




**ADVERTIMENT.** L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  [http://cat.creativecommons.org/?page\\_id=184](http://cat.creativecommons.org/?page_id=184)

**ADVERTENCIA.** El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

**WARNING.** The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

**The Combined effects of climate, environment, and socio-demographic factors on Human Health via the spread of emerging infectious diseases:  
Using big data methods to investigate macro-level determinants of disease transmission**

**Matthew Watts**

A thesis presented for the  
PhD programme in Environmental Science and Technology



Directors: Graham Mortyn, Victor Sarto Monteys, Panagiota Kotsila  
Tutor: Graham Mortyn  
Institut de Ciència i Tecnologia Ambientals (ICTA - UAB)  
Universitat Autònoma de Barcelona  
Barcelona  
Spain  
2022

## **Abstract**

One of the major health threats for European nations is the arrival and spread of neglected and emerging tropical diseases, brought about through factors such as climate change, environmental degradation and poverty. In this thesis, I take data-driven holistic approaches to examine some of the macro-level drivers of three emerging infectious diseases, dengue virus, West Nile virus, and SARS-CoV-2, making use of the huge pools of publicly available demographic, socio-economic, environmental and population health data. For each study, I constructed an analytical framework that mapped out local level drivers of disease, which was then used as a foundation to construct three unique spatial-temporal datasets for each study. Relationships were tested using a Generalised Additive Models (GAM), which could capture complex non-linear relationships and also account for spatial and temporal auto-correlation. A joint analysis of chapters two, three and four reveals that climate and environmental factors are major drivers of disease, but also societal factors such as poverty, occupation, and top-down political decision making also appears to be moderators of disease transmission. This work is relevant as it adds to the growing body of scientific literature focusing on infectious diseases since it tackles some of the broader and less explored areas of public health and epidemiology, such as analysing economic changes with environmental changes, examining the impacts of factors such as austerity on health, along with other factors such as political decision making and or lack of intervention by government authorities.

# Acknowledgements

Thanks to Panagiota Kotsila, Graham Mortyn, Victor Sarto Monteys and Cesira Urzi Brancati who provided advice and feedback on the research.

Panagiota Kotsila, Graham Mortyn and Victor Sarto Monteys also supervised the project and are responsible for formulating ideas for the umbrella project Impacts of Climate Change (CC) on Human Health (HH) at ICTA-UAB: Integrating socio-economic and policy studies with natural science studies to enhance consilience of climate policy science.

This research was funded by ICTA's Maria de Maeztu Unit of Excellence, awarded by the Spanish Ministry of Economy and Competitiveness. The award is the highest institutional recognition of scientific research in Spain. Thanks also to Patrizia Ziveri, Pedro Manuel Gonzalez Hernandez and Cristina Duran for supporting the project.

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Motivation for study . . . . .	5
<b>2</b>	<b>Influence of socio-economic, demographic and climate factors on the regional distribution of dengue in the United States and Mexico</b>	<b>10</b>
2.1	Introduction . . . . .	12
2.1.1	Motivation for study . . . . .	12
2.1.2	Conceptual framework . . . . .	13
2.2	Materials and Methods . . . . .	15
2.2.1	Species distribution models to estimate regional susceptibility	15
2.2.2	Data extraction and methods to assess the impact of climate, demographic and socio-economic factors on dengue . . . . .	17
2.2.3	Statistical Methods . . . . .	18
2.3	Results . . . . .	20
2.3.1	US/Mex analysis . . . . .	23
2.3.2	Mex analysis . . . . .	26
2.4	Discussion and Conclusions . . . . .	27



<b>3</b>	<b>The rise of West Nile Virus in Southern and Southeastern Europe: A spatial–temporal analysis investigating the combined effects of climate, land use and economic changes</b>	<b>34</b>
3.1	Introduction . . . . .	36
3.1.1	Conceptual framework . . . . .	37
3.2	Materials and methods . . . . .	39
3.2.1	Data sources . . . . .	39
3.2.2	Final data-set . . . . .	40
3.2.3	Statistical Methods . . . . .	41
3.3	Results . . . . .	42
3.4	Discussion . . . . .	49
3.4.1	Meteorological factors . . . . .	49
<b>4</b>	<b>Macro-Level drivers of SARS-CoV-2 transmission: A data-driven analysis of factors contributing to epidemic growth during the first wave of outbreaks in the United States</b>	<b>58</b>
4.1	Introduction . . . . .	60
4.1.1	Analytical framework . . . . .	61
4.2	Methods . . . . .	62
4.2.1	Data collection and processing . . . . .	63
4.2.2	Study design . . . . .	64
4.3	Results . . . . .	68
4.4	Discussion . . . . .	84
4.4.1	Containment measures to reduce disease spread . . . . .	84
4.4.2	Socio-economic, economic, and demographic factors . . . . .	84
4.4.3	Limitations . . . . .	86
4.5	Conclusions . . . . .	87
4.5.1	Abbreviations . . . . .	87
<b>5</b>	<b>Synthesis</b>	<b>95</b>
5.0.1	Caveats and Limitations . . . . .	97
5.0.2	Future Outlook . . . . .	98
5.0.3	Conclusions . . . . .	98
	<b>Appendices</b>	<b>102</b>
	<b>Appendix A Influence of socio-economic, demographic and climate factors on the regional distribution of dengue in the United States and Mexico</b>	<b>103</b>
A.1	Data availability . . . . .	103
A.2	Vector and Environmental Relationships: Regression Analysis (Step 1)	103
A.3	Vector and Environmental Relationships: regression analysis results	104
A.4	Diagnostics . . . . .	106
A.5	Data sources . . . . .	120
	<b>Appendix B The rise of West Nile Virus in Southern and South-eastern Europe: A spatial–temporal analysis investigating the combined effects of climate, land use and economic changes</b>	<b>122</b>
B.1	Data availability . . . . .	122

B.2	Code availability . . . . .	122
B.3	Extended data extraction and processing methods . . . . .	122
B.3.1	Aggregation . . . . .	122
B.3.2	Economic, Socio-Economic and Demographic Factors . . . . .	123
B.3.3	Climate Data . . . . .	123
B.3.4	Land-use data . . . . .	124
B.3.5	Surface Water data . . . . .	124
B.3.6	Extended Statistical methods . . . . .	124
B.4	Climate modeling . . . . .	125

<b>Appendix C Macro-Level drivers of SARS-CoV-2 transmission: A data-driven analysis of factors contributing to epidemic growth during the first wave of outbreaks in the United States</b>		<b>143</b>
C.1	Covid policy tracker . . . . .	144
C.2	Data availability . . . . .	144
C.3	Diagnostics: Infection mode . . . . .	144
C.4	Diagnostics: Mortality model . . . . .	148

# Chapter 1

## Introduction

Human civilisation is at a crossroads; continued economic growth at the expense of the natural environment is jeopardising its very existence on the planet through the deterioration of natural life-support systems such as clean water, a stable climate, fish stocks and the biodiversity of pollinating insects [2]. Humanity could be considered a victim of its own success and needs to turn a corner and act to limit its destructive influence on the natural world. However, this is no easy feat given the global economy is entrenched in destructive practices such as forest clearance, intensive animal rearing, and the burning of fossil fuels, that is also causing the climate to warm [20].

In general, rising temperatures and extreme heatwaves are expected to have direct consequences for human health through heat-related mortality and through impacts on respiratory, cardiovascular, or neurological health [17]; and secondary effects on ecological and agricultural systems which may lead to food shortages. We may also expect climate change to affect human health indirectly, inducing stress and mental health problems because of human displacement, failing economies, and declines in output from agriculture which will exacerbate poverty [13]. One of the current threats for European nations is the arrival of neglected tropical diseases, since rising temperatures are likely to facilitate their spread and establishment, evidence suggest that many pathogens and vectors are already making bio-geographic adjustments due to increasingly warmer global temperatures [10, 22].

Although there have been significant -albeit unequal- global increases in life expectancy over the past century as a result of improvements in diet, medical practices, improvements in public health measures to tackle diseases, and through the development of new treatments including antibiotics and vaccines [3, 19, 4, 18, 8], in many of the world's poorer nations, high morbidity and mortality caused specifically by infectious diseases remain a problem [4, 18, 8, 5]. Many of the diseases that plague poorer nations, such as malaria and dengue are often neglected, meaning that until they cause problems in richer nations, not enough money is spent on developing treatments for such diseases [4, 18, 8]. The consequence is that, because of climate change, neglected tropical diseases may spread to richer nations where conditions were previously too cold for such pathogens to proliferate. As we have witnessed during the COVID-19 pandemic, even when preventative treatments such as vaccines exist, they may not be fairly distributed and diseases are left to fester in poorer nations, which poses a risk to richer nations as a result of pathogen evolution, which may eventually render treatments ineffective [14].

Novel Emerging infectious diseases (EIDs) also present a significant threat to Europe, as we have recently witnessed with SARS-COV-2, pathogens can go under the radar of public health authorities but can quickly emerge with devastating consequences for global human health. It is estimated that around 631,000–827,000 zoonotic viruses remain undiscovered in the world today, most of which reside in tropical forests [15]. The number of EID events has been on the rise over the past century; around 335 diseases have emerged or re-emerged between the 1940s and early 2000s [11]. Many of these spill-over events are believed to be instigated by human activities that bring humans into close contact with wild infected animals, for example, expansion and construction of new settlements, agricultural expansion, or the harvesting and consumption of wild animals. Although such events have always occurred throughout human history, many outbreaks would have remained isolated reflecting the lifestyles of the communities in which they occurred. However, the world is now more connected than ever, and many of the global hotspots of pathogen diversity are now under intense pressure from extractive industries, such as mining, oil extraction, forestry, intensive animal rearing [1, 16, 6], such industries connect pathogens to the outside world through globalised transport networks, meaning they can potentially spread around the world in a matter of days, if not hours [21]. According to the United Nations, the human population will increase by around 2 billion in the next 30 years, which equates to around 9.3 billion people living on the planet by 2050. Given the current trajectory of human population growth accompanied by a dominant pattern of development, based on continuous economic growth coupled with resource extraction and depletion of natural forests, we are likely to see a continuation of human-induced environmental and ecological changes [7] which may signify continued growth in novel infectious disease emergence.

## 1.1 Motivation for study

Although research on emerging infectious diseases is extremely active, much of the work has traditionally focused on local-level and system-specific questions that seek to better understand fine-scale processes and mechanisms. However, recent initiatives such as One Health [12], Planetary Health [9], and EcoHealth [23] called for more holistic and interdisciplinary approaches to study infectious diseases that can provide critical insights into how human and natural systems are connected and can help us to better understand the contribution of environmental, biological, social, cultural, political, and economic factors on disease emergence, spread and the distribution at a global scale. Recent technological advances in modern computing, both hardware and software, and the development of open source economic and environmental databases are conducive to this, since they allow us to gather, process, merge and analyse vast amounts of heterogeneous data (which I will refer to as “big data”) to investigate large-scale ecological questions related to infectious diseases.

In this thesis, I set out to examine some of the macro-level drivers of disease spread and distribution of three emerging infectious diseases using “big data” approaches <sup>1</sup>: Dengue virus (DENV), West Nile virus (WNV), and SARS-CoV-2.

---

<sup>1</sup><https://www.ecdc.europa.eu/en/dengue-fever/facts>

1) Dengue virus (DENV) is a vector-borne disease; its transmission cycle involves humans and mosquitoes, that is, the virus is either transmitted or received by a biting mosquito.

2) West Nile virus (WNV) is also a vector-borne disease but is transmitted between mosquitoes and birds; humans are susceptible to the virus but are generally considered dead-end hosts i.e., they cannot transmit the virus to the mosquito.

3) SARS-CoV-2 is a virus that is transmitted between people and mammals mainly through respiratory particles (droplet and aerosols) and, to a lesser extent, indirect contact through fomite transmission (contact with contaminated surfaces).

Each chapter examines how climate and anthropogenic forces influence disease at the macro-level (chapters 2-4) in the hope of learning more about the diseases and the populations they affect, eventually seeking to generalise findings across these three studies by looking at the common factors that influence the health of communities in the areas studied (chapter 5). The three primary research articles in chapters 2, 3 and 4 share a common statistical methodology and comparable data. For each chapter, I constructed an analytical framework that mapped out local level drivers of disease; these relationships were then used as a foundation to construct the spatial-temporal datasets for each study, which merged heterogeneous data from several different sources and data types. I could then assess the importance of individual variables making up the study system and assess their contribution to the statistical models, all else equal.

## Overview of primary research articles

In chapter 2, I investigate the influence of socio-economic/demographic and climate factors on the regional distribution of dengue in the United States and Mexico, by analysing panel data on regional household income, education of the labour force, life expectancy at birth, and housing overcrowding, population growth, inter-regional migration, temperature and rainfall. Although this study focuses on dengue in regions in the Americas, it has implications for Europe, given the presence of the two vector mosquitoes (*A. aegypti* and *A. albopictus*). Indeed, over the past 10 years, autochthonous transmission has occurred in Croatia, France, Italy, and Spain <sup>2</sup>, so the virus was likely imported to Europe from infected regions by humans. Whether transmission can be sustained, and the virus becomes endemic in Europe is not yet apparent. In the chapter 3, I examine the possible drivers of the West Nile virus transmission in Europe by analysing time series data on land-use changes, temperature, rainfall, and government spending and regional gross domestic product. This disease is endemic to parts of central Africa and Asia and was until recently only occasionally reported in Europe where it was believed to be imported by infected migratory birds. However, it has recently emerged in Europe as a major health concern because its prevalence is increasing and gradually spreading in Europe from southerly regions to more northerly regions. This phenomenon also coincided with the European financial crisis and subsequent austerity, along with land use and climatic changes. In the chapter 4, I investigate the macro-level drivers of SARS-CoV-2 transmission during the first wave of the epidemic in the United States, analysing cross-sectional data including data on the age structure of regions, poverty levels, the health status of the population, climate factors, and population density. The reason

---

<sup>2</sup><https://www.ecdc.europa.eu/en/dengue-fever/facts>

for choosing the United States is that it provides us with a unique opportunity to study this phenomenon at a macro-scale, since it encompasses a diverse range of climate types over a vast geographical area, with a somewhat homogeneous political system, allowing us to disentangle the effects of the environment from other demographic and socio-economic conditions. Furthermore, the study analyses data on the implementation of public health measures, as captured by the Oxford COVID-19 Government Response Tracker (OxCGRT) "Stringency Index", to tackle spread and factors that may influence case reporting.

## Bibliography

- [1] Toph Allen, Kris A. Murray, Carlos Zambrana-Torrel, Stephen S. Morse, Carlo Rondinini, Moreno Di Marco, Nathan Breit, Kevin J. Olival, and Peter Daszak. Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications*, 8(1):1124, 2017.
- [2] John Bongaarts. Ipbes, 2019. summary for policymakers of the global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services. *Population and Development Review*, 45(3):680–681, 2019.
- [3] Eileen M. Crimmins. Lifespan and healthspan: Past, present, and promise. *The Gerontologist*, 55(6):901–911, 2015.
- [4] Peter Ndeboc Fonkwo. Pricing infectious disease. *EMBO reports*, 9(S1):S13–S17, 2008.
- [5] Gaëtan Gavazzi, Francois Herrmann, and Karl-Heinz Krause. Aging and infectious diseases in the developing world. *Clinical Infectious Diseases*, 39(1):83–91, 2004.
- [6] Jean-François Guégan, Ahidjo Ayoub, Julien Cappelle, and Benoît de Thoisy. Forests and emerging infectious diseases: unleashing the beast within. *Environmental Research Letters*, 15(8):083007, aug 2020.
- [7] A. Haines and A. Cassels. Can the millennium development goals be attained? *Bmj*, 329(7462):394–7, 2004.
- [8] Alan R. Hinman. Global progress in infectious disease control. *Vaccine*, 16(11):1116–1121, 1998.
- [9] Richard Horton, Robert Beaglehole, Ruth Bonita, John Raeburn, Martin McKee, and Stig Wall. From public to planetary health: a manifesto. *The Lancet*, 383(9920):847, 2014.
- [10] Peter J. Hotez. Southern europe's coming plagues: Vector-borne neglected tropical diseases. *PLOS Neglected Tropical Diseases*, 10(6):e0004243, 2016.
- [11] Kate E. Jones, Nikkita G. Patel, Marc A. Levy, Adam Storeygard, Deborah Balk, John L. Gittleman, and Peter Daszak. Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993, 2008.

- [12] Henrik Lerner and Charlotte Berg. The concept of health in one health and some practical implications for research and education: what is one health? *Infection ecology and epidemiology*, 5:25300–25300, 2015.
- [13] Anthony J. McMichael. Globalization, climate change, and human health. *New England Journal of Medicine*, 368(14):1335–1343, 2013.
- [14] John P. Moore and Paul A. Offit. SARS-CoV-2 Vaccines and the Growing Threat of Viral Variants. *JAMA*, 325(9):821–822, 03 2021.
- [15] Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES). Workshop Report on Biodiversity and Pandemics of the Intergovernmental Platform on Biodiversity and Ecosystem Services (IPBES), October 2020. Suggested citation: IPBES (2020) Workshop Report on Biodiversity and Pandemics of the Intergovernmental Platform on Biodiversity and Ecosystem Services. Daszak, P., das Neves, C., Amuasi, J., Hayman, D., Kuiken, T., Roche, B., Zambrana-Torrel, C., Buss, P., Dondarova, H., Feferholtz, Y., Foldvari, G., Igbinosa, E., Junglen, S., Liu, Q., Suzan, G., Uhart, M., Wannous, C., Woolaston, K., Mosig Reidl, P., O’Brien, K., Pascual, U., Stoett, P., Li, H., Ngo, H. T., IPBES secretariat, Bonn, Germany, DOI:10.5281/zenodo.4147317.
- [16] Jonathan A. Patz, Peter Daszak, Gary M. Tabor, A. Alonso Aguirre, Mary Pearl, Jon Epstein, Nathan D. Wolfe, A. Marm Kilpatrick, Johannes Foufopoulos, David Molyneux, David J. Bradley, Change Working Group on Land Use, and Emergence Disease. Unhealthy landscapes: Policy recommendations on land use change and infectious disease emergence. *Environmental health perspectives*, 112(10):1092–1098, 2004.
- [17] Rhea J. Rocque, Caroline Beaudoin, Ruth Ndjaboue, Laura Cameron, Louann Poirier-Bergeron, Rose-Alice Poulin-Rheault, Catherine Fallon, Andrea C. Tricco, and Holly O. Witteman. Health effects of climate change: an overview of systematic reviews. *BMJ Open*, 11(6):e046333, 2021.
- [18] John W. Sanders, Greg S. Fuhrer, Mark D. Johnson, and Mark S. Riddle. The epidemiological transition: The current status of infectious diseases in the developed world versus the developing world. *Science Progress*, 91(1):1–37, 2008.
- [19] Holger Strulik and Sebastian Vollmer. Long-run trends of human aging and longevity. *Journal of Population Economics*, 26(4):1303–1323, 2013.
- [20] Ian R. Swingland, Eric C. Bettelheim, John Grace, Ghilleen T. Prance, Lindsay S. Saunders, Yadvinder Malhi, Patrick Meir, and Sandra Brown. Forests, carbon and global climate. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 360(1797):1567–1591, 2002.
- [21] A. J. Tatem, D. J. Rogers, and S. I. Hay. Global transport networks and infectious disease spread. *Adv Parasitol*, 62:293–343, 2006.

- 
- [22] Rachel Tidman, Bernadette Abela-Ridder, and Rafael Ruiz de Castañeda. The impact of climate change on neglected tropical diseases: a systematic review. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 115(2):147–168, 2021.
- [23] Bruce A. Wilcox, A. Alonso Aguirre, Peter Daszak, Pierre Horwitz, Pim Martens, Margot Parkes, Jonathan A. Patz, and David Waltner-Toews. Ecohealth: A transdisciplinary imperative for a sustainable future. *EcoHealth*, 1(1):3–5, 2004.



## Chapter 2

# Influence of socio-economic, demographic and climate factors on the regional distribution of dengue in the United States and Mexico

RESEARCH

Open Access



# Influence of socio-economic, demographic and climate factors on the regional distribution of dengue in the United States and Mexico

Matthew J. Watts<sup>1\*</sup>, Panagiota Kotsila<sup>1,4</sup>, P. Graham Mortyn<sup>1,5</sup>, Victor Sarto i Monteys<sup>1,3</sup> and Cesira Urzi Brancati<sup>2</sup>

## Abstract

**Background:** This study examines the impact of climate, socio-economic and demographic factors on the incidence of dengue in regions of the United States and Mexico. We select factors shown to predict dengue at a local level and test whether the association can be generalized to the regional or state level. In addition, we assess how different indicators perform compared to per capita gross domestic product (GDP), an indicator that is commonly used to predict the future distribution of dengue.

**Methods:** A unique spatial-temporal dataset was created by collating information from a variety of data sources to perform empirical analyses at the regional level. Relevant regions for the analysis were selected based on their receptivity and vulnerability to dengue. A conceptual framework was elaborated to guide variable selection. The relationship between the incidence of dengue and the climate, socio-economic and demographic factors was modelled via a Generalized Additive Model (GAM), which also accounted for the spatial and temporal auto-correlation.

**Results:** The socio-economic indicator (representing household income, education of the labour force, life expectancy at birth, and housing overcrowding), as well as more extensive access to broadband are associated with a drop in the incidence of dengue; by contrast, population growth and inter-regional migration are associated with higher incidence, after taking climate into account. An ageing population is also a predictor of higher incidence, but the relationship is concave and flattens at high rates. The rate of active physicians is associated with higher incidence, most likely because of more accurate reporting. If focusing on Mexico only, results remain broadly similar, however, workforce education was a better predictor of a drop in the incidence of dengue than household income.

**Conclusions:** Two lessons can be drawn from this study: first, while higher GDP is generally associated with a drop in the incidence of dengue, a more granular analysis reveals that the crucial factors are a rise in education (with fewer jobs in the primary sector) and better access to information or technological infrastructure. Secondly, factors that were shown to have an impact of dengue at the local level are also good predictors at the regional level. These indices may help us better understand factors responsible for the global distribution of dengue and also, given a warming climate, may help us to better predict vulnerable populations on a larger scale.

**Keywords:** Dengue, Climate-change, Global-warming, Socio-economic, Mosquito-borne, Vector-borne-diseases, GDP, GAM

## Introduction

The dengue virus (DENV) is one of the most important mosquito-borne viral diseases in the world today. Two main arthropod vectors are responsible for transmission of dengue viruses: *Aedes aegypti* (commonly known as

\*Correspondence: matthew.watts@uab.cat

<sup>1</sup> Institute of Environmental Science and Technology (ICTA), Autonomous University of Barcelona (UAB), Bellaterra, Spain

Full list of author information is available at the end of the article



© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## 2.1 Introduction

The dengue virus (DENV) is one of the most important mosquito-borne viral diseases in the world today. Two main arthropod vectors are responsible for transmission of dengue viruses: *Aedes aegypti* (commonly known as yellow fever mosquito) and *Aedes albopictus* (commonly known as tiger mosquito). *A. aegypti* mainly feeds on humans and is highly adapted to human habitations and urban areas; *A. albopictus* feeds on animals and humans and is more prevalent in rural and peri-urban environments. While *A. albopictus* is also responsible for dengue transmission among humans, it is a less likely vector than *A. aegypti* since it is adapted to a wider range of environments and has less restrictive feeding habits [1]. Both *Aedes* mosquitoes are highly adapted to breeding in aquatic habitats like ponds and lakes, but also micro habitats, such as tree-holes, rock crevices and even leaf axils [2]. The latter behaviour in recent times has benefited both species by allowing them to exploit a range of man-made aquatic breeding habitats, where water can accumulate, like urban gardens, vases in cemeteries, discarded bottles and plant pots; therefore, both species can survive in drier climates than expected, by exploiting artificial water sources.

Dengue is a disease caused by any one of four closely related viruses: DENV 1, DENV 2, DENV 3, or DENV 4. Currently, all four dengue serotypes are in circulation in the Americas and can co-circulate within a region; the actual distribution of each serotype is difficult to establish for a number of reasons, such as inadequate surveillance, under reporting, high numbers of asymptomatic carriers, and so on, as laid out by [4]. DENV causes an acute flu-like illness that affects people of all age groups. Those who recover from a dengue infection can expect lifelong immunity against that serotype and some partial, but temporary, cross-immunity to the other serotypes, although secondary infections by other serotypes increase the risk of developing severe dengue, which may cause lethal complications, and sometimes death [5].

There is currently no specific antiviral therapy for dengue fever; once the disease is contracted, there is no way to combat it other than relying on the host's immune response. Several vaccines are currently in development; however, given the current cost-effectiveness, efficacy, safety and estimated impact of vaccination, the WHO's present recommendation is to introduce it only in geographic settings (national or sub national) where the disease is particularly problematic [6].

### 2.1.1 Motivation for study

Climate change, specifically rising temperatures, is likely playing a crucial role in dengue transmission, potentially driving its expansion across the globe, as predicted by several studies [8–12]. Socio-economic conditions in a given location can be vital for a disease to persist once local transmission has occurred [13–18]; however, research in this domain, generally, does not account for socio-economic factors other than Gross domestic product (GDP), which is a standard measure of the market value of all the final goods and services produced over a specific time period in a given location. Some studies have looked at the interaction between climate, socio-economic factors and demographics at a local level [19–23], focusing on factors specific to local areas, which means that their findings cannot be easily extrapolated

to the macro level. To get better estimates of where dengue may spread, there is a need to understand how climate factors, socio-economic factors and demographic factors interact over a greater geographic scale to reveal common global patterns.

The original contribution of this article is that it selects factors shown to predict dengue at a local level and tests whether the association can be generalised to the regional or state level. In addition, we propose a more comprehensive set of socio-economic predictors of dengue transmission, to disentangle the role of GDP from other measures. Although a useful and parsimonious indicator, GDP is a very broad measure and it is not necessarily reflective of population health and well-being, distribution of wealth, discrimination and spending on public welfare [24]. More importantly, GDP alone may not be able to capture cross-regional differences. The predominance of using GDP as an indicator has been largely questioned [25–28]; for some time now researchers in human health geography, critical public health, and social epidemiology have requested more careful consideration of the contextual social and economic conditions that shape diseases at the local level [29, 30].

To this end, this study investigates regional differences in the incidence of dengue by evaluating the impact of socio-economic and demographic factors such as household income, regional rates of education, housing overcrowding, life expectancy, medical resources, migration flows, age structure of the population (the proportion of people under 14 and over 65), and population density.

The study focuses on the occurrence and distribution of dengue in Mexico and southern regions of the United States (US) where dengue has been reported, as some US regions share very similar environmental conditions but have distinct socio-economic conditions. [15]. This study takes advantage of time series data between 2011 and 2019 and it is, therefore, able to exploit cross sectional variation between states, and variation over time for each state.

### 2.1.2 Conceptual framework

Dengue transmission is determined by interactions between host, vector and pathogen, and modulated by ecological, climatic and geographic factors, including socio-economic factors. Regions were selected for the empirical analysis if conditions were met in terms of their *receptivity* and *vulnerability*, based on principles laid out in the WHO's framework for malaria elimination [31].

Receptivity is defined as the ability of an ecosystem to allow transmission of a virus (dengue in this study). An ecosystem can be considered receptive if competent vectors, a suitable climate and a susceptible population are present; in other words, regions are selected if autochthonous virus transmission may occur because human populations and vector populations overlap/interact. Vulnerability occurs when either 1) a region was receptive and had regularly reported cases over the study period (endemic) or 2) bordered an endemic region and occasionally reported cases which (likely due to spread or importation from neighbouring regions). We defined modulating factors as variables that influence the transmission dynamics of dengue such as host population size, host density, climate factors and medical interventions.

#### Receptivity

Since dengue is a vector borne disease, understanding the key ecological requirements of its vectors is crucial to assessing the receptivity of a region. As explained below,

some of the main factors determining the receptivity of a region to dengue (due to the presence of its vectors) are: its physical environment (land use), the overlap with the human population, and its climate.

Both types of *Aedes* mosquitoes that transmit the dengue virus are ectothermic organisms and are highly sensitive to colder temperatures and extreme high temperatures. *A. albopictus* adults can survive in temperatures from 15 to 35°C and *A. aegypti* from 10 to 35°C [32], while their growth and development are severely inhibited in ambient temperatures below 13°C or above 35°C. *A. albopictus* eggs though, can go through diapause (suspended development) when exposed to extreme cold (down to -10°C). This adaptation allows them to inhabit environments with a wider annual temperature range, with more distinct seasonal changes than in tropical climates, where climate is more homogeneous. *A. aegypti* can endure a wider range of temperatures, but its survival at temperatures below 14-15°C is limited to short periods, since its mobility is severely restricted and its ability to imbibe blood impeded. *A. aegypti* is also highly sensitive to fluctuations in temperatures. As for most mosquito species, availability of freshwater habitat, humidity and precipitation are highly indicative of their distribution in the environment.

To account for this, we selected a range of humidity and temperature variables for analysis which would capture mosquitoes' living requirements.

## Vulnerability

As direct measures of vulnerability we include spatial effects (neighbourhood structures) in our models in order to explicitly account for spill over effects with infected neighbouring regions (for a more detailed description see the methods section). Indirect measures of vulnerability can be derived from traditional patterns of travel and population flow in the area; indeed, well connected areas, in terms of trade and transport with considerable human movement, can benefit both mosquito species and dengue, by facilitating their movement and spread [33–36].

## Modulating factors

Modulating factors can either speed up or slow down transmission. The transmission cycle of dengue is complex, since there are several key interactions at play between the virus, host and vector. Density of both the vector and host are fundamental factors in disease transmission, as contact between infected vectors and susceptible hosts is the source of new infections [37]. Mosquito Breeding habitat can be increased by precipitation and flooding [38], temperature heavily influences mosquito hatching rate, development time [39–40] and optimal temperature can shorten the extrinsic incubation period (EIP) [41]. While there are no datasets covering mosquito population abundance in all of our study regions, we selected meteorological variables that predict mosquito abundance and therefore are related to dengue transmission. Furthermore, there are several socio-economic risk factors of dengue including home water storage (rather than receiving piped water), poor sanitation, and poor public services (e.g. litter not collected) [15, 42–46]. Such factors can be responsible for creating breeding habitat for mosquitoes and bringing them into closer contact with humans, therefore increasing the risk of dengue. By contrast, use of mosquito nets, insect screens, and air-conditioning, can limit the chance of being bitten. Similarly, knowledge and education of mosquito ecology can also help people make personal in-

interventions and reduce risk of being bitten [47]. Because there are no direct measures of home water storage or the use of mosquito nets, we use a range of socio-economic indicators as proxies, capturing a latent variable that would represent vector risk. The rationale is that people living in locations with better socio-economic conditions can avoid contact with mosquitoes and restrict virus transmission, either from the bottom-up (e.g. personal interventions) or the top-down (e.g. Regional government pest control). However, it is important to note that factors associated with higher economic status can also bring humans into closer contact with mosquitoes, for example home owners with gardens and potted plants and ponds or having good access to recreational space where mosquitoes can breed [48]. In terms of post-infection factors that influence dengue transmission, access to health care, risk perception and access to information on dengue infection symptoms had positive effects on people's decision to seek medical help when presented with dengue infection symptoms [17, 47, 49]. To reflect this in the conceptual framework we selected variables that would proxy access to health care and variables which would represent access to information and personal knowledge. Finally, younger people are more likely to be infected by dengue [50], so we selected variables that represent the age structure of the population.

## 2.2 Materials and Methods

In this study, we compiled a spatial temporal data-set that would reflect the conceptual framework. By predicting the distribution of *A. albopictus* and *A. aegypti* in Mexico and the United States, we could determine which regions were receptive i.e. there was there an overlap between the vector distribution and the human population at risk. By combining these results with reported cases of dengue, we could determine which regions were vulnerable. We then went on to collect data on modulating factors of dengue transmission in vulnerable regions. Furthermore, our vector distribution maps allowed us to extract more accurate data on the host population at risk and climatic factors that contribute to disease transmission.

### 2.2.1 Species distribution models to estimate regional susceptibility

Because the exact distribution of vectors is unknown, we estimated the likelihood that a vector would occur in a region conditional on a set of covariates. More specifically, we estimated the distribution of the *Aedes* mosquitoes using a generalized additive logistic regression, with point location occurrence as dependent variable, and annual temperature range, mean temperature of the coldest quarter, precipitation during the driest quarter as covariates. Predictions were then used to select susceptible regions.

Point location occurrence data for *A. aegypti* and *A. albopictus* were obtained from a global geographic database of known occurrences between 1960 and 2014, compiled by members of the Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow [51]. Point occurrence data represent spatial geo-coordinates of a location in which a given individual organism was sampled or sighted. Many of the samples in this

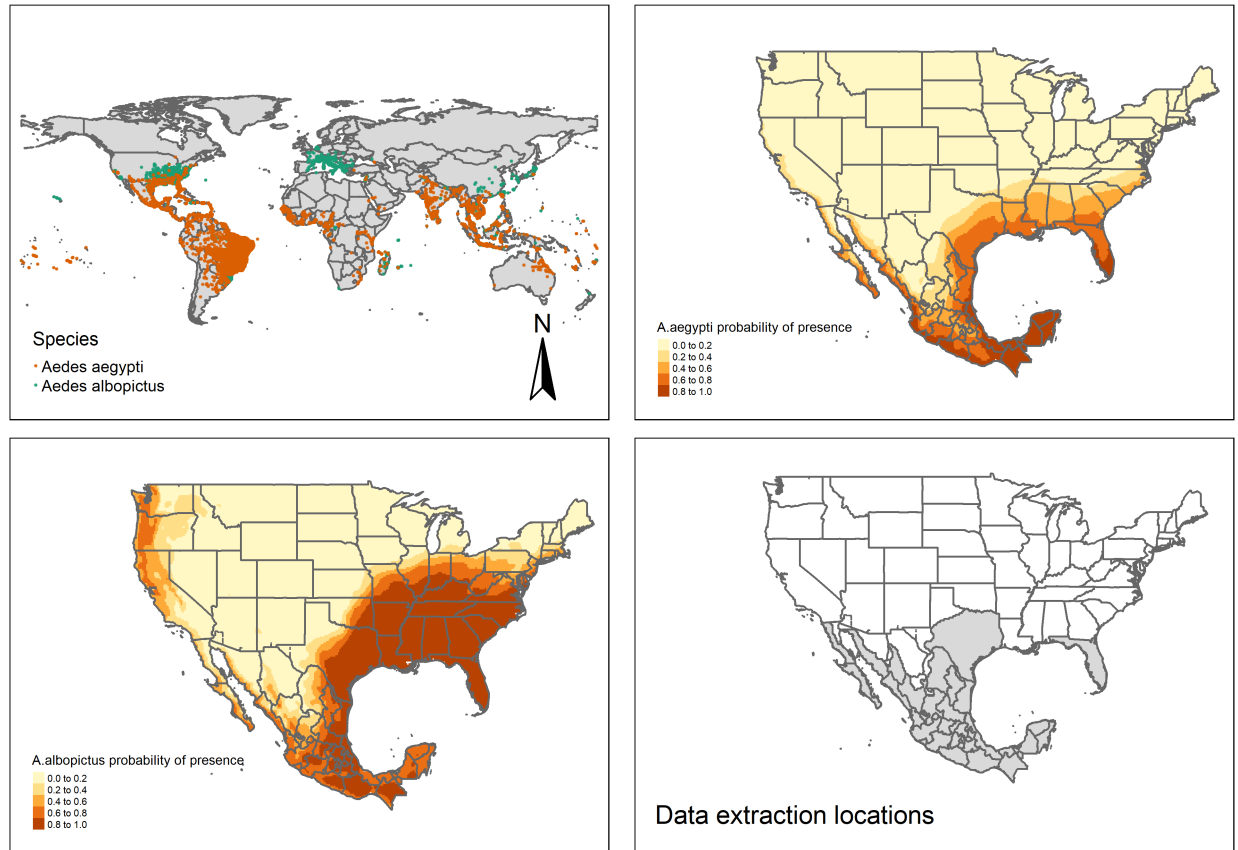


Figure 2.1: *Aedes* sample locations and SDM results: Top left: *Aedes* point locations. Top right: Results of *Aedes aegypti* SDM. Bottom left: Results of *Aedes albopictus* SDM. Bottom right: Receptive regions / data extraction locations.

data-set consists of museum records or unpublished studies including national entomological surveys. Since the data-set contained sparse information relating to the timings and frequency of each sample, we selected global observations from 1970 onward to capture the entire range of climatic conditions that species can survive in, and to limit potential sample bias caused through the selection of localised seasonal collections. We also removed any duplicate observations i.e. replicate coordinates. Climate data were extracted using R's DISMO package in all point locations where mosquitoes occurred. Climate data for the species distribution prediction modelling were sourced from the MERRAclim, a database compiled by members of the Department of Biology and Geology, Physics and Inorganic Chemistry, Rey Juan Carlos University [52]. This data-set was built using 2 metre above surface air temperature (Kelvin degrees) and 2 meter above surface specific humidity (kg of water/kg of air) satellite observations from the Modern Era Retrospective Analysis for Research and Applications Reanalysis.

Figure 2.1 shows the results of the modelling and *Aedes* sample locations. Tables providing summary statistics for the climate values at *Aedes* point locations can be found in Appendix A. More specific information on statistical methods and results from this analysis can also found in Appendix A.

## 2.2.2 Data extraction and methods to assess the impact of climate, demographic and socio-economic factors on dengue

The Global Administrative Unit Layers [54] data-set along with our *Aedes* distribution maps (results figure 2.1, bottom right) were combined using R's Sf package to create regional shape files that could spatially capture and process the human population and climate data for the main analysis. The GAUL data set contains geographic information in the form of shape files that lay out within country boundaries linked to a unique nomenclature. Countries are broken down into statistical subdivisions e.g., ~ADM0 representing data at country level (e.g. ~US), ADM1 at regional level (e.g. ~California).

Climate data for the main analysis i.e. measuring the impact of the climate variables on dengue transmission, were sourced from the Climate Prediction Center (CPC) of the National Centers for Environmental Prediction (NCEP), see [53]. These data represent a global summary of daily weather data. The CPC extracts surface synoptic weather observations from the Global Telecommunications System (GTS), which collects global data from a combination of weather station and satellite observations. Files were processed in R with the NetCDF, Raster and Dismo packages in order to create annual bio-climatic variables. The bio-climatic variables in this study were derived from daily maximum temperatures, daily minimum temperatures and total daily rainfall.

Population count data to predict the number of persons at risk in a region were sourced from the Socioeconomic Data and Applications Center's Gridded Population of the World data set [57]. This data set estimates population count for the years 2000, 2005, 2010, 2015, and 2020, consistent with national censuses and population registers. Data were extracted from areas where vector presence was predicted. R's Zoo package was used to replace values for missing years, by implementing a linear interpolation method that would predict trends between years. This way increases or decreases in human population were controlled for in the final model. All spatial data was aggregated to the state level.

### Dengue Case Data

Dengue case data for Mexico 2011-2019 were obtained from the Mexican Deputy General of Epidemiology web-page, which provided reports on all positive serious and non-serious cases of dengue ([www.gob.mx](http://www.gob.mx)). All data was provided at the regional level (ADM1 level). Case data for the United States were extracted from [www.cdc.gov/arboNet](http://www.cdc.gov/arboNet), since data are provided at the county level (ADM2 level) we needed to aggregate them to the state level (ADM1 level) in order to match them with the main data-set.

### OECD Socio-Economic and Demographic Data

Socio-economic and demographic data were extracted from the OECD's Regional Statistics and Indicators Database [56]. This database provides comparable statistics and indicators and is presented in yearly time series. To capture factors determining the vulnerability of a region, we selected the variables "Inter-regional migration



rate”, “Population density growth” and “Gross domestic product (GDP)”. For factors representing the socio-economic position of residents in a region we selected: “Household income”, “Life expectancy at birth”, and a measure of housing overcrowding “Number of rooms per person”. Furthermore, we selected “Secondary education” which would also help to capture areas where there is a higher proportion of manual labourer, e.g. agricultural workers or people working outdoors who may be more exposed to mosquitoes. We also selected “Perceived social network support”, “Self-evaluation of life satisfaction”, and “Perception of corruption” to try to capture additional features of a region, such as quality of life. Since these three variables yield some indication of how people perceive their surroundings and quality of life, we assume that poorer scores will capture poor infrastructure, poor public services, lack of basic provisions and lack of beneficial government intervention. To represent access to healthcare we selected “Active Physicians rate”, and variables which would represent access to information and personal knowledge i.e., “Broadband access” (however knowledge is also captured by “Secondary education”). Finally, younger people are more likely to be infected by dengue [50], so we selected variables that represent the age structure of the population i.e. “Percentage of Old Population Group (65+)” and “Percentage of Youth population group (0-14)”. Missing values were filled based on values for previous years or subsequent years, depending on their position in the data set.

All data were joined using the year of observation and region code, using R’s Dplyr package.

Table 2.1 provides summary statistics of all the collected data for the final models.

Statistic	Min	Max	Mean	St. Dev.
Primary Income of Private Households (USD per head)	2,541.8	1,159,750.0	68,486.2	229,793.6
Regional Gross Domestic Product per Capita	2,883.4	1,044,310.0	64,153.2	206,614.7
Share of labour force with at least secondary education	26.9	89.5	42.1	13.0
Share of households with broadband access	7.3	85.2	41.5	18.9
Self-evaluation of life satisfaction	6	9	7.2	0.6
Perceived social network support	59	96	79.4	9.6
Perception of corruption	36.5	90.1	63.3	11.1
Active Physicians Rate (physicians for 1000 population)	0.7	4.8	1.6	0.6
Life Expectancy at Birth	70.5	79.4	75.1	1.4
Number of rooms per person	0.7	2.5	1.1	0.3
Inter-regional migration rate, (% migrants over population)	0.5	7.0	2.0	1.3
Population density growth	97.8	179.6	122.4	12.9
Percentage of Old Population Group (65+)	3.1	20.5	7.2	2.5
Percentage of Youth Population Group (0-14)	16.3	34.4	27.5	2.9
DGE Mexico confirmed serious dengue cases	0	5,041	259.3	606.0
DGE Mexico confirmed non-serious dengue cases	0	9,195	627.2	1,161.6
CDC US confirmed dengue cases	0	53	0.3	3.5
Population in aedes infected regions	335,728.3	28,145,145.0	4,710,557.0	5,484,386.0
Mean (C) temperature of warmest quarter	16.2	32.3	25.6	3.7
Mean (C) Temperature of Coldest Quarter	9.0	25.3	18.0	4.3
Precipitation of warmest quarter	0.0	8.3	1.1	1.3
Precipitation of Warmest Quarter	0.1	24.9	9.1	4.7

Table 2.1: Final dataset 2012-2019

## 2.2.3 Statistical Methods

### Factor Analysis - Data Processing for Regional Analysis

A preliminary correlation analysis (see additional files S8-S9 - diagnostics) revealed how some of the socio-economic variables are strongly correlated with each other,

and if included in a regression would give rise to multi-collinearity issues. By over-inflating the standard errors, multi-collinearity makes some variables statistically insignificant when they should be significant. To address this issue, following similar methods to [58], a factor analysis by maximum likelihood (VARIMAX rotation) was performed on socio-economic variables.

