictions can be compared with the number of cases reported in 2022, 2023, and 2024. Still, this is difficult because of the fact that surveillance of respiratory infections changed in many countries after the pandemic. However, looking at W.H.O. influenza summary reports, it appears that overall the impact of influenza epidemics in years 2022, 2023, and 2024 was not greater than in the pre-COVID-19 seasons [271, 272], although some critical epidemics have occurred locally, as was the case for the last influenza season in Italy [273]). If on the one hand flu seasons were overall less severe than expected, on the other hand, many countries experienced out-of-season flu circulation, coinciding with the timing of the relaxation of the anti-COVID-19 measures. In Europe, the 2021-2022 influenza winter season did

not start until February 2022 and peaked in late March, probably delayed by the strong measures still in place in the previous weeks [274]. Similar dynamics were observed in Senegal [275] and Brazil [276]. In addition, the duration of the 2022 and 2023 epidemics changed in several regions of the Southern Hemisphere ([51]). This unprecedented situation provides a favorable context for studying the role of immunity and its waning on the spread of seasonal influenza [277].

Several studies have shown that non-pharmaceutical interventions (NPIs) played a key role in suppressing influenza and other diseases during the pandemic [6, 19, 278]. However, the effectiveness of interventions can vary widely from country to country, depending on the epidemiological situation and the combination of many socio-demographic, geographic, and meteorological factors. In Chapter 3, we also investigated this aspect by using regression tree-based methods to identify factors associated with influenza reduction in different countries and trimesters. We then classified the countries and trimesters into five groups with similar levels of influenza reduction. As a general trend, we found that countries with high reported COVID-19 impact and strong domestic restrictions (mainly from temperate regions) experienced a more pronounced decline in influenza, while tropical countries, characterized by younger populations, lower reported COVID-19 impact, and fewer restrictions, reported some influenza circulation, albeit limited. However, two particular groups of countries-trimesters stood out. On one side, the United States and European countries experienced a limited decline in influenza in the spring of 2020, despite the stringent domestic pandemic control measures. On the other side, a few isolated countries (New Zealand, Australia, Japan, and South Korea) have halted influenza circulation by massively reducing international travel, while applying mild domestic interventions.

In our analysis, we included data on changes in mobility (both local and international) as proxies of the impact of non-pharmaceutical interventions. These variables proved to be more informative than other variables simply encoding the implemented regulations (e.g. number of days of school closures), as the firsts are less prone to definition ambiguities and more suitable to capture actual people's behaviors. Our analysis suggests that flight interruption measures had a decisive effect, consistent with previous evidence of annual reseeding of flu in temperate regions driven by global virus circulation. Recently, Pendrey and colleagues [279] further investigated this aspect in an analysis of influenza resumption in Australia between November 2021 and April 2022. However, we also found that flight interruptions were effective only for the isolated countries and only when applied with unprecedented rigor (air travels reduced by more than 90% in the isolated countries, compared to the pre-pandemic situation). For instance, we found that a 94-97% reduction in flights in Vietnam in 2020-2021 did not prevent the introduction of influenza viruses from neighboring Cambodia, likely due to traveling flows through land borders. Investigating a different context, Bajardi and co-authors have already shown that a 40% reduction in flights in Mexico during the 2009 pandemic had a limited effect on slowing the spread of the virus. These findings underscore the importance of evaluating the impact of non-pharmaceutical interventions on the propagation of infectious diseases. A task that is nowadays more and more feasible, thanks to the increasing accessibility of data, even if the integration of such data into mathematical models remains challenging [280].

The statistical approaches based on regression trees proved to be particularly useful in our large ecological study of Chapter 3, which included a high number of variables, with possibly several confounding factors. Those methods help in summarizing complex scenarios into simple classifications. In our case, we obtain an effective country-trimester classification that describes well the trends in influenza decline observed during the COVID-19 pandemic.

In Chapters 4 and 5, we examined the coupled dynamics of influenza (sub)types (i.e. influenza A\H1N1, A\H3N2, and B). Even though the (sub)type composition affects the burden of seasonal epidemics, it is often neglected in epidemiological models and statistical analyses, especially in multi-country analyses. In our studies, we exploited data from FluNet to carry out a multi-country analysis. To enable comparisons between countries with different surveillance systems, we considered the relative proportions of infections by (sub)type rather than the absolute counts. Such percentage data were pre-processed by using proper log-ratio transformations from Compositional Data Analysis (CoDA), so as to obtain data in the Euclidean metric space. This eased the formulation of statistical analyses to answer relevant epidemiological questions. This pre-processing step was adopted in both Chapter 4 and Chapter 5.

In Chapter 4, we first investigated the change of (sub)type mixing worldwide from 2000 to 2022. We summarized those trends through visualizations and detected four years in which (sub)type mixing displayed an anomalous decrease. This happened in 2003-2004, as a consequence of the Fujian flu, in 2009-2010 due to the swine flu pandemic, and in 2020-2021 and 2021-2022, due to the COVID-19 pandemic. The *mixing score* defined for this analysis can be used to follow the evolution of (sub)type mixing in the first years after the COVID-19 pandemic. Our findings revealed a geographical segregation of (sub)types that started in 2020-2021 and continued in 2021-2022 accompanied by a global resurgence of A\H3N2. More recently, WHO reported a strong dominance of A\H3N2 throughout 2022, albeit with important differences for African and South Asian regions where also A\H1N1 and B circulated ([271, 272]). The mixing score provides a synthetic metric to quantify these changes and compare the pre- and post-COVID-19 pandemic situation.

We then analyzed in more detail the 2010-2019 period - enclosed between the swine flu pandemic and the COVID-19 pandemic, respectively - that corresponds to the longest period of stable influenza activity worldwide. We aimed to identify spatial patterns in the coupled dynamics of the (sub)types. In Chapter 4, we defined country trajectories of (sub)type relative abundances (again, defined within the CoDA framework). Past studies had already discovered some spatio-temporal trends, such as the different mixing of (sub)types in temperate and tropical regions [52], or the alternating dominance of A\H1N1 and A\H3N2 [54]. CoDA allowed us to further investigate these aspects: (i) through clustering of trajectories we identified regions of the world characterized by similar patterns of (sub)type alternation, and (ii) we integrated this information into a probabilistic prediction model for the forecast of (sub)type compositions one year in advance. This is a novel problem in the literature of influenza epidemiology. The statistical framework we employed allowed us to improve the predictions obtainable from naive methods (e.g., considering simple averages of past observations). Our predictions, although with wide confidence intervals and highly variable accuracy across countries, provide valuable information. In particular, for the group of countries that includes Europe and some neighboring countries, we were able to predict with relatively good precision whether B or A\H3N2 would have been negligible (<10% of infections) and whether A\H3N2 would have been dominant (>50% of infections). Results such as these can help identify at-risk age groups, target vaccine distribution, and organize hospital patient care.

In Chapter 5, we focused on the same data from 2010 to 2019 but explored different statistical techniques. In particular, we defined a conditional copula model, to describe the dependence between the (sub)type relative abundances, conditionally upon the year and the geographical region. We decided to adapt this type of model (i.e., conditional copulas) to allow the inclusion of categorical covariates, which is often useful

in epidemiology and many applied research settings. This was achieved by integrating regression trees with copulas. Our main objective was to formulate the model and study its asymptotic consistency. In addition, we applied it to (sub)type data and recovered some patterns in line with the findings of the previous analyses. Overall, both the studies presented in Chapters 4 and 5 suggest that (i) the relative proportions of the (sub)types vary more in time than in space, and (ii) the main spatial divide is between Europe, together with a few neighboring countries, and the rest of the world. The influenza literature often distinguishes between Northern and Southern Hemisphere countries, which differ in the timing of influenza epidemics (with countries in the intratropical belt showing more complex patterns). Our analyses suggest that other spatial patterns emerge regarding subtype composition, which only partially overlap with the North vs. South classification.

The flexibility of the tree-based conditional copula model can be improved by simultaneously exploring multiple families of copulas at each split of the tree estimation. This would allow a better approximation of different types of dependence for the different groups of observations. For example, this may occur in the scenario where, for specific values of the covariates, the dependence structure is dominated by the tails of the marginal distributions. In addition, regression trees are known to produce results that may be unstable even for small variations in the input data. Therefore, an extension to random forests could be defined to improve the stability of the model. Both proposals go in the direction of providing better performing estimators, but at the expense of model interpretability. In addition, it must be considered that the computational cost of the model may increase significantly and become prohibitive, especially for copula families for which direct parameter estimation (i.e., by inverse-tau method) is not possible.