Factor analysis is a method for investigating whether a number of variables of interest  $Y_1, Y_2, \dots, Y_n$ , are linearly related to a smaller number of latent (i.e. not directly measured) factors  $F_1, F_2, \dots, F_k$ . The basic concept of factor analysis is that multiple observed variables have similar patterns because they are all associated with a latent variable. The factors are constructed in such a way that they capture the maximum amount of common variance (correlation) of the original items; the eigenvalue is a measure of how much of the variance of the observed variables a factor explains. The factor analysis can be formalized as follows:

$$Y_1 = \beta_{10} + \beta_{11}F_1 + \beta_{12}F_2 + \dots + \beta_{1k}F_k + \epsilon$$

$$Y_2 = \beta_{20} + \beta_{21}F_1 + \beta_{22}F_2 + \dots + \beta_{2k}F_k + \epsilon$$

$$Y_N = \beta_{n0} + \beta_{n1}F_1 + \beta_{n2}F_2 + \dots + \beta_{nk}F_k + \epsilon$$

Before performing factor analysis, all variables had to be standardised to z-scores  $(x - \mu)/\sigma$  to ensure that they were on the same scale. After performing the factor analysis, the predicted values for the factors for any individual region can be estimated. These predictions, known as factor scores, are weighted sums of the values of the observed items. Roughly, items with a stronger correlation with a factor component (i.e. those with larger loadings) will receive higher weights in the calculation of a score for that factor.

### Quality of life index - Data Processing for Regional Analysis

We created a 'Quality of Life Index' by combining 3 variables from the OECD regional database: 'Self-evaluation of life satisfaction', 'Perceived social network support' and 'Perception of corruption'. The variables were standardised, harmonised and combined into a composite indicator, capturing a latent quality of life measure, because each element on its own is unlikely to have a direct relationship with dengue.

### General additive regression model to assess impact of independent variables on dengue case data at regional level

One of the main issues with our data-set is that it did not meet some basic assumptions for statistical inference, and specifically the data are not independent and identically distributed random variables (iid). More specifically, the data-set captured repeated measurements over the same regions, and observations were not independent because of spill over effects from neighbouring regions, therefore we needed to implement an appropriate statistical design to control for both temporal and spatial pseudo replication (lack of independence). We could deal with this in two ways, 1) either using a generalized linear mixed model (GLMM) approach, relaxing the assumption of independence and estimating the spatial/temporal correlation between residuals, or 2) model the spatial and temporal dependence in the systematic part of the model [59]. We opted to use a Generalized Additive Model (GAM) using R's *Mgcv* statistical package because of its versatility and ability to

fit complex models that would converge even with low numbers of observations and could capture potential complex non-linear relationships. One of the advantages of GAMs is that we do not need to determine the functional form of the relationship beforehand. In general, such models transform the mean response to an additive form so that additive components are smooth functions (e.g., splines) of the covariates, in which functions themselves are expressed as basis-function expansions. The spatial auto-correlation in the GAM model was approximated by a Markov random field (MRF) smoother, defined by the geographic areas and their neighbourhood structure. We used R's Spdep package to create a queen neighbours list (adjacency matrix) based on regions with contiguous boundaries i.e. those sharing one or more boundary point. We used a medium rank MRF, which represented roughly one coefficient for two areas. The local Markov property assumes that a region is conditionally independent of all other regions unless regions share a boundary. This feature allowed us to model the correlation between geographical neighbours and smooth over contiguous spatial areas, summarizing the trend of the response variable as a function of the predictors, for further information see section 5.4.2 of [59]. In order to account for variation in the response variable over time, not attributed to the other explanatory variables in our model, we used a saturated time effect for years, where a separate effect per time point is estimated.

We first tried to fit our model using a Poisson distribution. However, the mean of our dependent variable (dengue cases by region and year) was lower than its variance -  $E(Y) < \text{Var}(Y)$ , suggesting that the data are over-dispersed. We also tried to fit our models using the negative binomial, quasi poisson and tweedie distribution, all particularly suited when the variance is much larger than the mean. After several tests, we concluded that the tweedie distribution worked well with our data and allowed us to model the incident rate. Analysis of model diagnostic tests didn't reveal any major issues, in general residuals appeared to be randomly distributed (see additional files S10-S19 - diagnostics).

Tweedie distributions are defined as subfamily of (reproductive) exponential dispersion models (ED), with a special mean-variance relationship.

A random variable  $Y$  is Tweedie distributed  $TW_p(\mu, \sigma^2)$  if  $Y \sim ED(\mu, \sigma^2)$ , with mean  $\mu = E(Y)$ , positive dispersion parameter  $\sigma^2$  and  $\text{Var}(Y) = \mu\sigma^2$ .

The empirical model can then be written as:

$$E(Y) = f_1(X_{it}) + f_n(\text{Year}_t) + f_m(\text{Region}_i)$$

Where the  $f(\cdot)$  stands for smooth functions;  $E(Y)_{it}$  is equal to dengue incidence in region  $i$  at time  $t$ , which we assume to be Tweedie distributed;  $X_{it}$  - is a vector of socio-economic, demographic and climate variables.  $\text{Year}_t$  is a function of the time intercept and  $\text{Region}_i$  represents neighbourhood structure of region.

We run two separate sets of analyses: one comparing regions in the US and Mexico and another one looking at Mexico only, to check for robustness.

## 2.3 Results

Figures 2.2 and 2.3 provide a descriptive overview of the study regions, a characterisation of their environments and the reported disease incidence for those years. As we can observe, the majority of dengue cases are reported in tropical and sub tropical climates.

## Köppen-Geiger Climate Classification

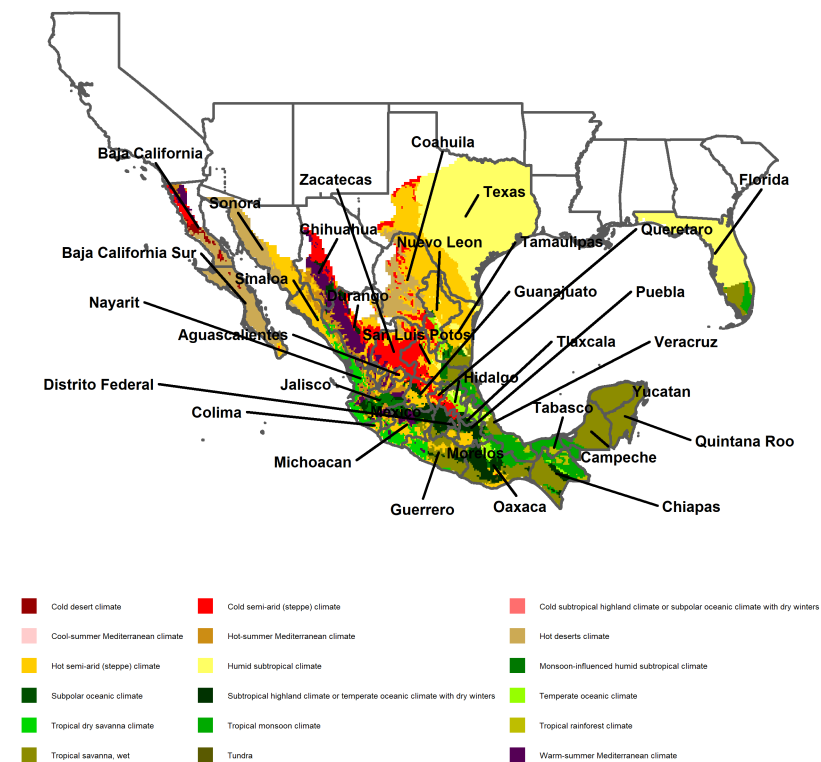


Figure 2.2: Köppen-Geiger Climate Classification in study regions (Source: [koeppen-geiger.vu-wien.ac.at](http://koeppen-geiger.vu-wien.ac.at))

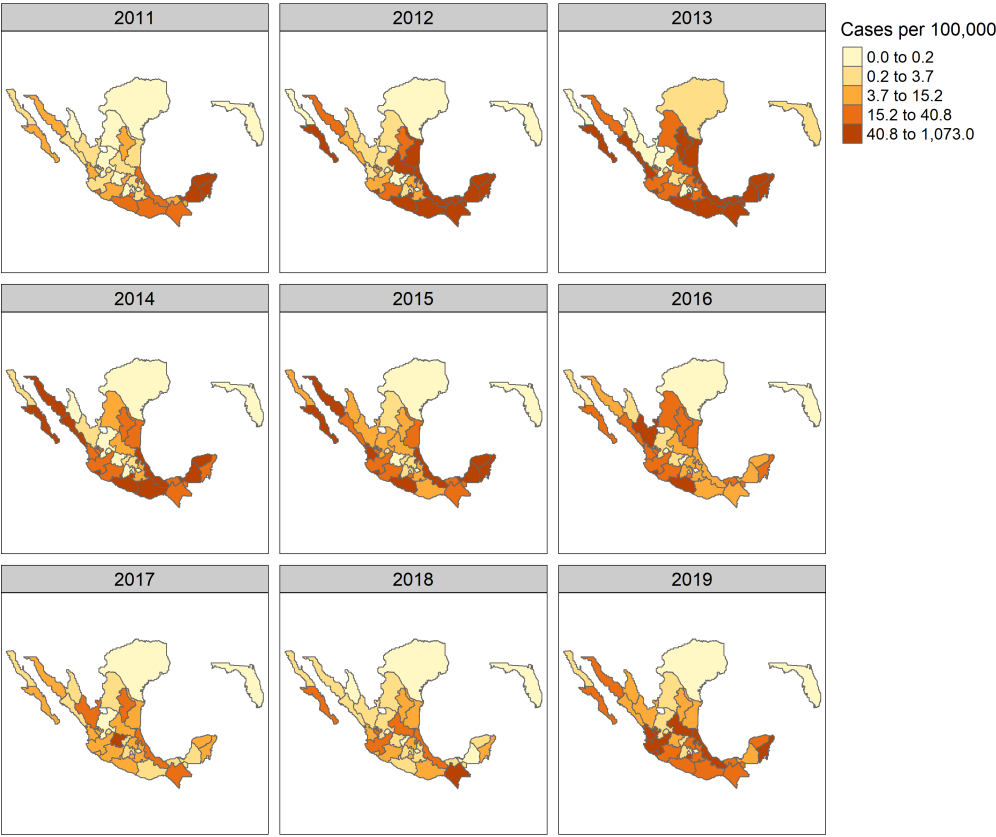


Figure 2.3: Crude incidence rates of dengue per 100,000 people

Tables 2.2 and 2.3 provide the results of the factor analysis i.e. the weighting of our socio-economic indicators.

	Factor1
Primary Income of Private Households (USD per head)	0.91
Share of labour force with at least secondary education	0.95
Life Expectancy at Birth	0.78
Number of rooms per person	0.97

Table 2.2: Socio-economic Factor Analysis Results US/MEX

	Factor1
Primary Income of Private Households (USD per head)	0.41
Share of labour force with at least secondary education	0.95
Life Expectancy at Birth	0.59
Number of rooms per person	0.65

Table 2.3: Socio-economic Factor Analysis Results Mexico

Table 2.4 shows the results for the regression model comparing confirmed dengue cases in the US and Mexico for 2011-2019. Table 2.5 restricts the analysis to Mexico only since we could exploit a better data-set in terms of case reporting, scale, and we could explore the impact of the socio-economic variables individually since, there was less correlation with this type of data between Mexican regions.

### 2.3.1 US/Mex analysis

#### Socio-economic and demographic indices Mexico/US

It was not possible to explore the individual impact of all of the variables in our data-set because of collinearity issues. Population density was found to be positively correlated with GDP and primary income. “Percentage of Old Population Group (65+)” was negatively correlated with “Percentage of Youth Population Group (0-14)” (see Appendix A1.3-4) diagnostics). For this reason, we performed a factor analysis to reduce the number of variables, as explained in more detail in the section on statistical methods. The Mexico/US factor analysis captured the variance in 4 highly correlated variables: higher share of labour force with at least secondary education, more rooms per inhabitant, life expectancy at birth, primary income of households, and yielded one composite indicator (see Table 2.2) , which we included as a regressor. A priori, the socio-economic indicator is expected to have a negative association with dengue.

We built our statistical model in a stepwise fashion so we could analyse it using the lowest Akaike Information Criterion (AIC) which would help us validate the quality of statistical models for our dataset. The first column of Table 2.4 (GDP Model) shows the association between regional GDP and dengue cases across the regions; the second column (SE Model) shows the association between regional dengue cases and the socio-economic indicator derived through factor analysis, plus other variables such as active physician rate, broadband access and the quality of life index. Column3 (Dem Model) includes demographic variables, such as inter-regional

	GDP Model	SE Model	Dem Model	Clim Model	GDP full model	Full Model
Intercept	3.19*** (0.26)	2.44*** (0.22)	2.53*** (0.23)	2.46*** (0.23)	2.59*** (0.22)	2.22*** (0.20)
Per capita GDP	-0.00*** (0.00)				-0.00*** (0.00)	
Socio-economic index		1.87*** (1.97)				1.78*** (1.92)
Active Physicians per 1000		1.72** (1.90)			1.87*** (1.97)	1.88*** (1.98)
Households (%) with broadband access		1.94*** (1.99)			1.95*** (1.99)	1.95*** (1.99)
Quality of Life index		1.78* (1.92)			1.00 (1.00)	1.00 (1.00)
Inter-regional migration rate			1.00*** (1.00)		1.65 (1.85)	1.61 (1.82)
Pop density growth			1.82 (1.96)		1.00*** (1.00)	1.00*** (1.01)
Percentage of Population (65+)			1.97*** (1.99)		1.93*** (1.96)	1.79*** (1.89)
Mean temp (C) of coldest quarter				1.90*** (1.98)	1.88*** (1.97)	1.83*** (1.95)
Prec of warmest quarter				1.57* (1.81)	1.63* (1.84)	1.53* (1.76)
Year	7.27*** (8.00)	6.75*** (8.00)	6.99*** (8.00)	7.28*** (8.00)	6.61*** (8.00)	6.61*** (8.00)
Region	13.13*** (13.91)	12.76*** (13.72)	13.04*** (13.69)	12.72*** (13.81)	10.46*** (12.24)	10.45*** (12.19)
AIC	2339.82	2322.32	2332.83	2285.47	2233.10	2231.57
Deviance	1288.59	1205.43	1257.24	1143.83	967.55	962.59
Deviance explained	0.51	0.55	0.53	0.59	0.66	0.66
Dispersion	3.52	3.37	3.47	3.20	2.84	2.83
R <sup>2</sup>	0.36	0.42	0.40	0.40	0.51	0.51
Num. obs.	306	306	306	306	306	306
Num. smooth terms	2	6	5	4	10	11

\*\*\* $p < 0.01$ ; \*\* $p < 0.05$ ; \* $p < 0.1$

Table 2.4: Final Regression Models US/MEX: EDF value is reported as the coefficient, and DF is included in parentheses (not standard errors because a chi-square test is used for the smooth terms).

migration rate, population density growth and the percentage of older population (65+). Column 4 (Clim Model) includes the climate variables mean temperature of the coldest quarter and precipitation in the warmest quarter. The “full model” in column 5 shows the relationship between dengue incidence and all explanatory variables in our final model. Table 2.1 also summarises the relevant statistics (AIC, Deviance, Adjusted R squared and so on) to compare the different specifications; the full model has the best fit (lower AIC and higher adjusted R squared), followed by the one in which we control only for the climate variables (as well as the year, regional effects); the first model, controlling for GDP alone, has the highest AIC and has a worse fit than the specification including the socio-economic indicators. When controlling for demographic and climate variables, the impact of the socio-economic indicators still remains statistically significant, as well as the impact of temperature.

Please note that as we are not estimating a standard regression model, the figures reported should not be read as coefficients, but degrees of freedom of the smooth terms. Given that we cannot interpret the coefficients to infer the sign and magnitude of the relationship, we visualise it by plot.

Figure 2.4 plots the partial effects - the relationship between a change in each of the covariates and a change in the fitted values in the full model; the first plot shows that the socio-economic index has linear negative impact, but the relationship becomes weaker at very high scores; given the weight of each variable in the factor analysis, the results can be interpreted as an increase higher share of labour force with at least secondary education, more rooms per inhabitant, life expectancy at birth, primary income of households are associated with fewer dengue cases. Regions with better broadband access tend to be those with lower incidence rates of dengue,

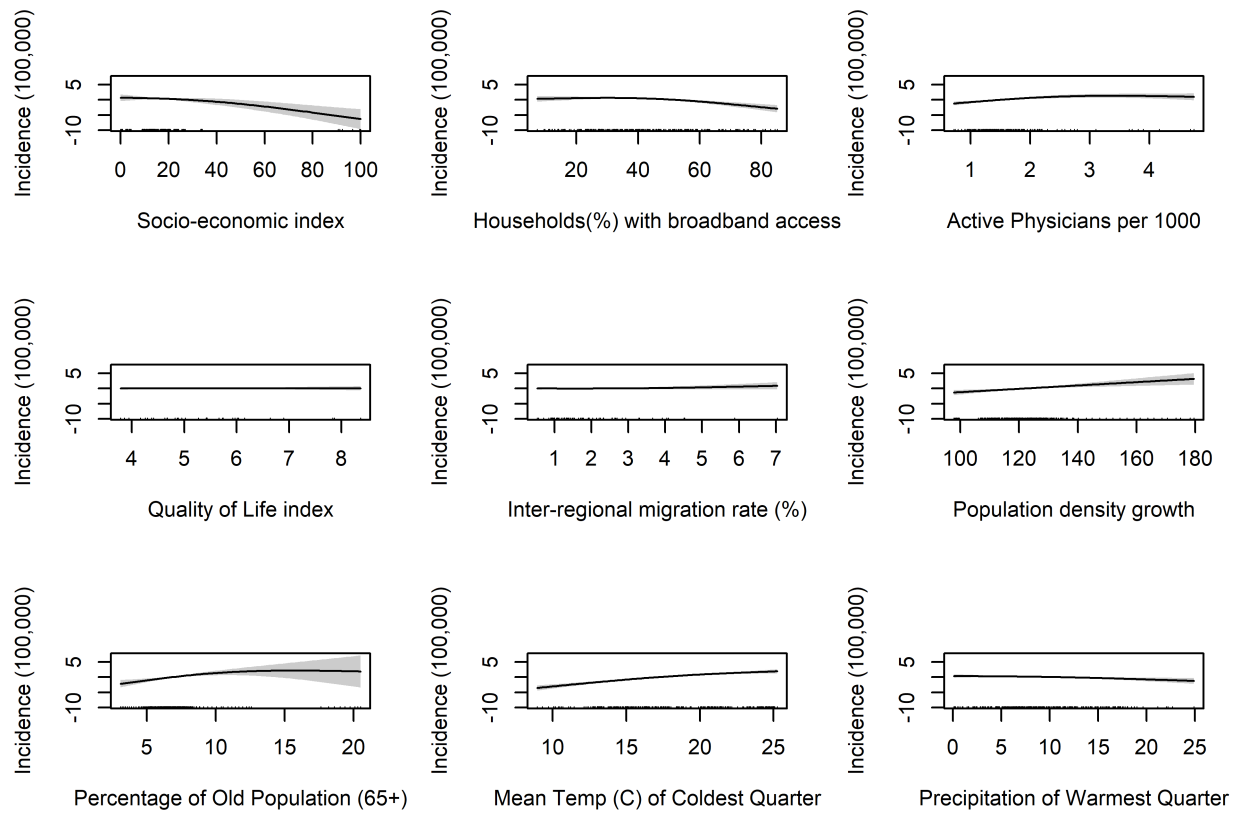


Figure 2.4: Partial effects of explanatory variables: GAM Mex/US model



however in this case the relationship is flat at low levels of broadband penetration (below 40 percent) and then turns negative and quadratic at higher levels of access. These results could suggest residents are more likely to search for information on dengue prevention measures consequently lowering transmission potential, or when suffering with symptoms may be more likely to seek medical advice, therefore breaking the transmission cycle; these results are consistent with findings by [63–64]. This result also could be an indicator of more advance and urbanised regions vs agricultural and less developed regions. It is reported that dengue tends to affect more those working in labour-intensive industries, such as agriculture or fishing [65–66]. The variable active physician rates has a positive impact on the incidence of dengue, in that regions with more active physicians tend to have higher incidence; however, this is likely due to more accurate reporting. Even in this case, the relationship is concave - positive up to 3 percent rate and flat afterwards.

The impact of the demographic variables on the incidence of dengue also follows the expected sign, with inter-regional migration rate and population density growth being associated with a linear increase in the incidence of dengue; the presence of an older population is associated with higher incidence of dengue up to a certain level - it peaks at around 14 percent - and then a reduction, as can be seen from Figure 2.4. One possible explanation for this is that a higher proportion of older people means a more vulnerable population, however very high rates are also associated with wealthier regions, which offset the main impact of age. Figure 2.4 also show the impact of the Mean temperature ( $^{\circ}\text{C}$ ) of coldest quarter variable is almost linear. We can see that most cases occur in regions which have particularly mild cold seasons. This is concurrent with the literature, we would expect to see more cases of dengue in regions with tropical climates, where there is a distinct absence of a cold season, during which low temperatures would kill the mosquitoes off or cause mosquitoes to overwinter effectively inhibiting disease transmission, instead such conditions allow the virus and mosquitoes to persist throughout the year.

The relationship between rainfall and dengue incidence in the full model is slightly negative and significant; even though this finding could appear counter intuitive, it is probably due to the fact that mosquito larvae can be washed away during intense rainfall [67]. Furthermore, both *Aedes* mosquitoes can survive in drier climates than expected, by exploiting artificial water sources and man-made habitats, as already mentioned in the introduction.

### 2.3.2 Mex analysis

For our second analysis focusing on differential diffusion of dengue within Mexican regions, we were able to analyse variables individually since there is significantly less correlation between the socio-economic variables. However, we could not select "population density" because of a correlation with "primary income of household"s and "GDP". "Percentage of Old Population Group (65+)" was negatively correlated with "population density growth" so was not included in the final model. Furthermore, "Percentage of population share (0-14)" was highly correlated with "access to broadband" and "workforce with secondary education" (and negatively correlated with population 65+), so we didn't include it in the study. We again built our second statistical model in a stepwise fashion so we could analyse it using the lowest Akaike Information Criterion (AIC) which would help us validate the quality

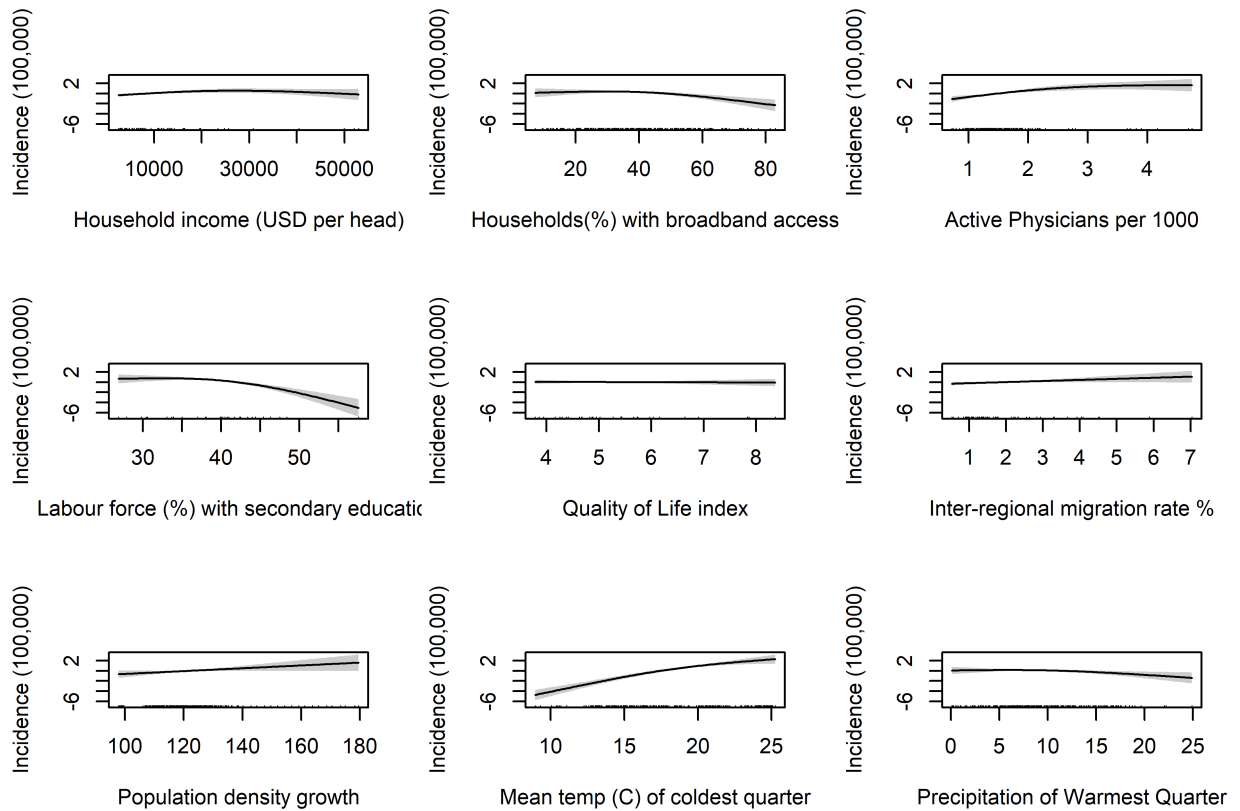


Figure 2.5: Partial effects of explanatory variables: GAM Mexico model

of statistical models for our data set.

Figure 2.5 and Table 2.5 present the results of our second analysis focusing only on Mexican regions.

Our findings for the second analysis are similar to the first: the most significant variables are "Share of households with internet access", "Active Physicians Rate (1000 pop)" and "Mean temperature (C) of coldest quarter". Our socio-economic indicator was a good predictor of dengue incidence, although when "GDP" was paired with other individual variables from the factor analysis (except primary income) it helped to create a very useful model. The best fit model was our final specification using our socio-economic variables individually; however, primary income of households is not a reliable predictor of dengue, since, by the concave relationship, it would appear that gains in economic activity may increase the spread of the virus (for instance because of movement of goods and people), but could also be correlated with higher reporting. One of the strongest predictors of dengue in our final specification is "Share of labour force with secondary education". As previously noted, this is consistent with other findings by [65–66] as dengue tends to affect more those working in labour-intensive industries, such as agriculture or fishing.

## 2.4 Discussion and Conclusions

The study investigated the impact of socio-economic, demographic and climate variables on the distribution of dengue. Its original contribution is that it selected factors

	SE Model	Dem Model	Clim Model	GDP full model	SEindex model	Full Model
Intercept	2.57*** (0.20)	2.89*** (0.21)	2.71*** (0.23)	2.20*** (0.24)	2.46*** (0.19)	2.44*** (0.18)
Per capita GDP				0.00* (0.00)		
Socio-economic index					1.96*** (1.99)	
Income of Private Households	1.95*** (2.00)					1.83** (1.96)
Share households with broadband	1.88*** (1.98)			1.93*** (1.99)	1.93*** (1.99)	1.92*** (1.99)
Active Physicians (1000 pop)	1.00*** (1.00)			1.81*** (1.95)	1.75*** (1.92)	1.81*** (1.95)
Number of rooms pp	1.95*** (1.99)			1.00 (1.00)		1.00 (1.00)
Labour force with secondary edu	1.97*** (2.00)			1.95*** (1.99)		1.95*** (1.99)
Quality of Life index	1.00 (1.00)			1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
Inter-regional migration rate		1.00*** (1.00)		1.00 (1.00)	1.00 (1.00)	1.00* (1.00)
Pop density growth		1.51 (1.74)		1.00* (1.00)	1.00 (1.00)	1.00* (1.00)
Mean temp (C) of coldest quarter			1.83*** (1.96)	1.89*** (1.98)	1.94*** (1.99)	1.89*** (1.98)
Precip of warmest quarter			1.53 (1.76)	1.71* (1.90)	1.68 (1.88)	1.71* (1.90)
Year	6.72*** (8.00)	6.90*** (8.00)	7.27*** (8.00)	6.45*** (8.00)	6.53*** (8.00)	6.44*** (8.00)
Region	13.18*** (13.89)	13.17*** (13.90)	12.07*** (13.57)	12.19*** (13.52)	12.32*** (13.59)	12.13*** (13.47)
AIC	2251.33	2340.28	2270.07	2207.77	2212.08	2204.83
Deviance	978.13	1242.10	1071.46	871.99	886.18	859.68
Deviance explained	0.61	0.47	0.57	0.66	0.66	0.67
Dispersion	3.00	3.57	3.17	2.76	2.78	2.73
R <sup>2</sup>	0.47	0.38	0.40	0.52	0.51	0.51
Num. obs.	288	288	288	288	288	288
Num. smooth terms	8	4	4	11	10	12

\*\*\* $p < 0.01$ ; \*\* $p < 0.05$ ; \* $p < 0.1$

Table 2.5: Final Regression Models Mexico:EDF value is reported as the coefficient, and DF is included in parentheses (not standard errors because a chi-square test is used for the smooth terms).

shown to predict dengue at a local level and tested whether the association could be generalised to the regional or state level. In addition, it showed the potential development of more sophisticated socio-economic indicators using regional and internationally available data. The study identified which regions are most at risk, by estimating where dengue vectors are likely to occur given their suitability to climate conditions in terms of receptivity and vulnerability. By estimating the chance of a vector occurring in a region, we could then assess the impact of socio-economic, demographic and climate factors on the incidence of dengue. The results confirmed a strong association between our novel indices of socio-economic factors and dengue cases per region. Such results are consistent with the findings reported by [15, 17, 43–46, 49, 62]. Two main lessons can be drawn from this study: first, while higher GDP is generally associated with a drop in the incidence of dengue, a more granular analysis revealed that the crucial factors are a rise in education (with fewer jobs in the primary sector) and better access to information or technological infrastructure. For this reason, the use of more sophisticated measures, aside from GDP, should be taken into account when building models that try to predict disease distribution. The use of more granular socio-economic indicators can explain with greater accuracy the differences in the spread of disease in places with similar physical geography and ecological characteristics. In addition, public health authorities should be aware of the presence of non-linearities in relationships between dengue and income. Secondly, factors that were shown to have an impact of dengue at the local level are also good predictors at the regional level. Given that data for these indicators are available at a sub-national scale for OECD countries and selected OECD non-member

economies, these indices may help us better understand factors responsible for the global distribution of dengue and also, given a warming climate, may help us to better predict vulnerable populations. Although the variables used in this study do not represent disease transmission mechanisms directly, understanding the relative impact of socio-economic, demographic and climate factors on disease outcomes can help risk assessors predict where diseases are likely to occur in the future, by identifying locations with vulnerabilities in public health systems and/or by identifying impoverished areas that tend to be susceptible to disease. Our findings are not only useful for public health, but also contribute to a wider scholarly debate on whether and to what extent can economic growth (measured via GDP) contribute to better outcomes of health and well-being. Finally, it is important to note that, with any analysis dealing with regional data, results should be taken with caution because of issues of scale and uncertainty introduced by the aggregation procedure. Further studies seeking to test the robustness of the indicators examined in this study should try to source data at a more refined scale, and test how these indicators can generalise across the different scales.

## Bibliography

1. Murray NEA, Quam MB, Wilder-Smith A. Epidemiology of dengue: Past, present and future prospects. *Clinical epidemiology*. 2013;5:299. <https://www.dovepress.com/getfile.php?fileID=17199>.
2. Anosike JC, Nwoke BE, Okere AN, Oku EE, Asor JE, Emmy-Egbe IO, et al. Epidemiology of tree-hole breeding mosquitoes in the tropical rainforest of imo state, south-east nigeria. *Ann Agric Environ Med*. 2007;14:31–8.
3. Philbert A. Preferred breeding habitats of aedes aegypti (diptera- culicidae) mosquito and its public health implications in dares salaam, tanzania anitha philbert1\* and jasper. N. Ijumba2. Book. 2013.
4. Gomez-Dantes H, Ramsey Willoquet J. Dengue in the americas: Challenges for prevention and control. *Cadernos De Saude Publica*. 2009;25:S19–31.
5. WHO. Dengue and severe dengue. 2020. <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>.
6. Bernardes Terzian AC, Mondini A, Moraes Bronzoni RV de, Drumond BP, Ferro BP, Sotello Cabrera EM, et al. Detection of saint louis encephalitis virus in dengue-suspected cases during a dengue 3 outbreak. *Vector-Borne and Zoonotic Diseases*. 2011;11:291–300.
7. WHO. Immunization, vaccines and biologicals. 2017;2018. [http://www.who.int/immunization/research/development/dengue\\_vaccines/en/](http://www.who.int/immunization/research/development/dengue_vaccines/en/).
8. Butterworth MK, Morin CW, Comrie AC. An analysis of the potential impact of climate change on dengue transmission in the southeastern united states. *Environmental health perspectives*. 2017;125:579–85.
9. Ryan SJ, Carlson CJ, Mordecai EA, Johnson LR. Global expansion and redistribution of aedes-borne virus transmission risk with climate change. *PLOS Neglected Tropical Diseases*. 2019;13:e0007213.
10. Messina JP, Brady OJ, Golding N, Kraemer MUG, Wint GRW, Ray SE, et al. The current and future global distribution and population at risk of dengue. *Nature Microbiology*. 2019;4:1508–15.
11. Xu Z, Bambrick H, Frentiu FD, Devine G, Yakob L, Williams G, et al. Projecting the future of dengue under climate change scenarios: Progress, uncertainties and research needs. *PLOS Neglected Tropical Diseases*. 2020;14:e0008118.
12. Ebi KL, Nealon J. Dengue in a changing climate. *Environmental Research*. 2016;151:115–23.
13. Bouzid M, Colon-Gonzalez FJ, Lung T, Lake IR, Hunter PR. Climate change and the emergence of vector-borne diseases in europe: Case study of dengue fever. *Bmc Public Health*. 2014;14.
14. Messina JP, Brady OJ, Golding N, Kraemer MUG, Wint GRW, Ray SE, et al. The current and future global distribution and population at risk of dengue. *Nature Microbiology*. 2019;4:1508–15.
15. Brunkard JM, Lopez JLR, Ramirez J, Cifuentes E, Rothenberg SJ, Hunsperger EA, et al. Dengue fever seroprevalence and risk factors, texas-mexico border, 2004. *Emerging Infectious Diseases*. 2007;13:1477–83.
16. Abelz A, Smith B, Fournier M, Betz T, Gaul L, Robles-Lopez JL, et al. Dengue hemorrhagic fever - us-mexico border, 2005. *Morbidity and Mortality Weekly Report*. 2007;56:785–9.
17. Ramos EF. Hemoterapia e febre dengue. *Revista Brasileira de Hematologia e Hemoterapia*. 2008;30:64–6.

18. Magori K, Drake JM. The population dynamics of vector-borne diseases. Book. 2013.
19. Vincenti-Gonzalez MF, Grillet ME, Velasco-Salas ZI, Lizarazo EF, Amarista MA, Sierra GM, et al. Spatial analysis of dengue seroprevalence and modeling of transmission risk factors in a dengue hyperendemic city of venezuela. *Plos Neglected Tropical Diseases*. 2017;11.
20. Toan DTT, Hoat LN, Hu W, Wright P, Martens P. Risk factors associated with an outbreak of dengue fever/dengue haemorrhagic fever in hanoi, vietnam. *Epidemiology and Infection*. 2015;143:1594–8.
21. Tipayamongkholgul M, Lisakulruk S. Socio-geographical factors in vulnerability to dengue in thai villages: A spatial regression analysis. *Geospat Health*. 2011;5.
22. Teurlai M, Menkès CE, Cavarero V, Degallier N, Descloux E, Grangeon J-P, et al. Socio-economic and climate factors associated with dengue fever spatial heterogeneity: A worked example in new caledonia. *PLoS Neglected Tropical Diseases*. 2015;9:e0004211.
23. Akter R, Naish S, Hu W, Tong S. Socio-demographic, ecological factors and dengue infection trends in australia. *PLoS One*. 2017;12:e0185551.
24. Robert C, Kubiszewski I, Giovannini E, Lovins H, McGlade J, Pickett K, et al. Time to leave gdp behind. *Nature*. 2014;505.
25. Stiglitz JE, Sen A, Fitoussi J-P. Mismeasuring our lives: Why gdp doesn't add up. Book. The New Press; 2010.
26. Bleys B. Beyond gdp: Classifying alternative measures for progress. *Social Indicators Research*. 2012;109:355–76.
27. Van den Bergh JC. The gdp paradox. *Journal of Economic Psychology*. 2009;30:117–35.
28. Costanza R, Kubiszewski I, Giovannini E, Lovins H, McGlade J, Pickett KE, et al. Development: Time to leave gdp behind. *Nature News*. 2014;505:283.
29. Navarro V. Assessment of the world health report 2000. *The Lancet*. 2000;356:1598–601.
30. Berkman LF, Kawachi I, Glymour MM. Social epidemiology. Book. Oxford University Press; 2014.
31. WHO. A framework for malaria elimination. Book. World Health Organization; 2017.
32. Brady OJ, Johansson MA, Guerra CA, Bhatt S, Golding N, Pigott DM, et al. Modelling adult aedes aegypti and aedes albopictus survival at different temperatures in laboratory and field settings. *Parasites & Vectors*. 2013;6:351.
33. Gubler DJ. Dengue, urbanization and globalization: The unholy trinity of the 21(st) century. *Tropical Medicine and Health*. 2011;39 4 Suppl:3–11.
34. Gubler D. Prevention and control of aedes aegypti-borne diseases: Lesson learned from past successes and failures. Book. 2013.
35. Lana RM, Costa Gomes MF da, Melo de Lima TF, Honorio NA, Codeco CT. The introduction of dengue follows transportation infrastructure changes in the state of acre, brazil: A network-based analysis. *Plos Neglected Tropical Diseases*. 2017;11.
36. Lana RM, Gomes MF da C, Lima TFM de, Honório NA, Codeço CT. The introduction of dengue follows transportation infrastructure changes in the state of acre, brazil: A network-based analysis. *PLOS Neglected Tropical Diseases*.

2017;11:e0006070.

37. Begon M. Ecological epidemiology. In: The princeton guide to ecology. Princeton University Press; 2009. <http://www.jstor.org/stable/j.ctt7t14n>.

38. Moore CG, Cline BL, Ruiz-Tibén E, Lee D, Romney-Joseph H, Rivera-Correa E. *Aedes aegypti* in puerto rico: Environmental determinants of larval abundance and relation to dengue virus transmission. *The American Journal of Tropical Medicine and Hygiene*. 1978;27:1225–31.

39. Mohammed A, Chadee DD. Effects of different temperature regimens on the development of *aedes aegypti* (l.)(Diptera: Culicidae) mosquitoes. *Acta tropica*. 2011;119:38–43.

40. Tun-Lin W, Lenhart A, Nam VS, Rebollar-Tellez E, Morrison AC, Barbazan P, et al. Reducing costs and operational constraints of dengue vector control by targeting productive breeding places: A multi-country non-inferiority cluster randomized trial. *Tropical Medicine & International Health*. 2009;14:1143–53.

41. Watts DM, Burke DS, Harrison BA, Whitmire RE, Nisalak A. Effect of temperature on the vector efficiency of *aedes aegypti* for dengue 2 virus. *The American journal of tropical medicine and hygiene*. 1987;36:143–52.

42. Thammapalo S, Chongsuwiatwong V, McNeil D, Geater A. The climatic factors influencing the occurrence of dengue hemorrhagic fever in thailand. *South-east Asian J Trop Med Public Health*. 2005;36.

43. Stewart Ibarra AM, Ryan SJ, Beltrán E, Mejía R, Silva M, Muñoz Á. Dengue vector dynamics (*aedes aegypti*) influenced by climate and social factors in ecuador: Implications for targeted control. *PLOS ONE*. 2013;8:e78263.

44. Thammapalo S, Chongsuwiatwong V, Geater A, Lim A, Choomalee K. Socio-demographic and environmental factors associated with *aedes* breeding places in phuket, thailand. *Southeast Asian J Trop Med Public Health*. 2005;36:426–33.

45. Qi X, Wang Y, Li Y, Meng Y, Chen Q, Ma J, et al. The effects of socioeconomic and environmental factors on the incidence of dengue fever in the pearl river delta, china, 2013. *PLOS Neglected Tropical Diseases*. 2015;9:e0004159.

46. Clark GG. Dengue and dengue hemorrhagic fever in northern mexico and south texas: Do they really respect the border? *Am J Trop Med Hyg*. 2008;78:361–2.

47. Khun S, Manderson L. Health seeking and access to care for children with suspected dengue in cambodia: An ethnographic study. *BMC Public Health*. 2007;7:262.

48. Unlu I, Farajollahi A, Strickman D, Fonseca DM. Crouching tiger, hidden trouble: Urban sources of *aedes albopictus* (diptera: Culicidae) refractory to source-reduction. *PloS one*. 2013;8:e77999–9.

49. Elsinga J, Lizarazo EF, Vincenti MF, Schmidt M, Velasco-Salas ZI, Arias L, et al. Health seeking behaviour and treatment intentions of dengue and fever: A household survey of children and adults in venezuela. *PLOS Neglected Tropical Diseases*. 2015;9:e0004237.

50. ACAPS. ACAPS briefing note: Mexico - dengue fever (16 september 2019). 2019. <https://reliefweb.int/report/mexico/acaps-briefing-note-mexico-dengue-fever-16>

51. Kraemer MUG, Sinka ME, Duda KA, Mylne A, Shearer FM, Brady OJ, et al. The global compendium of *aedes aegypti* and *ae. Albopictus* occurrence. *Scientific Data*. 2015;2:150035.

52. C. Vega G, Pertierra LR, Olalla-Tárraga MÁ. MERRAclim, a high-resolution

global dataset of remotely sensed bioclimatic variables for ecological modelling. *Scientific Data*. 2017;4:170078. doi:10.1038/sdata.2017.78 <https://www.nature.com/articles/sdata201778> information.