Bibliography

1. Azziz Baumgartner, E., Dao, C. N., Nasreen, S., *et al.* Seasonality, Timing, and Climate Drivers of Influenza Activity Worldwide. *The Journal of Infectious Diseases* **206,** 838–846. https://doi.org/10.1093/infdis/jis467 (2012).
2. Dave, K. & Lee, P. C. Global Geographical and Temporal Patterns of Seasonal Influenza and Associated Climatic Factors. *Epidemiologic Reviews* **41,** 51–68. https://doi.org/10.1093/epirev/mxz008 (2019).
3. Tamerius, J., Nelson, M. I., Zhou, S. Z., *et al.* Global Influenza Seasonality: Reconciling Patterns across Temperate and Tropical Regions. *Environmental Health Perspectives* **119,** 439–445. https://doi.org/10.1289/ehp.1002383 (2011).
4. Miller, M. A., Viboud, C., Balinska, M. & Simonsen, L. The Signature Features of Influenza Pandemics — Implications for Policy. *New England Journal of Medicine* **360,** 2595–2598. https://doi.org/10.1056/NEJMp0903906 (2009).
5. Olsen, S. J. Decreased Influenza Activity During the COVID-19 Pandemic — United States, Australia, Chile, and South Africa, 2020. *MMWR. Morbidity and Mortality Weekly Report* **69.** https://doi.org/10.15585/mmwr.mm6937a6 (2020).
6. Huang, Q. S., Wood, T., Jelley, L., *et al.* Impact of the COVID-19 Nonpharmaceutical Interventions on Influenza and Other Respiratory Viral Infections in New Zealand. *Nature Communications* **12,** 1001. https://doi.org/10.1038/s41467-021-21157-9 (2021).
7. Adlhoch, C., Mook, P., Lamb, F., *et al.* Very Little Influenza in the WHO European Region during the 2020/21 Season, Weeks 40 2020 to 8 2021. *Eurosurveillance* **26,** 2100221. https://doi.org/10.2807/1560-7917.ES.2021.26.11.2100221 (2021).
8. Paget, J., Spreeuwenberg, P., Charu, V., *et al.* Global Mortality Associated with Seasonal Influenza Epidemics: New Burden Estimates and Predictors from the GLaMOR Project. *Journal of Global Health* **9,** 020421. https://doi.org/10.7189/jogh.09.020421 (2019).
9. Flu scenario modeling hub. *Flu Scenario Modeling Hub* https://fluscenariomodelinghub.org/. 2023.
10. Caini, S., Andrade, W., Badur, S., *et al.* Temporal Patterns of Influenza A and B in Tropical and Temperate Countries: What Are the Lessons for Influenza Vaccination? *PLOS ONE* **11,** e0152310. https://doi.org/10.1371/journal.pone.0152310 (2016).
11. Caini, S., Spreeuwenberg, P., Kusznierz, G. F., *et al.* Distribution of Influenza Virus Types by Age Using Case-Based Global Surveillance Data from Twenty-Nine Countries, 1999-2014. *BMC Infectious Diseases* **18,** 269. https://doi.org/10.1186/s12879-018-3181-y (2018).
12. Gagnon, A., Acosta, E. & Miller, M. S. Age-Specific Incidence of Influenza A Responds to Change in Virus Subtype Dominance. *Clinical Infectious Diseases* **71,** e195–e198. https://doi.org/10.1093/cid/ciaa075 (2020).

13. Vieira, M. C., Donato, C. M., Arevalo, P., *et al.* Lineage-Specific Protection and Immune Imprinting Shape the Age Distributions of Influenza B Cases. *Nature Communications* **12,** 4313. https://doi.org/10.1038/s41467-021-24566-y (2021).

14. Trentini, F., Pariani, E., Bella, A., *et al.* Characterizing the transmission patterns of seasonal influenza in Italy: lessons from the last decade. en. *BMC Public Health* **22,** 19. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-021-12426-9 (2022).

15. Bedford, T., Riley, S., Barr, I. G., *et al.* Global Circulation Patterns of Seasonal Influenza Viruses Vary with Antigenic Drift. *Nature* **523,** 217–220. https://doi.org/10.1038/nature14460 (2015).

16. Findlater, A. & Bogoch, I. I. Human Mobility and the Global Spread of Infectious Diseases: A Focus on Air Travel. *Trends in Parasitology* **34,** 772–783. https://doi.org/10.1016/j.pt.2018.07.004 (2018).

17. GISRS. *FluNet Database - National Influenza Centres (NICs) of the Global Influenza Surveillance and Response System (GISRS) and World Health Organisation (WHO)* https://www.who.int/tools/flunet. 2022.

18. Flahault, A., Dias-Ferrao, V., Chaberty, P., *et al.* FluNet as a Tool for Global Monitoring of Influenza on the Web. *JAMA* **280,** 1330–1332. https://doi.org/10.1001/jama.280.15.1330 (1998).

19. Cowling, B. J., Ali, S. T., Ng, T. W. Y., *et al.* Impact Assessment of Non-Pharmaceutical Interventions against Coronavirus Disease 2019 and Influenza in Hong Kong: An Observational Study. *The Lancet Public Health* **5,** e279–e288. https://doi.org/10.1016/S2468-2667(20)30090-6 (2020).

20. Lee, H., Lee, H., Song, K.-H., *et al.* Impact of Public Health Interventions on Seasonal Influenza Activity During the COVID-19 Outbreak in Korea. *Clinical Infectious Diseases*. https://doi.org/10.1093/cid/ciaa672 (2020).

21. Kuo, S.-C., Shih, S.-M., Chien, L.-H. & Hsiung, C. A. Collateral Benefit of COVID-19 Control Measures on Influenza Activity, Taiwan - Volume 26, Number 8—August 2020 - Emerging Infectious Diseases Journal - CDC. https://doi.org/10.3201/eid2608.201192 (2020).

22. Soo, R. J. J., Chiew, C. J., Ma, S., Pung, R. & Lee, V. Decreased Influenza Incidence under COVID-19 Control Measures, Singapore. *Emerging Infectious Diseases* **26,** 1933–1935. https://doi.org/10.3201/eid2608.201229 (2020).

23. Belongia, E. A. & Osterholm, M. T. COVID-19 and Flu, a Perfect Storm. *Science* **368,** 1163–1163. https://doi.org/10.1126/science.abd2220 (2020).

24. Zipfel, C. M., Colizza, V. & Bansal, S. The Missing Season: The Impacts of the COVID-19 Pandemic on Influenza. *Vaccine* **39,** 3645–3648. https://doi.org/10.1016/j.vaccine.2021.05.049 (2021).

25. Dhanasekaran, V., Sullivan, S., Edwards, K. M., *et al.* Human Seasonal Influenza under COVID-19 and the Potential Consequences of Influenza Lineage Elimination. *Nature Communications* **13,** 1721. https://doi.org/10.1038/s41467-022-29402-5 (2022).

26. Koutsakos, M., Wheatley, A. K., Laurie, K., Kent, S. J. & Rockman, S. Influenza Lineage Extinction during the COVID-19 Pandemic? *Nature Reviews. Microbiology* **19,** 741–742. https://doi.org/10.1038/s41579-021-00642-4 (2021).

27. Caini, S., Meijer, A., Nunes, M. C., *et al. Is Influenza B/Yamagata Extinct and What Public Health Implications Could This Have? An Updated Literature Review and Comprehensive Assessment of Global Surveillance Databases* 2023. https://doi.org/10.1101/2023.09.25.23296068.

28. Aitchison, J. *The Statistical Analysis of Compositional Data* (Chapman & Hall, Ltd., GBR, 1986).

29. Nelsen, R. B. *An Introduction to Copulas* 2. ed (Springer, New York Berlin Heidelberg, 2006).

30. Bonacina, F., Boëlle, P.-Y., Colizza, V., *et al.* Global Patterns and Drivers of Influenza Decline during the COVID-19 Pandemic. *International Journal of Infectious Diseases* **128,** 132–139. https://doi.org/10.1016/j.ijid.2022.12.042 (2023).

31. Bonacina, F., Boëlle, P.-Y., Lopez, O., Thomas, M. & Poletto, C. Understanding the Coupled Dynamics of Influenza (Sub)Types: A Global Analysis Leveraging Compositional Data Analysis. *In preparation.*

32. Bonacina, F., Lopez, O. & Thomas, M. *Tree-Based Conditional Copula Estimation*

33. Kutter, J. S., Spronken, M. I., Fraaij, P. L., Fouchier, R. A. & Herfst, S. Transmission Routes of Respiratory Viruses among Humans. *Current Opinion in Virology. Emerging Viruses: Intraspecies Transmission • Viral Immunology* **28,** 142–151. https://doi.org/10.1016/j.coviro.2018.01.001 (2018).

34. Kackos, C. M. & Webby, R. J. in *Reference Module in Life Sciences* (Elsevier, 2023). https://doi.org/10.1016/B978-0-12-822563-9.00101-3. (2024).

35. Kessler, S., Harder, T. C., Schwemmle, M. & Ciminski, K. Influenza A Viruses and Zoonotic Events—Are We Creating Our Own Reservoirs? *Viruses* **13,** 2250. https://doi.org/10.3390/v13112250 (2021).

36. Olsen, B., Munster, V. J., Wallensten, A., *et al.* Global Patterns of Influenza A Virus in Wild Birds. *Science* **312,** 384–388. https://doi.org/10.1126/science.1122438 (2006).

37. Krauss, S., Stallknecht, D. E., Negovetich, N. J., *et al.* Coincident Ruddy Turnstone Migration and Horseshoe Crab Spawning Creates an Ecological 'Hot Spot' for Influenza Viruses. *Proceedings of the Royal Society B: Biological Sciences* **277,** 3373–3379. https://doi.org/10.1098/rspb.2010.1090 (2010).

38. Kim, Y., Biswas, P. K., Giasuddin, M., *et al.* Prevalence of Avian Influenza A(H5) and A(H9) Viruses in Live Bird Markets, Bangladesh. *Emerging Infectious Diseases* **24,** 2309–2316. https://doi.org/10.3201/eid2412.180879 (2018).

39. Mostafa, A., Abdelwhab, E. M., Mettenleiter, T. C. & Pleschka, S. Zoonotic Potential of Influenza A Viruses: A Comprehensive Overview. *Viruses* **10,** 497. https://doi.org/10.3390/v10090497 (2018).

40. Petrova, V. N. & Russell, C. A. The Evolution of Seasonal Influenza Viruses. *Nature Reviews Microbiology* **16,** 47–60. https://doi.org/10.1038/nrmicro.2017.118 (2018).

41. Jones, K. E., Patel, N. G., Levy, M. A., *et al.* Global Trends in Emerging Infectious Diseases. *Nature* **451,** 990–993. https://doi.org/10.1038/nature06536 (2008).

42. Kucharski, A. J., Lessler, J., Read, J. M., *et al.* Estimating the Life Course of Influenza A(H3N2) Antibody Responses from Cross-Sectional Data. *PLOS Biology* **13,** e1002082. https://doi.org/10.1371/journal.pbio.1002082 (2015).

43. Smith, A. J. & Davies, J. R. Natural Infection with Influenza A (H3N2). The Development, Persistence and Effect of Antibodies to the Surface Antigens. *Epidemiology & Infection* **77,** 271–282. https://doi.org/10.1017/S0022172400024712 (1976).

44. Lessler, J., Riley, S., Read, J. M., *et al.* Evidence for Antigenic Seniority in Influenza A (H3N2) Antibody Responses in Southern China. *PLOS Pathogens* **8,** e1002802. https://doi.org/10.1371/journal.ppat.1002802 (2012).

45. Russell, C. A., Jones, T. C., Barr, I. G., *et al.* The Global Circulation of Seasonal Influenza A (H3N2) Viruses. *Science* **320,** 340–346. https://doi.org/10.1126/science.1154137 (2008).

46. Agor, J. K. Models for Predicting the Evolution of Influenza to Inform Vaccine Strain Selection. *HUMAN VACCINES,* 7 (2018).

47. Biere, B., Bauer, B. & Schweiger, B. Differentiation of Influenza B Virus Lineages Yamagata and Victoria by Real-Time PCR. *Journal of Clinical Microbiology* **48,** 1425–1427. https://doi.org/10.1128/jcm.02116-09 (2010).

48. CDC. *Burden of Influenza* https://www.cdc.gov/flu/about/burden/index.html. 2023.

49. Tisa, V., Barberis, I., Faccio, V., *et al.* Quadrivalent Influenza Vaccine: A New Opportunity to Reduce the Influenza Burden. *Journal of Preventive Medicine and Hygiene* **57,** E28–E33 (2016).

50. Rouzine, I. M. & Rozhnova, G. Antigenic evolution of viruses in host populations. en. *PLOS Pathogens* **14** (ed Vignuzzi, M.) e1007291. https://dx.plos.org/10.1371/journal.ppat.1007291 (2018).

51. Zheng, L., Lin, Y., Yang, J., *et al.* Global Variability of Influenza Activity and Virus Subtype Circulation from 2011 to 2023. *BMJ Open Respiratory Research* **10,** e001638. https://doi.org/10.1136/bmjresp-2023-001638 (2023).

52. Zanobini, P., Bonaccorsi, G., Lorini, C., *et al.* Global Patterns of Seasonal Influenza Activity, Duration of Activity and Virus (Sub)Type Circulation from 2010 to 2020. *Influenza and Other Respiratory Viruses* **16,** 696–706. https://doi.org/10.1111/irv.12969 (2022).

53. An der Heiden, M. & Buchholz, U. Estimation of Influenza-Attributable Medically Attended Acute Respiratory Illness by Influenza Type/Subtype and Age, Germany, 2001/02-2014/15. *Influenza and Other Respiratory Viruses* **11,** 110–121. https://doi.org/10.1111/irv.12434 (2017).

54. Finkelman, B. S., Viboud, C., Koelle, K., *et al.* Global Patterns in Seasonal Activity of Influenza A/H3N2, A/H1N1, and B from 1997 to 2005: Viral Coexistence and Latitudinal Gradients. *PLoS ONE* **2** (ed Myer, L.) e1296. https://doi.org/10.1371/journal.pone.0001296 (2007).

55. Smith, D. J., Lapedes, A. S., de Jong, J. C., *et al.* Mapping the Antigenic and Genetic Evolution of Influenza Virus. *Science* **305,** 371–376. https://doi.org/10.1126/science.1097211 (2004).

56. Laurie, K. L., Horman, W., Carolan, L. A., *et al.* Evidence for Viral Interference and Cross-reactive Protective Immunity Between Influenza B Virus Lineages. *The Journal of Infectious Diseases* **217,** 548–559. https://doi.org/10.1093/infdis/jix509 (2018).

57. Nyirenda, M., Omori, R., Tessmer, H. L., Arimura, H. & Ito, K. Estimating the Lineage Dynamics of Human Influenza B Viruses. *PLoS ONE* **11,** e0166107. https://doi.org/10.1371/journal.pone.0166107 (2016).

58. Laurie, K. L., Guarnaccia, T. A., Carolan, L. A., *et al.* Interval Between Infections and Viral Hierarchy Are Determinants of Viral Interference Following Influenza Virus Infection in a Ferret Model. *The Journal of Infectious Diseases* **212,** 1701–1710. https://doi.org/10.1093/infdis/jiv260 (2015).

59. Yang, W., Lau, E. H. Y. & Cowling, B. J. Dynamic Interactions of Influenza Viruses in Hong Kong during 1998-2018. *PLOS Computational Biology* **16,** e1007989. https://doi.org/10.1371/journal.pcbi.1007989 (2020).

60. Goldstein, E., Cobey, S., Takahashi, S., Miller, J. C. & Lipsitch, M. Predicting the Epidemic Sizes of Influenza A/H1N1, A/H3N2, and B: A Statistical Method. *PLOS Medicine* **8,** e1001051. https://doi.org/10.1371/journal.pmed.1001051 (2011).

61. Gatti, L., Koenen, M. H., Zhang, J. D., *et al.* Cross-Reactive Immunity Potentially Drives Global Oscillation and Opposed Alternation Patterns of Seasonal Influenza A Viruses. *Scientific Reports* **12,** 8883. https://doi.org/10.1038/s41598-022-08233-w (2022).

62. WHO - FluNet. *Influenza Laboratory Influenza Information - Influenza Virus Detection Reported to FluNet* https://app.powerbi.com/view?r=eyJrIjoiNjViM2Y4NjktMjJmMC00Y2NjLW

63. Viboud, C., Grais, R. F., Lafont, B. A. P., Miller, M. A. & Simonsen, L. Multinational Impact of the 1968 Hong Kong Influenza Pandemic: Evidence for a Smoldering Pandemic. *The Journal of Infectious Diseases* **192,** 233–248. https://doi.org/10.1086/431150 (2005).

64. Launay, T., Souty, C., Vilcu, A.-M., *et al.* Common Communicable Diseases in the General Population in France during the COVID-19 Pandemic. *PLOS ONE* **16,** e0258391. https://doi.org/10.1371/journal.pone.0258391 (2021).

65. Caini, S., Alonso, W. J., Séblain, C. E.-G., Schellevis, F. & Paget, J. The Spatiotemporal Characteristics of Influenza A and B in the WHO European Region: Can One Define Influenza Transmission Zones in Europe? *Eurosurveillance* **22,** 30606. https://doi.org/10.2807/1560-7917.ES.2017.22.35.30606 (2017).

66. Alonso, W. J., Yu, C., Viboud, C., *et al.* A Global Map of Hemispheric Influenza Vaccine Recommendations Based on Local Patterns of Viral Circulation. *Scientific Reports* **5,** 17214. https://doi.org/10.1038/srep17214 (2015).

67. Tamerius, J. D., Shaman, J., Alonso, W. J., *et al.* Environmental Predictors of Seasonal Influenza Epidemics across Temperate and Tropical Climates. *PLOS Pathogens* **9,** e1003194. https://doi.org/10.1371/journal.ppat.1003194 (2013).

68. Lowen, A. C., Mubareka, S., Steel, J. & Palese, P. Influenza Virus Transmission Is Dependent on Relative Humidity and Temperature. *PLOS Pathogens* **3,** e151. https://doi.org/10.1371/journal.ppat.0030151 (2007).

69. Moriyama, M., Hugentobler, W. J. & Iwasaki, A. Seasonality of Respiratory Viral Infections. *Annual Review of Virology* **7,** 83–101. https://doi.org/10.1146/annurev-virology-012420-022445 (2020).

70. Worby, C. J., Chaves, S. S., Wallinga, J., *et al.* On the Relative Role of Different Age Groups in Influenza Epidemics. *Epidemics* **13,** 10–16. https://doi.org/10.1016/j.epidem.2015.04.003 (2015).

71. Cauchemez, S., Van Kerkhove, M. D., Archer, B. N., *et al.* School Closures during the 2009 Influenza Pandemic: National and Local Experiences. *BMC Infectious Diseases* **14,** 207. https://doi.org/10.1186/1471-2334-14-207 (2014).

72. Ewing, A., Lee, E. C., Viboud, C. & Bansal, S. Contact, Travel, and Transmission: The Impact of Winter Holidays on Influenza Dynamics in the United States. *The Journal of Infectious Diseases* **215,** 732–739. https://doi.org/10.1093/infdis/jiw642 (2017).

73. Luca, G. D., Kerckhove, K. V., Coletti, P., *et al.* The Impact of Regular School Closure on Seasonal Influenza Epidemics: A Data-Driven Spatial Transmission Model for Belgium. *BMC Infectious Diseases* **18,** 29. https://doi.org/10.1186/s12879-017-2934-3 (2018).

74. Ajelli, M., Poletti, P., Melegaro, A. & Merler, S. The role of different social contexts in shaping influenza transmission during the 2009 pandemic. *Scientific Reports* **4,** 7218. https://doi.org/10.1038/srep07218 (2014).

75. Lemey, P., Rambaut, A., Bedford, T., *et al.* Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2. *PLOS Pathogens* **10,** e1003932. https://doi.org/10.1371/journal.ppat.1003932 (2014).

76. He, D., Lui, R., Wang, L., *et al.* Global Spatio-temporal Patterns of Influenza in the Post-pandemic Era. *Scientific Reports* **5,** 11013. https://doi.org/10.1038/srep11013 (2015).

77. OECD & European Union. *Health at a Glance: Europe 2018: State of Health in the EU Cycle* https://doi.org/10.1787/health_glance_eur-2018-en. (2023) (OECD, 2018).

78. Antonova, E. N., Rycroft, C. E., Ambrose, C. S., Heikkinen, T. & Principi, N. Burden of Paediatric Influenza in Western Europe: A Systematic Review. *BMC Public Health* **12,** 968. https://doi.org/10.1186/1471-2458-12-968 (2012).