53. CPC/NCEP. National center for atmospheric research. 1987. <http://rda.ucar.edu/datasets/ds512.0/>.

54. FAO-UN. Global administrative unit layers (gaul). 2014. <http://www.fao.org/geonetwork/srv/en/metadata.show?id=12691>.

55. Healthmap. Dengue case reports 2011-2017. 2017. <http://www.healthmap.org/en/>.

56. OECD. Regional statistics and indicators database. 2018. [http://stats.oecd.org/Index.aspx?DataSetCode=REGION\\_DEMOGR](http://stats.oecd.org/Index.aspx?DataSetCode=REGION_DEMOGR).

57. SEDAC. Gridded population of the world, version 4 (gpwv4): Population count, revision 11. 2018. <https://doi.org/10.7927/H4JW8BX5>.

58. GFC. Spatial data analysis and modeling with r. 2018;2018. <http://rspatial.org/index.html>.

59. Aswi A, Cramb SM, Moraga P, Mengersen K. Bayesian spatial and spatio-temporal approaches to modelling dengue fever: A systematic review. *Epidemiology and infection*. 2018;147:1–14.

60. Wood SN. Generalized additive models: An introduction with r. Book. CRC press; 2017.

61. Gomez-Dantes H. Dengue in the americas. A problem of regional health. *Salud publica de Mexico*. 1991;33:347–55.

62. Brunkard JM, Cifuentes E, Rothenberg SJ. Assessing the roles of temperature, precipitation, and el nino in dengue re-emergence on the texas-mexico border region. *Salud Publica De Mexico*. 2008;50:227–34.

63. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of internet-based dengue query data: Google dengue trends. *Plos Neglected Tropical Diseases*. 2014;8.

64. Romero-Alvarez D, Parikh N, Osthus D, Martinez K, Generous N, Valle S del, et al. Google health trends performance reflecting dengue incidence for the brazilian states. *BMC Infectious Diseases*. 2020;20:252.

65. Nakano K. Future risk of dengue fever to workforce and industry through global supply chain. Mitigation and adaptation strategies for global change. 2018;23:433–49.

66. Jakobsen F, Nguyen-Tien T, Pham-Thanh L, Bui VN, Nguyen-Viet H, Tran-Hai S, et al. Urban livestock-keeping and dengue in urban and peri-urban hanoi, vietnam. *PLOS Neglected Tropical Diseases*. 2019;13:e0007774.

67. Cabrera M, Taylor G. Modelling spatio-temporal data of dengue fever using generalized additive mixed models. *Spatial and Spatio-temporal Epidemiology*. 2019;28:1–13.



## Chapter 3

**The rise of West Nile Virus in Southern and Southeastern Europe: A spatial–temporal analysis investigating the combined effects of climate, land use and economic changes**



# The rise of West Nile Virus in Southern and Southeastern Europe: A spatial–temporal analysis investigating the combined effects of climate, land use and economic changes

Matthew J. Watts<sup>a,\*</sup>, Victor Sarto i Monteys<sup>a,b</sup>, P. Graham Mortyn<sup>a,d</sup>, Panagiota Kotsila<sup>a,c</sup>

<sup>a</sup> Institute of Environmental Science and Technology (ICTA), Autonomous University of Barcelona (UAB), Bellaterra, Spain

<sup>b</sup> Departament d'Agricultura, Ramaderia, Pesca, Alimentació i Medi Natural, Generalitat de Catalunya, Avinguda Meridiana, Barcelona, Spain

<sup>c</sup> Barcelona Laboratory for Urban Environmental Justice and Sustainability (BCNEJ), Institute of Environmental Science and Technology (ICTA), Autonomous University of Barcelona (UAB), Bellaterra, Spain

<sup>d</sup> Department of Geography, Autonomous University of Barcelona (UAB), Bellaterra, Spain

## ARTICLE INFO

### Keywords:

West-Nile-virus  
Climate-change  
Economic-crisis  
Mosquito  
Austerity  
Vector-borne-disease  
Drought

## ABSTRACT

West Nile Virus (WNV) has recently emerged as a major public health concern in Europe; its recent expansion also coincided with some remarkable socio-economic and environmental changes, including an economic crisis and some of the warmest temperatures on record. Here we empirically investigate the drivers of this phenomenon at a European wide scale by constructing and analyzing a unique spatial–temporal data-set, that includes data on climate, land-use, the economy, and government spending on environmental related sectors. Drivers and risk factors of WNV were identified by building a conceptual framework, and relationships were tested using a Generalized Additive Model (GAM), which could capture complex non-linear relationships and also account for spatial and temporal auto-correlation. Some of the key risk factors identified in our conceptual framework, such as a higher percentage of wetlands and arable land, climate factors (higher summer rainfall and higher summer temperatures) were positive predictors of WNV infections. Interestingly, winter temperatures of between 2 °C and 6 °C were among some of the strongest predictors of annual WNV infections; one possible explanation for this result is that successful overwintering of infected adult mosquitoes (likely *Culex pipiens*) is key to the intensity of outbreaks for a given year. Furthermore, lower surface water extent over the summer is also associated with more intense outbreaks, suggesting that drought, which is known to induce positive changes in WNV prevalence in mosquitoes, is also contributing to the upward trend in WNV cases in affected regions. Our indicators representing the economic crisis were also strong predictors of WNV infections, suggesting there is an association between austerity and cuts to key sectors, which could have benefited vector species and the virus during this crucial period. These results, taken in the context of recent winter warming due to climate change, and more frequent droughts, may offer an explanation of why the virus has become so prevalent in Europe.

## 1. Introduction

Over the past few decades, new health risks have been emerging in Europe, particularly with the recent appearance of vector borne diseases (VBDs) such as Chikungunya, West Nile Virus (WNV), Dengue (DENV-1) and Crimean-Congo haemorrhagic fever [1–3]. Rising temperatures are likely increasing the transmission potential of tropical VBDs in Europe, by affecting the geographic spread, abundance, survival and feeding activity of vector species and benefiting pathogen development in infected vectors [4–9]. This, combined with other factors such as human

population growth, intensive animal rearing, global commerce, air travel, urbanization and land-use changes, is increasing the chances of novel diseases to enter and emerge in Europe [10–13].

In this study, we focus on WNV, a single-stranded RNA *Flavivirus* closely related to other *Flaviviridae* pathogens such as dengue, Japanese encephalitis and yellow fever viruses [14]. Although WNV is a zoonotic pathogen, infecting mammals, particularly humans and horses, the transmission cycle is believed to be driven mainly by mosquitoes and birds [15], although some wild mammals may serve as intermediate hosts for West Nile virus [16].

\* Corresponding author.

E-mail address: [matthewjohnwatts@googlemail.com](mailto:matthewjohnwatts@googlemail.com) (M.J. Watts).

<https://doi.org/10.1016/j.onehlt.2021.100315>

Received 15 June 2021; Received in revised form 18 August 2021; Accepted 18 August 2021

Available online 24 August 2021

2352-7714/© 2021 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 3.1 Introduction

Over the past few decades, new health risks have been emerging in Europe, particularly with the recent appearance of vector borne diseases (VBDs) such as Chikungunya, West Nile Virus (WNV), Dengue (DENV-1) and Crimean-Congo haemorrhagic fever [1, 2, 3]. Rising temperatures are likely increasing the transmission potential of tropical VBDs in Europe, by affecting the geographic spread, abundance, survival and feeding activity of vector species and benefiting pathogen development in infected vectors [4, 5, 6, 7, 8, 9]. This, combined with other factors such as human population growth, intensive animal rearing, global commerce, air travel, urbanization and land-use changes, is increasing the chances of novel diseases to enter and emerge in Europe [10, 11, 12, 13].

In this study, we focus on WNV, a single-stranded RNA *Flavivirus* closely related to other *Flaviviridae* pathogens such as dengue, Japanese encephalitis and yellow fever viruses [14]. Although WNV is a zoonotic pathogen, infecting mammals, particularly humans and horses, the transmission cycle is believed to be driven mainly by mosquitoes and birds [15], although some wild mammals may serve as intermediate hosts for West Nile virus [16].

West Nile Virus (WNV) has recently emerged as a major public health concern in Europe, its recent expansion also coincided with some remarkable socio-economic and environmental changes, including an economic crisis and some of the warmest temperatures on record. To date, there has been very little research investigating this phenomenon at a European wide scale and more work is required to reveal the key drivers of the disease. A better understanding of this phenomenon can help public health officials design health prevention measures and develop better predictive models for public health risk management. More specifically, little work has been done to explore the association between the rise of WNV in Europe and the economic crisis unfolding from 2008 on-wards. Although the study of physical factors is key in understanding disease transmission and distribution, few articles have considered links among societal factors, like changes in the economy and in policy making [17, 18], which we know can have wide and unintended effects on natural ecosystems and, eventually, on disease [19]. Although examining such factors presents certain challenges and uncertainties, given the scales involved and lack of data, the continuation of abrupt socio-economic changes (brought by the COVID-19 pandemic) and climate change impacts indicate an urgent need to examine such statistical relationships more closely, at the very least to open up scholarly debate and instigate further research on the topic.

Since 2010, WNV has been reported in 14 EU countries including Austria, Bulgaria, Croatia, Cyprus, Czechia, France, Greece, Hungary, Italy, the Netherlands, Portugal, Romania, Slovenia and Spain; and has also been reported in five neighbouring EU candidate countries including Albania, Montenegro, Serbia, Turkey and Kosovo [20, 21]. In 2010, major outbreaks hit Greece, Hungary, Romania, and Turkey. Since then, outbreaks have occurred annually in multiple regions, including more northerly regions that had not previously reported cases, like Germany and the Czech Republic (see Figure 3.1). This culminated in another major outbreak in 2018, that affected more regions than had been recorded in previous years.

Generally, WNV distribution is determined by the presence of suitable mosquito vectors and avian hosts, such as terrestrial and wetland birds. The spring migration

of birds from infected regions of Sub-Saharan Africa to temperate regions of Europe is considered to be one of the main drivers of the disease in Europe [22, 15]. In Europe, the main vectors are *Culex pipiens*, *Culex modestus*, and *Coquilleltidia richiardi*, although *Aedes* species can also transmit the disease [15] (see [23, 24, 25] for vector distribution maps).

Although WNV infections are more common of late, sporadic outbreaks have occurred in humans and equines in southern and eastern European countries over the last century; the majority of which have occurred in wetland areas and densely inhabited urban areas [15].

To examine the recent rise of WNV infections in Europe in more depth, we empirically investigate the combined effect of three sets of factors: (1) Climate/environmental factors including temperature, rainfall and surface water; (2) land-use factors including continuous / discontinuous urban fabric which represents physical characteristics of urban areas e.g. densely populated areas like cities or less dense areas like villages, regional coverage of wetlands and arable land; (3) socio-economic factors that capture the associations of the economic crisis and are proxied by GDP growth, central government spending on areas of the environment including agriculture, forestry and fisheries and waste water management. Our analysis focuses on regions in the 7 European countries where WNV has been regularly reported - Austria, Bulgaria, Croatia, Greece, Hungary Italy, and Romania. The time series data set captures the time period before and after the economic crisis (2007-2019).

### 3.1.1 Conceptual framework

WNV transmission requires the presence of competent vectors, a suitable climate and a susceptible host population. Studying WNV transmission at a macro scale presents significant challenges, since key data on the seasonal and annual abundance of competent mosquitoes and birds are not available; this is further complicated given the hundreds of potential host bird species in Europe [26]. To explain human infections, we therefore use environmental risk factors known to attract vector, and host. Furthermore, vector abundance for a given season is modulated by physical and environmental factors, such as temperature, rainfall and water resource availability [15, 27, 14]. We therefore use proxies that can predict mosquito abundance.

#### Modulating Factors

Typically, with most tropical and temperate mosquito species, elevated temperatures allow vector populations to increase their growth and reproduction rates, which in turn decreases blood meals intervals, accelerating transmission and virus evolution rates [28]. Furthermore, increasingly warmer winters allow mosquito vectors to expand their breeding seasons and survive during winter, either as eggs or as overwintering female mosquitoes. Weather conditions and climatic factors can also affect vector competence [14]. Viral replication rates and transmission of WNV are modulated by ambient temperature, affecting the length of the extrinsic incubation period (EIP), seasonal phenology of mosquito host populations and also times at which humans and mosquitoes come into contact [29]. Generally, higher rainfall in warmer weather can lead to higher mosquito abundance and disease transmission by increasing the potential habitat suitable for mosquito reproduction, e.g. standing

water [30, 15, 31]. Conversely, sometimes drought and shrinking water resources can also bring some species into closer contact, facilitating transmission and amplification of WNV within these locations [14]. To represent these points in our data-set / model, we selected mean winter temperature, mean summer temperature, number of days of rainfall in summer and summer surface water extent (a count of the number of satellite surface water observations per region represented as pixels in a geographic raster layer).

### Risk factors

West Nile virus circulation in Europe is usually confined to two different cycles and ecosystems: the sylvatic and the urban synanthropic cycle. Rural locations, including river deltas and floodplain areas, help create a sylvatic cycle, where wild, usually nesting wetland birds and ornithophilic mosquitoes *Culex pipiens*, *Culex modestus*, *Coquillettidia richiardii*) create the conditions for maintaining WNV transmission. In urban synanthropic cycles, mosquitoes, such as *Culex pipiens* or *Culex modestus*, feed on domestic birds and humans. However, these two cycles can overlap, so areas with wetlands close to human populations can be particularly vulnerable to the disease [15]. Irrigation from agriculture is also heavily linked to a greater incidence of human and veterinary WNV infections [32]. In order to represent these factors in our data set and final models, we selected land use variables (% cover) representing urban areas (metro areas), semi urban areas (lower density human settlements), wetlands and arable land.

### Economic crisis

We would expect the repercussions of an economic crisis to affect WNV transmission in several ways, at the individual level (bottom-up) and government level (top-down). Previous studies have shown that socio-economic factors tend to influence the distribution and intensity of mosquito-borne diseases both pre-infection and post-infection [33]. Poorer communities are less likely to have air-conditioned homes, tap water and adequate drainage, and therefore may be more exposed to biting mosquitoes. Several studies have demonstrated the link between WNV infections and a range of local-level socio-economic and demographic factors such as income, sanitation, and population density [22, 34, 35]. In general, we would expect to see a drop in living standards in regions experiencing an economic shock followed by sluggish economic growth. Those people most affected would find it more difficult to prevent mosquito infections through direct measures i.e. sprays and repellents and less likely to pay for things that indirectly influence WNV transmission, like the upkeep of homes and use of air conditioning. Factors associated with higher economic status can also bring humans into closer contact with mosquitoes, for example, home owners with gardens and potted plants, swimming pools and ponds or having good access to recreational space where mosquitoes can breed [36, 37]. However, neglect of such things through economic decline can have further unintended effects, even in wealthy neighborhoods [38, 39, 40].

Mosquito control (mosquito abatement) is regarded as an effective way to reduce the incidence of WNV in humans [14]. It is well documented, for example,

that during the European debt crisis, the Greek government cut mosquito abatement budgets, which may have led to a rise in vector borne disease outbreaks such as malaria and WNV [41]. WNV transmission is most likely to occur in places that favor the larval development of *Culex pipiens*, such as poorly drained low-lying areas, urban storm-water catch basins and manhole chambers, roadside ditches, sewage treatment lagoons, and man-made containers around houses, or other aquatic environments where mosquitoes deposit their eggs [14]. Additionally, during periods of austerity, governments can neglect hazard prevention efforts, such as spending on flood defences, as well as essential works like sanitation and up-keep and improvement of infrastructure [42]. Such degradation can lead to the creation of mosquito habitats [14, 15, 42]. Another critical component of preventing disease transmission is through public education programs and health promotion, educating the public on measures which can prevent being bitten can reduce risk of exposure. In general, we would expect to see a general deterioration in a government ability to run such programs during crisis and austerity. Other consequences of austerity can be expected in decreased disease detection because of cuts in public health services, prolonged periods between initial infection and treatment seeking due to dysfunctional healthcare systems, and reduced treatment of disease, all of which can lead to more intense outbreaks [43].

In order to represent the economic crisis in our model, we selected regional GDP and central government spending on healthcare; agriculture, forest and fisheries management; and wastewater management. Rather than using actual annual values, or year on year growth, we look at increases or decreases in growth using 2007 baseline levels, just before the crisis hit Europe. As a priori, we would expect WNV incidence to be associated with negative growth or very low growth in these sectors.

## 3.2 Materials and methods

In this study, we compiled a unique spatial-temporal data-set that captures the main drivers and risk factors of WNV infections in Europe, based on findings from the conceptual framework. Since WNV infection data is only available at the European NUTS 3 level (aggregated areal health data), our empirical strategy relies on aggregated areal health data. We selected regions for the study where autochthonous virus transmission had occurred at least once over the reporting period in the selected countries. By applying this criteria, we were left with 166 regions in total for the analysis. We assumed that all regions included in the study could be influenced by migratory birds that form part of the African and European flyways [44].

### 3.2.1 Data sources

The following subsection describes data sources for the study. For an extended description of data extraction and processing techniques, see Appendix B.

All data were aggregated annually at the European NUTS 3 country subdivision level (apart from central government spending data which was sourced at the country level), to produce a yearly panel data-set. The NUTS 3 classification represents small regions with a population ranging from 150,000 to 800,000 and is part of the Nomenclature of Territorial Units for Statistics (NUTS) classification system, used to divide economic territories of the EU into three hierarchical sub categories for the

purpose of data collection and statistical analysis (see [45] for further information). WNV case data were provided on request by the European Centre for Disease Prevention and Control ([www.ecdc.europa.eu](http://www.ecdc.europa.eu)). Case data are collected weekly by EU member states and affiliates. Positive cases were confirmed by at least one of the following techniques: 1) isolating WNV or WNV nucleic acid from blood or cerebrospinal fluid (CSF); 2) inducing a WNV-specific antibody response (either IgG / IgM) in a serological test. We should also bear in mind that the actual number of cases in Europe is likely to be much higher than reported, since most people infected with WNV will not develop symptoms (are asymptomatic). Around 20% of those infected with WNV will develop West Nile fever, a flu like illness, or severe West Nile disease [14]. All cases were aggregated yearly to create the annual panel data-set. Economic data were extracted from the Eurostat database (<https://ec.europa.eu/eurostat/data/database>), which provides comparable statistics and indicators and is presented in yearly time series. To capture factors determining the economic crisis, austerity and cuts to public spending, we selected regional Gross Domestic Product (GDP); country level agriculture, forestry, fisheries spending; country level waste water spending, and country level health spending. The “Agriculture, forestry, fisheries spending” variable captures spending in rural areas that help to improve the environment and agricultural development, that can benefit agricultural workers and/or mechanise production [46]. In order to represent spending before and after the economic crisis, we created a baseline index for each variable set at 2007 levels, which represented negative or positive growth from the point just before the economic crisis hit Europe.

Population count data to predict the number of people at risk in a region were sourced from the Socio-economic Data and Applications Center’s Gridded Population of the World data set [47]. This data-set estimates the population count, consistent with national censuses and population registers.

Climate data were sourced from the E-OBS Gridded Data-set [48]. This data-set was created using a series of daily temperature and rainfall observations at meteorological stations throughout Europe.

Land use statistics i.e., “Continuous urban fabric”, “Discontinuous urban fabric”, “Wetlands (fresh water)” and “Arable land” were captured sourced from the CORINE Land Cover (CLC) database [49], which provides data on the biophysical characteristics of the Earth’s surface.

Regional surface water data was sourced using the JRC Monthly Water History, v1.2 data set [50]. This data set contains maps of the location and temporal distribution of surface water from 1984 to 2019 and provides statistics on the extent and change of water surfaces.

### 3.2.2 Final data-set

Table 3.1 provides descriptive statistics of the final data-set. We did not include “Mean temp spring (°C)” in our final data set as it was correlated with winter and summer temperature variables; we concluded that we would capture more variation using the winter and summer variables which were not highly correlated. Healthcare spending was also not included in the final analysis as it was highly correlated with GDP (see Appendix B for data and model diagnostics).

Statistic	Min	Max	Mean	St. Dev.
WNV cases	0	100	1.649	6.012
Human population	6,254	10,534,640	443,384	513,000
Mean temp winter (C)	-6.072	14.564	3.190	3.623
Mean temp spring (C)	4.092	18.684	12.433	2.157
Mean temp summer (C)	13.109	28.011	22.465	2.285
Days of rain in winter	0	68	30.156	12.546
Days of rain in spring	0	71	31.723	12.918
Days of rain in summer	0	65	26.021	14.374
Spring surface water extent Z-score (30m2)	-2.876	2.404	0.000	0.958
Summer surface water extent Z-score (30m2)	-3.301	3.080	-0.000	0.958
Continuous urban fabric % cover	0.000	45.056	1.336	6.754
Discontinuous urban fabric (% cover)	0.534	60.457	5.511	7.044
Wetlands (% cover)	0.000	25.460	0.569	2.026
Arable land (% cover)	0.000	86.307	33.893	22.172
Regional GDP growth (2007=100%)	57	217	106.752	24.587
Agri, forest + fish spending (2007=100%)	27	251	80.202	35.165
Waste water mngmnt spending (2007=100%)	5	352	93.476	53.933
Health spending (2007=100%)	60	212	114.802	33.307

Table 3.1: Summary statistics of variables selected for statistical analysis - 2007-2019

### 3.2.3 Statistical Methods

The relationship between the incidence of WNV infections (per 100,000) and the climate, land-use, and economic factors was modelled via a Generalised Additive Model (GAM), which also accounted for the spatial and temporal auto-correlation. One of the main issues with our data-set is that it does not meet some basic assumptions for statistical inference, and specifically the data are not independent and identically distributed random variables (iid). More specifically, the data-set captured repeated measurements over the same regions, and observations were not independent because of spill over effects from neighbouring regions. Therefore, spatial auto-correlation in the GAM model was approximated by a Markov random field (MRF) smoother, defined by the geographic areas and their neighbourhood structure. We used R's Spdep package [51] to create a queen neighbours list (adjacency matrix) based on regions with contiguous boundaries i.e. those sharing one or more boundary point. We used a full rank MRF, which represented roughly one coefficient for each area. The local Markov property assumes that a region is conditionally independent of all other regions unless regions share a boundary. This feature allowed us to model the correlation between geographical neighbours and smooth over contiguous spatial areas, summarising the trend of the response variable as a function of the predictors (see section 5.4.2 of [52]). In order to account for variation in the response variable over time, not attributed to the other explanatory variables in our model, we used a saturated time effect for years, where a separate effect per time point is estimated.

Since not all regions report cases every year, we fit the our main model using the Tweedie distribution, which can handle excess zeros [53]. This distribution also allowed us to model the non-negative, right-skewed integer case data as the incident



rate per 100,000.

The empirical model can then be written as:

$$E(Y) = f_1(X_{it}) + f_n(\text{Year}_t) + f_m(\text{Region}_i)$$

Where the  $f(.)$  stands for smooth functions;  $E(Y)_{it}$  is equal to the WNV infection incidence per 100,000 in region  $i$  at time  $t$ , which we assume to be Tweedie distributed;  $X_{it}$  - is a vector of economic, demographic, environmental and climate variables.  $\text{Year}_t$  is a function of the time intercept and  $\text{Region}_i$  represents neighbourhood structure of region.

We built the statistical model in a step-wise fashion using the lowest Akaike Information Criterion (AIC) to help us assess the different specifications. The AIC allows us to measure model performance accounting for model complexity and reflects how well the model fits the data.

We selected relevant variables in each specification according to their category, i.e. climate, land-use and economic. All variables were included in the final specification to ascertain the contribution of each driver, all else equal.

### 3.3 Results

Figure 3.1 characterises the climate in the study regions. As we can observe, WNV infections occur in regions with climates that can be described as “Hot-summer Mediterranean”, “Humid subtropical”, “Temperature oceanic” and “Warm-summer humid continental” or “Temperate oceanic”.

Figure 3.2 shows the WNV infection incidence rates over the study period. From 2007 to 2009, very few regions were affected by WNV, however, 2010 saw an outbreak that spread far and wide. Since 2010, the number of regions that have been affected by WNV increased. In 2018, a massive outbreak affected almost all of the regions in our study.

Table 3.2 shows the results of our statistical analysis and summarises the relevant statistics (AIC, BIC and Deviance explained and so on) to compare the different specifications. We find that our final model (Full model) has the best fit in terms of the AIC, followed by the economic model, the climate model and land use model, as shown in Table 3.2. Note that as we are not estimating a standard regression model, the figures reported should not be read as coefficients, but degrees of freedom of the smooth terms. Given that we cannot interpret the coefficients to infer the sign and magnitude of the relationship, we visualise it by plot. Figures (3.3-3.5) plot the partial effects—the relationship between a change in each of the covariates and a change in the fitted values in the full model.

### Köppen-Geiger Climate Classification

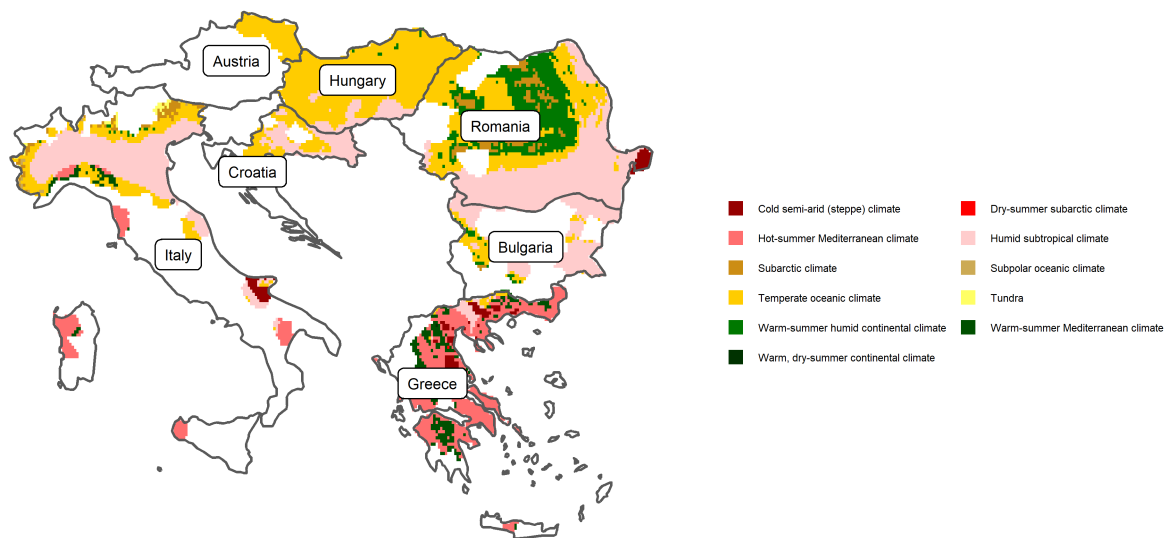


Figure 3.1: Köppen-Geiger (CG) Climate Classification in study regions. Coloured areas correspond to the overlap between the known WNV distribution and the CG classification in those areas. Areas highlighted in white represent places where human WNV infections have not been reported. (Data source: [koeppen-geiger.vu-wien.ac.at](http://koeppen-geiger.vu-wien.ac.at)).

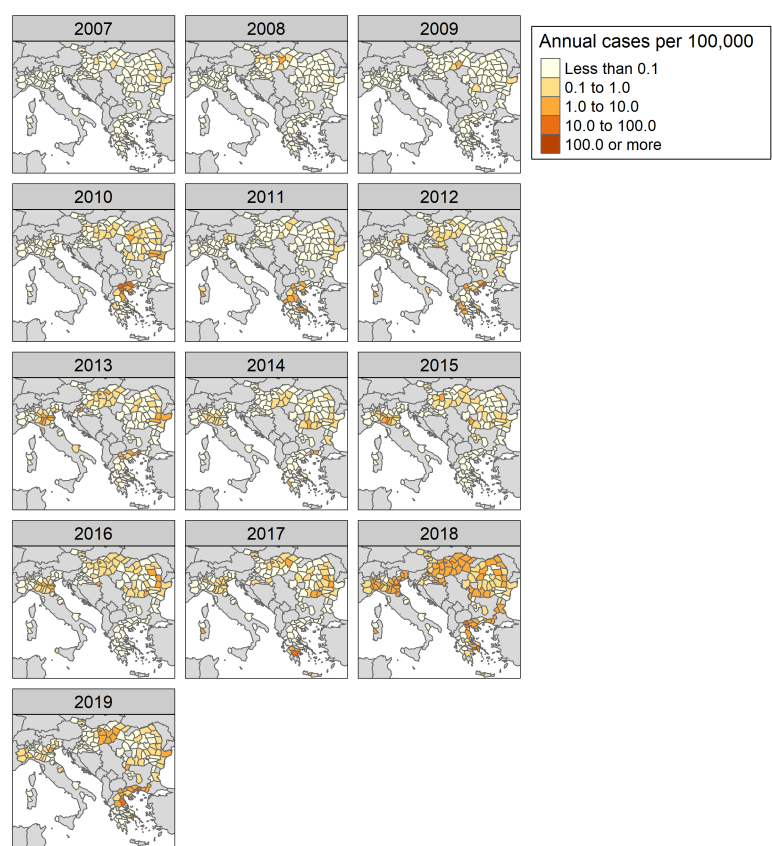


Figure 3.2: Distribution of regional West Nile virus infections per 100,000 in humans from 2006 to 2019: (Data source: ECDC).

Table 3.2: Generalized additive regression model for assessing associations between climate, land use and socio-economic factors on regional WNV incidence per 100,000 people

	Clim model	Land-use model	Econ model	Full model
Intercept	-2.21*** (0.47)	-2.13*** (0.45)	-2.25*** (0.33)	-2.35*** (0.40)
Mean temp summer (C)	1.00*** (1.00)			1.00* (1.00)
Mean temp winter (C)	1.96*** (1.99)			1.94*** (1.99)
Days of rain in summer	1.00** (1.00)			1.00** (1.00)
Summer surface water extent (30m2)	1.60** (1.84)			1.02*** (1.03)
Continuous urban fabric %		1.00 (1.00)		1.00 (1.00)
Discontinuous urban fabric %		1.00 (1.00)		1.00 (1.00)
Wetlands %		1.00** (1.00)		1.00 (1.00)
Arable land %		1.81*** (1.89)		1.74** (1.84)
Regional GDP index (2007=100%)			1.00* (1.00)	1.00 (1.00)
Agri, forest + fish spending (2007=100%)			1.95*** (1.99)	1.93*** (1.99)
Waste water management spending (2007=100%)			1.60*** (1.83)	1.10*** (1.19)
Spatial lag	78.44*** (109.60)	80.54*** (111.42)	88.90*** (121.08)	76.19*** (106.52)
Year	11.73*** (12.00)	11.75*** (12.00)	11.48*** (12.00)	11.56*** (12.00)
AIC	3952.99	3992.59	3931.25	3907.56
BIC	4526.68	4569.26	4560.01	4538.85
Log Likelihood	-1875.44	-1894.71	-1854.87	-1842.57
Deviance	3747.15	3895.52	3592.25	3520.85
Deviance explained	0.63	0.62	0.64	0.65
Dispersion	2.85	2.95	2.76	2.73
R <sup>2</sup>	0.26	0.36	0.32	0.25
GCV score	1900.89	1927.27	1889.38	1871.90
Num. obs.	2158	2158	2158	2158
Num. smooth terms	6	6	5	13

\*\*\*\* $p < 0.001$ ; \*\*\* $p < 0.01$ ; \*\* $p < 0.05$ ; \* $p < 0.1$

All climate variables in both model specifications are statistically significant (“Clim model” & “Full model”). Mean summer temperatures (Figure 3.3) of above 22°C are positively associated with WNV infections; the relationship is linear and strong in the first specification, however, becomes less significant in the “Full model” specification after controlling for all other variables.

Mean winter temperature (Figure 3.3) has a quadratic relationship with WNV. Temperatures of between 2 °C and 6°C have a positive association with WNV infections, whereas colder and warmer temperatures outside of this range are negatively associated with WNV infections. The number of rain days per summer (Figure 3.3) is also a strong predictor of WNV infections and has a linear positive relationship. Higher surface water (Figure 3.3) in the summer is negatively correlated with WNV infections. The relationship is fairly strong considering its complexity i.e., the variable complexity has been reduced to standard deviation scores to standardise it across regions and seasons.

As for the land use variables (Figure 3.4), the percentage of arable land and wetlands in a region (“Land-use model”) is positively correlated with the incidence of WNV and highly significant. However, the wetlands variable loses significance in the final model. The percentage of “Continuous” and “Discontinuous urban fabric” variables, which represent metropolitan and built up areas (residential suburbs, villages), are not statistically significant in any of the specifications. Although the partial effect plot for “Continuous urban fabric” (Figure 3.5) reveals that it has a slightly negative relationship with WNV infections and “Discontinuous urban fabric” has a positive association with WNV infections.

The economic indicators (Figure 3.5) are negatively correlated with WNV infec-

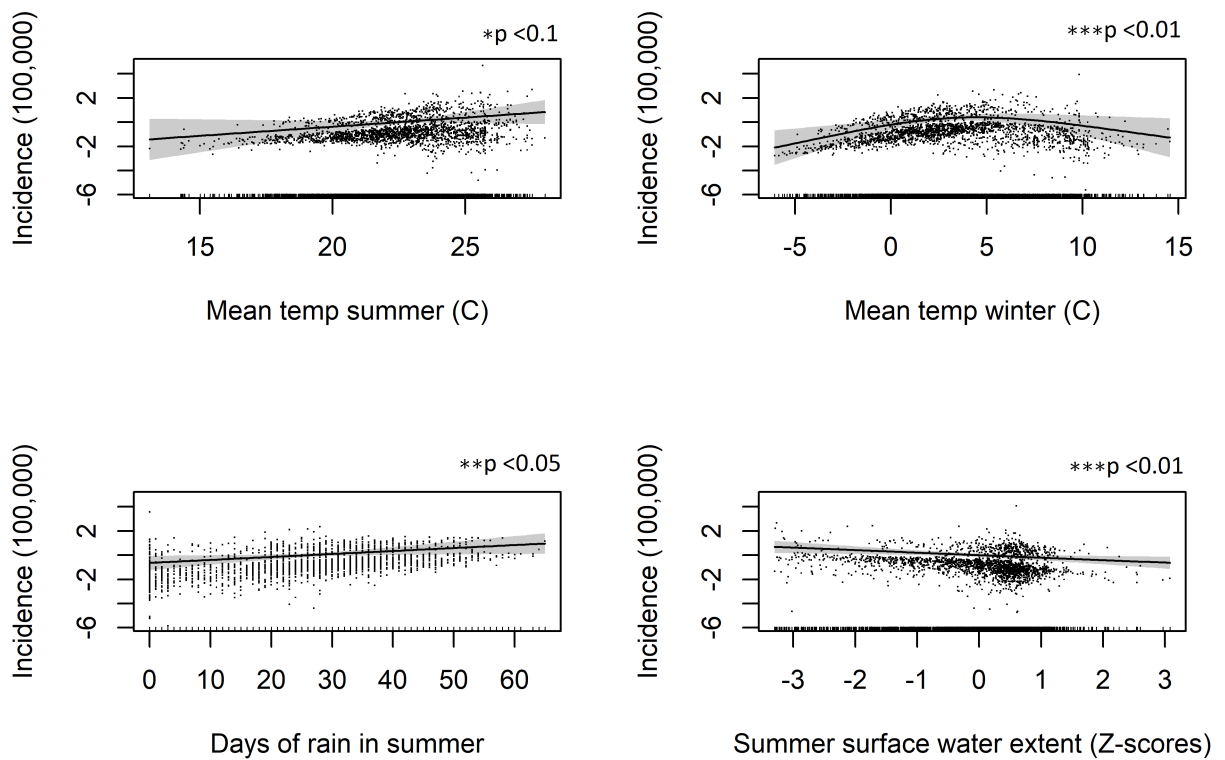


Figure 3.3: Generalised additive model (GAM) plots showing the partial effects of the explanatory variables on the incidence of WNV per 100,000. The tick marks on the x-axis are observed data points. The y-axis represents the partial effect of each variable. The dots represent partial residuals. The shaded areas indicate the 95% confidence intervals.

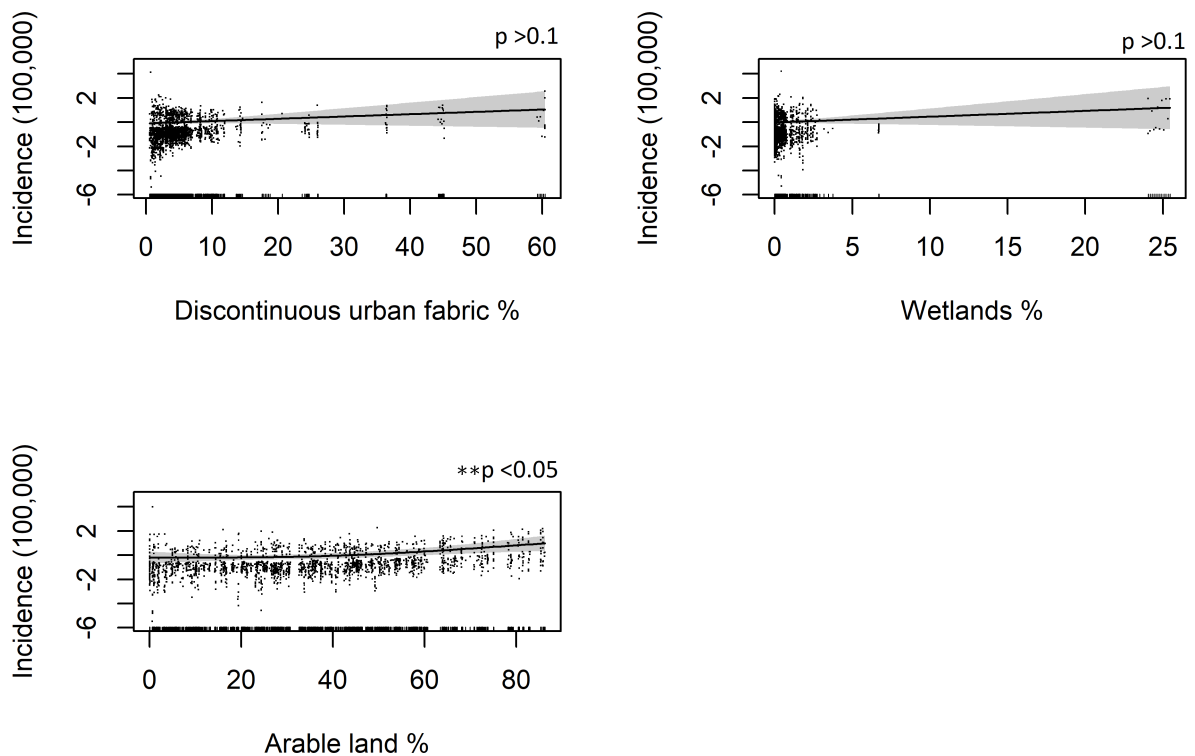


Figure 3.4: Generalised additive model (GAM) plots showing the partial effects of the explanatory variables on the incidence of WNV per 100,000. The tick marks on the x-axis are observed data points. The y-axis represents the partial effect of each variable. The dots represent partial residuals. The shaded areas indicate the 95% confidence intervals.

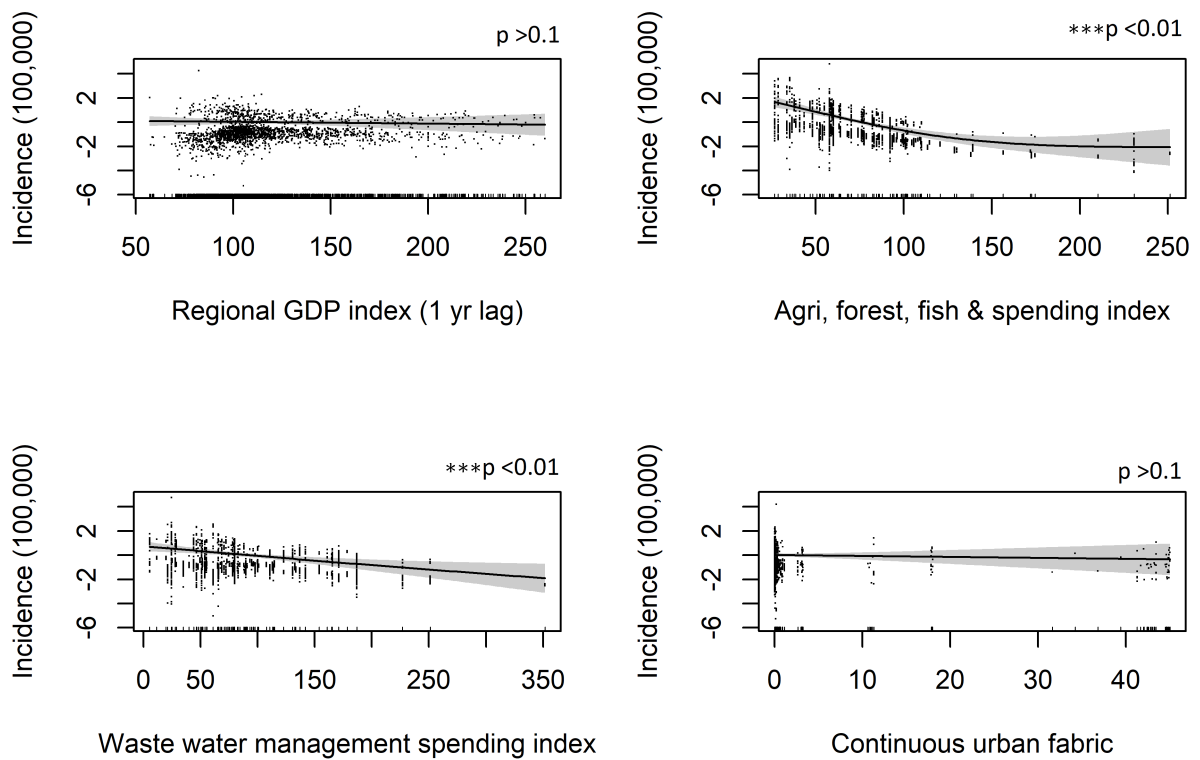


Figure 3.5: Generalised additive model (GAM) plots showing the partial effects of the explanatory variables on the incidence of WNV per 100,000. The tick marks on the x-axis are observed data points. The y-axis represents the partial effect of each variable. The dots represent partial residuals. The shaded areas indicate the 95% confidence intervals.

tions. These variables represent 2007 baseline regional GDP and central government spending growth. Regional GDP has a gentle negative association, but statistically significant relationship with WNV infections in the "Econ model specification", but loses significance in the final model. The two indicators that directly represent central government spending on areas of the environment, such as agriculture, forest & fisheries spending; and waste water management, have highly statistically significant negative associations with WNV infections.