79. Putri, W. C. W. S., Muscatello, D. J., Stockwell, M. S. & Newall, A. T. Economic Burden of Seasonal Influenza in the United States. *Vaccine* **36,** 3960–3966. https://doi.org/10.1016/j.vaccine.2018.05.057 (2018).

80. Buchy, P., Buisson, Y., Cintra, O., *et al.* COVID-19 Pandemic: Lessons Learned from More than a Century of Pandemics and Current Vaccine Development for Pandemic Control. *International Journal of Infectious Diseases* **112,** 300–317. https://doi.org/10.1016/j.ijid.2021.09.045 (2021).

81. Barry, J. M. *The Great Influenza: The Story of the Deadliest Pandemic in History* (Penguin UK, 2020).

82. Kelly, H., Peck, H. A., Laurie, K. L., *et al.* The Age-Specific Cumulative Incidence of Infection with Pandemic Influenza H1N1 2009 Was Similar in Various Countries Prior to Vaccination. *PLOS ONE* **6,** e21828. https://doi.org/10.1371/journal.pone.0021828 (2011).

83. Simonsen, L., Spreeuwenberg, P., Lustig, R., *et al.* Global Mortality Estimates for the 2009 Influenza Pandemic from the GLaMOR Project: A Modeling Study. *PLOS Medicine* **10,** e1001558. https://doi.org/10.1371/journal.pmed.1001558 (2013).

84. Taylor, S. in *Health in a Post-COVID World* 1–10 (Policy Press, 2023). (2024).

85. Liu, Y. The Impact of Non-Pharmaceutical Interventions on SARS-CoV-2 Transmission across 130 Countries and Territories, 12 (2021).

86. Li, Y., Campbell, H., Kulkarni, D., *et al.* The Temporal Association of Introducing and Lifting Non-Pharmaceutical Interventions with the Time-Varying Reproduction Number (R) of SARS-CoV-2: A Modelling Study across 131 Countries. *The Lancet Infectious Diseases* **21,** 193–202. https://doi.org/10.1016/S1473-3099(20)30785-4 (2021).
87. WHO. *Preparing for Pandemics* https://www.who.int/westernpacific/activities/preparing-for-pandemics.
88. Trifonov, V., Khiabanian, H. & Rabadan, R. Geographic Dependence, Surveillance, and Origins of the 2009 Influenza A (H1N1) Virus. *New England Journal of Medicine* **361,** 115–119. https://doi.org/10.1056/NEJMp0904572 (2009).
89. Marani, M., Katul, G. G., Pan, W. K. & Parolari, A. J. Intensity and Frequency of Extreme Novel Epidemics. *Proceedings of the National Academy of Sciences* **118,** e2105482118. https://doi.org/10.1073/pnas.2105482118 (2021).
90. Williams, B. A., Jones, C. H., Welch, V. & True, J. M. Outlook of Pandemic Preparedness in a Post-COVID-19 World. *npj Vaccines* **8,** 1–12. https://doi.org/10.1038/s41541-023-00773-0 (2023).
91. Carlson, C. J., Albery, G. F., Merow, C., *et al.* Climate Change Increases Cross-Species Viral Transmission Risk. *Nature* **607,** 555–562. https://doi.org/10.1038/s41586-022-04788-w (2022).
92. Vora, N. M., Hannah, L., Lieberman, S., *et al.* Want to Prevent Pandemics? Stop Spillovers. *Nature* **605,** 419–422. https://doi.org/10.1038/d41586-022-01312-y (2022).
93. Krammer, F. & Schultz-Cherry, S. We Need to Keep an Eye on Avian Influenza. *Nature Reviews Immunology* **23,** 267–268. https://doi.org/10.1038/s41577-023-00868-8 (2023).
94. Adlhoch, C. & Baldinelli, F. Avian Influenza, New Aspects of an Old Threat. *Eurosurveillance* **28,** 2300227. https://doi.org/10.2807/1560-7917.ES.2023.28.19.2300227 (2023).
95. WHO TEAM - Global Influenza Programme (GIP). *A Checklist for Respiratory Pathogen Pandemic Preparedness Planning* in (World Health Organization, 2023), 56.
96. WHO & PRET. *Preparedness and Resilience for Emerging Threats Module 1: Planning for Respiratory Pathogen Pandemics. Version 1.0* 2023.
97. Sentinelles, R. *Reseau Sentinelles - Bilan d'activité 2022* tech. rep. (IPLESP, 2023). https://doi.org/10.4000/books.iheal.1740. (2024).
98. Mook, P., Meerhoff, T., Olsen, S. J., *et al.* Alternating Patterns of Seasonal Influenza Activity in the WHO European Region Following the 2009 Pandemic, 2010-2018. *Influenza and Other Respiratory Viruses* **14,** 150–161. https://doi.org/10.1111/irv.12703 (2020).
99. Bahl, J., Nelson, M. I., Chan, K. H., *et al.* Temporally Structured Metapopulation Dynamics and Persistence of Influenza A H3N2 Virus in Humans. *Proceedings of the National Academy of Sciences* **108,** 19359–19364. https://doi.org/10.1073/pnas.1109314108 (2011).
100. Le, M. Q., Lam, H. M., Cuong, V. D., *et al.* Migration and Persistence of Human Influenza A Viruses, Vietnam, 2001–2008 - Volume 19, Number 11—November 2013 - Emerging Infectious Diseases Journal - CDC. https://doi.org/10.3201/eid1911.130349 (2013).

101. McGowan, C. J., Biggerstaff, M., Johansson, M., *et al.* Collaborative Efforts to Forecast Seasonal Influenza in the United States, 2015–2016. *Scientific Reports* **9,** 683. https://doi.org/10.1038/s41598-018-36361-9 (2019).

102. Viboud, C. & Vespignani, A. The Future of Influenza Forecasts. *Proceedings of the National Academy of Sciences* **116,** 2802–2804. https://doi.org/10.1073/pnas.1822167116 (2019).

103. Biggerstaff, M., Johansson, M., Alper, D., *et al.* Results from the Second Year of a Collaborative Effort to Forecast Influenza Seasons in the United States. *Epidemics* **24,** 26–33. https://doi.org/10.1016/j.epidem.2018.02.003 (2018).

104. Reich, N. G., Brooks, L. C., Fox, S. J., *et al.* A Collaborative Multiyear, Multimodel Assessment of Seasonal Influenza Forecasting in the United States. *Proceedings of the National Academy of Sciences* **116,** 3146–3154. https://doi.org/10.1073/pnas.1812594116 (2019).

105. RespiCast, R. |. E. R. D. F. *RespiCast | European Respiratory Diseases Forecasting Hub* https://respicast.ecdc.europa.eu.

106. Sherratt, K., Gruson, H., Grah, R., *et al.* Predictive Performance of Multi-Model Ensemble Forecasts of COVID-19 across European Nations. *eLife* **12** (eds Wesolowski, A., Ferguson, N. M., Shaman, J. L. & Pei, S.) e81916. https://doi.org/10.7554/eLife.81916 (2023).

107. Cramer, E. Y., Ray, E. L., Lopez, V. K., *et al.* Evaluation of Individual and Ensemble Probabilistic Forecasts of COVID-19 Mortality in the United States. *Proceedings of the National Academy of Sciences* **119,** e2113561119. https://doi.org/10.1073/pnas.2113561119 (2022).

108. Viboud, C., Sun, K., Gaffey, R., *et al.* The RAPIDD Ebola Forecasting Challenge: Synthesis and Lessons Learnt. *Epidemics. The RAPIDD Ebola Forecasting Challenge* **22,** 13–21. https://doi.org/10.1016/j.epidem.2017.08.002 (2018).

109. Ajelli, M., Gonçalves, B., Balcan, D., *et al.* Comparing Large-Scale Computational Approaches to Epidemic Modeling: Agent-based versus Structured Metapopulation Models. *BMC Infectious Diseases* **10,** 190. https://doi.org/10.1186/1471-2334-10-190 (2010).

110. Apolloni, A., Poletto, C., Ramasco, J. J., Jensen, P. & Colizza, V. Metapopulation Epidemic Models with Heterogeneous Mixing and Travel Behaviour. *Theoretical Biology and Medical Modelling* **11,** 3. https://doi.org/10.1186/1742-4682-11-3 (2014).

111. Meloni, S., Perra, N., Arenas, A., *et al.* Modeling Human Mobility Responses to the Large-Scale Spreading of Infectious Diseases. *Scientific Reports* **1,** 62. https://doi.org/10.1038/srep00062 (2011).

112. Merler, S. & Ajelli, M. The Role of Population Heterogeneity and Human Mobility in the Spread of Pandemic Influenza. *Proceedings of the Royal Society B: Biological Sciences* **277,** 557–565. https://doi.org/10.1098/rspb.2009.1605 (2009).

113. Poletto, C., Pelat, C., Lévy-Bruhl, D., *et al.* 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,** 20824. https://doi.org/10.2807/1560-7917.ES2014.19.23.20824 (2014).

114. Pullano, G., Pinotti, F., Valdano, E., *et al.* Novel Coronavirus (2019-nCoV) Early-Stage Importation Risk to Europe, January 2020. *Eurosurveillance* **25,** 2000057. https://doi.org/10.2807/1560-7917.ES.2020.25.4.2000057 (2020).

115. Choi, S. B., Kim, J. & Ahn, I. Forecasting Type-Specific Seasonal Influenza after 26 Weeks in the United States Using Influenza Activities in Other Countries. *PloS One* **14,** e0220423. https://doi.org/10.1371/journal.pone.0220423 (2019).