## 3.4 Discussion

To investigate the rise in WNV infections in Europe over the last 13 years (2007-2019), we compiled a unique spatial-temporal data-set including variables identified in the conceptual framework, following a thorough review of the literature. By taking this approach, we were able to carefully evaluate and adjust for environmental and economic factors that may have contributed to the recent rise in infections over the past decade or so. This study focuses on geographical factors which tend to influence the spread of the disease at the regional level, rather than trying to infer the determinants of the disease at the individual level.

### 3.4.1 Meteorological factors

Over the past 70 years, the countries analysed in this study have been experiencing increasingly warmer temperatures throughout the year and according to our initial analysis, the last decade has been the warmest (see Appendix B: figures B1-4). The results of our final model (figures 3.3-3.5) show that average summer temperatures above 22 °C are positively associated with an increase in WNV incidence. This finding is consistent with the literature, according to which warmer temperatures influence the hatching rate and development time of mosquitoes, and shorten the extrinsic incubation period (EIP) of WNV and related viruses, therefore representing a key driver of WNV transmission (especially in summer months) [54, 55, 56, 57, 58, 29]. However, this variable lost significance in our final model specification, suggesting it is not one of the main drivers associated with transmission in our study locations and that other factors may be at play. In particular, average summer temperatures lose their explanatory power once economic and land use factors are taken into account.

Our analysis of the mean winter temperature (Dec-Feb) reveals a quadratic relationship with WNV and is one of the main predictors of annual WNV infections in the model. Temperatures below 2 °C and above 6 °C have a negative association with WNV infections. These results are consistent with findings by Koenraadt et al., 2019 [59], who found that diapausing *Culex pipiens* mosquitoes do not necessarily do better under warmer conditions and there is a temperature range in which they can successfully diapause. Furthermore, other authors suggest [60, 61, 20] that outbreaks may be more intense following winters with optimal temperatures for diapausing mosquitoes. Since a larger number of mosquitoes can successfully survive the winter, and those that are infected with WNV can transmit it earlier on in the year, leading to increased disease prevalence in mosquitoes and reservoir bird species than in years when winter conditions are not optimal. It is also important to note



that it is not currently clear at what temperatures *Culex modestus* and *Coquillettidia richiardii* overwinter as adults since our literature search did not yield any findings.

The number of rain days per summer is also positively correlated with WNV infections. This result is consistent with the literature, i.e., a steady flow of aquatic resources for mosquitoes has a positive association with their abundance, and therefore an increase in disease transmission [31, 30]. Rainfall patterns have also been shifting since the 1950s (see Appendix B: Figures B1-B4) although unlike climate, there is no a clear trend and results are more difficult to interpret. In general, Austria, Croatia, and Italy are seeing less intense rainfall in the summer months, but higher in autumn, whereas Bulgaria, Greece, Hungary and Romania are receiving more rainfall in summer months.

Our results also show that higher regional summer surface water extent for a given year, is negatively associated with WNV and is one of the strongest predictors of WNV incidence. This was not an expected finding, since we would expected higher levels of surface water to be positively correlated with WNV incidence because of the extra water resources available to mosquitoes. However, it may be explained by the fact that sometimes desiccation of water resources can bring mosquito and bird hosts closer together, increasing transmission potential and therefore the prevalence of the virus [15, 14]. This was also a major finding in a recent study by Paull et al., 2017 [62], who reported that drought was closely linked to the intensity of outbreaks for a given year in the United States. Another explanation for this result is that with higher surface water extent, there may be more flooding and fast water movement, which may wash away mosquito eggs and larvae [63], and also may inhibit contact between birds, mosquitoes and humans [15, 14]. It is important to note that this variable probably does not capture the creation of short-term water resources created by rainfall (e.g. pools, puddles), which can be used as breeding habitat by mosquitoes. It rather captures long term and large water surface such as deltas, lakes, and flood plains.

## Land-use

As for the land-use variables, as expected, regions with a larger proportion of arable land and wetlands are associated with higher WNV incidence. This is consistent with other literature, according to which humans are particularly at risk in areas close to rice paddies, irrigated agriculture and wetlands, since these areas tend to attract susceptible mosquitoes and birds [15, 32]. The percentage of discontinuous urban fabric, that represents populated areas of low to medium density that tend to have gardens, parks, ponds, such as residential suburbs and villages [64], is not statistically significant in our model, although it is often cited as a driver of WNV infections in humans.

## Economic-factors

In terms of economic factors associated with WNV infections, higher GDP growth, higher spending, growth on environmental factors - such as agriculture, forest, fisheries - and waste water management are negatively associated with WNV incidence, consistently with concepts laid out in the conceptual framework. In other words, populations living in locations harder hit by economic slowdown and austerity could have been more exposed to mosquitoes, for instance drops in income make it difficult

to afford mosquito repellents, air conditioning and upkeep of homes leading to the creation of mosquito habitats. General cuts to waste water management and hazard prevention efforts, such as spending on flood defences, essential works like sanitation and upkeep of infrastructure, could have also led to the creation of mosquito breeding habitats, e.g. potholes, blocked drains [14, 15, 42]. Furthermore, many studies report strong associations between agriculture [65, 32, 66, 67] and WNV incidence. In general, cuts and lower spending in this sector, may have led to degradation on farms and the wider environment which may have benefited mosquitoes through the creation of habitat or lack of measures to control their abundance. The literature is scarce on this topic which makes it very difficult to compare our findings with other sources of information, so our interpretations of such results can only be speculative.

## Limitations

Some of the limitations of the study are as follows. Since we were limited to using aggregated data at the NUTS-3 regional level, we cannot make inference about individual-level associations and could not adjust for individual-level risk factors e.g. age, gender, race, and occupation. However, that would be outside the scope of this study, since we were interested in macro ecological and socio-economic trends and drivers. Additionally, we cannot draw causal inference as the methodology we applied only reveals adjusted correlations. Indeed, we would have also liked to include further explanatory variables on avian host and mosquito abundance but were restricted by the availability of data. It is also important to note that data quality issues arise owing to the under-reporting of cases through under-diagnosis, lack of diagnostic tests and a lack of resources/time to carry out and implement mass testing. Another factor we did not consider is bird immunity, which may influence WNV incidence following a major outbreak, although this was not considered an important factor in explaining the rise in WNV infections in Europe, but may have influenced the results. Furthermore, a growing body of literature reports that mammals can serve as intermediate hosts for West Nile virus [16] and more research needs to be done to determine if wild mammals act as reservoirs and contribute significantly to the transmission cycle. We also realise that the economic analysis is limited, in part because of a lack of refined data and in part because of scale issues, i.e., the amount of work required to look at individual local level policies and spending was not feasible for 166 regions.

## Conclusions

In this study, we set out to investigate why WNV outbreaks have become so frequent in Europe over the past decade. If we consider the findings of this work together with other important research in this area, we can start to build a picture of why the virus has become so prevalent in Europe. We hypothesise that:

- 1) Rising winter temperatures, or rather the creation of optimal temperature conditions allowed the virus to overwinter with *Culex pipiens*. Given current trends, we can also expect to see regions that have previously been too cold for *Culex pipiens* to survive over winter become viable locations and cause further havoc in regions that are currently experiencing just a few annual cases. On the contrary, regions which currently have optimal conditions for overwintering mosquitoes may become too warm in the future.

2) Warmer summer temperatures are benefiting mosquitoes, influencing their hatching rate and development time, and shortening the extrinsic incubation period (EIP) of WNV; and

3) Shrinking water resources are increasing WNV prevalence in birds and mosquitoes during some seasons. It may be the case that this phenomenon is also acting at a macro-scale in Europe and is a significant driver of recent outbreaks, especially given that meteorological and hydrological droughts are becoming more frequent and extreme [68]. These changes also occurred during an economic crisis and subsequent austerity, where government institutions were severely weakened and had to limit spending on key sectors, and segments of the human population were exposed to increased financial hardship.

We hope this study will spur further research into this topic, especially in areas less explored, such as the impacts of the European debt crisis on health, and the long-term trade-offs and unintended consequences austerity can have on the environment and human health. This is an especially important topic when considering we are facing multiple threats brought about by global warming and other anthropogenic induced changes that can benefit emerging diseases, i.e., global trade in wild animals, intensive agriculture / animal rearing and land use conversion.

## Bibliography

- [1] P. J. Hotez. Southern europe's coming plagues: Vector-borne neglected tropical diseases. *PLoS Negl Trop Dis*, 10(6):e0004243, 2016.
- [2] Ole F. Olesen and Marit Ackermann. Increasing european support for neglected infectious disease research. *Computational and Structural Biotechnology Journal*, 15:180–184, 2017.
- [3] Guido Calleri, Andrea Angheben, and Marco Albonico. Neglected tropical diseases in europe: rare diseases and orphan drugs? *Infection*, 47(1):3–5, 11 2018.
- [4] J. A. Patz, P. R. Epstein, T. A. Burke, and J. M. Balbus. Global climate change and emerging infectious diseases. *JAMA*, 275, 1996.
- [5] Maha Bouzid, Felipe J. Colón-González, Tobias Lung, Iain R. Lake, and Paul R. Hunter. Climate change and the emergence of vector-borne diseases in europe: case study of dengue fever. *BMC Public Health*, 14(1):781, 2014.
- [6] W. J. Tabachnick. Climate change and the arboviruses: Lessons from the evolution of the dengue and yellow fever viruses. *Annual review of Virology*, 2016.
- [7] C. N. Mweya, S. I. Kimera, G. Stanley, G. Misinzo, and L. E. G. Mboera. Climate change influences potential distribution of infected aedes aegypti co-occurrence with dengue epidemics risk areas in tanzania. *Plos One*, 11(9), 2016.
- [8] R. Shope. Global climate change and infectious diseases. *Environ Health Perspect*, 96, 1991.

- [9] Amy Morrison, Helvio Astete, Claudio Rocha, Victor Lopez, Jim Olson, Tacleuz Kochel, Moises Sihuincha, and Jeff Stancil. Impact of the dengue vector control system (dvcs) on aedes aegypti populations in iquitos, peru 2004-2005. *American Journal of Tropical Medicine and Hygiene*, 73(6):326–326, 2005.
- [10] J. A. Patz, W. J. Martens, D. A. Focks, and T. H. Jetten. Dengue fever epidemic potential as projected by general circulation models of global climate change. *Environ Health Perspect*, 106, 1998.
- [11] S. Naish, P. Dale, J. S. Mackenzie, J. McBride, K. Mengersen, and S. L. Tong. Climate change and dengue: a critical and systematic review of quantitative modelling approaches. *Bmc Infectious Diseases*, 14, 2014.
- [12] Duane J. Gubler. Dengue, urbanization and globalization: The unholy trinity of the 21(st) century. *Tropical Medicine and Health*, 39(4 Suppl):3–11, 2011.
- [13] K. Magori and J. M. Drake. The population dynamics of vector-borne diseases, 2013.
- [14] Shlomit Paz and Jan C. Semenza. Environmental drivers of west nile fever epidemiology in europe and western asia—a review. *International journal of environmental research and public health*, 10(8):3543–3562, 2013.
- [15] Z. Hubálek and J. Halouzka. West nile fever a reemerging mosquito borne viral disease in europe. *Emerging infectious diseases*, 5(5):643–650, 1999.
- [16] J. Jeffrey Root and Angela M. Bosco-Lauth. West nile virus associations in wild mammals: An update. *Viruses*, 11(5):459, 05 2019.
- [17] M. Karanikolos, P. Mladovsky, J. Cylus, S. Thomson, S. Basu, D. Stuckler, J. P. Mackenbach, and M. McKee. Financial crisis, austerity, and health in europe. *Lancet*, 381(9874):1323–31, 2013.
- [18] A. Baumbach and G. Gulis. Impact of financial crisis on selected health outcomes in europe. *The European Journal of Public Health*, 24(3):399–403, 04 2014.
- [19] Sandra Crouse Quinn and Supriya Kumar. Health inequalities and infectious disease epidemics: A challenge for global health security. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, 12(5):263–273, 09 2014.
- [20] Johanna J Young, Joana M Haussig, Stephan W Aberle, Danai Pervanidou, Flavia Riccardo, Nebojša Sekulić, Tamás Bakonyi, and Céline M Gossner. Epidemiology of human west nile virus infections in the european union and european union enlargement countries, 2010 to 2018. *Eurosurveillance*, 26(19), 05 2021.
- [21] Reina S Sikkema, Maarten Schrama, Tijs van den Berg, Jolien Morren, Emmanuelle Munger, Louie Krol, Jordy G van der Beek, Rody Blom, Irina Chesnakova, Anne van der Linden, Marjan Boter, Tjomme van Mastrigt, Richard Molenkamp, Constantianus JM Koenraadt, Judith MA van den Brand, Bas B

- Oude Munnink, Marion PG Koopmans, and Henk van der Jeugd. Detection of west nile virus in a common whitethroat (*curruca communis*) and culex mosquitoes in the netherlands, 2020. *Eurosurveillance*, 25(40), 2020.
- [22] William J. Landesman, Brian F. Allan, R. Brian Langerhans, Tiffany M. Knight, and Jonathan M. Chase. Inter-annual associations between precipitation and human incidence of west nile virus in the united states. *Vector-Borne and Zoonotic Diseases*, 7(3):337–343, 2007.
- [23] ECDC. Confirmed culex modestus distribution], 2021.
- [24] ECDC. Confirmed culex pipiens distribution, 2021.
- [25] ECDC. Confirmed coquillettidia richiardii distribution, 2021.
- [26] Benoit Durand, Annelise Tran, Gilles Balança, and Véronique Chevalier. Geographic variations of the bird-borne structural risk of west nile virus circulation in europe. *PLOS ONE*, 12(10):e0185962, 10 2017.
- [27] ECDC. West nile virus infection, 2020.
- [28] R. P. Meyer, J. L. Hardy, and W. K. Reisen. Diel changes in adult mosquito microhabitat temperatures and their relationship to the extrinsic incubation of arboviruses in mosquitoes in kern county, california. *J Med Entomol*, 27(4):607–14, 1990.
- [29] W. K. Reisen, Y. Fang, and V. M. Martinez. Effects of temperature on the transmission of west nile virus by culex tarsalis (diptera: Culicidae). *J Med Entomol*, 43(2):309–17, 2006.
- [30] J. E. Soverow, G. A. Wellenius, D. N. Fisman, and M. A. Mittleman. Infectious disease in a warming world: how weather influenced west nile virus in the united states (2001-2005). *Environ Health Perspect*, 117(7):1049–52, 2009.
- [31] T. Takeda, C. A. Whitehouse, M. Brewer, A. D. Gettman, and T. N. Mather. Arbovirus surveillance in rhode island: assessing potential ecologic and climatic correlates. *J Am Mosq Control Assoc*, 19(3):179–89, 2003.
- [32] M. C. Gates and R. C. Boston. Irrigation linked to a greater incidence of human and veterinary west nile virus cases in the united states from 2004 to 2006. *Prev Vet Med*, 89(1-2):134–7, 2009.
- [33] Joan Marie Brunkard, Jose Luis Robles Lopez, Josue Ramirez, Enrique Cifuentes, Stephen J. Rothenberg, Elizabeth A. Hunsperger, Chester G. Moore, Regina M. Brussolo, Norma A. Villarreal, and Brent M. Haddad. Dengue fever seroprevalence and risk factors, texas-mexico border, 2004. *Emerging Infectious Diseases*, 13(10):1477–1483, 2007.
- [34] J. P. DeGroot and R. Sugumaran. National and regional associations between human west nile virus incidence and demographic, landscape, and land use conditions in the coterminous united states. *Vector Borne Zoonotic Dis*, 12(8):657–65, 2012.

- [35] Julie Tackett, Richard Charnigo, and Glyn Caldwell. Relating west nile virus case fatality rates to demographic and surveillance variables. *Public health reports (Washington, D.C. : 1974)*, 121(6):666–673, 2006.
- [36] Z. Dowling, S. L. Ladeau, P. Armbruster, D. Biehler, and P. T. Leisnham. Socioeconomic status affects mosquito (diptera: Culicidae) larval habitat type availability and infestation level. *J Med Entomol*, 50(4):764–72, 2013.
- [37] Isik Unlu, Ary Farajollahi, Daniel Strickman, and Dina M. Fonseca. Crouching tiger, hidden trouble: urban sources of aedes albopictus (diptera: Culicidae) refractory to source-reduction. *PloS one*, 8(10):e77999–e77999, 2013.
- [38] William K. Reisen, Richard M. Takahashi, Brian D. Carroll, and Rob Quiring. Delinquent mortgages, neglected swimming pools, and west nile virus, california. *Emerging Infectious Diseases*, 14(11):1747–1749, 11 2008.
- [39] Minh Kim, James B. Holt, Rebecca J. Eisen, Kerry Padgett, William K. Reisen, and Janet B. Croft. Detection of swimming pools by geographic object-based image analysis to support west nile virus control efforts. *Photogrammetric Engineering and Remote Sensing*, 77(11):1169–1179, 11 2011.
- [40] William K. Reisen, Brian D. Carroll, Richard Takahashi, Ying Fang, Sandra Garcia, Vincent M. Martinez, and Rob Quiring. Repeated west nile virus epidemic transmission in kern county, california, 2004–2007. *Journal of Medical Entomology*, 46(1):139–157, 01 2009.
- [41] Elias Kondilis, Stathis Giannakopoulos, Magda Gavara, Ioanna Ierodiakonou, Howard Waitzkin, and Alexis Benos. Economic crisis, restrictive policies, and the population’s health and health care: the greek case. *American journal of public health*, 103(6):973–979, 2013.
- [42] Rodrick Wallace, Luis Fernando Chaves, Luke R Bergmann, Constância Ayres, Lenny Hogerwerf, Richard Kock, and Robert G Wallace. *Clear-cutting disease control: capital-led deforestation, public health austerity, and vector-borne infection*. Springer, 2018.
- [43] Lindsay Steele, Emma Orefuwa, and Petra Dickmann. Drivers of earlier infectious disease outbreak detection: a systematic literature review. *International Journal of Infectious Diseases*, 53:15–20, 2016.
- [44] UNEP. A review of migratory bird flyways and priorities for management. *CMS Technical Series*, 2014.
- [45] Eurostat. Nuts - nomenclature of territorial units for statistics, 2020.
- [46] Eurostat. Agriculture, forestry and fishery statistics. Report, Eurostat, 2019.
- [47] SEDAC. Gridded population of the world, version 4 (gpwv4): Population count, revision 11, 2018.
- [48] Richard C. Cornes, Gerard van der Schrier, Else J. M. van den Besselaar, and Philip D. Jones. An ensemble version of the e-obs temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, 123(17):9391–9409, 2018.

- [49] EU. Copernicus land monitoring service 2018, 2018.
- [50] Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S. Belward. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633):418–422, 2016.
- [51] Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013.
- [52] Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.
- [53] Christoph F. Kurz. Tweedie distributions for fitting semicontinuous health care utilization cost data. *BMC Medical Research Methodology*, 17(1), 12 2017.
- [54] Shi Chen, Justine I. Blanford, Shelby J. Fleischer, Michael Hutchinson, Michael C. Saunders, and Matthew B. Thomas. Estimating west nile virus transmission period in pennsylvania using an optimized degree-day model. *Vector-Borne and Zoonotic Diseases*, 13(7):489–497, 07 2013.
- [55] Tsukushi Kamiya, Megan A. Greischar, Kiran Wadhawan, Benjamin Gilbert, Krijn Paaijmans, and Nicole Mideo. Temperature-dependent variation in the extrinsic incubation period elevates the risk of vector-borne disease emergence. *Epidemics*, 30:100382, 03 2020.
- [56] Azad Mohammed and Dave D Chadee. Effects of different temperature regimens on the development of aedes aegypti (l.)(diptera: Culicidae) mosquitoes. *Acta tropica*, 119(1):38–43, 2011.
- [57] W. Tun-Lin, A. Lenhart, V. S. Nam, E. Rebollar-Tellez, A. C. Morrison, P. Barbazan, M. Cote, J. Midega, F. Sanchez, P. Manrique-Saide, A. Kroeger, M. B. Nathan, F. Meheus, and M. Petzold. Reducing costs and operational constraints of dengue vector control by targeting productive breeding places: a multi-country non-inferiority cluster randomized trial. *Tropical Medicine and International Health*, 14(9):1143–1153, 2009.
- [58] Douglas M Watts, Donald S Burke, Bruce A Harrison, Richard E Whitmire, and Ananda Nisalak. Effect of temperature on the vector efficiency of aedes aegypti for dengue 2 virus. *The American journal of tropical medicine and hygiene*, 36(1):143–152, 1987.
- [59] Constantianus J. M. Koenraadt, Tim W. R. Mohlmann, Niels O. Verhulst, Jeroen Spitzen, and Chantal B. F. Vogels. Effect of overwintering on survival and vector competence of the west nile virus vector culex pipiens. *Parasites and Vectors*, 12(1):147, 2019.
- [60] Flavia Riccardo, Federica Monaco, Antonino Bella, Giovanni Savini, Francesca Russo, Roberto Cagarelli, Michele Dottori, Caterina Rizzo, Giulietta Venturi, Marco Di Luca, Simonetta Pupella, Letizia Lombardini, Patrizio Pezzotti, Patrizia Parodi, Francesco Maraglino, Alessandro Nanni Costa, Giancarlo Maria Liumbruno, Giovanni Rezza, and the working group. An early start of west nile virus seasonal transmission: the added value of one health surveillance in

- detecting early circulation and triggering timely response in Italy, June to July 2018. *Eurosurveillance*, 23(32), 2018.
- [61] I. Rudolf, L. Betasova, H. Blazejova, K. Venclíkova, P. Strakova, O. Vebesta, J. Mendel, T. Bakonyi, F. Schaffner, N. Nowotny, and Z. Hubalek. West Nile virus in overwintering mosquitoes, Central Europe. *Parasites & Vectors*, 10, 2017.
- [62] Sara H. Paull, Daniel E. Horton, Moetasim Ashfaq, Deeksha Rastogi, Laura D. Kramer, Noah S. Diffenbaugh, and A. Marm Kilpatrick. Drought and immunity determine the intensity of West Nile virus epidemics and climate change impacts. *Proceedings of the Royal Society B: Biological Sciences*, 284(1848):20162078, 02 2017.
- [63] C. J. M. Koenraadt and L. C. Harrington. Flushing effect of rain on container-inhabiting mosquitoes *Aedes aegypti* and *Culex pipiens* (Diptera: Culicidae). *Journal of Medical Entomology*, 45(1):28–35, January 2008.
- [64] ETC/ULS. Updated CLC illustrated nomenclature guidelines. Report, European Environment Agency, 2019.
- [65] David W. Crowder, Elizabeth A. Dykstra, Jo Marie Brauner, Anne Duffy, Caitlin Reed, Emily Martin, Wade Peterson, Yves Carrière, Pierre Dutilleul, and Jeb P. Owen. West Nile virus prevalence across landscapes is mediated by local effects of agriculture on vector and host communities. *PLoS ONE*, 8(1):e55006, 01 2013.
- [66] Lars Eisen, Christopher M. Barker, Chester G. Moore, W. John Pape, Anna M. Winters, and Nicholas Cheronis. Irrigated agriculture is an important risk factor for West Nile virus disease in the hyperendemic Larimer-Boulder-Weld area of North Central Colorado. *Journal of Medical Entomology*, 47(5):939–951, 09 2010.
- [67] Roque Miramontes, William E. Lafferty, Bonnie K. Lind, and Mark W. Oberle. Is agricultural activity linked to the incidence of human West Nile virus? *American Journal of Preventive Medicine*, 30(2):160–163, 02 2006.
- [68] Martin Hanel, Oldřich Rakovec, Yannis Markonis, Petr Máca, Luis Samaniego, Jan Kyselý, and Rohini Kumar. Revisiting the recent European droughts from a long-term perspective. *Scientific Reports*, 8(1), 06 2018.



## Chapter 4

# Macro-Level drivers of SARS-CoV-2 transmission: A data-driven analysis of factors contributing to epidemic growth during the first wave of outbreaks in the United States

# Macro-level drivers of SARS-CoV-2 transmission: A data-driven analysis of factors contributing to epidemic growth during the first wave of outbreaks in the United States

Matthew J Watts\*

Correspondence: matthewjohn-watts@googlemail.com  
Institute of Environmental Science and Technology (ICTA),  
Autonomous University of Barcelona (UAB), Bellaterra,  
Spain, Barcelona, Spain  
Full list of author information is available at the end of the article  
\*Equal contributor

## Abstract

**Background:** Many questions remain unanswered about how SARS-CoV-2 transmission is influenced by aspects of the economy, environment, and health. A better understanding of how these factors interact can help us to design early health prevention and control strategies, and develop better predictive models for public health risk management of SARS-CoV-2. This study examines the associations between COVID-19 epidemic growth and macro-level determinants of transmission such as demographic factors, socio-economic factors, climate and population health, during the first wave of outbreaks in the United States.

**Methods:** A spatial-temporal data-set was created from a variety of relevant data sources. A unique data-driven study design was implemented to assess the relationship between COVID-19 case and death epidemic doubling times and explanatory variables using a Generalized Additive Model (GAM).

**Results:** The main factors associated with case doubling times are higher population density, home overcrowding, manufacturing, and recreation industries. Poverty was also an important predictor of faster epidemic growth perhaps because of factors associated with in-work poverty-related conditions, although poverty is also a predictor of poor population health which is likely driving case and death reporting. Air pollution and diabetes were other important drivers of case reporting. Warmer temperatures are associated with slower epidemic growth, which is most likely explained by human behaviors associated with warmer locations i.e. ventilating homes and workplaces. and socializing outdoors. The main factors associated with death doubling times were population density, poverty older age, diabetes, and air pollution. Temperature was also slightly significant slowing death doubling times.

**Conclusions:** Such findings help underpin current understanding of the disease epidemiology and also supports current policy and advice recommending ventilation of homes, work-spaces, and schools, along with social distancing and mask-wearing. Given the strong associations between doubling times and the stringency index, it is likely that those states that responded to the virus more quickly by implementing a range of measures such as school closing, workplace closing, restrictions on gatherings, close public transport, restrictions on internal movement, international travel controls, and public information campaigns, did have some success slowing the spread of the virus.

**Keywords:** SARS-CoV-2; COVID-19; Epidemic-growth; Doubling-time; United States

## 4.1 Introduction

The current COVID-19 pandemic is posing severe challenges to health systems, societies, and economies worldwide. At the time of writing, the SARS-CoV-2 virus has already infected more than 175 million people globally and caused 3.7 M deaths. In addition, the long-term health impacts on those who have recovered from the SARS-CoV-2 infection are still unknown [1]. Approximately a sixth of the total deaths - more than 600,000 - occurred in the United States (US), the country that currently stands with the highest number of fatalities.

In the US and in other European countries like the the United Kingdom, governments and public health systems were initially caught off guard by the sudden and rapid spread of the virus. This was partly due to a lack of political preparedness and a coherent strategy; lack of public health resources after years of cuts to public health budgets; or to the adoption of the wrong or no policy in terms of mask-wearing, contact tracing, border controls, or lack of testing to detect community transmission [2, 3, 4, 5, 6, 7]. Furthermore, the scientific community took some time to reach a general consensus regarding the modes of transmission of the virus; in particular, airborne dispersal was not considered a major pathway at the beginning of the pandemic, and this inhibited control and containment strategies [8]. Even though thousands of papers have been written on COVID-19 related topics in the past year or so, many questions still remain unanswered, especially in terms of how SARS-CoV-2 transmission is influenced by aspects of the economy, environment, and health. A better understanding of how these factors interact can help us to design timely health prevention and control strategies, and to develop better predictive models for public health risk management of SARS-CoV-2 and other novel coronaviruses [9].

This study explores how some of the macro-level drivers of epidemic growth in the United States are associated with COVID-19 case and death doubling times during the first wave of the pandemic (in early 2020). The reason for selecting the United States is not only that it is one of the hardest-hit countries, but also that it provides us with a unique opportunity to study this phenomenon at a macro-scale, since it encompasses a diverse range of climate types over a vast geographical area, with a somewhat homogeneous political system, allowing us to disentangle the effects of the environment from other demographic and socio-economic conditions. Furthermore, the scientific institutions of the United States offer a vast quantity of high-quality data which allows us to investigate our research question rigorously. By focusing on the first wave of the pandemic, it is possible to better isolate the effects of the environment and socio-economic and demographic factors, since it took some time for the population to adopt self-protective behaviours like vaccination, social distancing and mask-wearing; it also took some time for state governments to apply containment measures, like school closures, limits on gathering and non-essential business closures [7, 10, 11].

The empirical strategy for this study relies on county-level morbidity and mortality data as the main unit of analysis, which consists of counts of individual cases and deaths, aggregated per county. The use of data aggregated at the county level means we cannot make individual level inferences and adjust for individual-level risk factors e.g. age, gender, and occupation. Nevertheless, this type of empirical investigation maintains high merit, as it enables a quick exploration of geographic

associations between the disease and the predictor variables, which can instigate further debate on this topic and may trigger more refined channels of research. The next subsection presents a short analytical framework, explaining how demographic factors, socio-economic factors, climate and population health, as well as containment measures, are expected to influence the spread of the disease, and describes the variables selected to measure such factors.

### 4.1.1 Analytical framework

SARS-CoV-2 transmission takes place through 4 major pathways including exchange of saliva and mucus through human to human physical contact, indirect contact via fomites, or inhalation of large droplets and fine aerosols [12, 8]. Social distancing can be one of the most effective measures to limit transmission, but this can be rendered ineffective in closed spaces with poor ventilation since the virus can transmit through long-distance airborne dispersal [13, 14, 8]. This study emphasises demographic factors, socio-economic factors and climate factors that can influence human to human contact and proximity, and can therefore modulate SARS-CoV-2 transmission [15]. Data on government containment measures will also be analysed since they can moderate SARS-CoV-2 transmission and morbidity and mortality reporting.

#### Economic / demographic / health factors

Given the transmission pathways of SARS-CoV-2, as a priori, we would expect to see more infections occur in locations with higher population densities (e.g., metropolitan areas, cities) with high public transport usage, overcrowded living spaces, and industries where business takes place indoors - all of which naturally bring people into closer contact, allowing airborne transmission to take place. To represent this in the models, variables were selected representing population density, public transport usage and household overcrowding. We would also expect areas with a higher number of new residents arriving from abroad or out of state, to have had a larger number of outbreaks during the early stages of the pandemic through importation of the virus from infected areas. To represent this in the model, a variable was built that captured the annual rate of new residents arriving to a county from abroad or a different state.

At the beginning of the pandemic, it took some time before a consensus was reached about airborne transmission [13, 2, 12, 8, 16], which had major implications for early policy and practice, like improving ventilation in workspaces and adoption of behavioural changes like mask-wearing. We would expect the adoption of self-protective health behaviours (e.g., social distancing, work from home) that can reduce the chance of contracting and spreading the virus [10, 17, 18] to be harder for low skilled workers or those working in specific economic sectors (like manufacturing). Moreover, the inability to self protect may be accentuated for those who suffer from in-work poverty or precariousness since they may also be obliged to work, even when suffering with symptoms, because of a lack of sick pay, fear of losing a day's salary and top-down pressures [19, 20, 21]. These factors are represented in the empirical models using variables that capture unemployment rates, employment levels in key economic sectors, education of the labour force, and poverty.

### **Environmental factors**

Meteorological factors may affect SARS-CoV-2 transmission by altering human behaviour; a basic assumption is people are likely to stay indoors on days with very low or very high temperatures, and/or high rainfall. Furthermore, as a priori, we would expect people to better ventilate their homes/workspaces in places with warmer climates (e.g., leave their windows open, use wall and ceiling fans), which could have an observable overall effect on disease transmission. To represent these factors in the models, variables were selected representing average rainfall, temperate and relative humidity. Meteorological factors can also change the transmission potential and decay rate of the virus in air and on surfaces by altering its stability. [22, 23]. Strong UV light can also inactivate SARS-CoV-2; however, this was not considered a significant predictor for COVID-19 infections and mortality since most transmission takes place indoors [24].

### **Population health**

Initial reports from the ECDC [25], the WHO [26] and the CDC [27] suggest that those most at risk of serious morbidity and mortality are older people and people with underlying health conditions such as diabetes, obesity, respiratory diseases, cancer, and cardiovascular diseases; poverty is a major risk factor of poor population health and is correlated with such conditions [28, 29, 30, 31, 32]. As a priori, we would expect locations with higher proportions of residents with underlying health conditions to report more infections and deaths. To represent this in the models, variables were selected that capture the age structure of the population, poverty rates, long term air pollution to proxy underlying pulmonary health conditions and the prevalence of diabetes.

### **Containment measures**

State governments implemented a wide range of measures to tackle COVID-19 outbreaks such as school closures, workplace closures, restrictions on gatherings, close public transport, stay at home requirements, restrictions on internal movement, international travel controls and public information campaigns, all of which could have had some success in suppressing the spread of the disease [33], such containment measures would moderate the effects of the risk factors and drivers of disease transmission. To account for this in the models a “Stringency Index” measure was selected that reflects the level of a state government’s response to COVID-19 outbreaks, by quantifying how many measures were implemented and to what degree they were applied. The equations used to construct the “Stringency Index” will be further explained in the next section. Compulsory stay at home orders (lock-downs) were not included in the “Stringency Index”, since they were used to determine the temporal cut off points of the study window, this is also explained in the section.

## **4.2 Methods**

All data were aggregated at the county level, apart from some data on containment measures which are presented at state level. Below a detailed description of the data sources.

## 4.2.1 Data collection and processing

### Morbidity and mortality data

SARS-CoV-2 morbidity and mortality data were sourced from Johns Hopkins University’s Centre for Systems Science and Engineering’s (CSSE) GitHub repository [34]. In general, during the first wave of outbreaks in the US, testing was conducted only on those reporting more serious symptoms (see Appendix 3 - COVID policy tracker). Almost all diagnostic testing for COVID-19 was done with the PCR-based methods, using nasopharyngeal or oropharyngeal specimens (nose or throat swabs).

### Economic, demographic and population health data

Data on county population, public transport usage, population age structure, health insurance coverage, immigration, disabilities, and household overcrowding were sourced from the United States Census Bureau using 2015-2019 ACS 5-year estimates [35]. To standardise data across counties, all appropriate variables were converted to percentages/averages of the total county population. A household was considered overcrowded if the number of rooms was less than the number of inhabitants (above 1.01 people per room), this figure included all rooms in a household (not just bedrooms). The disabilities measure captured various health conditions such as difficulty seeing or hearing, restricted movement, learning disabilities, cerebral palsy or other developmental disabilities, or intellectual or mental health disabilities [36].

Population density per km<sup>2</sup> was calculated using R’s SF package and the United States Census Bureau Cartographic county-level shape-files. Because the range of population density values was very wide, all values above 2500 km<sup>2</sup> were capped to this value. This modification was tested in the final models and did not affect the results and allowed for better interpretability of the results.

County-level data on unemployment (%), median household income (\$), and poverty % were sourced from the USDA Economic Research Service [37]. The “Poverty %” indicator represents the percentage of people/families whose earnings are less than the threshold designated by the Census Bureau’s set of money income thresholds. Data on diabetes prevalence were sourced from the CDC’s diabetes atlas [38]. Economic dependence of a county was represented using the ERS county-level typology data-set [37]; this breaks down a county into one of 6 major economic typologies: farming, mining, manufacturing, federal/state government, recreation, and non-specialized.

### Environmental data

Temperature (°C), precipitation (1/100”), and relative humidity data were sourced from the Global Surface Summary of the Day (GSOD) data provided by the US National Climatic Data Center (NCDC)[39]. This data-set provides daily GPS observations from all weather stations situated in the US. To join county data with the GSOD weather observations, centroids were created for each county using R’s SF package and the United States Census Bureau’s county shape-files. The K-nearest neighbor join function in R’s SF package was used to create a spatial join between the weather stations (GPS coordinates) and the county centroids. Mean climate values were created for a county-based on data from a maximum of 10 nearest weather stations within a 100km radius of each county centroid.

Data on air quality was sourced from the United States Environmental Protection Agency [40]. Annual maximum reported Air Quality Index (AQI) values were used, taken over a 20-year average. This indicator was derived from data from EPA's AQS (Air Quality System) database. The EPA establishes an AQI based on five major air pollutants including ground-level ozone ( $O_3$ ), particle pollution (also known as particulate matter, including PM2.5 and PM10), carbon monoxide ( $CO$ ) sulfur dioxide ( $SO_2$ ) and nitrogen dioxide ( $NO_2$ ). The U.S. AQI index runs from 0 to 500. The higher the AQI value, the greater the level of air pollution and the greater the health concern. The AQI is divided into 6 categories, each corresponding to a different level of health concern; generally, they represent 0 to 50 - good; 51 to 100 - moderate; 101 to 150 - unhealthy for sensitive groups; 151 to 200 - unhealthy; 201 to 300 - very unhealthy; 301 and higher - hazardous.

### Containment measures

Data on county-level stay-at-home orders (lock-down) were extracted from the CDC's "U.S. State, Territorial, and County Stay-At-Home Orders" dataset [41]. This dataset provides information on county-level executive orders, administrative orders, resolutions, and proclamations and can be used to determine the date of county-level stay-at-home orders (lock-down).

Data on state-level control measures were sourced from the Oxford COVID-19 Government Response Tracker (OxCGRT) data set [42]. The "Stringency Index" variable from this dataset was used to account for the application of state-level control measures in our final models. The composite time-series measure, ranging from 0 to 100 (100 = strictest) is based on 9 response indicators including data on school closing, workplace closing, restrictions on gatherings, close public transport, stay at home requirements, restrictions on internal movement, international travel controls, and public information campaigns. The indicator reflects the level of a state government's response to COVID-19 outbreaks and quantifies how many measures were implemented, and to what degree they were implemented. The index cannot ascertain whether a government's policy has been implemented effectively nor the effectiveness of an individual measure [33]. To get an estimate of a government's response leading up to the first lock-down (compulsory stay at home order), the average stringency index value was calculated using a time window: from the day the first 5 cases were reported the day before the first lock-down. Arkansas, Iowa, Nebraska, North Dakota, and South Dakota did not implement state-wide lock-downs. In these states, the average score was calculated from the day the first cases were reported to the last lock-down date in our sample (2020-07-04) to make this value comparable to other states.

### 4.2.2 Study design

The spread of the disease (epidemic growth) is modelled by calculating COVID-19 case and death doubling times; these measures were then used as dependent variables to explore associations between epidemic growth, socio-economic, demographic, and environmental factors, and population health. Doubling times capture exponential growth, in this instance, the number of days taken for cases and deaths to double. This measure has several advantages: first, it provides a way of standardising differences in sampling effort between different locations and health authorities; second,

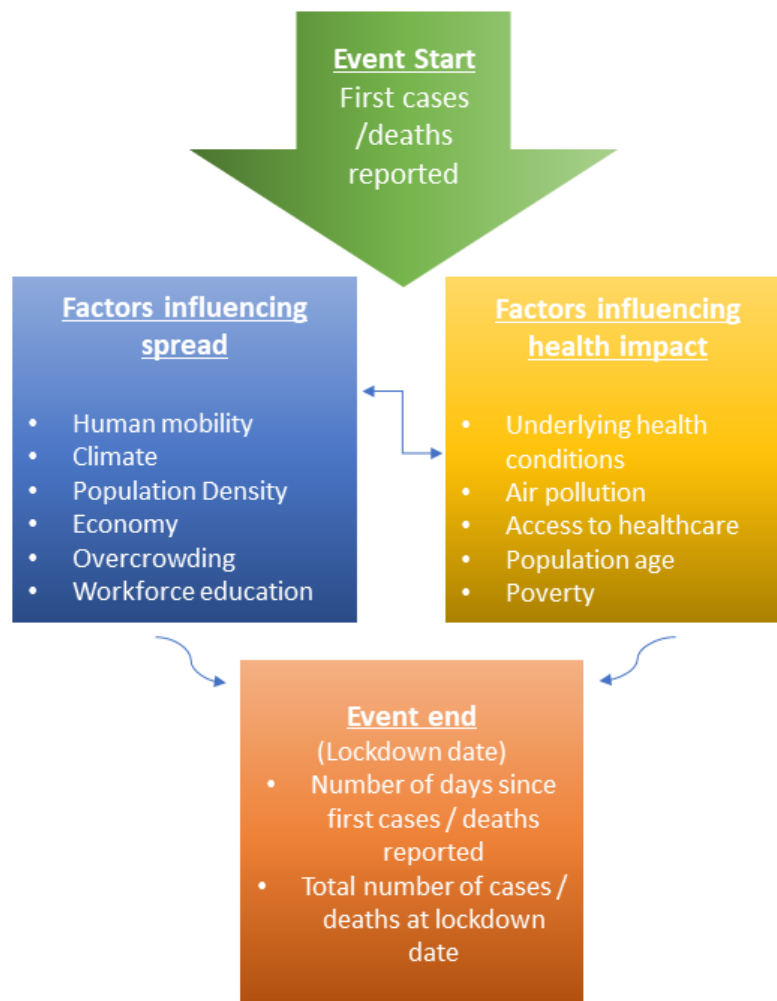


Figure 4.1: Study design: capturing epidemic growth

because it provides us with a time determinant measure to facilitate understanding of the spread of the virus. In other words, this metric not only has the advantage of accounting for population size but also incorporates a time dimension. Therefore, COVID-19 transmission is measured by calculating doubling times for infections and mortality, at the county level [43, 44, 45, 46].

### Calculating case and death doubling times

Doubling times were calculated by capturing a window of infection opportunity, which started on the date a minimum number of cases/deaths were detected in a county, to the date of the first major state or county level intervention was implemented i.e. compulsory stay-at-home orders, otherwise known as a lock-down (see Figure 4-1). A time lag was also applied to the doubling times in order to account for the time infection or mortality events took place, since there is a lag between the date an event is reported (a case or death) and the date the transmission event took place. Therefore all case and mortality data was lagged by a maximum incubation period (onset of symptoms) or a maximum time from final infection to death; these



are further described below.

For the calculation of the infection doubling times, the count was started when the county reached a minimum of 50 confirmed cases, over a minimum 7 day reporting period. Any county that did not meet this requirement was excluded from the study.