116. Kandula, S., Yang, W. & Shaman, J. Type- and Subtype-Specific Influenza Forecast. *American Journal of Epidemiology* **185,** 395–402. https://doi.org/10.1093/aje/kww211 (2017).

117. Nickbakhsh, S., Mair, C., Matthews, L., *et al.* Virus–Virus Interactions Impact the Population Dynamics of Influenza and the Common Cold. *Proceedings of the National Academy of Sciences* **116,** 27142–27150. https://doi.org/10.1073/pnas.1911083116 (2019).

118. Waterlow, N. R., Flasche, S., Minter, A. & Eggo, R. M. Competition between RSV and Influenza: Limits of Modelling Inference from Surveillance Data. *Epidemics* **35,** 100460. https://doi.org/10.1016/j.epidem.2021.100460 (2021).

119. Zhang, L., Xiao, Y., Xiang, Z., *et al.* Statistical Analysis of Common Respiratory Viruses Reveals the Binary of Virus-Virus Interaction. *Microbiology Spectrum* **0,** e00019–23. https://doi.org/10.1128/spectrum.00019-23 (2023).

120. Bajardi, P., Poletto, C., Ramasco, J. J., *et al.* Human Mobility Networks, Travel Restrictions, and the Global Spread of 2009 H1N1 Pandemic. *PLoS ONE* **6** (ed Perc, M.) e16591. https://doi.org/10.1371/journal.pone.0016591 (2011).

121. Pang, X., Zhu, Z., Xu, F., *et al.* Evaluation of Control Measures Implemented in the Severe Acute Respiratory Syndrome Outbreak in Beijing, 2003. *JAMA* **290,** 3215–3221. https://doi.org/10.1001/jama.290.24.3215 (2003).

122. Svoboda, T., Henry, B., Shulman, L., *et al.* Public Health Measures to Control the Spread of the Severe Acute Respiratory Syndrome during the Outbreak in Toronto. *New England Journal of Medicine* **350,** 2352–2361. https://doi.org/10.1056/NEJMoa032111 (2004).

123. Ciavarella, C., Fumanelli, L., Merler, S., Cattuto, C. & Ajelli, M. School closure policies at municipality level for mitigating influenza spread: a model-based evaluation. *BMC Infectious Diseases* **16,** 576. https://doi.org/10.1186/s12879-016-1918-z (2016).

124. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification And Regression Trees* 1st ed. https://doi.org/10.1201/9781315139470. (2022) (Chapman and Hall/CRC, 1984).

125. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32. https://doi.org/10.1023/A:1010933404324 (2001).

126. Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D. & Rakowski, W. Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison with Logistic Regression. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine* **26,** 172–181. https://doi.org/10.1207/S15324796ABM2603_02 (2003).

127. De'ath, G. & Fabricius, K. E. Classification and Regression Trees: A Powerful yet Simple Technique for Ecological Data Analysis. *Ecology* **81,** 3178–3192. https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2 (2000).

128. Cutler, D. R., Edwards Jr., T. C., Beard, K. H., *et al.* Random Forests for Classification in Ecology. *Ecology* **88,** 2783–2792. https://doi.org/10.1890/07-0539.1 (2007).

129. Athey, S. & Imbens, G. W. Machine Learning Methods That Economists Should Know About. *Annual Review of Economics* **11,** 685–725. https://doi.org/10.1146/annurev-economics-080217-053433 (2019).

130. Vaart, A. W. v. d. & Wellner, J. A. in *Weak Convergence and Empirical Processes: With Applications to Statistics* (eds van der Vaart, A. W. & Wellner, J. A.) 127–384 (Springer International Publishing, Cham, 2023). https://doi.org/10.1007/978-3-031-29040-4_2.

131. Van der Vaart, A. W. *Asymptotic Statistics* https://doi.org/10.1017/CBO9780511802256. (2024) (Cambridge University Press, Cambridge, 1998).

132. Therneau, T. & Clinic, M. User Written Splitting Functions for RPART (2022).

133. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. VSURF: An R Package for Variable Selection Using Random Forests. *The R Journal* **7,** 19–33 (2015).

134. Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics* **26,** 1340–1347. https://doi.org/10.1093/bioinformatics/btq134 (2010).

135. Greenwell, B. M., Boehmke, B. C. & McCarthy, A. J. A Simple and Effective Model-Based Variable Importance Measure. *ArXiv* (2018).

136. Antoniadis, A., Lambert-Lacroix, S. & Poggi, J.-M. Random Forests for Global Sensitivity Analysis: A Selective Review. *Reliability Engineering & System Safety* **206,** 107312. https://doi.org/10.1016/j.ress.2020.107312 (2021).

137. Filzmoser, P., Hron, K. & Templ, M. *Applied Compositional Data Analysis: With Worked Examples in R* https://doi.org/10.1007/978-3-319-96422-5. (2023) (Springer International Publishing, Cham, 2018).

138. Pearson, K. Mathematical Contributions to the Theory of Evolution.—On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs. *Proceedings of the Royal Society of London* **60,** 489–498. https://doi.org/10.1098/rspl.1896.0076 (1997).

139. Chayes, F. On Correlation between Variables of Constant Sum. *Journal of Geophysical Research (1896-1977)* **65,** 4185–4193. https://doi.org/10.1029/JZ065i012p04185 (1960).

140. Aitchison, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44,** 139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x (1982).

141. Aitchison, J. & J. Egozcue, J. Compositional Data Analysis: Where Are We and Where Should We Be Heading? *Mathematical Geology* **37,** 829–850. https://doi.org/10.1007/s11004-005-7383-7 (2005).

142. Filzmoser, P., Hron, K. & Reimann, C. Univariate Statistical Analysis of Environmental (Compositional) Data: Problems and Possibilities. *Science of The Total Environment* **407,** 6100–6108. https://doi.org/10.1016/j.scitotenv.2009.08.008 (2009).

143. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* **8.** https://doi.org/10.3389/fmicb.2017.02224 (2017).

144. Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barceló-Vidal, C. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology* **35,** 279–300. https://doi.org/10.1023/A:1023818214614 (2003).

145. Lu, F., Zheng, Y., Cleveland, H., Burton, C. & Madigan, D. Bayesian Hierarchical Vector Autoregressive Models for Patient-Level Predictive Modeling. *PLOS ONE* **13,** e0208082. https://doi.org/10.1371/journal.pone.0208082 (2018).

146. Bracher, J., Ray, E. L., Gneiting, T. & Reich, N. G. Evaluating Epidemic Forecasts in an Interval Format. *PLOS Computational Biology* **17,** e1008618. https://doi.org/10.1371/journal.pcbi.1008618 (2021).

147. Held, L., Meyer, S. & Bracher, J. Probabilistic Forecasting in Infectious Disease Epidemiology: The 13th Armitage Lecture. *Statistics in Medicine* **36,** 3443–3460. https://doi.org/10.1002/sim.7363 (2017).

148. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis: With ... 36 Tables* 1. ed., corr. 2. print (Springer, Berlin Heidelberg, 2007).

149. Frees, E. W. & Valdez, E. A. Understanding Relationships Using Copulas. *North American Actuarial Journal* **2,** 1–25. https://doi.org/10.1080/10920277.1998.10595667 (1998).

150. Genest, C. & Rivest, L.-P. Statistical Inference Procedures for Bivariate Archimedean Copulas. *Journal of the American Statistical Association* **88,** 1034–1043. https://doi.org/10.1080/01621459.1993.10476372 (1993).

151. Acar, E. F., Craiu, R. V. & Yao, F. Dependence Calibration in Conditional Copulas: A Nonparametric Approach. *Biometrics* **67,** 445–453. JSTOR: 41242482 (2011).

152. Abegaz, F., Gijbels, I. & Veraverbeke, N. Semiparametric Estimation of Conditional Copulas. *Journal of Multivariate Analysis. Special Issue on Copula Modeling and Dependence* **110,** 43–73. https://doi.org/10.1016/j.jmva.2012.04.001 (2012).

153. Sabeti, A., Wei, M. & Craiu, R. V. Additive Models for Conditional Copulas. *Stat* **3,** 300–312. https://doi.org/10.1002/sta4.64 (2014).

154. Grazian, C., Dalla Valle, L. & Liseo, B. Approximate Bayesian Conditional Copulas. *Computational Statistics & Data Analysis* **169,** 107417. https://doi.org/10.1016/j.csda.2021.107417 (2022).

155. Bhatti, M. I. & Do, H. Q. Recent Development in Copula and Its Applications to the Energy, Forestry and Environmental Sciences. *International Journal of Hydrogen Energy* **44,** 19453–19473. https://doi.org/10.1016/j.ijhydene.2019.06.015 (2019).

156. Wang, Q. Multivariate Kernel Smoothing and Its Applications. *Journal of the American Statistical Association* **115,** 486–486. https://doi.org/10.1080/01621459.2020.1721247 (2020).

157. Gijbels, I., Veraverbeke, N. & Omelka, M. Conditional Copulas, Association Measures and Their Applications. *Computational Statistics & Data Analysis* **55,** 1919–1932. https://doi.org/10.1016/j.csda.2010.11.010 (2011).

158. Shih, J. H. & Louis, T. A. Inferences on the Association Parameter in Copula Models for Bivariate Survival Data. *Biometrics* **51,** 1384–1399. https://doi.org/10.2307/2533269. JSTOR: 2533269 (1995).

159. Genest, C. & Werker, B. J. M. in *Distributions With Given Marginals and Statistical Modelling* (eds Cuadras, C. M., Fortiana, J. & Rodriguez-Lallena, J. A.) 103–112 (Springer Netherlands, Dordrecht, 2002). https://doi.org/10.1007/978-94-017-0061-0_12. (2024).

160. Kojadinovic, I. & Yan, J. Comparison of Three Semiparametric Methods for Estimating Dependence Parameters in Copula Models. *Insurance: Mathematics and Economics* **47,** 52–63. https://doi.org/10.1016/j.insmatheco.2010.03.008 (2010).