Since the mortality data-set contained fewer observations than the cases data set, the count was set when the county reported a minimum of 20 deaths over a minimum 7 day reporting period. Although these values yielded enough observations to carry out the study on mortality doubling times, the doubling times may be less stable than that of the case data-set.

Again, any county that did not meet this requirement was excluded from the study.

To calculate the case and death doubling times for each county, the following formulas were applied:

$$r = \frac{E_{end} - E_{start}}{E_{start}} \times 100$$

Where:

$r$  = growth rate;

$E_{start}$  = Start of the event - when the 50 cases / 20 deaths are detected

$E_{end}$  = End of the event - cumulative cases / deaths per county at the lock-down date;

Next, the doubling time is calculated using the following formula:

$$T_d = t \frac{\ln(2)}{\ln(1 + \frac{r}{100})}$$

Where:

$T_d$  = doubling time in days

$t$  = time in days (Estart to Eend)

$r$  = growth rate

Arkansas, Iowa, Nebraska, North Dakota, and South Dakota did not implement a state-wide lock-down (stay at home order), so an artificial date was set to calculate doubling times, mirroring the latest lock-down date in our sample (2020-07-04).

### Time lags - disease progression

Disease progression was also considered when calculating the doubling times; a time lag was applied to account for the discrepancy between the date an event was reported (a case or death) and the date the transmission event is likely to have took place.

For data on confirmed COVID-19 cases, a lag of 21 days was set which considers a maximum 14-day incubation period based on findings from cohort studies by Lauer 2020 [47], with an extra 7 days to account for any reporting delays. The implication here is that case data for anything up to 21 days post lock-down was used to calculate doubling times.

For the mortality data set, a lag of 42 days was set days which includes the maximum 14-day incubation period based on findings from cohort studies by Lauer

et al., 2020 [47] and a maximum of 21 days from the first onset of symptoms to death based on findings from cohort studies by Verity et al., 2020 [48], plus an extra 7 days to account for any reporting delays. The implication here is that mortality data for anything up to 42 days post lock-down was used to calculate doubling times.

Data on environmental factors were also joined to the lagged county doubling time variables, meaning that they were linked to the date when a disease event is likely to have took place, rather than when reported.

### **General additive regression model to assess the impact of independent variables on doubling times at the county level**

One of the main issues with the data-set is that it did not meet some basic assumptions for statistical inference, that is the data are not independent and identically distributed random variables (iid). More specifically, observations cannot be considered independent because of spillover effects from neighbouring counties, therefore an appropriate statistical design was needed to control for a lack of independence between neighbouring counties. A Generalised Additive Model (GAM) using R's MgcV statistical package because of its versatility and ability to fit complex models that would converge even with low numbers of observations and could capture potential complex non-linear relationships. One of the advantages of GAMs is that we do not need to determine the functional form of the relationship beforehand. In general, such models transform the mean response to an additive form so that additive components are smooth functions (e.g., splines) of the covariates, in which functions themselves are expressed as basis-function expansions. The spatial auto-correlation in the GAM was approximated by a Markov random field (MRF) smoother, which represents the spatial dependence structure in the data. R's Spdep package was used to create a queen neighbours list (adjacency matrix) based on counties with contiguous boundaries i.e., those sharing one or more boundary points. The local Markov property assumes that a county is conditionally independent of all other counties unless they share a boundary. This feature allows us to model the correlation between geographical neighbours and smooth over contiguous spatial areas, summarising the trend of the response variable as a function of the predictors [49]. Models were fit using a gamma distribution; after inspecting the data, it was concluded that the a gamma distribution worked well with the shape of our response variable, which was positively skewed (i.e., non-normal, with a long tail on the right). The gamma distribution is a two-parameter distribution, where the parameters are traditionally known as shape and rate. Its density function is:

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta},$$

where  $\alpha$  is the shape parameter and  $\beta-1$  is the rate parameter (alternatively,  $\beta$  is known as the scale parameter).

The empirical model can then be written as:

$$E(Y) = f_1(X_i) + f_m(\text{County}_i)$$

Where the  $f(.)$  stands for smooth functions;  $E(Y)_i$  is equal to infection or death doubling time in county  $i$ , which we assume to be gamma-distributed;  $X_i$  - is a

vector of economic, demographic, environmental and climate variables (as described in the previous section).  $county_i$  represents neighbourhood structure of the county.

Analysis of model diagnostic tests didn't reveal any major issues, in general residuals appeared to be randomly distributed. For robustness, models were also fit using the Gaussian and Tweedie distributions, and also fit using a non-additive-GLM (see Appendix 3).

## 4.3 Results

To carry out the empirical analysis, a unique spatial data-set was compiled that captured potential drivers of human-to-human SARS-CoV-2 transmission and risk factors of serious infections and mortality due to COVID-19 in US counties.

### Descriptive statistics

Two sources of information were analysed, data on confirmed cases and deaths. Tables 4-1 and 4-2 provide summary statistics for our final data-sets.

To calculate doubling times, counties were only selected that had reported at least 50 cases or 20 deaths over a minimum 7-day period before the first lock-down. Both sources of information were chosen as they allow us to explore and compare different features and characteristics of the epidemic. Figures 4-2 and 4-3 map the geographical distribution for case and death doubling times in counties that met our inclusion criteria (coloured from red to yellow). Major cities with populations > 250,000 people are highlighted on each map. The counties first affected by SARS-CoV-2 during the first wave of the epidemic tended to be located around major cities and metropolitan areas on the east coast, mid west, and south of the United States, with high population density and presumably higher numbers of international and domestic travellers.

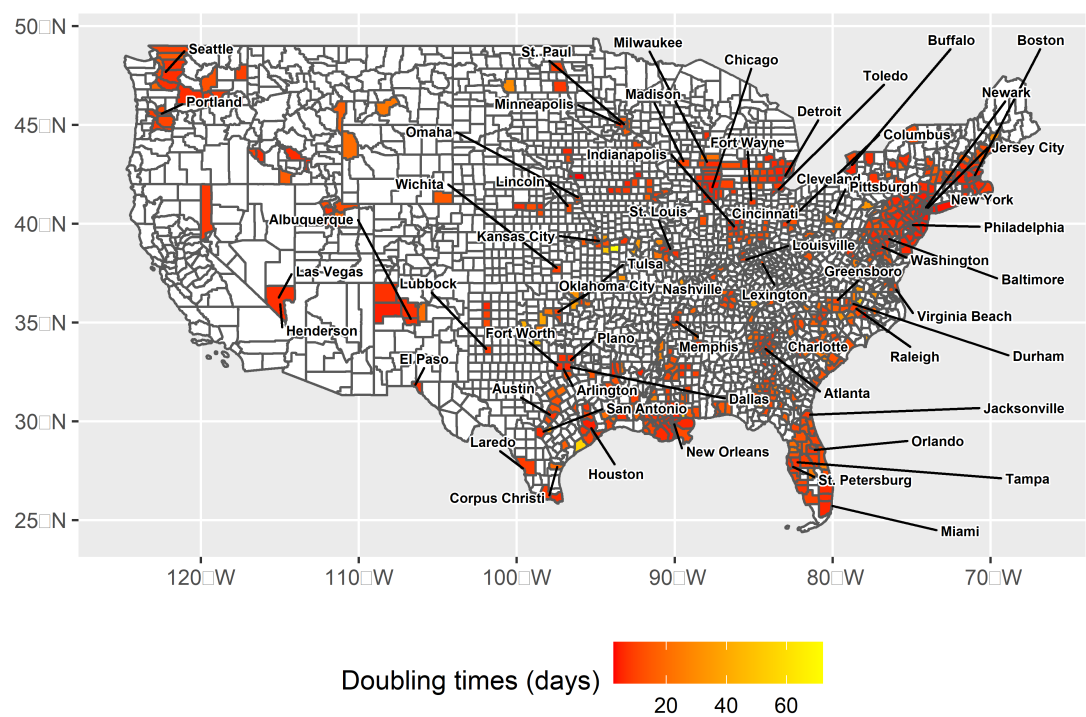


Figure 4.2: COVID-19 case doubling times in US Counties and major cities with over 250,000 people. Counties highlighted in white not selected for study (Date range: 2020-03-05 to 2020-04-29, data source: Johns Hopkins University)

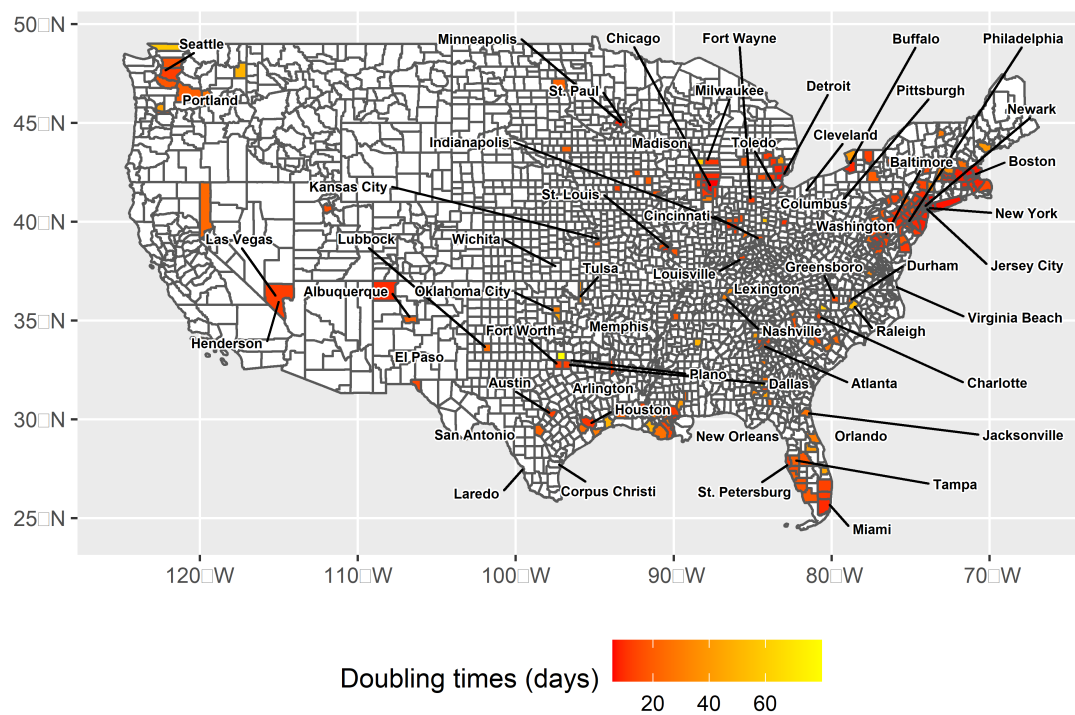


Figure 4.3: COVID-19 death doubling times in US Counties and major cities with over 250,000 people. Counties highlighted in white not selected for study (Date range: 2020-03-10 to 2020-05-19, data source: Johns Hopkins University)

Table 4.1: Cases data-set - summary statistics. N = 640 (number of counties selected for the study which met the inclusion criteria laid out in study design section)

Statistic	Min	Max	Mean	St. Dev.
Confirmed cases	54	35,571	950.9	2,953.7
Incidence per 100,000	15.8	5,611.3	277.5	434.1
Study period days	7	40	18.7	7.5
Growth rate (over period)	8	62,305	1,503.5	5,040.1
Case doubling times	2.5	72.0	11.6	8.4
County population	5,861	10,081,570	373,339.6	657,728.9
Population density per km2	2.0	2,500.0	245.3	415.8
Stringency index	11.0	64.1	34.1	10.5
Public transport usage (pop %)	0.0	32.0	1.3	3.1
Median household income (\$)	26,348	151,806	65,642.1	18,915.3
Unemployment %	1.8	12.0	3.8	1.2
Population % 65+	7.9	41.1	16.5	4.2
Poverty %	2.7	36.6	13.1	5.9
Health insurance coverage %	62.2	99.9	98.0	3.0
Annual new arrivals / by pop (%)	0.0	3.1	0.5	0.4
Population % with disabilities	5.0	25.7	13.1	3.3
Population % with diabetes	2.2	23.1	10.1	2.9
Household overcrowding %	0.1	7.0	1.3	1.0
Population % with degree or higher	3.6	29.0	12.6	4.6
Temperature (°C)	−3.4	24.7	12.7	6.4
Precipitation (1/100")	0.0	9.2	2.1	1.6
Relative humidity	30.7	86.1	67.0	8.6
Air quality index (AQI)	32.5	347.4	130.4	31.3

Table 4.2: Mortality dataset - summary statistics. N = 263 (number of counties selected for the study which met the inclusion criteria laid out in study design section)

Statistic	Min	Max	Mean	St. Dev.
Confirmed cases	22	4,902	233.4	541.3
Incidence per 100,000	1.4	284.6	39.9	45.4
Study period days	7	56	24.4	10.2
Growth rate (over period)	9.5	24,410.0	906.5	2,346.6
Cases doubling time	5.7	80.0	18.5	12.5
County population	8,737	10,081,570	677,179.5	929,606.3
Population density per km2	2.9	2,500.0	442.2	570.3
Stringency index	11.0	64.1	32.4	11.6
Public transport usage (pop %)	0.0	32.0	2.4	4.5
Median household income (\$)	36,894	151,806	71,804.3	20,582.3
Unemployment %	1.8	9.6	3.7	1.0
Population % 65+	9.5	41.1	16.1	4.0
Poverty %	2.7	30.7	12.0	5.1
Health insurance coverage %	89.9	99.6	98.7	1.2
Annual new arrivals / by pop (%)	0.01	2.2	0.6	0.4
Population % with disabilities	5.8	23.5	12.1	2.9
Population % with diabetes	5.2	22.3	9.3	2.3
Household overcrowding %	0	6	1.4	1.0
Population % with degree or higher	3.9	27.6	14.2	4.4
Temperature (°C)	0.2	25.3	11.6	5.9
Precipitation (1/100")	0.0	7.4	2.0	1.4
Relative humidity	25.5	81.6	64.9	8.8
Air quality index (AQI)	41.5	347.4	143.8	31.5

## Regression results

It was not possible to explore the individual impact of all the variables in our dataset because of collinearity issues (see Appendix C). Public transport was positively correlated with population density so therefore removed from the analysis. Median income was also removed from the analysis because it was positively correlated with education, and negatively correlated with poverty, disabilities and diabetes.

Tables 4-3 and 4-4 show the results of the statistical analysis for both data sets and summarise the relevant statistics (AIC, Deviance, Adjusted R squared ( $R^2$  and so on) to compare the different specifications. Both statistical models were built in a step-wise fashion using the lowest Akaike Information Criterion (AIC) and  $R^2$  to help us assess the different specifications. Variables were included in each specification according to their category i.e., spatial, socio-economic, and environmental. All variables were included in the final specification to ascertain the contribution of each driver or risk factor, all else equal. Note that, as we are not estimating a standard regression model, the figures reported should not be read as coefficients, but degrees of freedom of the smooth terms. Given that we cannot interpret the coefficients to infer the sign and magnitude of the relationship, we visualise it by plot. Figures 4-3 to 4-11 plot the partial effects—the relationship between a change

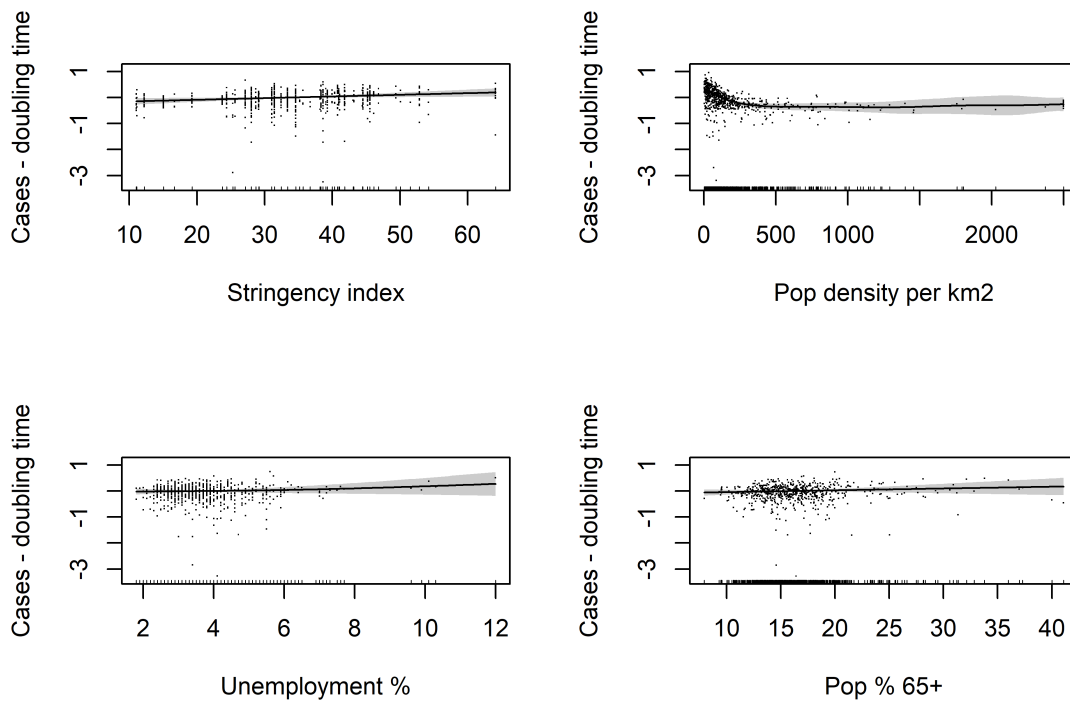


Figure 4.4: Generalised additive model (GAM) plots showing the partial effects of the explanatory variables on the doubling times of COVID-19 infections. The tick marks on the x-axis are observed data points. The y-axis represents the partial effect of each variable. The dots represent partial residuals. The shaded areas indicate the 95% confidence intervals. Higher values on the y-axis represent slower case doubling times.

in each of the covariates and a change in the fitted values in the full model. Standard errors on the plots show the 95% confidence interval for the mean shape of the effect.

### Case data model

Table 4-3 and figures 4.4 to 4.7 show the results of the model fit using case data. The “Spatial” model was fit first to estimate the contribution of the spatial lag component against the other specifications. A high proportion of the variance is explained just by controlling for spatial correlation between counties ( $R^2$  0.35). The “Full model” has the best fit in terms of the AIC and adjusted  $R^2$ , followed by the socio-economic model, and finally the environmental model. The adjusted  $R^2$  in the final model is 0.56, indicating that 56% of the variance in our model is explained by the explanatory variables.

As for the contribution of individual variables on case doubling times, counties with manufacturing and recreation as their predominant economic activity were associated with faster case doubling times although the confidence intervals are fairly large so the sample does not provide a precise representation of the population mean. The stringency index variable, which captures the number of containment measures adopted by states, and the degree to which they were implemented, is also statistically significant ( $p < 0.05$ ); and has a positive relationship with the case doubling



	Spatial	Socio-econ	Envir	Full
Intercept	2.34*** (0.02)	2.45*** (0.12)	2.34*** (0.02)	2.48*** (0.12)
Industry: Manufacturing		-0.27* (0.13)		-0.32* (0.13)
Industry: Government		-0.02 (0.13)		-0.04 (0.12)
Industry: Recreational		-0.22 (0.14)		-0.23 (0.13)
Industry: Non-specialised		-0.13 (0.12)		-0.17 (0.12)
Industry: Agricultural		0.09 (0.24)		0.04 (0.23)
Stringency index		1.00* (1.00)		1.00* (1.00)
Pop density per km2		6.30*** (7.56)		5.81*** (7.01)
Unemployment %		1.61 (1.82)		1.31 (1.50)
Population % 65+		1.00* (1.00)		1.00 (1.00)
Poverty %		1.00* (1.00)		1.00** (1.00)
Health insurance coverage %		1.62 (1.83)		1.57 (1.80)
New arrivals into county population %		1.40 (1.63)		1.55 (1.78)
Population % with disabilities		1.00** (1.00)		1.00** (1.00)
Population % with diabetes		1.00* (1.00)		1.00** (1.00)
Population % living in overcrowded homes		1.00* (1.00)		1.00** (1.00)
Temperature °C			3.54** (4.10)	3.71** (4.25)
Precipitation			2.21 (2.75)	1.00 (1.00)
Relative humidity			1.00 (1.00)	3.54 (4.15)
Air quality index (AQI)			1.00*** (1.00)	1.67** (1.87)
County	137.13*** (169.51)	135.30*** (167.57)	132.98*** (165.05)	135.55*** (167.11)
AIC	3772.29	3499.61	3705.13	3459.43
BIC	4393.02	4210.00	4341.94	4212.15
Log Likelihood	-1747.01	-1590.58	-1709.83	-1561.00
Deviance	91.15	56.41	81.36	51.50
Deviance explained	0.61	0.76	0.65	0.78
Dispersion	0.21	0.13	0.19	0.11
R <sup>2</sup>	0.35	0.54	0.41	0.56
GCV score	0.23	0.16	0.21	0.15
Num. obs.	640	640	640	640
Num. smooth terms	1	11	5	15

\*\*\* $p < 0.001$ ; \*\* $p < 0.01$ ; \* $p < 0.05$ ;  $p < 0.1$

Table 4.3: COVID-19 Infection model - Generalised additive regression model for assessing associations between the demographic, socio-economic, climate and population health factors on county level case doubling times. Note that as we are not estimating a standard regression model, the figures reported should not be read as coefficients, but degrees of freedom of the smooth terms. Given that we cannot interpret the coefficients to infer the sign and magnitude of the relationship, we visualise it by plot

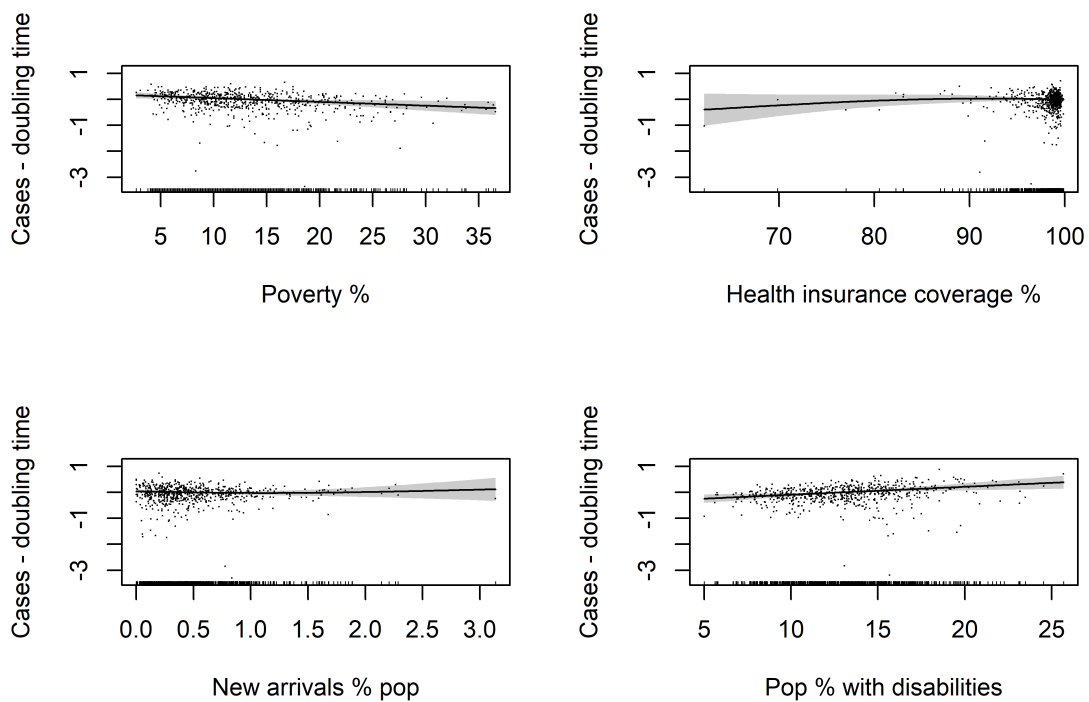


Figure 4.5: Generalised additive model (GAM) plots showing the partial effects of the explanatory variables on the doubling times of COVID-19 infections. The tick marks on the x-axis are observed data points. The y-axis represents the partial effect of each variable. The dots represent partial residuals. The shaded areas indicate the 95% confidence intervals. Higher values on the y-axis represent slower case doubling times.

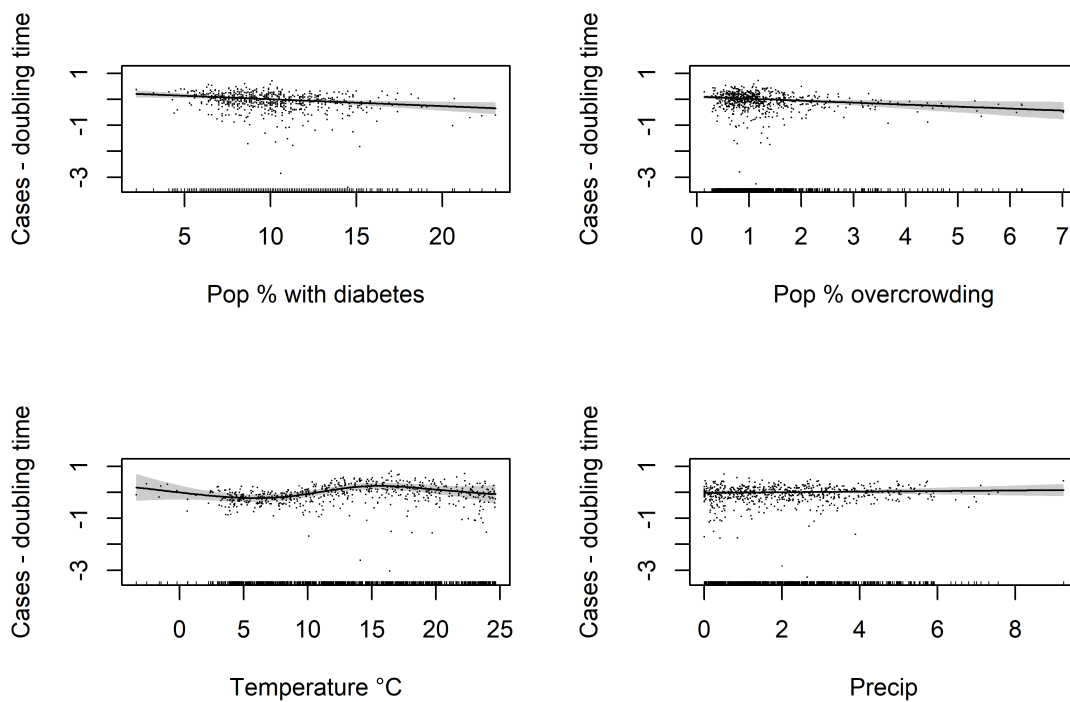


Figure 4.6: Generalised additive model (GAM) plots showing the partial effects of the explanatory variables on the doubling times of COVID-19 infections. The tick marks on the x-axis are observed data points. The y-axis represents the partial effect of each variable. The dots represent partial residuals. The shaded areas indicate the 95% confidence intervals. Higher values on the y-axis represent slower case doubling times.

times, suggesting that measures had some success in suppressing the virus. Human population density per  $km^2$  is highly significant ( $p < 0.001$ ), higher densities are associated with faster case doubling times, although the relationship is not linear and flattens out at higher population densities. Although the slope is gentle, “Poverty %” is a highly significant ( $p < 0.01$ ) predictor of case doubling times, the relationship is negative which means doubling times are faster with higher levels of poverty (in other words, the infection spreads faster). On the contrary, the variable “Pop % with disabilities” ( $p < 0.01$ ) has a positive relationship with case doubling times, meaning it is a predictor of slower doubling times. The prevalence of diabetes (Pop % with diabetes) in a county, an indicator that not only represents the disease itself, but also a range of other conditions such as obesity, poor diet, lack of exercise was also a significant ( $p < 0.01$ ) predictor of faster case doubling times. “Population % home overcrowding”, which represents the percentage of households in a county where there is less than one room per inhabitant ( $> 1.01$  people per room) is highly significant ( $< 0.01$ ) and is associated with faster case doubling times. Temperature is also a good predictor of case doubling times; higher temperatures appear to slow case doubling times. ( $p < 0.01$ ), although this relationship breaks down at lower temperatures given there are few observations, the confidence intervals are much larger meaning the results are less accurate. “Max AQI”, which represents the maximum air quality index values averaged over 20 years, is also highly significant and is associated with faster case doubling times in locations with poor air quality ( $p < 0.01$ ).

### **Mortality model**

	Spatial	Socio-econ	Envir	Full
Intercept	2.83*** (0.03)	2.65*** (0.25)	2.81*** (0.03)	2.78*** (0.25)
Industry: Manufacturing		0.14 (0.28)		-0.02 (0.28)
Industry: Government		0.18 (0.26)		0.03 (0.26)
Industry: Recreational		0.18 (0.28)		0.13 (0.28)
Industry: Non-specialised		0.13 (0.26)		-0.02 (0.26)
Stringency index		1.00* (1.00)		1.00* (1.00)
Population density per km2		7.00*** (8.35)		5.84*** (7.05)
Unemployment %		1.00 (1.00)		1.00 (1.00)
Population % 65+		1.00* (1.00)		1.00*** (1.00)
Poverty %		1.00* (1.00)		1.00** (1.00)
Health insurance coverage %		1.00 (1.00)		1.00 (1.00)
New arrivals into county population %		1.00 (1.00)		1.00 (1.00)
Population % with disabilities		1.00 (1.00)		1.00** (1.00)
Population % with diabetes		1.00 (1.00)		1.00* (1.00)
Population % living in overcrowded homes		1.00 (1.00)		1.00 (1.00)
Temperature °C			2.51 (3.07)	2.71 (3.28)
Precipitation			3.03 (3.66)	4.61 (4.91)
Relative humidity			1.00 (1.00)	1.00 (1.00)
Air quality index (AQI)			1.04** (1.07)	1.00* (1.00)
County	56.42*** (67.05)	56.15*** (66.65)	59.31* (69.50)	57.32*** (67.52)
AIC	1854.52	1742.72	1824.06	1724.01
BIC	2063.22	2021.91	2070.15	2036.51
Log Likelihood	-868.84	-793.20	-843.14	-774.52
Deviance	46.09	26.26	38.10	22.83
Deviance explained	0.50	0.72	0.59	0.75
Dispersion	0.27	0.15	0.22	0.13
R <sup>2</sup>	0.23	0.44	0.31	0.48
GCV score	0.29	0.20	0.26	0.19
Num. obs.	263	263	263	263
Num. smooth terms	1	11	5	15

\*\*\* $p < 0.001$ ; \*\* $p < 0.01$ ; \* $p < 0.05$ ;  $p < 0.1$

Table 4.4: COVID-19 mortality model- Generalised additive regression model for assessing associations between the demographic, socio-economic, climate and population health factors on county level death doubling times. Note that as we are not estimating a standard regression model, the figures reported should not be read as coefficients, but degrees of freedom of the smooth terms. Given that we cannot interpret the coefficients to infer the sign and magnitude of the relationship, we visualise it by plot

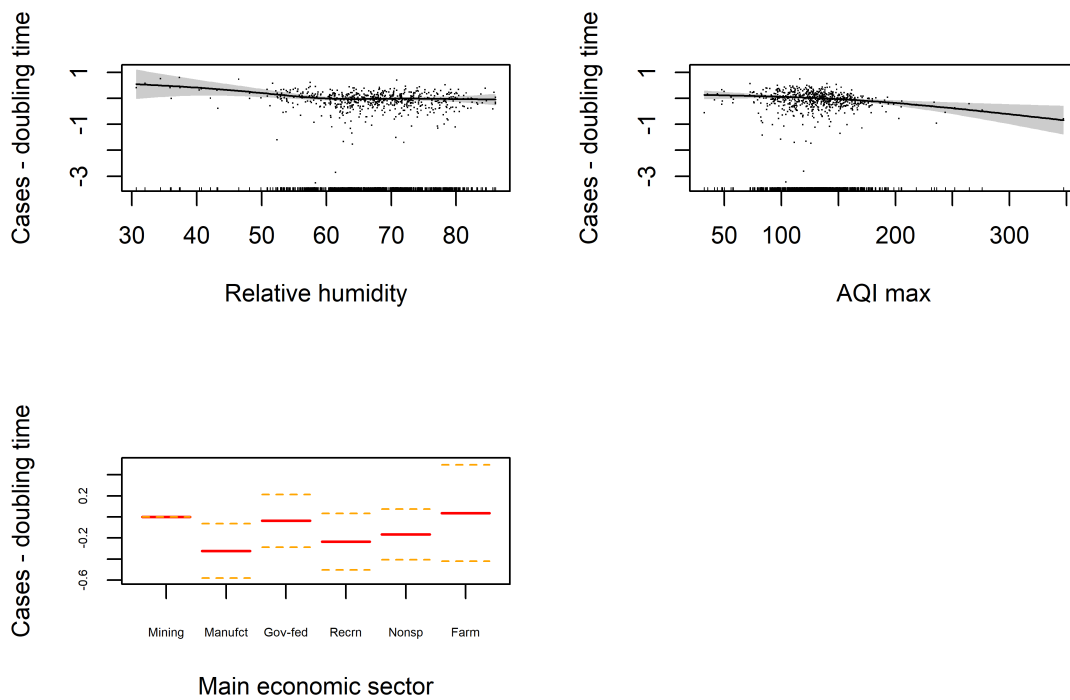


Figure 4.7: Generalised additive model (GAM) plots showing the partial effects of the explanatory variables on the doubling times of COVID-19 infections. The tick marks on the x-axis are observed data points. The y-axis represents the partial effect of each variable. The dots represent partial residuals. The shaded areas indicate the 95% confidence intervals. Higher values on the y-axis represent slower case doubling times. Bottom left: Categorical variables - dashed horizontal lines on the categorical variables represent the confidence intervals and solid red lines represents the mean value

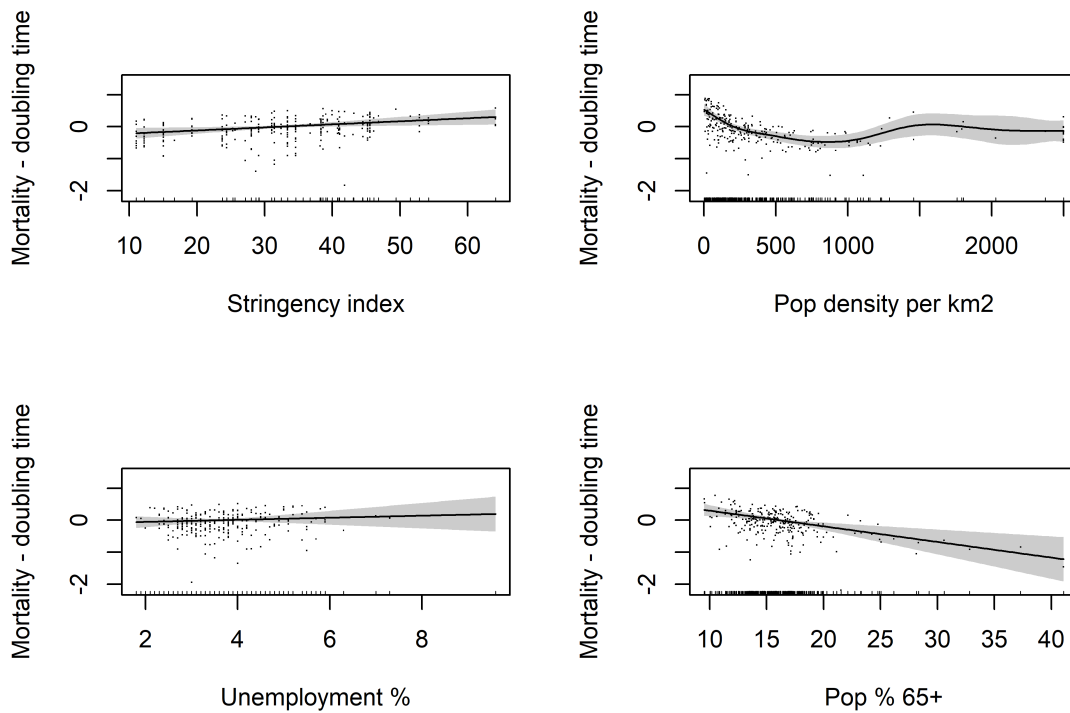


Figure 4.8: Generalised additive model (GAM) plots showing the partial effects of the explanatory variables on the doubling times of COVID-19 deaths. The tick marks on the x-axis are observed data points. The y-axis represents the partial effect of each variable. The dots represent partial residuals. The shaded areas indicate the 95% confidence intervals. Higher values on the y-axis represent slower death doubling times.

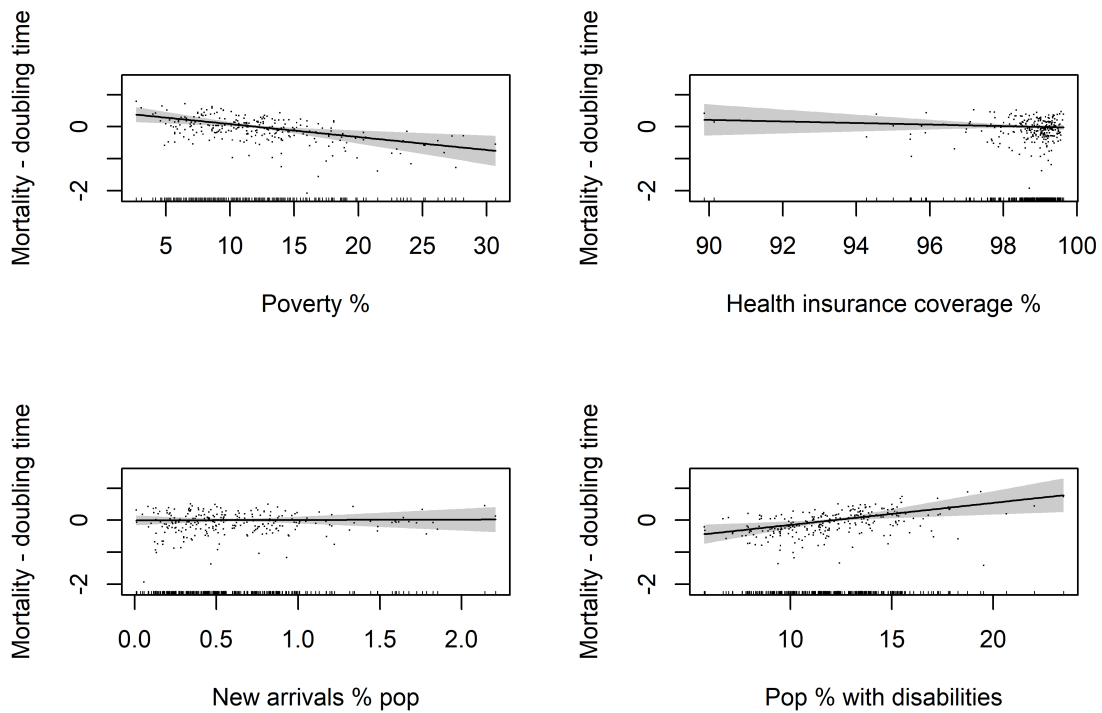


Figure 4.9: Generalised additive model (GAM) plots showing the partial effects of the explanatory variables on the doubling times of COVID-19 deaths. The tick marks on the x-axis are observed data points. The y-axis represents the partial effect of each variable. The dots represent partial residuals. The shaded areas indicate the 95% confidence intervals. Higher values on the y-axis represent slower death doubling times.



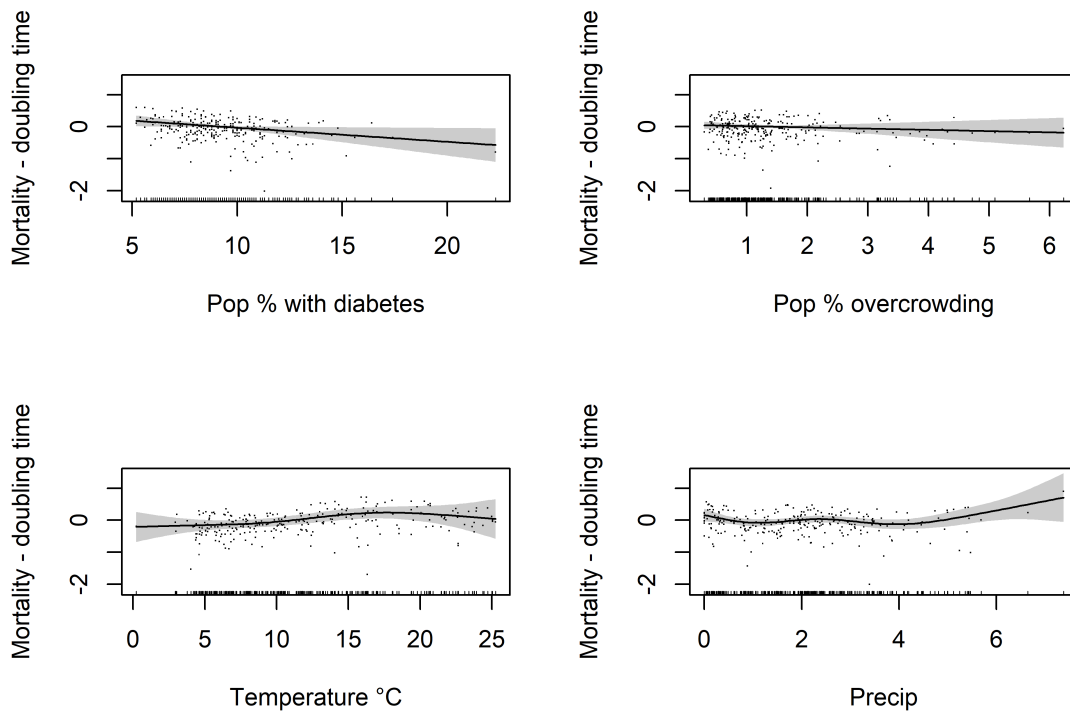


Figure 4.10: Generalised additive model (GAM) plots showing the partial effects of the explanatory variables on the doubling times of COVID-19 deaths. The tick marks on the x-axis are observed data points. The y-axis represents the partial effect of each variable. The dots represent partial residuals. The shaded areas indicate the 95% confidence intervals. Higher values on the y-axis represent slower death doubling times.