161. Derumigny, A. & Fermanian, J.-D. About Tests of the "Simplifying" Assumption for Conditional Copulas. *Working Papers* (2017).

162. Patton, A. J. MODELLING ASYMMETRIC EXCHANGE RATE DEPENDENCE*. *International Economic Review* **47,** 527–556. https://doi.org/10.1111/j.1468-2354.2006.00387.x (2006).

163. Veraverbeke, N., Omelka, M. & Gijbels, I. Estimation of a Conditional Copula and Association Measures. *Scandinavian Journal of Statistics* **38,** 766–780. https://doi.org/10.1111/j.1467-9469.2011.00744.x (2011).

164. Emborg, H.-D., Carnahan, A., Bragstad, K., *et al.* Abrupt Termination of the 2019/20 Influenza Season Following Preventive Measures against COVID-19 in Denmark, Norway and Sweden. *Eurosurveillance* **26,** 2001160. https://doi.org/10.2807/1560-7917.ES.2021.26.22.2001160 (2021).

165. Qi, Y., Shaman, J. & Pei, S. Quantifying the Impact of COVID-19 Nonpharmaceutical Interventions on Influenza Transmission in the United States. *The Journal of Infectious Diseases* **224,** 1500–1508. https://doi.org/10.1093/infdis/jiab485 (2021).

166. Baker, R. E., Park, S. W., Yang, W., *et al.* The Impact of COVID-19 Nonpharmaceutical Interventions on the Future Dynamics of Endemic Infections. *Proceedings of the National Academy of Sciences* **117,** 30547–30553. https://doi.org/10.1073/pnas.2013182117 (2020).

167. Chow, E. J., Uyeki, T. M. & Chu, H. Y. The Effects of the COVID-19 Pandemic on Community Respiratory Virus Activity. *Nature Reviews Microbiology*, 1–16. https://doi.org/10.1038/s41579-022-00807-9 (2022).

168. Ullrich, A., Schranz, M., Rexroth, U., *et al.* Impact of the COVID-19 Pandemic and Associated Non-Pharmaceutical Interventions on Other Notifiable Infectious Diseases in Germany: An Analysis of National Surveillance Data during Week 1–2016 – Week 32–2020. *The Lancet Regional Health - Europe* **6,** 100103. https://doi.org/10.1016/j.lanepe.2021.100103 (2021).

169. Davis, W. W., Mott, J. A. & Olsen, S. J. The Role of Non-Pharmaceutical Interventions on Influenza Circulation during the COVID-19 Pandemic in Nine Tropical Asian Countries. *Influenza and Other Respiratory Viruses* **16,** 568–576. https://doi.org/10.1111/irv.12953 (2022).

170. Qiu, Z., Cao, Z., Zou, M., *et al.* The Effectiveness of Governmental Nonpharmaceutical Interventions against COVID-19 at Controlling Seasonal Influenza Transmission: An Ecological Study. *BMC infectious diseases* **22,** 331. https://doi.org/10.1186/s12879-022-07317-2 (2022).

171. Siegers, J. Y., Dhanasekaran, V., Xie, R., *et al.* Genetic and Antigenic Characterization of an Influenza A(H3N2) Outbreak in Cambodia and the Greater Mekong Subregion during the COVID-19 Pandemic, 2020. *Journal of Virology* **95,** e01267–21. https://doi.org/10.1128/JVI.01267-21 (2021).

172. Karlsson, E. A. *Review of Global Influenza Circulation, Late 2019 to 2020, and the Impact of the COVID-19 Pandemic on Influenza Circulation* Weekly Epidemiological Record 96(25) (World Health Organization, 2021), 241–265. (2023).

173. Ritchie, H., Mathieu, E., Rodés-Guirao, L., *et al.* Coronavirus Pandemic (COVID-19). *Our World in Data* (2020).

174. United Nations Department of Economic and Social Affairs. *World Population Prospects 2017 - Volume I: Comprehensive Tables* https://doi.org/10.18356/9789210001014. (2022) (United Nations, 2021).

175. United Nations, Department of Economic and Social Affairs, Population Division. *World Urbanization Prospects: The 2018 Revision* https://population.un.org/wup/Download/. 2018.

176. Hersbach, H., Bell, B., Berrisford, P., *et al.* The ERA5 Global Reanalysis. *Quarterly Journal of the Royal Meteorological Society* **146,** 1999–2049. https://doi.org/10.1002/qj.3803 (2020).

177. Moore, M., Gelfeld, B., Okunogbe, A. & Paul, C. *Identifying Future Disease Hot Spots: Infectious Disease Vulnerability Index* https://doi.org/10.7249/RR1605. (2021) (RAND Corporation, 2016).

178. Dong, E., Du, H. & Gardner, L. An Interactive Web-Based Dashboard to Track COVID-19 in Real Time. *The Lancet Infectious Diseases* **20,** 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1 (2020).

179. *COVID-19 Community Mobility Report* https://www.google.com/covid19/mobility?hl=en.

180. *IATA* https://www.iata.org/en/. 2021.

181. Hale, T., Angrist, N., Goldszmidt, R., *et al.* A Global Panel Database of Pandemic Policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour* **5,** 529–538. https://doi.org/10.1038/s41562-021-01079-8 (2021).

182. Team, R. C. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing. Vienna, Austria, 2022.

183. Mott, J. A., Fry, A. M., Kondor, R., Wentworth, D. E. & Olsen, S. J. Re-emergence of Influenza Virus Circulation during 2020 in Parts of Tropical Asia: Implications for Other Countries. *Influenza and Other Respiratory Viruses* **15,** 415–418. https://doi.org/10.1111/irv.12844 (2021).

184. Pablos-Méndez, A., Vega, J., Aranguren, F. P., Tabish, H. & Raviglione, M. C. Covid-19 in Latin America. *BMJ* **370,** m2939. https://doi.org/10.1136/bmj.m2939 (2020).

185. Salyer, S. J., Maeda, J., Sembuche, S., *et al.* The First and Second Waves of the COVID-19 Pandemic in Africa: A Cross-Sectional Study. *The Lancet* **397,** 1265–1275. https://doi.org/10.1016/S0140-6736(21)00632-2 (2021).

186. Pullano, G., Di Domenico, L., Sabbatini, C. E., *et al.* Underdetection of Cases of COVID-19 in France Threatens Epidemic Control. *Nature* **590,** 134–139. https://doi.org/10.1038/s41586-020-03095-6 (2021).

187. Di Domenico, L., Sabbatini, C. E., Boëlle, P.-Y., *et al.* Adherence and Sustainability of Interventions Informing Optimal Control against the COVID-19 Pandemic. *Communications Medicine* **1,** 1–13. https://doi.org/10.1038/s43856-021-00057-5 (2021).

188. Weitz, J. S., Park, S. W., Eksin, C. & Dushoff, J. Awareness-Driven Behavior Changes Can Shift the Shape of Epidemics Away from Peaks and toward Plateaus, Shoulders, and Oscillations. *Proceedings of the National Academy of Sciences* **117,** 32764–32771. https://doi.org/10.1073/pnas.2009911117 (2020).

189. Opatowski, L., Baguelin, M. & Eggo, R. M. Influenza Interaction with Cocirculating Pathogens and Its Impact on Surveillance, Pathogenesis, and Epidemic Profile: A Key Role for Mathematical Modelling. *PLOS Pathogens* **14,** e1006770. https://doi.org/10.1371/journal.ppat.1006770 (2018).

190. Piret, J. & Boivin, G. Viral Interference between Respiratory Viruses - Volume 28, Number 2—February 2022 - Emerging Infectious Diseases Journal - CDC. https://doi.org/10.3201/eid2802.211727 (2022).

191. Levin, A. T., Owusu-Boaitey, N., Pugh, S., *et al.* Assessing the Burden of COVID-19 in Developing Countries: Systematic Review, Meta-Analysis and Public Policy Implications. *BMJ Global Health* **7,** e008477. https://doi.org/10.1136/bmjgh-2022-008477 (2022).

192. Gostic, K. M., Bridge, R., Brady, S., *et al.* Childhood Immune Imprinting to Influenza A Shapes Birth Year-Specific Risk during Seasonal H1N1 and H3N2 Epidemics. *PLOS Pathogens* **15,** e1008109. https://doi.org/10.1371/journal.ppat.1008109 (2019).

193. Messacar, K., Baker, R. E., Park, S. W., *et al.* Preparing for Uncertainty: Endemic Paediatric Viral Illnesses after COVID-19 Pandemic Disruption. *The Lancet* **400,** 1663–1665. https://doi.org/10.1016/S0140-6736(22)01277-6 (2022).

194. Emborg, H.-D., Vestergaard, L. S., Botnen, A. B., *et al.* A Late Sharp Increase in Influenza Detections and Low Interim Vaccine Effectiveness against the Circulating A(H3N2) Strain, Denmark, 2021/22 Influenza Season up to 25 March 2022. *Eurosurveillance* **27,** 2200278. https://doi.org/10.2807/1560-7917.ES.2022.27.15.2200278 (2022).

195. Paget, J., Caini, S., Riccio, M. D., van Waarden, W. & Meijer, A. Has Influenza B/Yamagata Become Extinct and What Implications Might This Have for Quadrivalent Influenza Vaccines? *Eurosurveillance* **27,** 2200753. https://doi.org/10.2807/1560-7917.ES.2022.27.39.2200753 (2022).

196. Therneau, T. M., Atkinson, B. & Ripley, B. *Rpart : Recursive Partitioning and Regression Trees* https://cran.r-project.org/web/packages/rpart/rpart.pdf. 2015.

197. *Supplementary Data* https://docs.google.com/spreadsheets/d/1PirC3iJ_yrlw9CoNL0_dkmXEI7DV 2022.

198. Palese, P. & Wang, T. T. Why Do Influenza Virus Subtypes Die Out? A Hypothesis. *mBio* **2,** e00150–11. https://doi.org/10.1128/mBio.00150-11 (2011).

199. Suzuki, A., Mizumoto, K., Akhmetzhanov, A. R. & Nishiura, H. Interaction Among Influenza Viruses A/H1N1, A/H3N2, and B in Japan. *International Journal of Environmental Research and Public Health* **16,** 4179. https://doi.org/10.3390/ijerph16214179 (2019).

200. Arevalo, P., McLean, H. Q., Belongia, E. A. & Cobey, S. Earliest Infections Predict the Age Distribution of Seasonal Influenza A Cases. *eLife* **9** (eds Cooper, B. S., Ferguson, N. M., Cooper, B. S. & Baguelin, M.) e50060. https://doi.org/10.7554/eLife.50060 (2020).

201. Puzelli, S., Di Martino, A., Facchini, M., *et al.* Co-Circulation of the Two Influenza B Lineages during 13 Consecutive Influenza Surveillance Seasons in Italy, 2004-2017. *BMC infectious diseases* **19,** 990. https://doi.org/10.1186/s12879-019-4621-z (2019).

202. Zhang, X.-S. & De Angelis, D. Construction of the Influenza A Virus Transmission Tree in a College-Based Population: Co-Transmission and Interactions between Influenza A Viruses. *BMC infectious diseases* **16,** 38. https://doi.org/10.1186/s12879-016-1373-x (2016).

203. Truscott, J., Fraser, C., Cauchemez, S., *et al.* Essential Epidemiological Mechanisms Underpinning the Transmission Dynamics of Seasonal Influenza. *Journal of The Royal Society Interface* **9,** 304–312. https://doi.org/10.1098/rsif.2011.0309 (2011).

204. Zhang, X.-S. Strain Interactions as a Mechanism for Dominant Strain Alternation and Incidence Oscillation in Infectious Diseases: Seasonal Influenza as a Case Study. *PloS One* **10,** e0142170. https://doi.org/10.1371/journal.pone.0142170 (2015).

205. Buccianti, A., Lima, A., Albanese, S., *et al.* Exploring Topsoil Geochemistry from the CoDA (Compositional Data Analysis) Perspective: The Multi-Element Data Archive of the Campania Region (Southern Italy). *Journal of Geochemical Exploration* **159,** 302–316. https://doi.org/10.1016/j.gexplo.2015.10.006 (2015).

206. Chong, F. & Spencer, M. Analysis of Relative Abundances with Zeros on Environmental Gradients: A Multinomial Regression Model. *PeerJ* **6,** e5643. https://doi.org/10.7717/peerj.5643 (2018).

207. Jackson, D. A. COMPOSITIONAL DATA IN COMMUNITY ECOLOGY: THE PARADIGM OR PERIL OF PROPORTIONS? **78** (1997).

208. Tolosana-Delgado, R., Mueller, U. & van den Boogaart, K. G. Geostatistics for Compositional Data: An Overview. *Mathematical Geosciences* **51,** 485–526. https://doi.org/10.1007/s11004-018-9769-3 (2019).

209. European Centre for Disease Prevention and Control. *Seasonal Influenza 2021-2022 - Annual Epidemiological Report* tech. rep. (2021).

210. European Centre for Disease Prevention and Control. *Seasonal Influenza 2022-2023 Annual Epidemiological Report* tech. rep. (2023).

211. CDC. *Update: Influenza Activity — United States and Worldwide, 2003–04 Season, and Composition of the 2004–05 Influenza Vaccine* tech. rep. 53(25) (2004), 547–552. (2022).

212. Ghedin, E., Sengamalay, N. A., Shumway, M., *et al.* Large-Scale Sequencing of Human Influenza Reveals the Dynamic Nature of Viral Genome Evolution. *Nature* **437,** 1162–1166. https://doi.org/10.1038/nature04239 (2005).

213. Boëlle, P. Y., Bernillon, P. & Desenclos, J. C. A Preliminary Estimation of the Reproduction Ratio for New Influenza A(H1N1) from the Outbreak in Mexico, March-April 2009. *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* **14,** 19205. https://doi.org/10.2807/ese.14.19.19205-en (2009).

214. Fraser, C., Donnelly, C. A., Cauchemez, S., *et al.* Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings. *Science* **324,** 1557–1561. https://doi.org/10.1126/science.1176062 (2009).

215. Chinazzi, M., Davis, J. T., Ajelli, M., *et al.* The Effect of Travel Restrictions on the Spread of the 2019 Novel Coronavirus (COVID-19) Outbreak. *Science* **368,** 395–400. https://doi.org/10.1126/science.aba9757 (2020).

216. CDC. *Preliminary Flu Burden Estimates, 2021-22 Season* tech. rep. (2023). (2023).

217. Organization, W. H. *Influenza_transmission_zones20180914.Pdf* 2018.

218. Earth, N. *Natural Earth - Free Vector and Raster Map Data* https://www.naturalearthdata.com/.

219. Gneiting, T. & Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **102,** 359–378. https://doi.org/10.1198/016214506000001437 (2007).

220. Scheuerer, M. & Hamill, T. M. Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities. *Monthly Weather Review* **143,** 1321–1334. https://doi.org/10.1175/MWR-D-14-00269.1 (2015).

221. Xie, H., Wan, X.-F., Ye, Z., *et al.* H3N2 Mismatch of 2014–15 Northern Hemisphere Influenza Vaccines and Head-to-head Comparison between Human and Ferret Antisera Derived Antigenic Maps. *Scientific Reports* **5,** 15279. https://doi.org/10.1038/srep15279 (2015).

222. Ali, S. T. & Cowling, B. J. Influenza Virus: Tracking, Predicting, and Forecasting. *Annual Review of Public Health* **42,** 43–57. https://doi.org/10.1146/annurev-publhealth-010720-021049 (2021).

223. CDC. *About Flu Forecasting | CDC* https://www.cdc.gov/flu/weekly/flusight/how-flu-forecasting.htm. 2023.

224. Chretien, J.-P., George, D., Shaman, J., Chitale, R. A. & McKenzie, F. E. Influenza Forecasting in Human Populations: A Scoping Review. *PLOS ONE* **9,** e94130. https://doi.org/10.1371/journal.pone.0094130 (2014).

225. Kramer, S. C. & Shaman, J. Development and Validation of Influenza Forecasting for 64 Temperate and Tropical Countries. *PLOS Computational Biology* **15,** e1006742. https://doi.org/10.1371/journal.pcbi.1006742 (2019).

226. Nsoesie, E. O., Brownstein, J. S., Ramakrishnan, N. & Marathe, M. V. A Systematic Review of Studies on Forecasting the Dynamics of Influenza Outbreaks. *Influenza and Other Respiratory Viruses* **8,** 309–316. https://doi.org/10.1111/irv.12226 (2014).

227. Łuksza, M. & Lässig, M. A Predictive Fitness Model for Influenza. *Nature* **507,** 57–61. https://doi.org/10.1038/nature13087 (2014).

228. Neher, R. A., Russell, C. A. & Shraiman, B. I. Predicting Evolution from the Shape of Genealogical Trees. *eLife* **3** (ed McVean, G.) e03568. https://doi.org/10.7554/eLife.03568 (2014).

229. Neher, R. A., Bedford, T., Daniels, R. S., Russell, C. A. & Shraiman, B. I. Prediction, Dynamics, and Visualization of Antigenic Phenotypes of Seasonal Influenza Viruses. *Proceedings of the National Academy of Sciences* **113,** E1701–E1709. https://doi.org/10.1073/pnas.1525578113 (2016).

230. Steinbrück, L., Klingen, T. R. & McHardy, A. C. Computational Prediction of Vaccine Strains for Human Influenza A (H3N2) Viruses. *Journal of Virology* **88,** 12123–12132. https://doi.org/10.1128/jvi.01861-14 (2014).

231. Paul, M. & Held, L. Predictive Assessment of a Non-Linear Random Effects Model for Multivariate Time Series of Infectious Disease Counts. *Statistics in Medicine* **30,** 1118–1136. https://doi.org/10.1002/sim.4177 (2011).

232. Wang, J., Teng, Z., Cui, X., *et al.* Epidemiological and Serological Surveillance of Hand-Foot-and-Mouth Disease in Shanghai, China, 2012-2016. *Emerging Microbes & Infections* **7,** 8. https://doi.org/10.1038/s41426-017-0011-z (2018).

233. Xu, M., Su, L., Cao, L., *et al.* Genotypes of the Enterovirus Causing Hand Foot and Mouth Disease in Shanghai, China, 2012-2013. *PLOS ONE* **10,** e0138514. https://doi.org/10.1371/journal.pone.0138514 (2015).

234. Zeng, H., Lu, J., Zheng, H., *et al.* The Epidemiological Study of Coxsackievirus A6 Revealing Hand, Foot and Mouth Disease Epidemic Patterns in Guangdong, China. *Scientific Reports* **5,** 10550. https://doi.org/10.1038/srep10550 (2015).

235. Martín-Fernández, J.-A., Hron, K., Templ, M., Filzmoser, P. & Palarea-Albaladejo, J. Bayesian-Multiplicative Treatment of Count Zeros in Compositional Data Sets. *Statistical Modelling* **15,** 134–158. https://doi.org/10.1177/1471082X14535524 (2015).

236. Palarea-Albaladejo, J. & Martín-Fernández, J. A. zCompositions — R Package for Multivariate Imputation of Left-Censored Data under a Compositional Approach. *Chemometrics and Intelligent Laboratory Systems* **143,** 85–96. https://doi.org/10.1016/j.chemolab.2015.02.019 (2015).