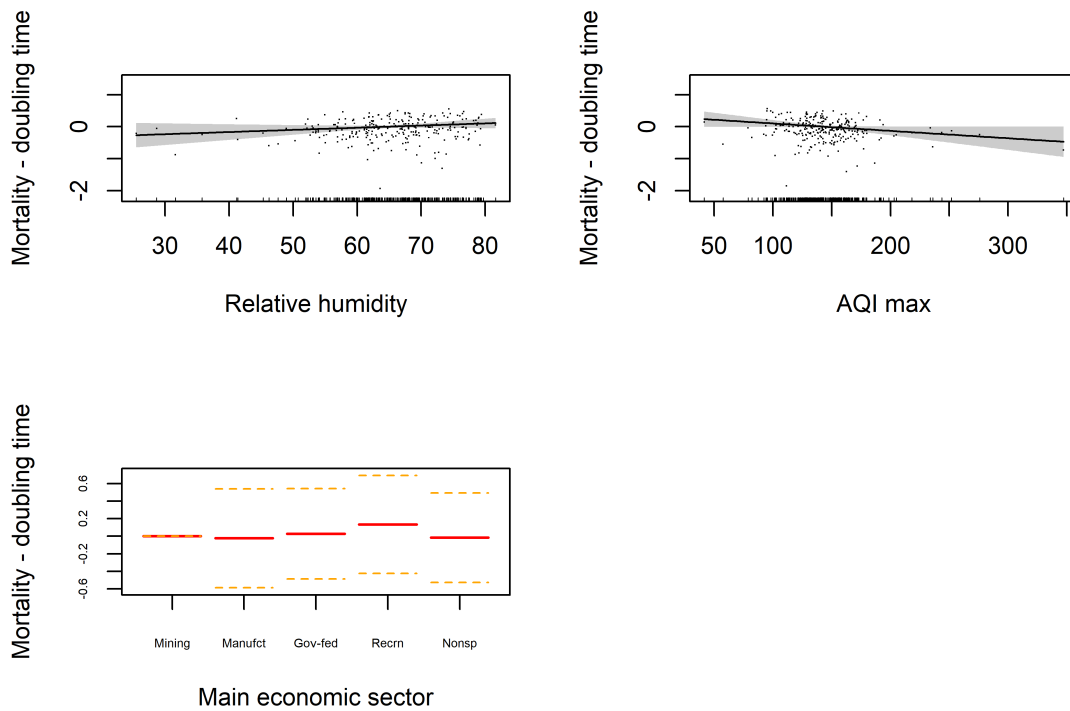


Figure 4.11: Generalised additive model (GAM) plots showing the partial effects of the explanatory variables on the doubling times of COVID-19 deaths. The tick marks on the x-axis are observed data points. The y-axis represents the partial effect of each variable. The dots represent partial residuals. The shaded areas indicate the 95% confidence intervals. Higher values on the y-axis represent longer death doubling times. Bottom left: Categorical variables - dashed horizontal lines on the categorical variables represent the confidence intervals and solid red lines represents the mean value.

Table 4-4 and figures 4-8 to 4-11 show the results of our model fit using mortality data. A high proportion of the variance is explained just by controlling for spatial correlation between counties ( $R^2$  0.22). The “Full model” has the best fit in terms of the AIC and adjusted  $R^2$  0.48, followed by the socio-economic model (0.44) and the environmental model (0.31). The “Stringency index” indicator is statistically significant ( $p < 0.05$ ) and is associated with slower death doubling times; that is more stringent containment measures are associated with slower COVID-19 death doubling times. “Population density per  $km^2$ ” ( $< 0.001$ ) is also an important predictor: generally, higher population density is associated with faster death doubling times, however, this trend reverses at around 1400 inhabitants per  $km^2$  and levels off. “Population % 65+” ( $< 0.001$ ) is highly significant; higher values are associated with faster death doubling times. Again, as with the case data analysis, “Poverty %” is also a highly significant predictor of death doubling times ( $< 0.001$ ), that is higher levels of poverty are associated with mortality. “Pop % with disabilities” ( $< 0.01$ ) is also highly significant; as with the case data model, this predictor is associated with slower death doubling times. The prevalence of diabetes (Pop % with diabetes) in a county is also a significant predictor ( $p < 0.05$ ) of faster death doubling times, as is the air quality index (“Max AQI”), which is highly statistically

significant ( $p < 0.01$ ). Temperature and precipitation are slightly significant ( $p < 0.1$ ) and appear to slow down death doubling times at higher values.

## 4.4 Discussion

In this study, I examined which socio-economic, demographic, and environmental factors are associated with SARS-CoV-2 epidemic growth. To explain biases in reporting, I included health risk factors that can contribute to serious SARS CoV-2 infections and deaths. We would expect case reporting to be a function of all these factors since testing policy during this phase of the epidemic was aimed at those with symptoms (see Appendix 3 - COVID-19 policy tracker).

We can also assume that, during this wave of the epidemic in the US, only one strain of SARS-CoV-2 (although always evolving) was in circulation and therefore the variation in infection and death rates across space can be attributed to external factors i.e., testing differences, aspects of the population and environment, rather than variation in viral traits/strains. Furthermore, no vaccines were yet in circulation.

### 4.4.1 Containment measures to reduce disease spread

During the first wave of the epidemic in the US, governments, and public health systems were initially caught off guard by the rapid spread of the virus. Some of the states did apply more rigorous control measures than others, attempting to suppress the spread of the virus early on e.g., by restricting gatherings, closure of public spaces, creating public awareness campaigns and contact tracing (see Appendix 3). Stringency index scores in both our models are associated with slower doubling times and can be interpreted as, the more stringent the measures applied by state governments early on, the more success they had in suppressing the virus.

### 4.4.2 Socio-economic, economic, and demographic factors

Results show that human population density is one of the strongest predictors of case and death doubling times, the relationship is negatively linear to a point, where higher population densities are associated with faster doubling times, but this trend tends to level off at population densities of above 400 people per  $km^2$ , and reverses slightly for death doubling times at densities above 1000 people per  $km^2$ . Perhaps because of features relating to the built environment i.e. building types, age structures, demographic or socio-economic conditions associated with wealthier city dwellers. However, in general, the relationship between population density and COVID-19 transmission is logical given the virus mainly transmits when humans are in close proximity to one another. Human population density also captures other important features of the built environment; for example, locations with high population density are cities or metropolitan areas, usually with high public transport usage, more recreational businesses like restaurants and bars, and indoor work-spaces like offices. All of which naturally bring people into closer contact and encourages airborne transmission of SARS-CoV-2.

Results also show that counties that rely on manufacturing or recreation as their main economic activity, also tend to have faster case doubling times. Again, this

is likely due to aspects of the work environment like the lack of proper physical distancing and ventilation. These findings are corroborated by studies [50, 51] that report many SARS-CoV-2 clusters were linked to a variety of indoor settings including households, hospitals, elderly care homes, and food processing plants (classed as factories). This concept is also further supported by our indicator representing household overcrowding, which is another strong predictor of case reporting doubling time. However, these variables are only significant in the infections model and not the mortality model. One possible explanation is that they represent transmission among younger people of working age, students, and younger families, who are less likely to die from COVID-19.

In terms of age population structure, having a higher proportion over 65-year-olds was also a significant predictor of faster death doubling times, concurrent with the literature and common understanding about the disease; age is one of the major risk factors. Major outbreaks have occurred in care homes [50] suggesting that some of the counties most affected by COVID-19 in the first wave of the epidemic was in locations with a higher proportion of retirees and care homes.

In terms of other socio-economic factors affecting the disease, poverty was also a significant predictor of faster doubling times in both case and mortality models. As mentioned in the conceptual framework, this can be explained since those who suffer from in-work poverty are likely to be doing jobs where it is difficult to work from home or adopt self-protective health behaviours such as social distancing [10]. Furthermore, even when suffering from symptoms, many low skilled workers and precarious workers may have been obliged to work because of a lack of sick pay, fear of losing a day's salary and pressures from bosses [19, 20, 21]. Poverty is also a risk factor of poor population health and is correlated with a multitude of underlying health conditions believed to lead to adverse outcomes for those suffering from COVID-19 [28]. This is further supported by the results of our final models; higher diabetes prevalence is also associated with faster case and death doubling times. Again, those suffering from diabetes are likely to suffer from comorbidities such as obesity and heart problems [52]. These results are also concurrent with work conducted by Williamson et al., 2020 [53], who found that greater age, deprivation, diabetes, severe asthma, and various other medical conditions were at higher risk of death due to COVID-19 infection. For both data-sets "Pop % with disabilities" tended to be correlated with slower doubling times. Although this group may be vulnerable to COVID-19 infections, they can often suffer from social isolation which provides some explanation. Furthermore, these groups are more likely to self-isolate [54, 55] to avoid infections.

### Environmental factors

Although a broad measure, the air quality index ("Max AQI") provides us with a way to proxy for counties with poor air quality and population-level pulmonary health conditions, caused by long-term exposure to harmful pollutants such as PM 2.5, PM10,  $NO_2$ ,  $SO_2$  and  $NO_x$ . This indicator is strongly correlated with COVID-19 infections and death doubling times, where higher AQI tends to speed up case and death reporting. This result is consistent with other observational studies [56, 57]. Some authors propose that air pollution increases infectivity, as SARS-CoV-2 binds with airborne particulate matter [58, 59, 60] allowing the disease to persist for longer in the air. Although this should not be ruled out, as mentioned, air quality

indicators also tend to proxy poor pulmonary health, which may increase death and case reporting, that is people with lung problems induced by air pollution are more likely to have symptomatic infections. It is well documented that long term exposure to certain pollutants has knock-on effects for people suffering from pulmonary viral infections [61, 62, 63, 64]. For example, a study by Soukup et al. [65] found that regulated inflammatory responses to viral infections are altered by exposure to PM10, potentially increasing the spread of infection and therefore increasing viral pneumonia-related hospital admissions.

In general, case and death reporting doubling times were negatively associated with temperature. There is increasing evidence that COVID-19 is a seasonal disease [66, 67], especially in temperate climates where there are distinct seasonal phases i.e. summer and winter, with distinct temperature ranges, distinct levels of ultra-violet radiation (UV) and seasonal differences in air moisture carrying capacity. Although, it is important not to rule out physical factors influencing transmission, especially for long-distance transmission, given the nature of the disease (transmission mainly takes place over short distances in closed spaces), the influence of weather on human behaviour is likely one of the major drivers of SARS-CoV-2 transmission. Weather is widely considered to influence people's behaviour [68] but research on this topic is surprisingly scant. According to Daniel et al., 2014 [69], people living in warmer / hotter locations, or during periods of warmer weather are more likely to employ a range of adaptive behaviours in response to warm and hot conditions i.e., keeping windows and doors open, use of wall and ceiling fans, air conditioning, which in turn may initiate a range of self-protective behaviours against SARS-CoV-2 transmission. Furthermore, warmer weather is also associated with recreational time spent outdoors [70] where SARS-CoV-2 transmission risk is likely to be lower. Although temperature also exhibited similar patterns for the death data model, it was only weakly statistically significant.

### 4.4.3 Limitations

Some of the limitations of the study are as follows. Since the study is limited to using aggregated data at the county level, we cannot make inferences about individual-level associations and cannot not adjust for individual-level risk factors e.g. age, gender, race, and occupation. However, that would be outside the scope of this study, since we were interested in macro ecological and socio-economic trends and drivers. Additionally, we cannot draw causal inference as the methodology we applied only reveals adjusted correlations. Therefore, results were carefully evaluated from individual-level and clinical-based studies to draw conclusions. The use of further explanatory variables would have surely improved the study i.e. on homelessness, availability of Intensive Care Units (ICU), quality of medical facilities, and ratio of medical staff per person, but these data were not available. It is also important to note that given the unprecedented nature and scale of COVID-19 outbreaks, data quality issues arise owing to the under-reporting of cases i.e., through under-diagnosis, lack of diagnostic tests and a lack of resources/time to carry out and implement mass testing. If data collection methods remained constant across counties over the time frame of this study, the calculation of doubling times can be a reliable measure. However, doubling times can be inflated by improving test-

ing procedures i.e., better detection and reporting through the availability of better diagnostic tests, better sampling techniques, resource allocation, and increased awareness of the disease.

## 4.5 Conclusions

This paper investigated drivers of epidemic growth during the first wave of outbreaks in US counties, by assessing the association between COVID-19 epidemic doubling times with socio-economic, demographic, environmental factors, and government containment measures. Results suggest that the main drivers of new infections are higher population density, home overcrowding, manufacturing and recreation industries and poverty. By contrast, warmer temperatures slowed epidemic growth which was likely to be the result of human behaviour responses to temperature. The main factors associated with death doubling times were age, poverty, air pollution and diabetes prevalence. Such findings help underpin current understanding of the disease epidemiology and also support current policy and advice recommending ventilation of homes, work-spaces and schools, along with social distancing and mask-wearing.

The results also suggest that states which adopted more stringent containment measures early on, did have some success at slowing the spread of the virus. There are numerous reports that there were huge failures at local level i.e. in care homes and business owners failing to protect residents and staff, by acting too slow or failing to implement control measures such as mask wearing and creating better ventilation in closed spaces [71, 72, 73]. The results also show that those counties with the highest percentages of people with certain underlying health conditions, age, and poverty were also those which had higher death doubling times. Protecting these groups early on with income support schemes could have allowed the working vulnerable to stay at home and avoid infection [74, 75]. Furthermore, home overcrowding was also a very important factor in case doubling times and a policy of providing a quarantine location for those infected with SAR-CoV-2 would have surely slowed epidemic growth [76].

Finally, while it is not clear where the next threat will come from, anthropogenic activity like deforestation, wildlife trade, and intensive animal rearing, that encourages spillover from wild reservoirs, and influences the emergence and evolution of novel coronaviruses [9, 77, 78] will continue to present risks globally until better controls and regulations can be implemented [79]. If new coronaviruses emerge, with similar modes of transmission, we should hope that governments can quickly apply top-down measures to suppress the virus before more sophisticated measures can be implemented i.e. rapid community testing to isolate the infected. I hope this work will contribute to the scholarly debate and can shed light on some of the environmental and socio-economic factors driving SAR-COV-2 transmission.

### 4.5.1 Abbreviations

GDP: Gross Domestic Product; US: United States of America.

## Bibliography

- [1] Elisabeth Mahase. Covid-19: What do we know about “long covid”? *BMJ*, page m2815, 07 2020.
- [2] Drew Altman. Understanding the us failure on coronavirus—an essay by drew altman. *BMJ*, page m3417, 09 2020.
- [3] A.C.K. Lee, P. English, B. Pankhania, and J.R. Morling. Where england’s pandemic response to covid-19 went wrong. *Public Health*, 192:45–48, 03 2021.
- [4] Clare Dyer. Covid-19: Uk government response was overcentralised and poorly communicated, say peers. *BMJ*, page m4445, 11 2020.
- [5] Chris Ham. The uk’s poor record on covid-19 is a failure of policy learning. *BMJ*, page n284, 02 2021.
- [6] Armin Nowroozpoor, Esther K. Choo, and Jeremy S. Faust. Why the united states failed to contain covid-19. *Journal of the American College of Emergency Physicians Open*, 1(4):686–688, 06 2020.
- [7] Dyani Lewis(a). Why many countries failed at covid contact-tracing — but some got it right. *Nature*, 588(7838):384–387, 12 2020.
- [8] Dyani Lewis(b). Is the coronavirus airborne? experts can’t agree. *Nature*, 580(7802):175–175, 04 2020.
- [9] Robert Barouki, Manolis Kogevinas, Karine Audouze, Kristine Belesova, Ake Bergman, Linda Birnbaum, Sandra Boekhold, Sebastien Denys, Celine Desseille, Elina Drakvik, Howard Frumkin, Jeanne Garric, Delphine Destoumieux-Garzon, Andrew Haines, Anke Huss, Genon Jensen, Spyros Karakitsios, Jana Klanova, Iida-Maria Koskela, Francine Laden, Francelyne Marano, Eva Franziska Matthies-Wiesler, George Morris, Julia Nowacki, Riikka Paloniemi, Neil Pearce, Annette Peters, Aino Rekola, Denis Sarigiannis, Katerina Šebková, Remy Slama, Brigit Staatsen, Cathryn Tonne, Roel Vermeulen, and Paolo Vineis. The covid-19 pandemic and global environmental change: Emerging research needs. *Environment International*, 146:106272, 01 2021.
- [10] Nicholas W. Papageorge, Matthew V. Zahn, Michèle Belot, Eline van den Broek-Altenburg, Syngjoo Choi, Julian C. Jamison, and Egon Tripodi. Socio-demographic factors associated with self-protecting behavior during the covid-19 pandemic. *Journal of Population Economics*, 34(2):691–738, 01 2021.
- [11] Jürgen Margraf, Julia Brailovskaia, and Silvia Schneider. Behavioral measures to fight covid-19: An 8-country study of perceived usefulness, adherence and their predictors. *PLOS ONE*, 15(12):e0243523, 12 2020.
- [12] Nancy H. L. Leung. Transmissibility and transmission of respiratory viruses. *Nature Reviews Microbiology*, 03 2021.

- [13] Renyi Zhang, Yixin Li, Annie L. Zhang, Yuan Wang, and Mario J. Molina. Identifying airborne transmission as the dominant route for the spread of covid-19. *Proceedings of the National Academy of Sciences*, 117(26):14857–14863, 06 2020.
- [14] Karolina Nissen, Janina Krambrich, Dario Akaberi, Tove Hoffman, Jiaxin Ling, Åke Lundkvist, Lennart Svensson, and Erik Salaneck. Long-distance airborne dispersal of sars-cov-2 in covid-19 wards. *Scientific Reports*, 10(1), 11 2020.
- [15] CDC. Social-distancing. 2020.
- [16] Julian W Tang, Linsey C Marr, Yuguo Li, and Stephanie J Dancer. Covid-19 has redefined airborne transmission. *BMJ*, page n913, 04 2021.
- [17] Simon Mongey, Laura Pilossoph, and Alex Weinberg. Which workers bear the burden of social distancing? Working Paper 27085, National Bureau of Economic Research, May 2020.
- [18] Marta Fana, Sergio Torrejón Pérez, and Enrique Fernández-Macías. Employment impact of covid-19 crisis: from short term effects to long terms prospects. *Journal of Industrial and Business Economics*, 47(3):391–410, 07 2020.
- [19] Margaret Whitehead, David Taylor-Robinson, and Ben Barr. Poverty, health, and covid-19. *BMJ*, page n376, 02 2021.
- [20] J.A. Patel, F.B.H. Nielsen, A.A. Badiani, S. Assi, V.A. Unadkat, B. Patel, R. Ravindrane, and H. Wardle. Poverty, inequality and covid-19: the forgotten vulnerable. *Public Health*, 183:110–111, 06 2020.
- [21] W. Holmes Finch and Maria E. Hernández Finch. Poverty and covid-19: Rates of incidence and deaths in the united states during the first 10 weeks of the pandemic. *Frontiers in Sociology*, 5, 06 2020.
- [22] Michael Schuit, Shanna Ratnesar-Shumate, Jason Yolitz, Gregory Williams, Wade Weaver, Brian Green, David Miller, Melissa Krause, Katie Beck, Stewart Wood, Brian Holland, Jordan Bohannon, Denise Freeburger, Idris Hooper, Jennifer Biryukov, Louis A Altamura, Victoria Wahl, Michael Hevey, and Paul Dabisch. Airborne sars-cov-2 is rapidly inactivated by simulated sunlight. *The Journal of Infectious Diseases*, 222(4):564–571, 06 2020.
- [23] K. H. Chan, J. S. Malik Peiris, S. Y. Lam, L. L. M. Poon, K. Y. Yuen, and W. H. Seto. The effects of temperature and relative humidity on the viability of the sars coronavirus. *Advances in Virology*, 2011:1–7, 2011.
- [24] P. M. de Oliveira, L. C. C. Mesquita, S. Gkantonas, A. Giusti, and E. Mastorakos. Evolution of spray and aerosol from respiratory releases: theoretical estimates for insight on viral transmission. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2245):20200584, 01 2021.
- [25] ECDC. Risk factors and risk groups. 2020.
- [26] WHO. Covid-19 and ncd risk factors. 2020.



- [27] CDC. Assessing risk factors for severe covid-19 illness. 2020.
- [28] CDC. People with certain medical conditions. 2021.
- [29] Adam Drewnowski and SE Specter. Poverty and obesity: the role of energy density and energy costs. *The American Journal of Clinical Nutrition*, 79(1):6–16, 01 2004.
- [30] Nathaniel M. Hawkins, Pardeep S. Jhund, John J.V. McMurray, and Simon Capewell. Heart failure and socioeconomic status: accumulating evidence of inequality. *European Journal of Heart Failure*, 14(2):138–146, 2012.
- [31] Juliet Addo, Luis Ayerbe, Keerthi M. Mohan, Siobhan Crichton, Anita Sheldenkar, Ruoling Chen, Charles D.A. Wolfe, and Christopher McKeivitt. Socioeconomic status and stroke. *Stroke*, 43(4):1186–1191, 04 2012.
- [32] E. Ward, A. Jemal, V. Cokkinides, G. K. Singh, C. Cardinez, A. Ghafoor, and M. Thun. Cancer disparities by race/ethnicity and socioeconomic status. *CA: A Cancer Journal for Clinicians*, 54(2):78–93, 03 2004.
- [33] Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, and Helen Tatlow. A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nature Human Behaviour*, 5(4):529–538, 03 2021.
- [34] CSSE. Johns hopkins university’s centre for systems science and engineering’s (csse) github repository. 2020.
- [35] USCB. United states census bureau. 2020.
- [36] Danielle; Taylor. Americans with disabilities: 2014. *Household Economic Studies*, 2014.
- [37] ERS. U.s. department of agriculture. 2020.
- [38] CDC. Diabetes data and statistics. 2019.
- [39] NCEI. Noaa’s national centers for environmental information (ncei). 2021.
- [40] EPA. Air data: Air quality data collected at outdoor monitors across the us. 2020.
- [41] DATA.CDC.GOV. U.s. state, territorial, and county stay-at-home orders: March 15-may 5 by county by day. 2020.
- [42] OxCGRT. Covid policy tracker. 2020.
- [43] Martin KrÄ¶ger and Reinhard Schlickeiser. Gaussian doubling times and reproduction factors of the covid-19 pandemic disease. *Frontiers in Physics*, 8, 07 2020.

- [44] Mark N Lurie, Joe Silva, Rachel R Yorlets, Jun Tao, and Philip A Chan. Coronavirus disease 2019 epidemic doubling time in the united states before and during stay-at-home restrictions. *The Journal of Infectious Diseases*, 222(10):1601–1606, 08 2020.
- [45] Kamalich Muniz-Rodriguez, Gerardo Chowell, Chi-Hin Cheung, Dongyu Jia, Po-Ying Lai, Yiseul Lee, Manyun Liu, Sylvia K. Ofori, Kimberly M. Roosa, Lone Simonsen, Cecile Viboud, and Isaac Chun-Hai Fung. Doubling time of the covid-19 epidemic by chinese province, 02 2020.
- [46] Lorenzo Pellis, Francesca Scarabel, Helena B. Stage, Christopher E. Overton, Lauren H. K. Chappell, Elizabeth Fearon, Emma Bennett, Katrina A. Lythgoe, Thomas A. House, Ian Hall, and . Challenges in control of covid-19: short doubling time and long delay to effect of interventions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1829):20200264, 05 2021.
- [47] Stephen A. Lauer, Kyra H. Grantz, Qifang Bi, Forrest K. Jones, Qulu Zheng, Hannah R. Meredith, Andrew S. Azman, Nicholas G. Reich, and Justin Lessler. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9):577–582, 05 2020.
- [48] Robert Verity, Lucy C Okell, Ilaria Dorigatti, Peter Winskill, Charles Whitaker, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick G T Walker, Han Fu, Amy Dighe, Jamie T Griffin, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Anne Cori, Zulma Cucunubá, Rich FitzJohn, Katy Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley, Daniel Laydon, Gemma Nedjati-Gilani, Steven Riley, Sabine van Elsland, Erik Volz, Haowei Wang, Yuanrong Wang, Xiaoyue Xi, Christl A Donnelly, Azra C Ghani, and Neil M Ferguson. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet Infectious Diseases*, 20(6):669–677, 06 2020.
- [49] Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.
- [50] Quentin J. Leclerc, Naomi M. Fuller, Lisa E. Knight, Sebastian Funk, Gwenan M. Knight, and . What settings have been linked to sars-cov-2 transmission clusters? *Wellcome Open Research*, 5:83, 06 2020.
- [51] John Middleton, Ralf Reintjes, and Henrique Lopes. Meat plants—a new front line in the covid-19 pandemic. *BMJ*, page m2716, 07 2020.
- [52] Herbert F Jelinek, Wael M Osman, Ahsan H Khandoker, Kinda Khalaf, Sungmun Lee, Wael Almahmeed, and Habiba S Alsafar. Clinical profiles, comorbidities and complications of type 2 diabetes mellitus in patients from united arab emirates. *BMJ Open Diabetes Research and Care*, 5(1):e000427, 08 2017.
- [53] Elizabeth J. Williamson, Alex J. Walker, Krishnan Bhaskaran, Seb Bacon, Chris Bates, Caroline E. Morton, Helen J. Curtis, Amir Mehrkar, David Evans, Peter Inglesby, Jonathan Cockburn, Helen I. McDonald, Brian MacKenna, Laurie Tomlinson, Ian J. Douglas, Christopher T. Rentsch, Rohini Mathur, Angel

- Y. S. Wong, Richard Grieve, David Harrison, Harriet Forbes, Anna Schultze, Richard Croker, John Parry, Frank Hester, Sam Harper, Rafael Perera, Stephen J. W. Evans, Liam Smeeth, and Ben Goldacre. Factors associated with covid-19-related death using opensafely. *Nature*, 584(7821):430–436, 07 2020.
- [54] Stephen J. Macdonald, Lesley Deacon, Jackie Nixon, Abisope Akintola, Anna Gillingham, Jacqueline Kent, Gillian Ellis, Debbie Mathews, Abolaji Ismail, Sylvia Sullivan, Samouka Dore, and Liz Highmore. ‘the invisible enemy’: disability, loneliness and isolation. *Disability and Society*, 33(7):1138–1159, 08 2018.
- [55] Eric Emerson, Nicola Fortune, Gwynnyth Llewellyn, and Roger Stancliffe. Loneliness, social support, social isolation and wellbeing among working age adults with and without disability: Cross-sectional study. *Disability and Health Journal*, 14(1):100965, 01 2021.
- [56] Matthew A. Cole, Ceren Ozgen, and Eric Strobl. Air pollution exposure and covid-19 in dutch municipalities. *Environmental and Resource Economics*, 76(4):581–610, 08 2020.
- [57] Marco Travaglio, Yizhou Yu, Rebeka Popovic, Liza Selley, Nuno Santos Leal, and Luis Miguel Martins. Links between air pollution and covid-19 in england. *Environmental Pollution*, 268:115859, 01 2021.
- [58] Norefrina Shafnaz Md Nor, Chee Wai Yip, Nazlina Ibrahim, Mohd Hasni Jaafar, Zetti Zainol Rashid, Norlaila Mustafa, Haris Hafizal Abd Hamid, Kuhan Chandru, Mohd Talib Latif, Phei Er Saw, Chin Yik Lin, Kemal Maulana Alhasa, Jamal Hisham Hashim, and Mohd Shahrul Mohd Nadzir. Particulate matter (pm<sub>2.5</sub>) as a potential sars-cov-2 carrier. *Scientific Reports*, 11(1), 01 2021.
- [59] Simone Lolli, Ying-Chieh Chen, Sheng-Hsiang Wang, and Gemine Vivone. Impact of meteorological conditions and air pollution on covid-19 pandemic transmission in italy. *Scientific Reports*, 10(1), 10 2020.
- [60] Angelo Solimini, F. Filipponi, D. Alunni Fegatelli, B. Caputo, C. M. De Marco, A. Spagnoli, and A. R. Vestri. A global association between covid-19 cases and airborne particulate matter at regional level. *Scientific Reports*, 11(1), 03 2021.
- [61] Anoop J Chauhan and Sebastian L Johnston. Air pollution and infection in respiratory illness. *British Medical Bulletin*, 68(1):95–112, 12 2003.
- [62] Daniel P Croft, Wangjian Zhang, Shao Lin, Sally W Thurston, Philip K Hopke, Mauro Masiol, Stefania Squizzato, Edwin van Wijngaarden, Mark J. Utell, and David Q Rich. The association between respiratory infection and air pollution in the setting of air quality policy and economic change. *Annals of the American Thoracic Society*, 11 2018.
- [63] Jonathan Grigg. Air pollution and respiratory infection: An emerging and troubling association. *American Journal of Respiratory and Critical Care Medicine*, 198(6):700–701, 09 2018.

- [64] Kipruto Kirwa, Carly M Eckert, Sverre Vedal, Anjum Hajat, and Joel D Kaufman. Ambient air pollution and risk of respiratory infection among adults: evidence from the multiethnic study of atherosclerosis (mesa). *BMJ Open Respiratory Research*, 8(1):e000866, 03 2021.
- [65] Susanne Becker, Joleen M. Soukup. Exposure to urban air particulates alters the macrophage-mediated inflammatory response to respiratory viral infection. *Journal of Toxicology and Environmental Health, Part A*, 57(7):445–457, 07 1999.
- [66] Adam Kaplin, Caesar Junker, Anupama Kumar, Mary Anne Ribeiro, Eileen Yu, Michael Wang, Ted Smith, Shesh N. Rai, and Aruni Bhatnagar. Evidence and magnitude of the effects of meteorological changes on sars-cov-2 transmission. *PLOS ONE*, 16(2):e0246167, 02 2021.
- [67] Yeon-Woo Choi, Alexandre Tuel, and Elfatih A. B. Eltahir. On the environmental determinants of covid-19 seasonality. *GeoHealth*, 5(6), 05 2021.
- [68] C. R. de Freitas. Weather and place-based human behavior: recreational preferences and sensitivity. *International Journal of Biometeorology*, 59(1):55–63, 04 2014.
- [69] Lyrian Daniel. ‘we like to live in the weather’: Cooling practices in naturally ventilated dwellings in darwin, australia. *Energy and Buildings*, 158:549–557, 01 2018.
- [70] Mathieu Bélanger, Katherine Gray-Donald, Jennifer O’loughlin, Gilles Paradis, and James Hanley. Influence of weather conditions and season on physical activity in adolescents. *Annals of Epidemiology*, 19(3):180–186, 03 2009.
- [71] Desmond O’Neill. Covid-19 in care homes: the many determinants of this perfect storm. *BMJ*, page m2096, 05 2020.
- [72] Susan Chapman and Charlene Harrington. Policies matter! factors contributing to nursing home outbreaks during the covid-19 pandemic. *Policy, Politics, and Nursing Practice*, 21(4):191–192, 10 2020.
- [73] David C. Grabowski and Vincent Mor. Nursing home care in crisis in the wake of covid-19. *JAMA*, 324(1):23, 07 2020.
- [74] Sharoda Dasgupta, Virginia B. Bowen, Andrew Leidner, Kelly Fletcher, Trieste Musial, Charles Rose, Amy Cha, Gloria Kang, Emilio Dirlikov, Eric Pevzner, Dale Rose, Matthew D. Ritchey, Julie Villanueva, Celeste Philip, Leandris Liburd, and Alexandra M. Oster. Association between social vulnerability and a county’s risk for becoming a covid-19 hotspot — united states, june 1–july 25, 2020. *MMWR. Morbidity and Mortality Weekly Report*, 69(42):1535–1541, 10 2020.
- [75] Chiao-yu Yang, Katharine Briar-Lawson, and Jildyz Urbaeva. Employment and income support policies during the early phases of covid-19: Lessons from the u.s., denmark, and taiwan. *Greenwich Social Work Review*, 1(2):97–108, 12 2020.

- 
- [76] Shamil Haroon, Joht Singh Chandan, John Middleton, and Kar Keung Cheng. Covid-19: breaking the chain of household transmission. *BMJ*, page m3181, 08 2020.
  - [77] Toph Allen, Kris A. Murray, Carlos Zambrana-Torrel, Stephen S. Morse, Carlo Rondinini, Moreno Di Marco, Nathan Breit, Kevin J. Olival, and Peter Daszak. Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications*, 8(1), 10 2017.
  - [78] Maya Wardeh, Matthew Baylis, and Marcus S. C. Blagrove. Predicting mammalian hosts in which novel coronaviruses can be generated. *Nature Communications*, 12(1), 02 2021.
  - [79] Andrew P. Dobson, Stuart L. Pimm, Lee Hannah, Les Kaufman, Jorge A. Ahumada, Amy W. Ando, Aaron Bernstein, Jonah Busch, Peter Daszak, Jens Engelmann, Margaret F. Kinnaird, Binbin V. Li, Ted Loch-Temzelides, Thomas Lovejoy, Katarzyna Nowak, Patrick R. Roehrdanz, and Mariana M. Vale. Ecology and economics for pandemic prevention. *Science*, 369(6502):379–381, 07 2020.

# Chapter 5

## Synthesis

In this thesis, I set out to investigate some of the broader questions in epidemiology, which have only recently become possible because of advances in modern computing. In particular, I examined how aspects of climate, environment, socio-demographic conditions, and political factors act together to influence the spread and establishment of infectious diseases on a macro-scale.

The research carried out for this thesis shows that climate is one of the main factors affecting disease distribution and intensity. However, the impact of increasing temperatures differs across diseases: while warmer climates are positively associated with the spread of the two tropical vector-borne diseases investigated in this thesis, which is a fairly well established scientific fact and concurrent with studies such as [10, 22], higher temperatures appear to act as a limiting factor when it comes to the spread of SARS-CoV-2 [15, 21, 8]. More specifically, the work in Chapter 2, examining the distribution of dengue in Mexico and the United States, shows how the disease is most prevalent in tropical areas of Mexico, where temperatures are generally milder during the coldest parts of the year compared to more temperate regions, and there is less seasonal variation and more consistent rainfall. The results also suggest that winter temperatures may be limiting the spread of dengue into more northerly US states since the *Aedes* mosquito's species, and the virus do not overwinter in very cold temperatures. However, we could expect dengue to shift further north with increasing winter temperatures due to climate change. In regions where dengue is endemic in lower-lying areas (particularly in Mexico), increasingly warmer weather at higher altitudes may also allow the virus and vectors to push up into mountainous areas where the disease is not currently endemic. Results from Chapter 3 also find that climate is also an important predictor of West Nile infections. Generally, warmer temperatures and more consistent rainfall in the summer months affect the disease positively. Increasingly warmer temperatures can in part explain why the virus has expanded in Europe, especially to more northerly regions. Drought and the shrinking of freshwater sources, also influenced by global climate change also appears to be affecting the intensity of summer outbreaks, by increasing contact between wetland bird species and mosquitoes and therefore increasing prevalence and potential spillover to the human population. This finding is also corroborated by local-level studies [20, 5, 12], and one macro-scale [19] study. The results from Chapter 4 also found that there is some temperature dependence in the human-to-human transmission of SARS-CoV-2, although warmer appears to slow down transmission. These results are supported by the findings from other empirical

studies [15, 21, 8], although results there are studies that report little to no impact of temperature on SARS-COV-2 transmission( [11], UJIIE2020301 ). Currently, the predominant theory is that warmer temperatures allow humans to adopt intentional or unintentional behaviours that protect them against SARS-COV-2 infection, through better ventilating buildings and carrying out certain activities outdoor, e.g., socialising, cultural events. Although evidence for this is scant and more mechanistic and behavioural modelling (physical, human behaviour) needs to be conducted to validate these theories along with the empirical findings just mentioned. When focusing on the impact of socio-economic factors on the spread of the diseases investigated, results from chapters 2 and 4 suggest that poverty and social inequality, along with demographic and human behavioural factors, may also influence the intensity of transmission of dengue viruses (DENVs) and SARS-COV-2. For DENVs, infections are more prevalent in poorer regions with low income and low education. These findings are in line with studies such as [13, 2] which look at the relationship between population health, income and government spending. Furthermore, areas with higher population growth and more mobility were also predictors of a higher incidence of dengue. Such findings are particularly relevant since dengue has recently re-emerged in Europe and can help identify which communities may be most at risk, for instance, those doing manual outdoor work such as agricultural workers. As for socio-economic/demographic factors influencing SARS-COV-2, metropolitan areas with major transport hubs were hit hardest by the virus during the first wave of the epidemic in the US. This can be explained as the virus is likely to have first entered the US via these locations which tend to attract international travellers and import goods from abroad, as explored by [18]. The virus then tended to spread in areas where work is carried out indoors, such as factories, hospitals, eateries, and where the population is denser, especially in cities, which tends to attract more people for work and pleasure, hence bringing them into close contact. Socio-economic factors like poverty and social inequality created through job insecurity were also likely to play a role in disease transmission. Certainly, those people in lower-skilled jobs such as cleaners, bus drivers, shop keepers, factory workers may have been more exposed to the virus since they were unable to work from home. Workers without proper workers' rights were also less likely to break the transmission cycle because of lack of sick pay, therefore increasing the chance of transmitting the virus to colleagues [16, 14, 6, 24, 17, 7]. However, it is important to note that all front-line workers, even those in better socio-economic situations such as medical workers and teachers would have been more exposed to the virus. Perhaps more importantly, as revealed in Chapter 4, death doubling times in the poorest areas were much higher than in more affluent areas, which is largely due to the health problems that occur in poorer populations [3, 4, 9, 1, 23]. Housing overcrowding also appeared to be a driver of new infections; those living in more cramped conditions may have found it difficult to isolate when infected and raised the likelihood of spreading the virus to members of their household. Another aspect considered is how political will and decisions (or lack of them) can affect the spread of diseases, in this case, WNV and SARS-COV-2. Results in Chapter 3 revealed strong associations between places hardest hit by the economic crisis and those that had the highest prevalence of WNV infections. Indeed, WNV became a serious public health issue in Europe during a period of severe economic decline and cuts to specific sectors of government. Government spending in areas, such as wastewater management, environment, and health, suffered se-

vere budget cuts, and these political decisions are likely to be the drivers and/or moderators of WNV infections. Similarly, I investigated whether control measures implemented by state governments were able to reduce SARS-COV-2 transmission. Although due to the rapidness and scale of outbreaks, public health systems were initially overwhelmed, broke down and unable to function at an adequate capacity to contain the virus; states that applied stricter measures early on had some success in suppressing the virus. These results show how top-down measures can improve or worsen disease outcomes, demonstrating the importance of how political decisions made by a few can have serious impacts on the wider population.

### 5.0.1 Caveats and Limitations

Since all three studies relied on a similar study design, statistical methods and data, the major limitations of each study are comparable:

- Sourcing, merging, and analysing of data at relevant scales: Although environmental data drawn from satellite images are generally provided at spatial-temporal resolutions suitable for studying disease transmission mechanisms, health data are only provided as aggregated areal data, which represents the number of infections and deaths per geographical (political) boundary. By using geographical boundaries as our unit of analysis we lose important location-based information which would allow us to better assess the contribution of environmental factors (e.g. land use, housing type) to disease transmission risk. Important information is also lost on individual variability such as health status, age, or any genetic predispositions and we cannot make inferences about individuals based on aggregated areal data.
- Variation in sampling effort across regions since data for each variable can be collected from several sources. Knowledge gaps in the geographical range of hosts, vectors and human cases of the disease can also overstate or understate the real drivers of disease. Therefore, findings need to be carefully evaluated with local-level studies, especially those that look at mechanisms and causation.
- Sourcing data that represented certain ecological, environmental, political, and socio-demographic factors identified in the conceptual frameworks was not possible, and therefore I had to use proxies to represent these factors in the models, this meant making strong assumptions about certain relationships. For example, for chapters 2 and 3 data, since regional mosquito population abundance data was not available, I assumed that mosquito abundance directly influences disease transmission, I also consulted secondary literature to define proxies that could be used to represent drivers of mosquito abundance. Similarly, since no data was readily available on mosquito control policies at the scales studied, I made the assumption that poorer regions or those regions suffering some form of austerity would have had fewer control measures in place and/or would have had the poor or declining infrastructure, such as potholes and poor sanitation which would have thus benefited the spread of mosquitoes. To gather such data on mosquito population abundance or regional mosquito control measures would require an effort far beyond the object of this thesis.



### 5.0.2 Future Outlook

One of the most effective ways to further this research would be to upscale the unit of analysis for each study, by incorporating individually geo-referenced health outcome data into the models. The level of aggregation in the studies makes it harder to infer causation and elucidate diseases transmission mechanisms; similarly, the lack of individual-level data does not allow to control for individual factors affecting the likelihood to get the disease, such as age, and underlying health conditions. Such datasets are not publicly available or are very difficult to obtain for researchers not connected to universities that have access to governmental data. In my personal experience, data requests were either denied or ignored. Given the potential social and economic damage caused by infectious diseases (not to mention human suffering), foresight is needed by governments and health authorities to implement better data collection strategies to help improve research. One initial solution would be for science funders to create new initiatives and harmonise data collection protocols and develop specialised global databases for important diseases, this would require a closer integration between the academic sector and health and government authorities. Data could easily be presented at more refined and biologically relevant scales without revealing personal information about study subjects. For example, geo-coordinates or postcodes could be obfuscated at a scale that does not identify street names but provides more refined spatial information. The contact tracing data collection initiatives that have been recently set up by governments and health authorities to keep track of COVID-19 cases could be used as infrastructure to build on and extended to other infectious diseases. In general, a focus on providing more refined data would open up more useful lines of enquiry and could help us understand how disease transmission is dynamically affected through interactions between environmental and human socio-demographic and behavioural factors at macro-scales.

### 5.0.3 Conclusions

This work is relevant as it adds to the growing body of scientific literature focusing on infectious diseases, taking more holistic approaches and harnessing big data to understand under which environmental and social conditions some populations become more exposed and burdened by infectious diseases than others. The research tackles some of the broader and less explored areas of public health and epidemiology, such as analysing economic changes with environmental changes, examining the impacts of factors such as austerity on health, along with other factors such as political decision making and or lack of intervention by government authorities. Such work is especially important when considering the multiple threats brought about by climate change and other anthropogenic-induced changes that can benefit emerging diseases, i.e., global trade in wild animals, intensive agriculture/animal rearing, and land-use conversion.

#### General recommendations

To better tackle infectious diseases, we need to look more closely at how human activity interacts with the natural environment, and how general neglect of the population, environment and infrastructure at a political level can consequently jeopardise

human health across the globe. Particular attention should be paid by governments when making spending reductions to health and environmental protection, particularly during economic crises or economic downturns. More collaboration between academic sectors and governments internationally needs to take place to provide better databases for researchers, enabling them to investigate disease transmission mechanisms and subsequently develop effective control measures and interventions strategies. The public debate needs to take place on how to allocate more resources to tackle infectious diseases in all parts of the world: there needs recognition that infectious disease in one part of the world is potentially a problem for everyone. Finally, we need to immediately act towards lowering carbon emissions to restrict extreme changes in the climate given this may erode some of the gains we have made over the past century in terms of poverty reduction, human health, welfare, and food security. More work needs to go into educating the public on such issues, showing the link between climate change, environmental degradation, and infectious disease emergence, and spread. This could then push these issues up the political agenda, which may lead to more inter-governmental investment in initiatives that tackle neglected and emerging infectious diseases worldwide.