237. Kaufman, L. & Rousseeuw, P. J. in *Finding Groups in Data* 68–125 (John Wiley & Sons, Ltd, 1990). https://doi.org/10.1002/9780470316801.ch2. (2023).

238. Van der Laan, M., Pollard, K. & Bryan, J. A New Partitioning around Medoids Algorithm. *Journal of Statistical Computation and Simulation* **73,** 575–584. https://doi.org/10.1080/0094965031000136012 (2003).

239. Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* **20,** 53–65. https://doi.org/10.1016/0377-0427(87)90125-7 (1987).

240. Templ, M., Hron, K. & Filzmoser, P. in *Compositional Data Analysis* 341–355 (John Wiley & Sons, Ltd, 2011). https://doi.org/10.1002/9781119976462.ch25. (2024).

241. Harper, M., Weinstein, B., Simon, C., *et al. Python-Ternary: Ternary Plots in Python* Zenodo. 2015. https://doi.org/10.5281/zenodo.34938.

242. Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.* Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

243. Jordan, A., Krüger, F. & Lerch, S. Evaluating Probabilistic Forecasts with scoringRules. *Journal of Statistical Software* **90,** 1–37. https://doi.org/10.18637/jss.v090.i12 (2019).

244. *SciPy API — SciPy v1.12.0 Manual* https://docs.scipy.org/doc/scipy/reference/.

245. Genest, C., Ghoudi, K. & Rivest, L.-P. A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions. *Biometrika* **82,** 543–552. http://www.jstor.org/stable/2337532 (1995).

246. Tsukahara, H. Semiparametric Estimation in Copula Models. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **33,** 357–375. JSTOR: 25046185 (2005).

247. Charpentier, A., Fermanian, J.-D. & Scaillet, O. in *Copulas: From Theory to Application in Finance* 35 (Risk Books, 2007). (2024).

248. Segers, J. Asymptotics of Empirical Copula Processes under Non-Restrictive Smoothness Assumptions. *Bernoulli* **18.** https://doi.org/10.3150/11-BEJ387. arXiv: 1012.2133 [math, stat] (2012).

249. Alquier, P., Chérief-Abdellatif, B.-E., Derumigny, A. & Fermanian, J.-D. Estimation of Copulas via Maximum Mean Discrepancy. *Journal of the American Statistical Association* **118,** 1997–2012 (2023).

250. Dupuis, D. J. & Jones, B. L. Multivariate Extreme Value Theory And Its Usefulness In Understanding Risk. *North American Actuarial Journal* **10,** 1–27. https://doi.org/10.1080/10920277.2006.10597411 (2006).

251. Lopez, O. A Censored Copula Model for Micro-Level Claim Reserving. *Insurance: Mathematics and Economics* **87,** 1–14 (2019).

252. Farkas, S. & Lopez, O. *Semiparametric Copula Models Applied to the Decomposition of Claim Amounts* 2023. (2024).

253. *Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009* in (eds Jaworski, P., Durante, F., Härdle, W. K. & Rychlik, T.) **198** (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010). https://doi.org/10.1007/978-3-642-12465-5. (2024).

254. Czado, C. & Nagler, T. Vine Copula Based Modeling. *Annual Review of Statistics and Its Application* **9,** 453–477. https://doi.org/10.1146/annurev-statistics-040220-101153 (2022).

255. Gijbels, I., Omelka, M. & Veraverbeke, N. Multivariate and Functional Covariates and Conditional Copulas. *Electronic Journal of Statistics* **6,** 1273–1306. https://doi.org/10.1214/12-EJS712 (2012).

256. Fermanian, J.-D. & Lopez, O. Single-Index Copulas. *Journal of Multivariate Analysis* **165,** 27–55. https://doi.org/10.1016/j.jmva.2017.11.004 (2018).

257. Gocheva-Ilieva, S. G., Voynikova, D. S., Stoimenova, M. P., Ivanov, A. V. & Iliev, I. P. Regression Trees Modeling of Time Series for Air Pollution Analysis and Forecasting. *Neural Computing and Applications* **31,** 9023–9039. https://doi.org/10.1007/s00521-019-04432-1 (2019).

258. Farkas, S., Heranval, A., Lopez, O. & Thomas, M. *Generalized Pareto Regression Trees for Extreme Events Analysis* 2021. https://doi.org/10.48550/arXiv.2112.10409. arXiv: 2112.10409 [math, stat].

259. Loh, W.-Y. Fifty Years of Classification and Regression Trees. *International Statistical Review* **82,** 329–348. https://doi.org/10.1111/insr.12016 (2014).

260. Kurz, M. S. & Spanhel, F. Testing the Simplifying Assumption in High-Dimensional Vine Copulas. *Electronic Journal of Statistics* **16.** https://doi.org/10.1214/22-EJS2051. arXiv: 1706.02338 [stat] (2022).

261. Sklar, M. Fonctions de Répartition à N Dimensions et Leurs Marges. *Annales de l'ISUP* **VIII,** 229–231 (1959).

262. Einmahl, U. & Mason, D. M. Uniform in Bandwidth Consistency of Kernel-Type Function Estimators. *The Annals of Statistics* **33,** 1380–1403. https://doi.org/10.1214/009053605000000129 (2005).

263. Nelder, J. A. & Wedderburn, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)* **135,** 370–384. https://doi.org/10.2307/2344614. JSTOR: 2344614 (1972).

264. Wellner, J. A. Limit Theorems for the Ratio of the Empirical Distribution Function to the True Distribution Function. *Zeitschrift fr Wahrscheinlichkeitstheorie und Verwandte Gebiete* **45,** 73–88. https://doi.org/10.1007/BF00635964 (1978).

265. Omelka, M., Hudecová, Š. & Neumeyer, N. Maximum Pseudo-Likelihood Estimation Based on Estimated Residuals in Copula Semiparametric Models. *Scandinavian Journal of Statistics* **48,** 1433–1473. https://doi.org/10.1111/sjos.12498 (2021).

266. Kularatne, T. D., Li, J. & Pitt, D. On the Use of Archimedean Copulas for Insurance Modelling. *Annals of Actuarial Science* **15,** 57–81. https://doi.org/10.1017/S1748499520000147 (2021).

267. Hennessy, D. A. & Lapan, H. E. The Use of Archimedean Copulas to Model Portfolio Allocations. *Mathematical Finance* **12,** 143–154 (2002).

268. Lei, H., Yang, L., Yang, M., *et al.* Quantifying the Rebound of Influenza Epidemics after the Adjustment of Zero-COVID Policy in China. *PNAS Nexus* **2,** pgad152. https://doi.org/10.1093/pnasnexus/pgad152 (2023).

269. Ali, S. T., Lau, Y. C., Shan, S., *et al.* Prediction of Upcoming Global Infection Burden of Influenza Seasons after Relaxation of Public Health and Social Measures during the COVID-19 Pandemic: A Modelling Study. *The Lancet Global Health* **10,** e1612–e1622. https://doi.org/10.1016/S2214-109X(22)00358-8 (2022).

270. Boudewijns, B., Paget, J., Del Riccio, M., Coudeville, L. & Crépey, P. Preparing for the Upcoming 2022/23 Influenza Season: A Modelling Study of the Susceptible Population in Australia, France, Germany, Italy, Spain and the United Kingdom. *Influenza and Other Respiratory Viruses* **17,** e13091. https://doi.org/10.1111/irv.13091 (2023).

271. WHO Regional Office for Europe and European Centre for Disease Prevention and Control. *Influenza Virus Characterization Summary Report, Europe, February 2023* tech. rep. (2023). (2023).

272. WHO Regional Office for Europe and Stockholm: European Centre for Disease Prevention and Control. *Influenza Virus Characterization - Summary Report, Europe, October 2023* tech. rep. (2023). (2023).

273. 24h, I. S. *Influenza in Italia, tutti i dati aggiornati | Il Sole 24 Ore* https://lab24.ilsole24ore.com/influenz dati-italia/?refresh_ce=1.

274. ECDC. Seasonal Influenza 2021-2022 - Annual Epidemiological Report (2021).

275. Lampros, A., Talla, C., Diarra, M., *et al.* Shifting Patterns of Influenza Circulation during the COVID-19 Pandemic, Senegal - Volume 29, Number 9—September 2023 - Emerging Infectious Diseases Journal - CDC. https://doi.org/10.3201/eid2909.230307.

276. Nott, R., Fuller, T. L., Brasil, P. & Nielsen-Saines, K. Out-of-Season Influenza during a COVID-19 Void in the State of Rio de Janeiro, Brazil: Temperature Matters. *Vaccines* **10,** 821. https://doi.org/10.3390/vaccines10050821 (2022).

277. De Jong, S. P. J., Garza, Z. C. F., Gibson, J. C., *et al. Potential Impacts of Prolonged Absence of Influenza Virus Circulation on Subsequent Epidemics* 2022. https://doi.org/10.1101/2022.02.05.22270494.

278. Chiu, N.-C., Chi, H., Tai, Y.-L., *et al.* Impact of Wearing Masks, Hand Hygiene, and Social Distancing on Influenza, Enterovirus, and All-Cause Pneumonia During the Coronavirus Pandemic: Retrospective National Epidemiological Surveillance Study. *Journal of Medical Internet Research* **22,** e21257. https://doi.org/10.2196/21257 (2020).

279. Pendrey, C. G., Strachan, J., Peck, H., *et al.* The Re-Emergence of Influenza Following the COVID-19 Pandemic in Victoria, Australia, 2021 to 2022. *Eurosurveillance* **28,** 2300118. https://doi.org/10.2807/1560-7917.ES.2023.28.37.2300118 (2023).

280. Kretzschmar, M. E., Ashby, B., Fearon, E., *et al.* Challenges for modelling interventions for future pandemics. *Epidemics* **38,** 100546. https://www.sciencedirect.com/science/article/pii/S1755436522000081 (2022).