## Bibliography

- [1] Juliet Addo, Luis Ayerbe, Keerthi M. Mohan, Siobhan Crichton, Anita Sheldenkar, Ruoling Chen, Charles D.A. Wolfe, and Christopher McKevitt. Socioeconomic status and stroke. *Stroke*, 43(4):1186–1191, 04 2012.
- [2] A. Baumbach and G. Gulis. Impact of financial crisis on selected health outcomes in europe. *The European Journal of Public Health*, 24(3):399–403, 04 2014.
- [3] CDC. People with certain medical conditions. 2021.
- [4] Adam Drewnowski and SE Specter. Poverty and obesity: the role of energy density and energy costs. *The American Journal of Clinical Nutrition*, 79(1):6–16, 01 2004.
- [5] Paul R Epstein and Caroline Defilippo. West nile virus and drought. *Global change and Human health*, 2(2):105–107, 2001.
- [6] Marta Fana, Sergio Torrejón Pérez, and Enrique Fernández-Macías. Employment impact of covid-19 crisis: from short term effects to long terms prospects. *Journal of Industrial and Business Economics*, 47(3):391–410, 07 2020.
- [7] W. Holmes Finch and Maria E. Hernández Finch. Poverty and covid-19: Rates of incidence and deaths in the united states during the first 10 weeks of the pandemic. *Frontiers in Sociology*, 5, 06 2020.
- [8] Syed Emdadul Haque and Mosiur Rahman. Association between temperature, humidity, and covid-19 outbreaks in bangladesh. *Environmental Science and Policy*, 114:253–255, 2020.

- [9] Nathaniel M. Hawkins, Pardeep S. Jhund, John J.V. McMurray, and Simon Capewell. Heart failure and socioeconomic status: accumulating evidence of inequality. *European Journal of Heart Failure*, 14(2):138–146, 2012.
- [10] Peter J. Hotez. Southern europe’s coming plagues: Vector-borne neglected tropical diseases. *PLOS Neglected Tropical Diseases*, 10(6):e0004243, 2016.
- [11] Tahira Jamil, Intikhab Alam, Takashi Gojobori, and Carlos M. Duarte. No evidence for temperature-dependence of the covid-19 epidemic. *Frontiers in Public Health*, 8, 2020.
- [12] B. J. Johnson and M.V.K. Sukhdeo. Drought-Induced Amplification of Local and Regional West Nile Virus Infection Rates in New Jersey. *Journal of Medical Entomology*, 50(1):195–204, 01 2013.
- [13] M. Karanikolos, P. Mladovsky, J. Cylus, S. Thomson, S. Basu, D. Stuckler, J. P. Mackenbach, and M. McKee. Financial crisis, austerity, and health in europe. *Lancet*, 381(9874):1323–31, 2013.
- [14] Simon Mongey, Laura Pilossoph, and Alex Weinberg. Which workers bear the burden of social distancing? Working Paper 27085, National Bureau of Economic Research, May 2020.
- [15] Alessio Notari. Temperature dependence of covid-19 transmission. *Science of The Total Environment*, 763:144390, 2021.
- [16] Nicholas W. Papageorge, Matthew V. Zahn, Michèle Belot, Eline van den Broek-Altenburg, Syngjoo Choi, Julian C. Jamison, and Egon Tripodi. Socio-demographic factors associated with self-protecting behavior during the covid-19 pandemic. *Journal of Population Economics*, 34(2):691–738, 01 2021.
- [17] J.A. Patel, F.B.H. Nielsen, A.A. Badiani, S. Assi, V.A. Unadkat, B. Patel, R. Ravindrane, and H. Wardle. Poverty, inequality and covid-19: the forgotten vulnerable. *Public Health*, 183:110–111, 06 2020.
- [18] Rohan Patil, Raviraj Dave, Harsh Patel, Viraj M Shah, Deep Chakrabarti, and Udit Bhatia. Assessing the interplay between travel patterns and sars-cov-2 outbreak in realistic urban setting. *Applied network science*, 6(1):1–19, 2021.
- [19] Sara H. Paull, Daniel E. Horton, Moetasim Ashfaq, Deeksha Rastogi, Laura D. Kramer, Noah S. Dittenbaugh, and A. Marm Kilpatrick. Drought and immunity determine the intensity of west nile virus epidemics and climate change impacts. *Proceedings of the Royal Society B: Biological Sciences*, 284(1848):20162078, 02 2017.
- [20] Jeffrey Shaman, Jonathan F. Day, and Marc Stieglitz. Drought-Induced Amplification and Epidemic Transmission of West Nile Virus in Southern Florida. *Journal of Medical Entomology*, 42(2):134–141, 03 2005.
- [21] Peng Shi, Yinqiao Dong, Huanchang Yan, Chenkai Zhao, Xiaoyang Li, Wei Liu, Miao He, Shixing Tang, and Shuhua Xi. Impact of temperature on the dynamics of the covid-19 outbreak in china. *Science of The Total Environment*, 728:138890, 2020.

- [22] Rachel Tidman, Bernadette Abela-Ridder, and Rafael Ruiz de Castañeda. The impact of climate change on neglected tropical diseases: a systematic review. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 115(2):147–168, 2021.
- [23] E. Ward, A. Jemal, V. Cokkinides, G. K. Singh, C. Cardinez, A. Ghafoor, and M. Thun. Cancer disparities by race/ethnicity and socioeconomic status. *CA: A Cancer Journal for Clinicians*, 54(2):78–93, 03 2004.
- [24] Margaret Whitehead, David Taylor-Robinson, and Ben Barr. Poverty, health, and covid-19. *BMJ*, page n376, 02 2021.

# Appendices

# Appendix A

## Influence of socio-economic, demographic and climate factors on the regional distribution of dengue in the United States and Mexico

### A.1 Data availability

Availability of data and materials The R project folder, main spatial dataset and R code for the project is available from <https://doi.org/10.5281/zenodo.887909>.

### A.2 Vector and Environmental Relationships: Regression Analysis (Step 1)

To assess the relationships between vector and environment, techniques were adapted from [58]. Since our species distribution data set only consisted of presence data, absence data was substituted with background data. Background data points were sampled randomly from the study area, matching roughly the number of observations in the presence data. (Mex/US). Using background data allows us to characterise environments in the study region, which establishes the environmental domain of the study, whilst presence data should represent the conditions a species is more likely to be present than on average. Since vector presence is indicated by a binary variable equal to 1 and absence equal to 0, a relationship between environmental and vector presence was estimated with a logit model by maximum likelihood. Logit estimation techniques are based on the assumption that there is a latent variable  $y$  and that this latent variable is a linear function of all the explanatory variables. Climate data for the species distribution prediction modelling were sourced from MERRAclim [52]. This data-set was built using 2m air temperature (Kelvin degrees) and 2 m specific humidity (kg of water/kg of air) hourly data derived from satellite observations from the Modern Era Retrospective Analysis for Research and Applications Reanalysis. Tables A1.1 and A1.2 provide summary statistics for these data sets.

In order to predict if a vector was present in a given location  $i$ , the following equations were used

$$Pr(Aedes.aegypti = 1) = \beta_1 Tvar1_i + \beta_2 Tvar2_i + \beta_3 Pvar1_i + \epsilon$$

$$Pr(Aedes.albopictus = 1) = \beta_1 Tvar1_i + \beta_2 Tvar2_i + \beta_3 Pvar1_i + \epsilon$$

where  $X_i$  is a vector of regressors/independent variables (in this case: Tvar 1 represents temperature annual range; Tvar2 represents mean temperature of the coldest quarter; Pvar1 represents precipitation of the driest month and *epsilon* is the error term.

### A.3 Vector and Environmental Relationships: regression analysis results

The impact of climate on the probability of a vector being present was assessed by running a GAM logit regression. Results are reported in Table A1.3 and model diagnostics in figures A1.1 and A1.2 .

All three predictors were highly significant for both *A. aegypti* and *A. albopictus*. The results from the GAM logistic regressions the study confirmed a positive and highly significant association between some climatic factors and vector presence.

*A. albopictus* seems to be distributed in environments that are warmer and wetter during the hotter months, *A. aegypti* seems to be sensitive to colder temperatures and temperature range.

For both species minimum temperature is a major limiting factor affecting distribution and this is conclusive with our results.

The models were validated using an assessment of the residuals and using a 5-fold cross-validation of the operating characteristic (ROC) curve, which compared false and true positives (see figures A1.1-A1.2 ). The data set was divided into 5 (k-5) subsets, and the model run 5 times. For each run, one of the k subsets is used as the test the rate of false positives and true positives. Both models demonstrated high accuracy in terms of predicting true positives with values not falling under 0.87 for *A. aegypti* and 0.89 for *A. albopictus*.

Table A1.1: *A. aegypti* climate variable summary

Statistic	N	Min	Max	Mean	St. Dev.
Temp.Annual.Range	8,584	4.600	55.200	23.880	8.032
Mean.Temp.Coldest.Quarter	8,584	-7.600	30.700	23.470	3.438
Precip.of.Driest.Month	8,584	348	2,493	1,488.795	321.253

Table A1.2: *A. albopictus* climate variable summary

Statistic	N	Min	Max	Mean	St. Dev.
Temp.Annual.Range	7,425	46	588	289.262	96.656
Mean.Temp.Coldest.Quarter	7,425	−83	307	191.324	65.688
Precip.of.Driest.Month	7,425	267	2,493	1,326.908	317.151

	<i>A.aegypti</i>	<i>A.albopictus</i>
Intercept	−2.14*** (0.51)	−6.56*** (0.67)
Temperature annual range	2.04** (2.26)	2.78*** (2.96)
Mean temperature of coldest quarter	2.12*** (2.32)	2.99*** (3.00)
Precipitation of driest month	2.83*** (2.97)	2.98*** (3.00)
AIC	13478.91	11510.16
BIC	13541.08	11584.66
Log Likelihood	−6731.47	−5745.33
Deviance	13462.94	11490.65
Deviance explained	0.45	0.46
Dispersion	1.00	1.00
R <sup>2</sup>	0.52	0.52
GCV score	−0.24	−0.25
Num. obs.	17798	15326
Num. smooth terms	3	3

\*\*\*  $p < 0.001$ ; \*\*  $p < 0.01$ ; \*  $p < 0.05$ Table A1.3: *Aedes* GAM SDM results

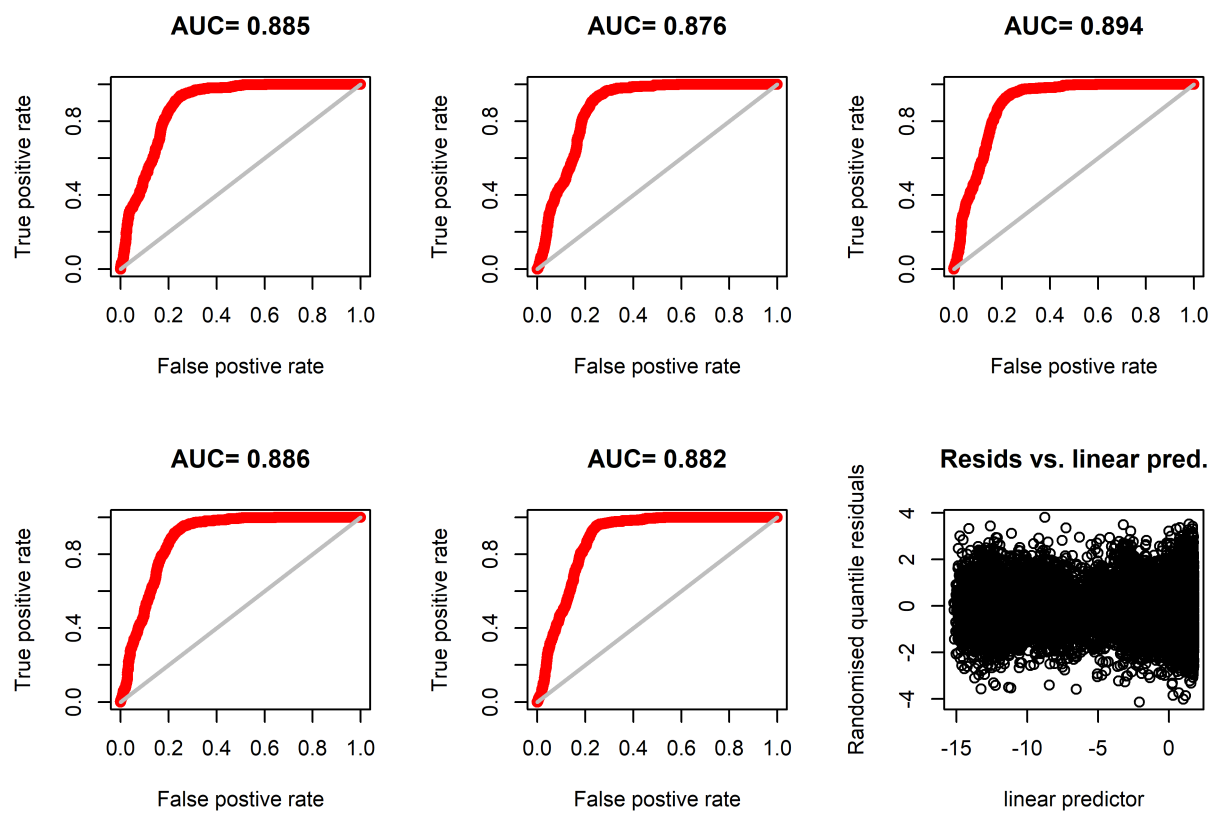


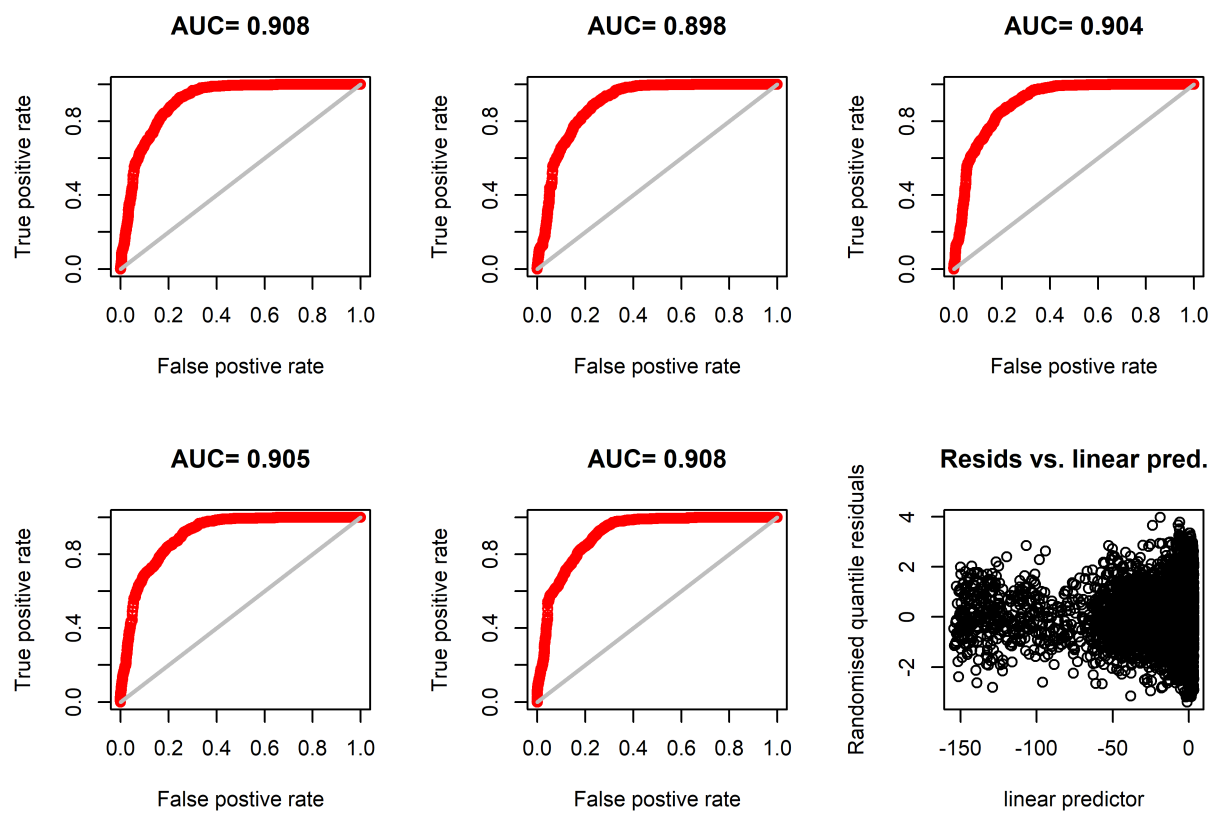
	NB US/Mex	TW US/Mex	QP US/Mex	NB Mex	TW Mex	QP Mex
(Intercept)	-9.44*** (0.26)	2.22*** (0.20)	-9.12*** (0.26)	-9.20*** (0.18)	2.44*** (0.18)	-8.75*** (0.21)
EDF: s(socio_economic_index_norm1)	1.65*** (1.82)	1.78** (1.92)	1.92 (1.97)			
EDF: s(BB_ACC)	1.95*** (1.99)	1.95*** (1.99)	1.88** (1.97)	1.92*** (1.99)	1.92*** (1.99)	1.80 (1.94)
EDF: s(DOC_RA)	1.68* (1.88)	1.88*** (1.98)	1.95*** (1.99)	1.34** (1.54)	1.81*** (1.95)	1.95*** (1.99)
EDF: s(FLOWMOB_ALL_RA)	1.00 (1.00)	1.61 (1.82)	1.57 (1.77)	1.00* (1.00)	1.00 (1.00)	1.00 (1.00)
EDF: s(QOL_index)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
EDF: s(POP_DEN_GR)	1.84*** (1.95)	1.00** (1.01)	1.28* (1.44)	1.00 (1.00)	1.00 (1.00)	1.01 (1.02)
EDF: s(Y65_MAX)	1.92*** (1.96)	1.79** (1.89)	1.15* (1.24)			
EDF: s(bio11)	1.77*** (1.92)	1.83*** (1.95)	1.68*** (1.86)	1.87*** (1.97)	1.89*** (1.98)	1.00*** (1.00)
EDF: s(bio18)	1.65* (1.86)	1.53* (1.76)	1.00 (1.00)	1.86* (1.97)	1.71 (1.90)	1.00 (1.00)
EDF: s(fyear)	6.79*** (8.00)	6.61*** (8.00)	6.67*** (8.00)	<b>5.97***</b> (8.00)	6.44*** (8.00)	6.46*** (8.00)
EDF: s(factor(ADM1_CODE))	11.54*** (12.97)	10.45*** (12.19)	9.30*** (11.22)	12.35*** (13.56)	12.13*** (13.47)	10.32*** (12.31)
EDF: s(INCOME_PRIM)				1.84* (1.96)	1.83* (1.96)	1.93** (1.99)
EDF: s(ROOMS_PC)				1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
EDF: s(EDU38_SH)				1.97*** (2.00)	1.95*** (1.99)	1.78 (1.92)
AIC	3989.14	2231.57	—	3866.02	2204.83	—
BIC	4130.53	2373.21	—	4006.05	2350.39	—
Log Likelihood	-1956.60	-1077.75	—	-1894.78	-1062.68	—
Deviance	332.45	962.59	179965.91	303.55	859.68	169568.80
Deviance explained	0.61	0.66	0.75	0.62	0.67	0.70
Dispersion	1.00	2.83	881.70	1.00	2.73	957.09
R <sup>2</sup>	0.09	0.51	0.57	0.09	0.51	0.58
GCV score	2011.57	1127.23	1180.27	1950.50	1119.74	1117.09
Num. obs.	306	306	306	288	288	288
Num. smooth terms	11	11	11	12	12	12

\*\*\* $p < 0.001$ ; \*\* $p < 0.01$ ; \* $p < 0.05$ 

Table A1.4: Model diagnostics - distributions

## A.4 Diagnostics

Figure A1.1: ROC + residual check *A.egypti* model

Figure A1.2: ROC + residual check *A.albopictus* model

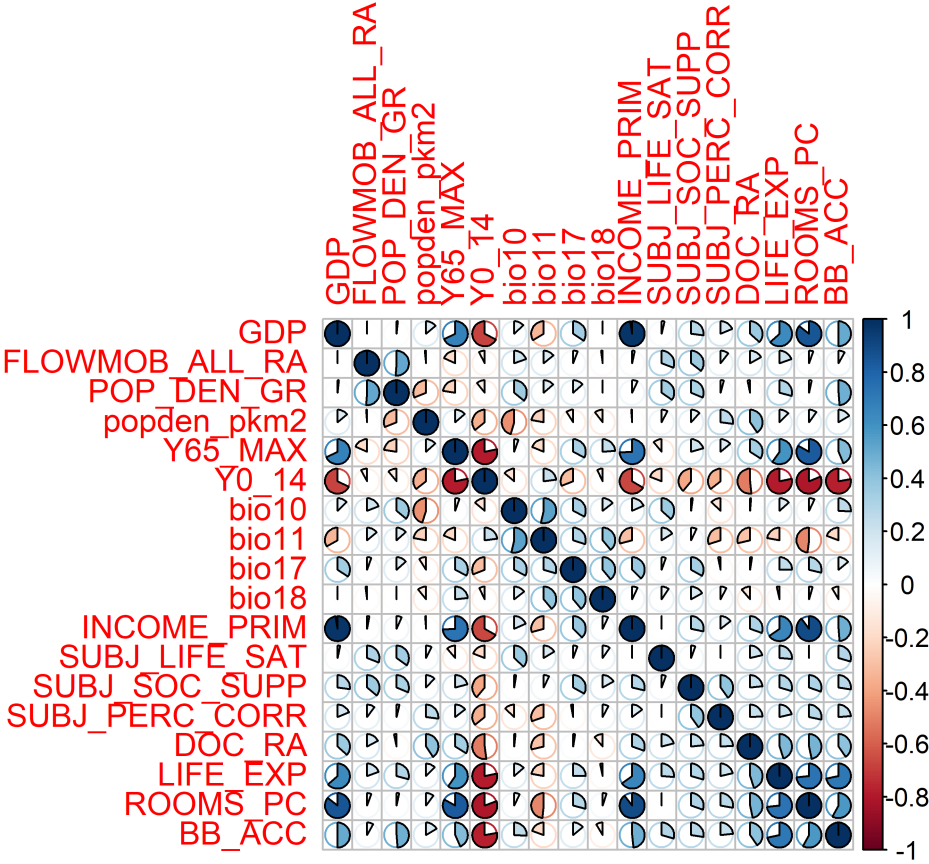


Figure A1.3: Correlation matrix: US/Mex dataset

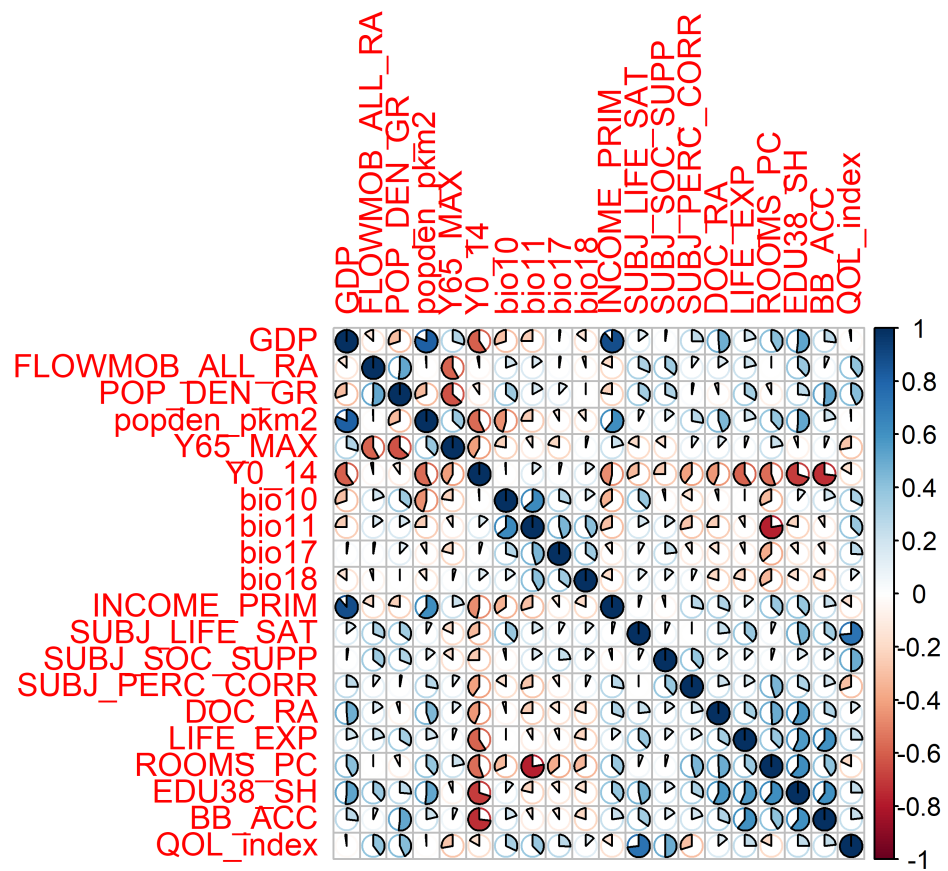


Figure A1.4: Correlation matrix: US/Mex dataset

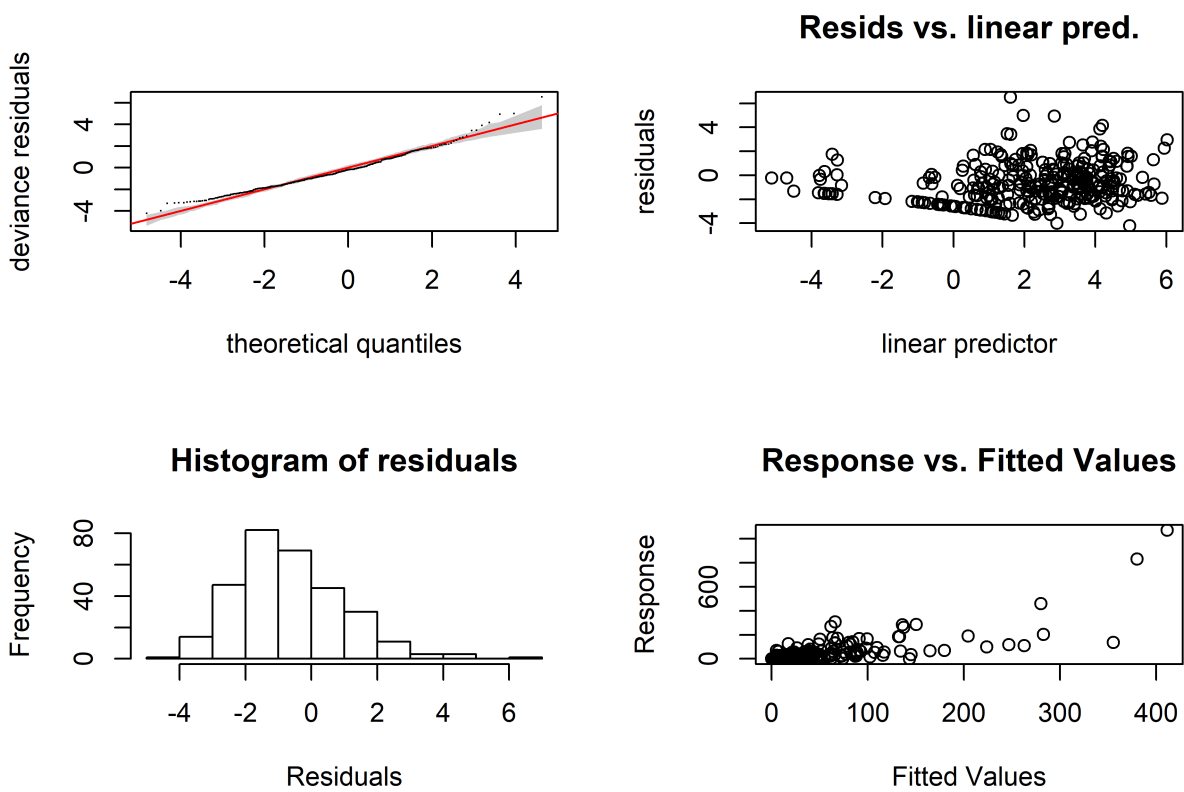


Figure A1.5: Diagnostics: Mex/US main model

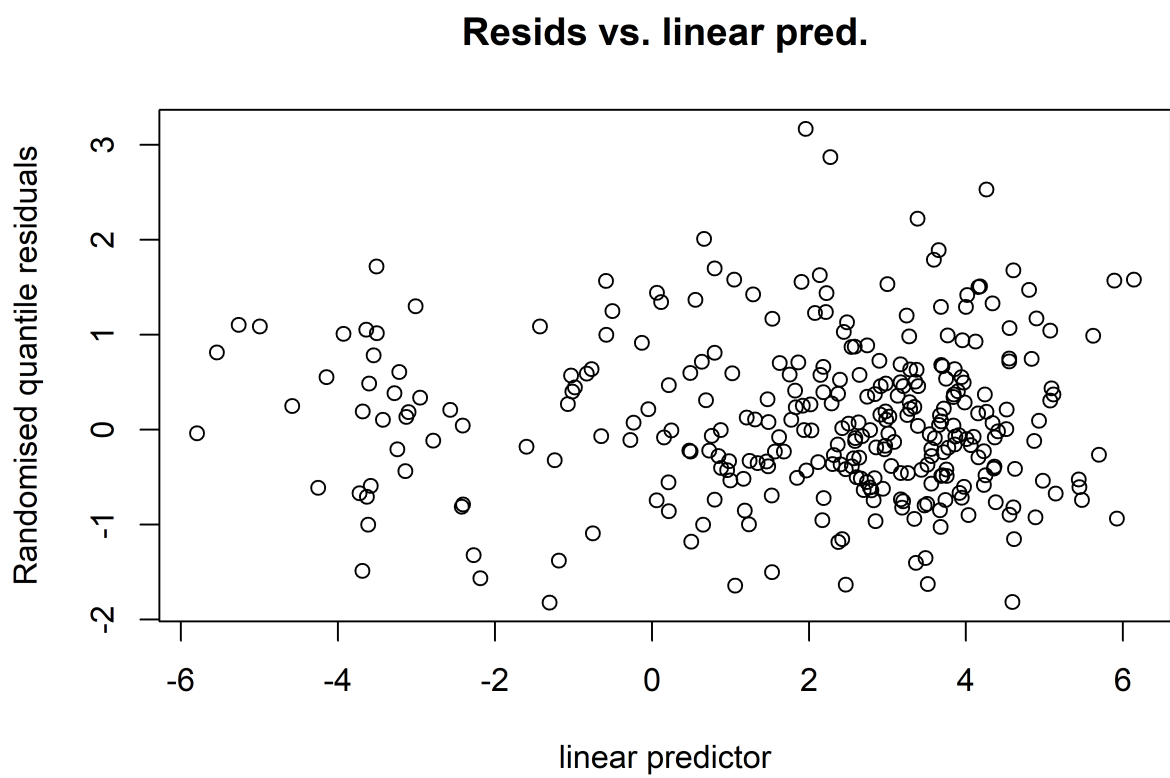


Figure A1.6: Diagnostics (cont): Mex/US main model

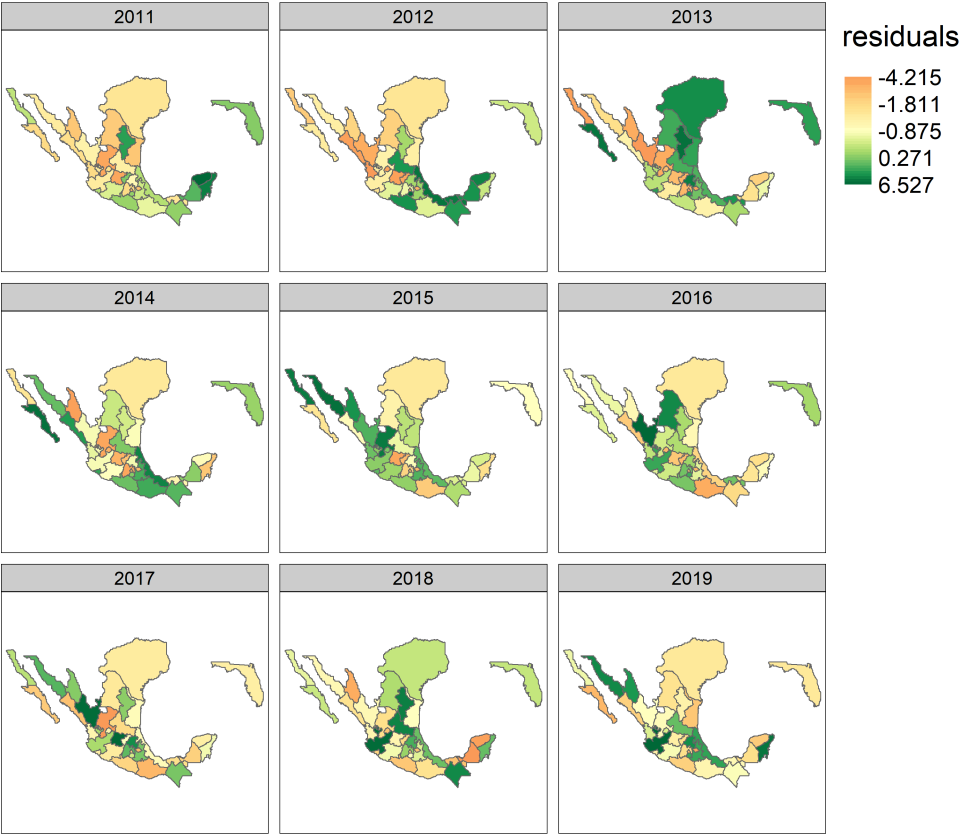


Figure A1.7: Plotted residuals: Mex/US main model



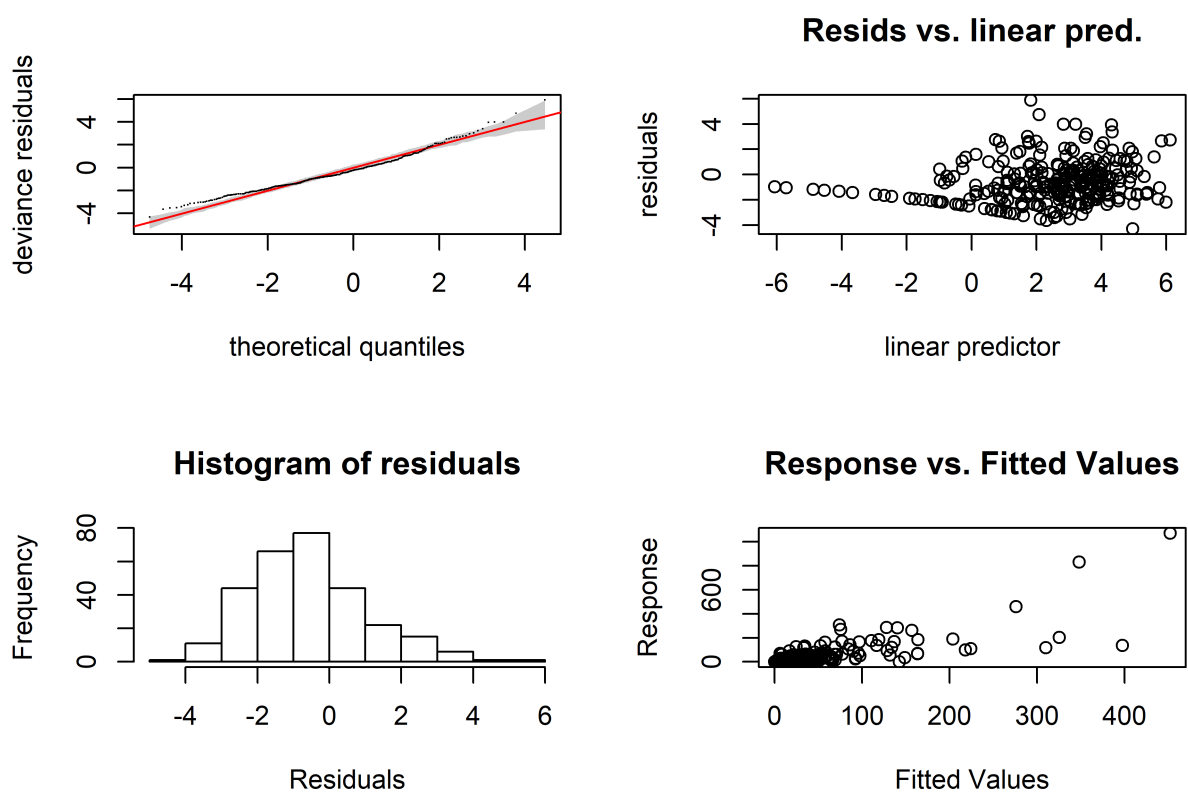


Figure A1.8: Diagnostics: Mex main model

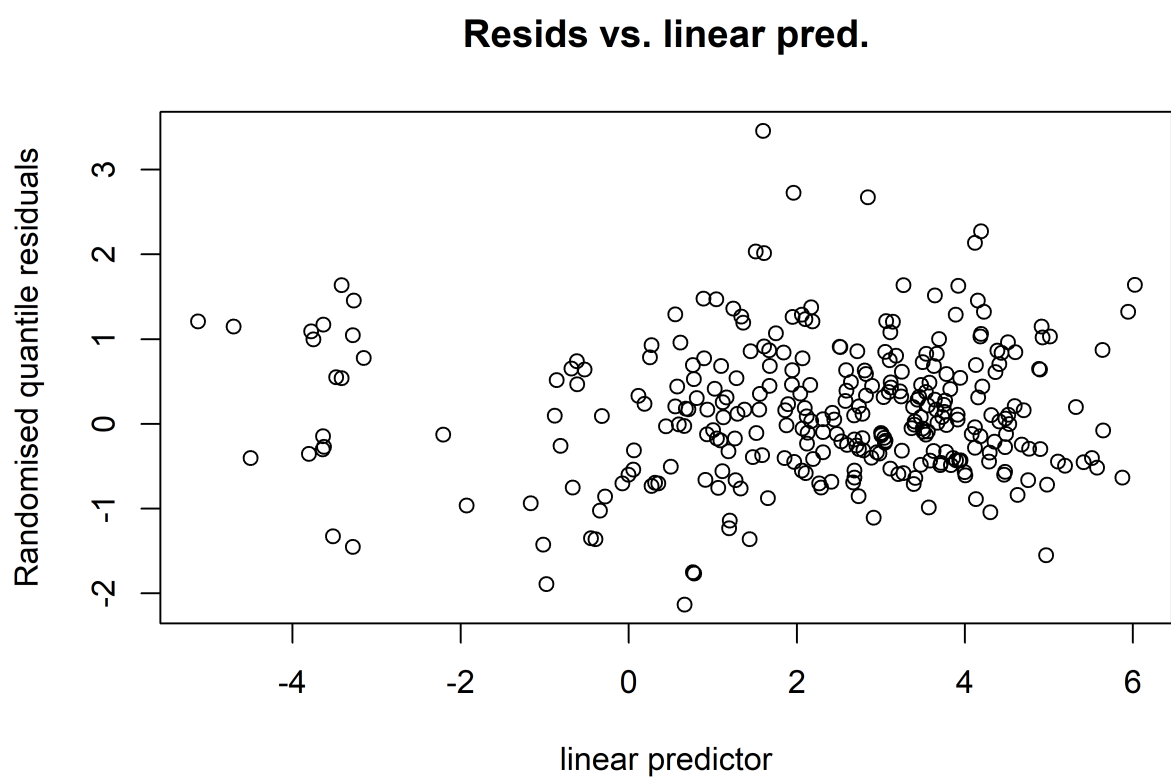


Figure A1.9: Diagnostics(cont): Mex main model

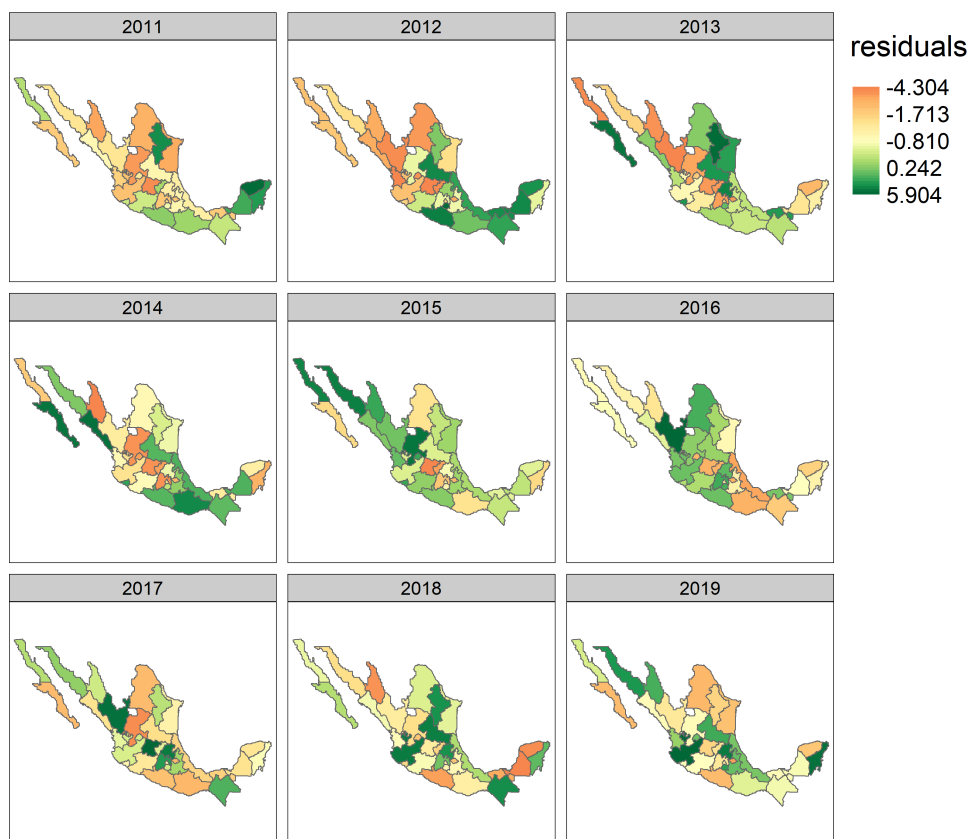


Figure A1.10: Diagnostics: Plotted residuals - Mex main model

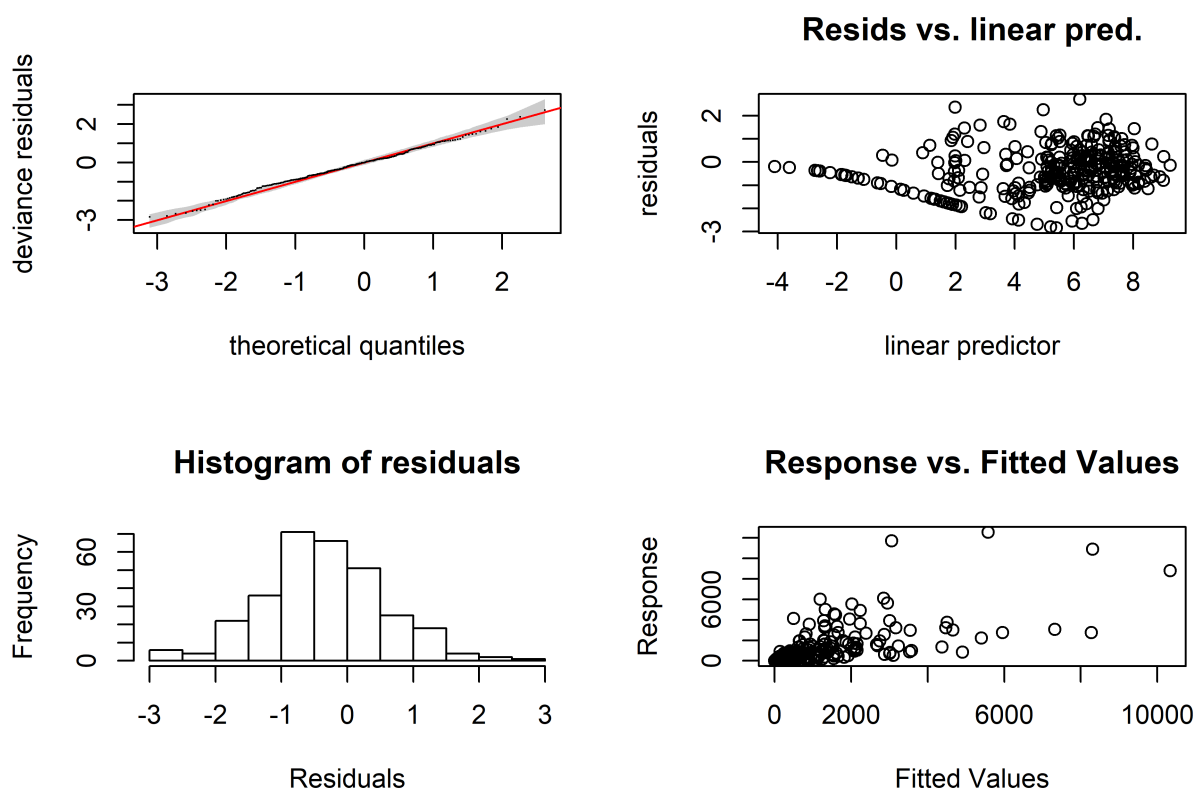


Figure A1.11: Diagnostics - Negbin Mex/US main model

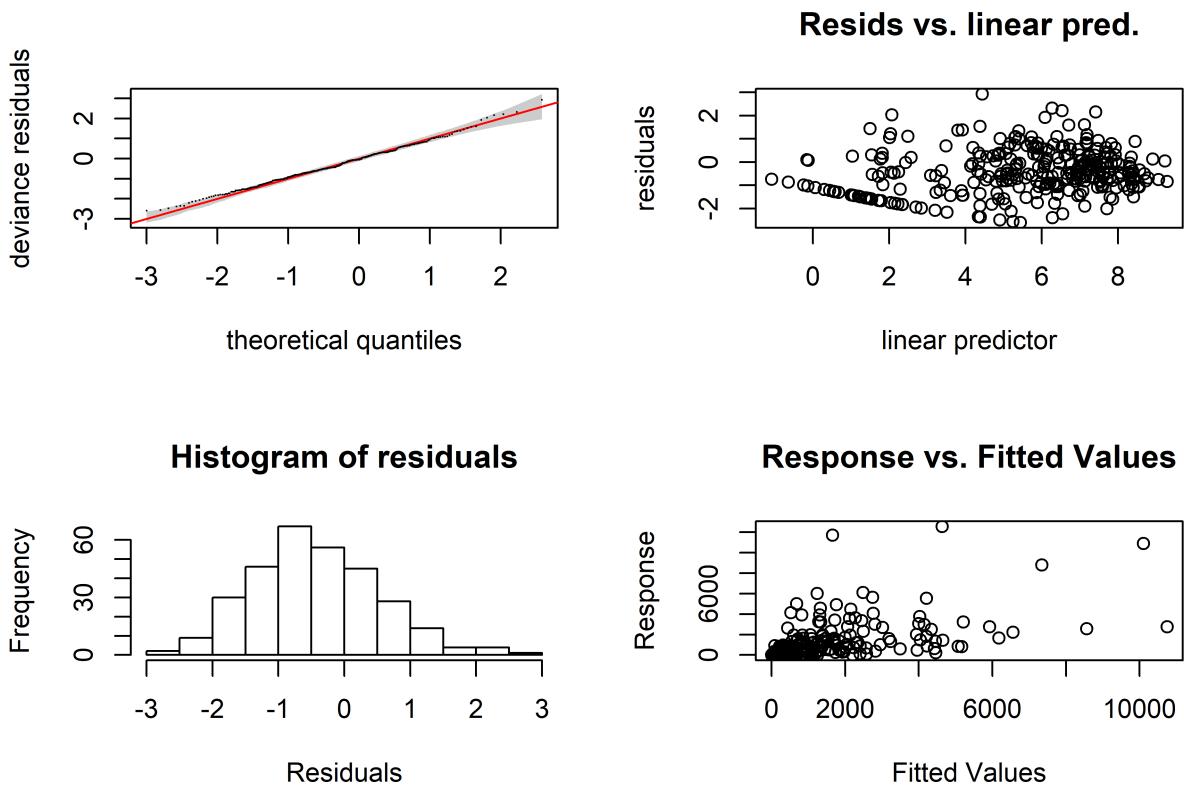


Figure A1.12: Diagnostics: Mex main model

Figure A1.13: Diagnostics: Quasi-poisson Mex/US main model

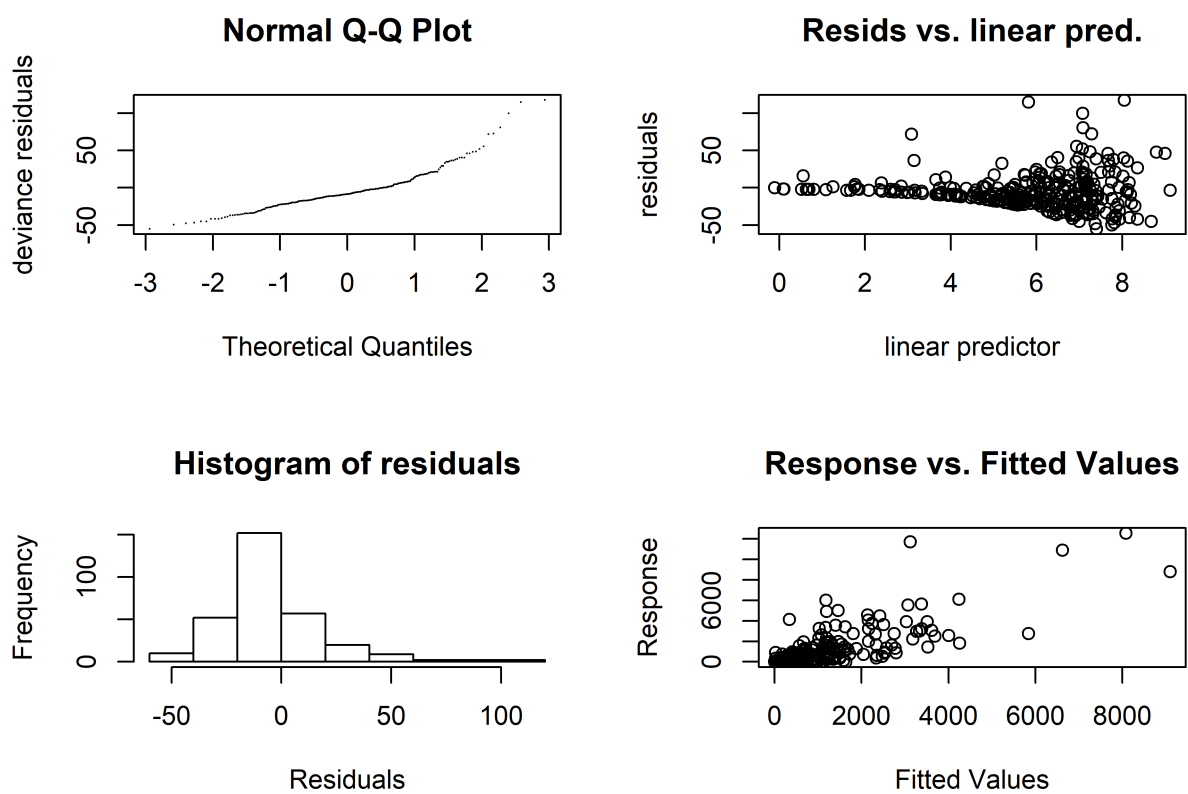


Figure A1.14: Diagnostics: Quasi-poisson Mex main model

## A.5 Data sources

	VAR	Unit	Indicator	Source
1	bio11	degrees C	Mean_Temperature_of_Coldest_Quarter	Climate Prediction Center
2	bio18	Precip mm	Precipitation_of_Warmest_Quarter	Climate Prediction Center
3	FLOWMIG_ALL_RA	Ratio	Inter-regional migration rate, (% migrants over population)	OECD regional database
4	GDP	USD per head	Regional GDP	OECD regional database
5	POP_DEN_GR	Index	Population density growth index (2001=100)	OECD regional database
6	ROOMS_PC		Average number of rooms per inhabitant (rooms per capita)	OECD regional database
7	Y0_14	Persons	Youth Population Group (0-14)	OECD regional database
8	Y65_MAX	Persons	Old Population Group (65+)	OECD regional database
9	DOC_RA		Active Physicians Rate (physicians for 1000 population)	OECD regional database
10	LIFE_EXP	Years	Life Expectancy at Birth	OECD regional database
11	INCOME_PRIM	USD per head	Primary Income of Private Households	OECD regional database
12	BB_ACC	Percentage	Share of households with internet broadband access	OECD regional database
13	EDU38_SH	Percentage	Share of labour force with at least secondary education	OECD regional database
14	INCOME_DISP	US Dollar	Disposable income per capita	OECD regional database
15	ROOMS_PC	Ratio	Number of rooms per person	OECD regional database
16	SUBJ_LIFE_SAT	Index	Self-evaluation of life satisfaction	OECD regional database
17	SUBJ_PERC_CORR	Percentage	Perception of corruption	OECD regional database
18	SUBJ_SOC_SUPP	Percentage	Perceived social network support	OECD regional database
19	dengue_non_serious	cases	number of non-serious cases reported in mexico	www.gob.mex
20	dengue_serious	cases	number of non-serious cases reported in mexico	www.gob.mex
21	dengue_cdc	cases	number of cases reported in USA	www.cdc.gov/arboNet
22	pop	population	human population in aedes infected areas	<a href="https://sedac.ciesin.columbia.edu">https://sedac.ciesin.columbia.edu</a>

Table A1.5: Main analysis variable description and codes

## Bibliography

52. C. Vega G, Pertierra LR, Olalla-Tárraga MÁ. MERRAclim, a high-resolution global dataset of remotely sensed bioclimatic variables for ecological modelling. *Scientific Data*. 2017;4:170078.
58. GFC. Spatial data analysis and modeling with r. 2018;2018. <http://rspatial.org/index.html>.



# Appendix B

## The rise of West Nile Virus in Southern and Southeastern Europe: A spatial–temporal analysis investigating the combined effects of climate, land use and economic changes

### B.1 Data availability

An R project containing all data that supports the findings of this study are available in .Rdata format from <https://doi.org/10.5281/zenodo.4656902>).

### B.2 Code availability

An R project containing all code used to set-up the models is available here is available from <https://doi.org/10.5281/zenodo.4656902>.

### B.3 Extended data extraction and processing methods

#### B.3.1 Aggregation

All data were aggregated annually to produce the final yearly panel data-set and aggregated at the NUTS 3 country subdivision level, apart from central government spending data which was sourced at the country level. All spatial information was captured at the NUTS 3 level using shapefiles (polygons) sourced from R’s ‘eurostat’ package [9].

The Nomenclature of territorial units for statistics (NUTS) is a classification system used to divide economic territories of the EU into three hierarchical sub categories for the purpose of data collection and and statistical analysis:

- NUTS 1: Major socio-economic regions with a population ranging from 3 to 7 million.
- NUTS 2: Basic regions are generally used for the application of regional policies with a population ranging from 800,000 to 3 million.
- NUTS 3: Small regions for specific diagnoses with a population ranging from 150,000 to 800,000.

For further details see <https://ec.europa.eu/eurostat/web/nuts/background>.

## WNV Case Data

WNV case data were provided at request by the European Centre for Disease Prevention and Control ([www.ecdc.europa.eu](http://www.ecdc.europa.eu)). Case data are collected weekly by EU member states and affiliates. Data were aggregated at NUTS 3 country subdivisions [6]. Positive cases were confirmed by at least one of the following techniques: 1); isolating WNV or WNV nucleic acid from blood or cerebrospinal fluid (CSF); 2) inducing a WNV-specific antibody response (either IgG / IgM) in a serological test. All cases were aggregated yearly to create the annual panel data-set.

## B.3.2 Economic, Socio-Economic and Demographic Factors

Economic data were extracted from the Eurostat database (<https://ec.europa.eu/eurostat/data/database>), which provides comparable statistics and indicators and is presented in yearly time series. To capture factors determining the economic crisis, austerity and cuts to public spending we selected NUTS3 regional level Gross Domestic Product (GDP); and country level agriculture, forestry, fisheries spending, waste water spending Health spending. The “Agriculture, forestry, fisheries spending” variable captures spending in rural areas that help to improve the environment and agricultural development, that can benefit agricultural workers and/or mechanize production [5]. In order to represent spending before and after the economic crisis, we created a baseline index for each variable set at 2007 levels, which represented negative or positive growth from the point just before the economic crisis hit Europe.

## B.3.3 Climate Data

Climate data were sourced from the E-OBS Gridded Data-set [3]. This data-set was created using a series of daily temperature and rainfall observations at meteorological stations throughout Europe. R’s “Raster Extract” function from the “Raster” package [7] was used to extract and aggregate cell values to each NUTS 3 region. The subsequent regional values were then processed further to create regional seasonal variables: “Mean temp winter (°C)”, “Mean temp spring (°C)”, “Mean temp summer (°C)”, “Days of rain in winter”, “Days of rain in spring” and “Days of rain in summer”. Winter was designated as December to March, Spring as March to June, and summer June to September.

### B.3.4 Land-use data

Land use statistics were captured at the NUTS3 level using the CORINE Land Cover (CLC) 2006, 2012 and 2018 data-sets [4]. These data-sets provide information on the biophysical characteristics of the Earth’s surface in the form of categorical raster data. For each region, we calculated percentage land cover for each of the land-use risk factors identified in our conceptual framework, i.e., “Continuous urban fabric”, “Discontinuous urban fabric”, “Wetlands (fresh water)” and “Arable land”. R’s SF and Raster packages [10, 7] were used to extract information for each available year (2006, 2012, 2018). R’s Zoo package [14] was used to calculate values for missing years, by implementing a linear interpolation method that would predict trends between years, apart from 2019 where 2018 values were used.

### B.3.5 Surface Water data

Regional surface water data was sourced using the JRC Monthly Water History, v1.2 data set [11] via Google Earth Engine at a 30 meter pixel resolution. This data set contains maps of the location and temporal distribution of surface water from 1984 to 2019 and provides statistics on the extent and change of water surfaces. Data were generated using scenes from Landsat 5, 7, and 8. Each pixel was individually classified into water / non-water using an expert system and the results were collated into a monthly history. Water / non-water count observations were extracted and aggregated by each NUTS 3 region. The sum of the Water / non-water observations were then used to create a % water surface water indicator, which was averaged by season and converted to Z-scores to standardise values. This would help determine if the seasonal water extent was average, below the mean (low), or above the mean (high) for a given year.

### B.3.6 Extended Statistical methods

#### General additive regression model to assess associations of independent variables on WNV case data at regional level

One of the main issues with our data-set is that it did not meet some basic assumptions for statistical inference, and specifically the data are not independent and identically distributed random variables (iid). More specifically, the data-set captured repeated measurements over the same regions, and observations were not independent because of spill over effects from neighbouring regions, therefore we needed to implement an appropriate statistical design to control for both temporal and spatial pseudo replication (lack of independence). We could deal with this in two ways, 1) either using a generalised linear mixed model (GLMM) approach, relaxing the assumption of independence and estimating the spatial/temporal correlation between residuals, or 2) model the spatial and temporal dependence in the systematic part of the model [1]. We opted to use a Generalised Additive Model (GAM) using R’s Mgecv statistical package [12] because of its versatility and ability to fit complex models that would converge even with low numbers of observations, and could capture potential complex non-linear relationships. One of the advantages of GAMs is that we do not need to determine the functional form of the relationship beforehand. In general, such models transform the mean response to an additive form so

that additive components are smooth functions (e.g., splines) of the covariates, in which functions themselves are expressed as basis-function expansions. The spatial auto-correlation in the GAM model was approximated by a Markov random field (MRF) smoother, defined by the geographic areas and their neighborhood structure. We used R's Spdep package [2] to create a queen neighbors list (adjacency matrix) based on regions with contiguous boundaries i.e. those sharing one or more boundary point. We used a full rank MRF, which represented roughly one coefficient for each area. The local Markov property assumes that a region is conditionally independent of all other regions unless regions share a boundary. This feature allowed us to model the correlation between geographical neighbors and smooth over contiguous spatial areas, summarizing the trend of the response variable as a function of the predictors (see section 5.4.2 of [13]). In order to account for variation in the response variable over time, not attributed to the other explanatory variables in our model, we used a saturated time effect for years, where a separate effect per time point is estimated.

We first tried to fit our model using a Poisson distribution. However, the mean of our dependent variable (WNV cases by region and year) was lower than its variance -  $E(Y) < \text{Var}(Y)$ , suggesting that the data are over-dispersed. We also tried to fit our models using the negative binomial, quasi-Poisson and Tweedie distribution, all particularly suited when the variance is much larger than the mean. After several tests, we concluded that the Tweedie distribution worked well with our data since it can handle excess zeros [8], and allows us to model the incident rate, although results were comparative across all distributions (note that WNV infection count data, offset by a log of population at risk was used for the neg bin and quasi-Poisson models). Analysis of model diagnostic tests did not reveal any major issues; in general residuals appeared to be randomly distributed (see additional information - Figures S10-S11 and Table S1 for diagnostics).

Tweedie distributions are defined as subfamily of (reproductive) exponential dispersion models (ED), with a special mean-variance relationship. A random variable  $Y$  is Tweedie distributed if:

$TW_p(\mu, \sigma^2)$  if  $Y \sim ED(\mu, \sigma^2)$ , with mean  $= \mu = E(Y)$ , positive dispersion parameter  $\sigma^2$  and  $\text{Var}(Y) = \mu\sigma^2$ .

The empirical model can then be written as:

$$E(Y) = f_1(X_{it}) + f_n(\text{Year}_t) + f_m(\text{Region}_i)$$

Where the  $f(\cdot)$  stands for smooth functions;  $E(Y)_{it}$  is equal to the WNV infection incidence per 100,000 in region  $i$  at time  $t$ , which we assume to be Tweedie distributed;  $X_{it}$  - is a vector of economic, demographic, environmental and climate variables.  $\text{Year}_t$  is a function of the time intercept and  $\text{Region}_i$  represents neighborhood structure of region.

## B.4 Climate modeling

In order to model long term seasonal climate trends, we fit a GAM model using the following equation.

$$y = \beta_0 + f(x_1, x_2) + \varepsilon$$

where  $y =$  is either the mean of the monthly regional temperatures ( $^{\circ}\text{C}$ ) or regional sum precipitation (mm).

$B_0$  is the intercept, month is represented by  $x_1$  and  $x_2$  is the series of years in the entire time period i.e. within-year and between year.

$f$  is a smooth function interaction that accounts for variation in, or interaction between, the trend and seasonal features of the data.

Temperature models were fit using the Gaussian distribution and precipitation models fit using the Tweedie distribution.

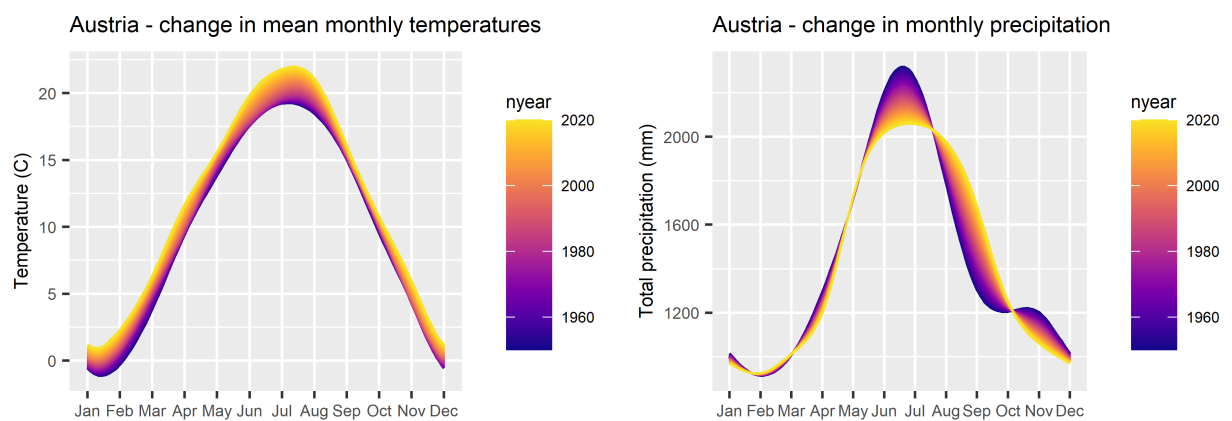


Figure B1: Austria - seasonal climate trends (Data source: E-OBS version 22.0e).

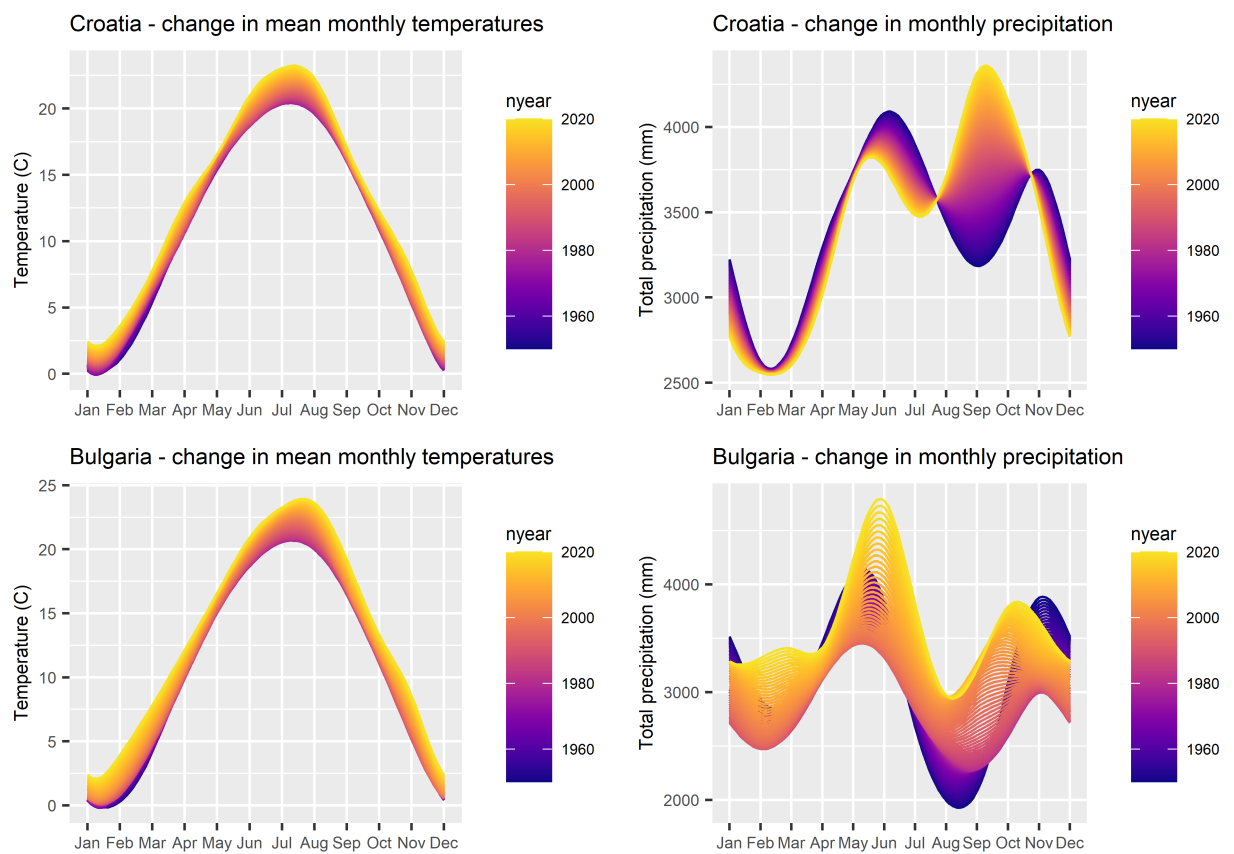


Figure B2: Bulgaria / Croatia - seasonal climate trends (Data source: E-OBS version 22.0e).

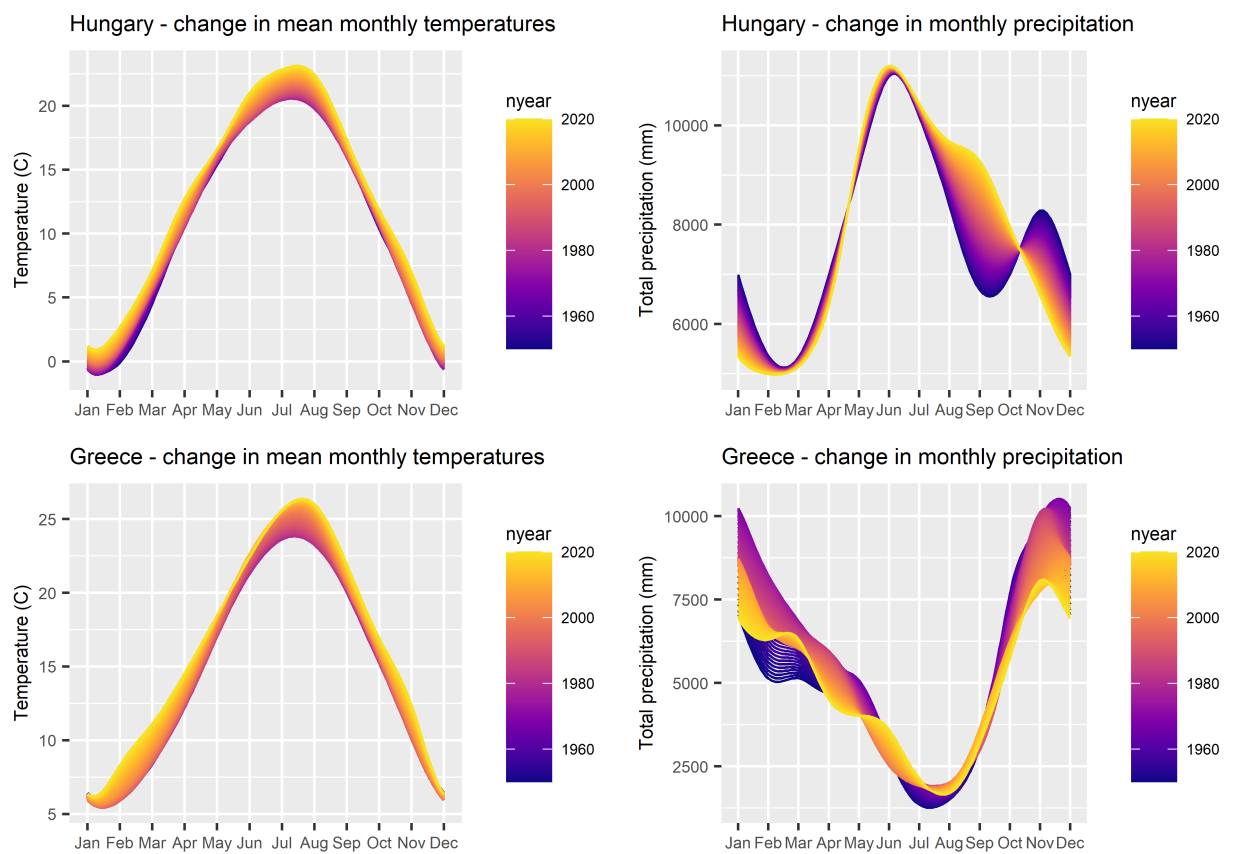


Figure B3: Greece / Hungary- seasonal climate trends (Data source: E-OBS version 22.0e).



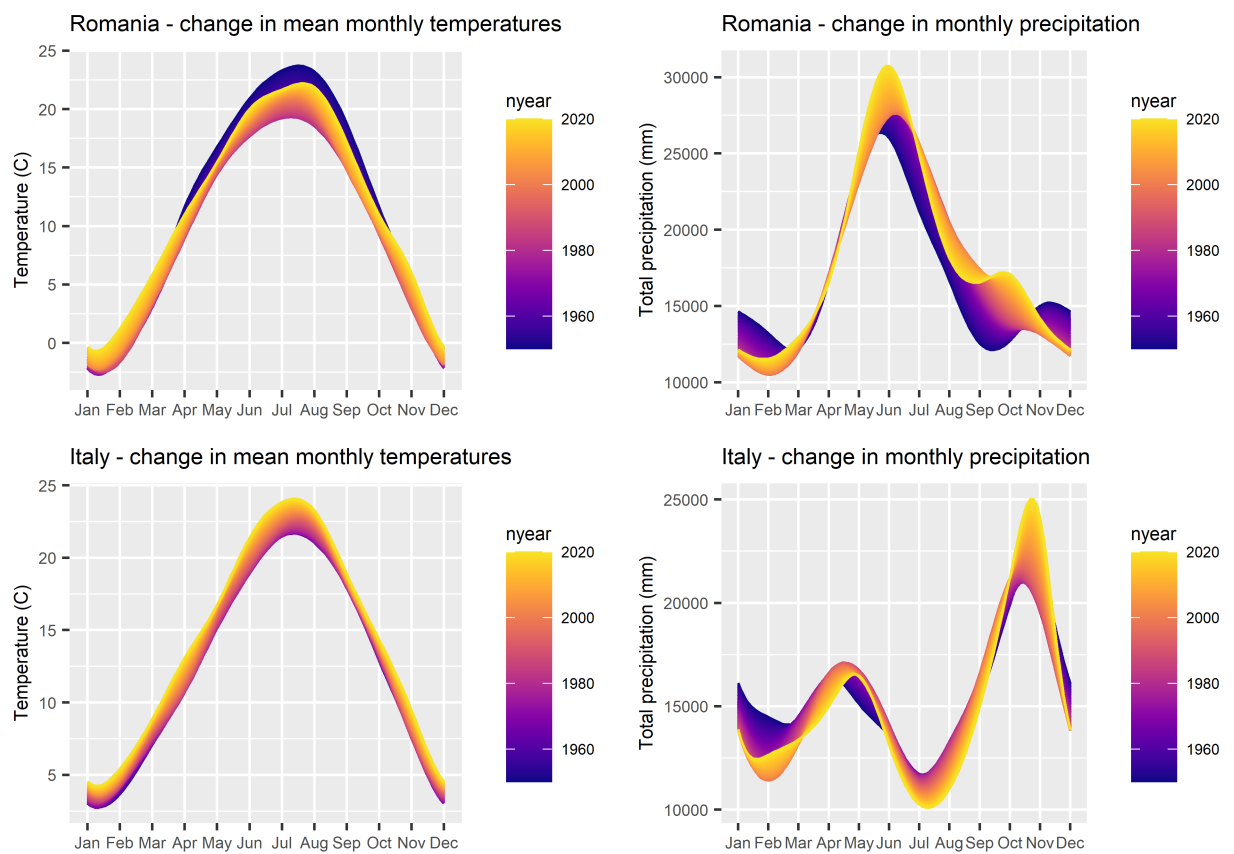


Figure B4: Romania / Italy - seasonal climate trends (Data source: E-OBS version 22.0e).



Figure B5: Land-use: 1 = Discontinuous Urban Fabric, 2-4 = Arable land break-down (Source: CORINE Land Cover)

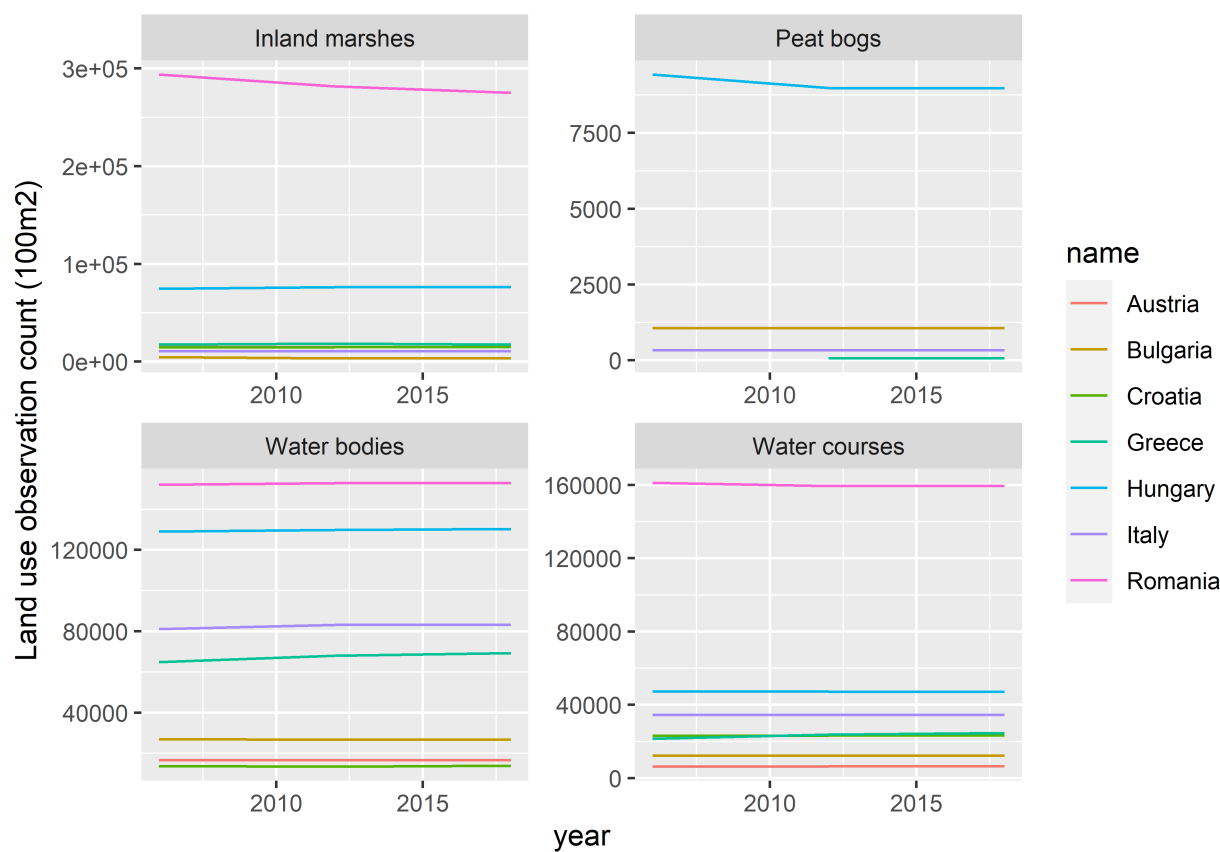


Figure B6: Land-use: Fresh water bodies break-down (Source: CORINE Land Cover)

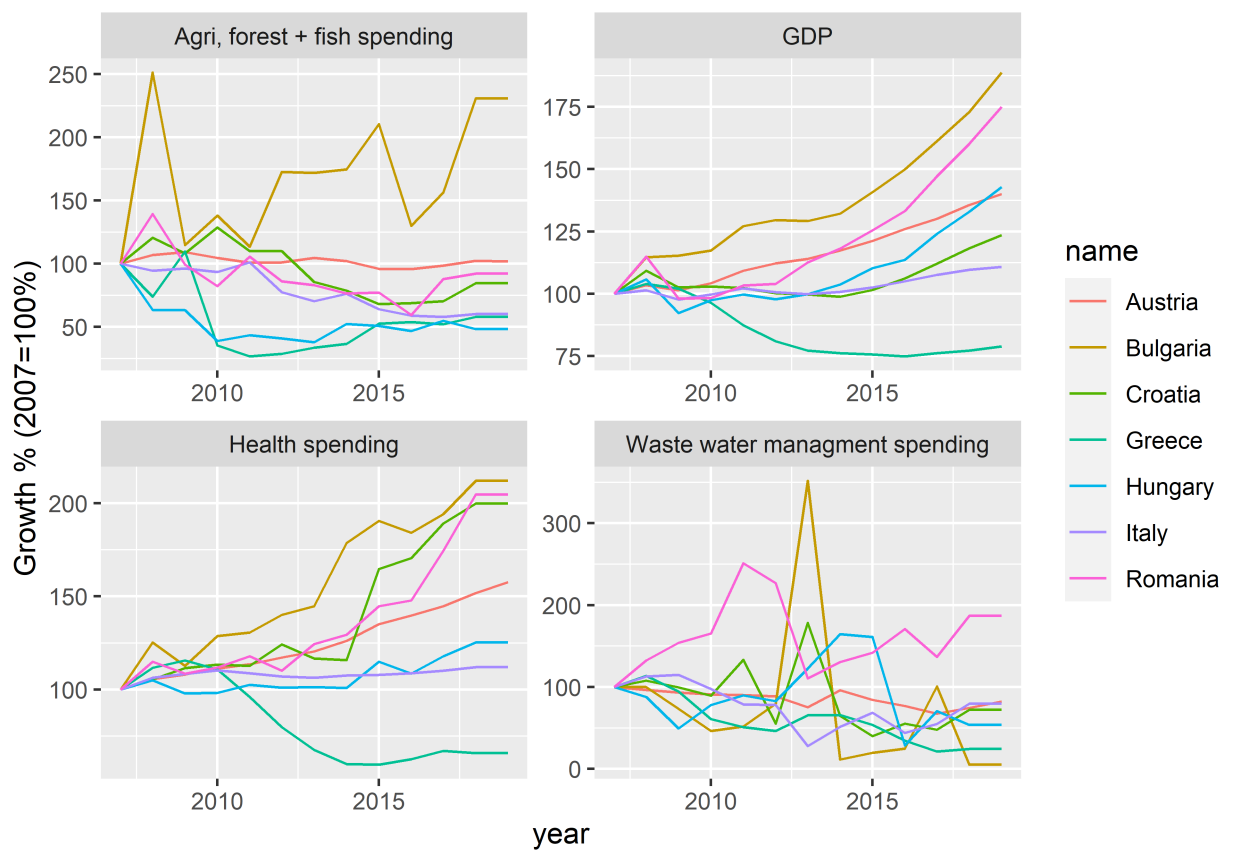


Figure B7: Government spending growth, GDP growth and unemployment 2007-2019 (2007=100%) (Source: Eurostat)

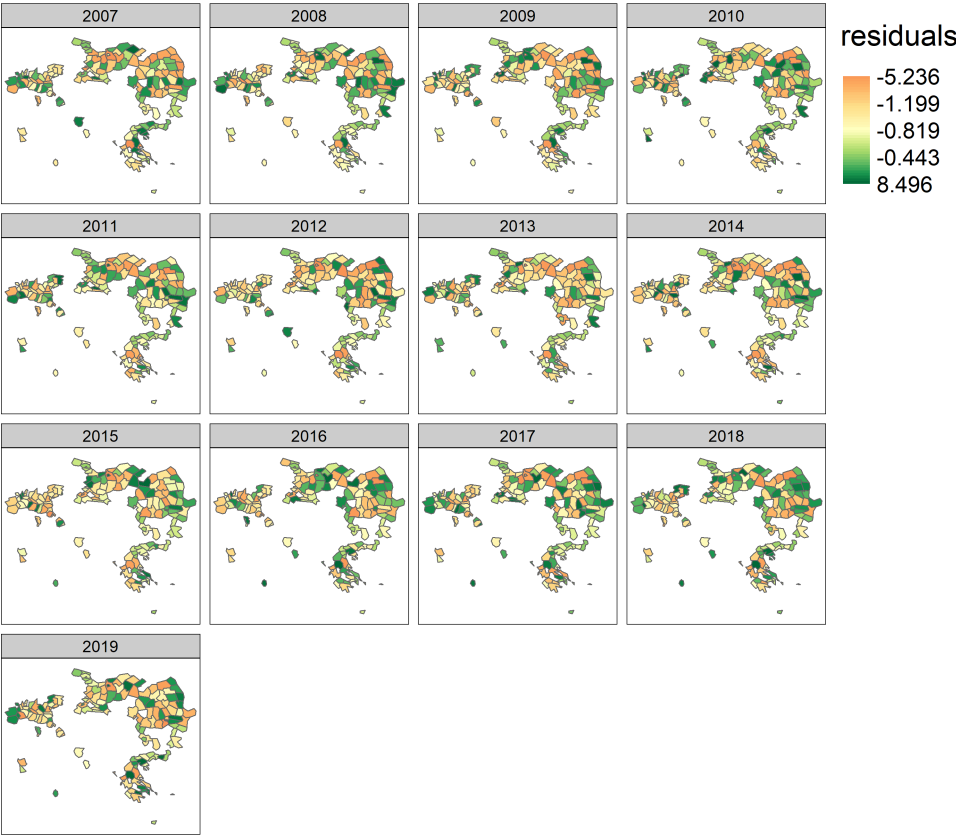


Figure B8: Variable correlation plot.

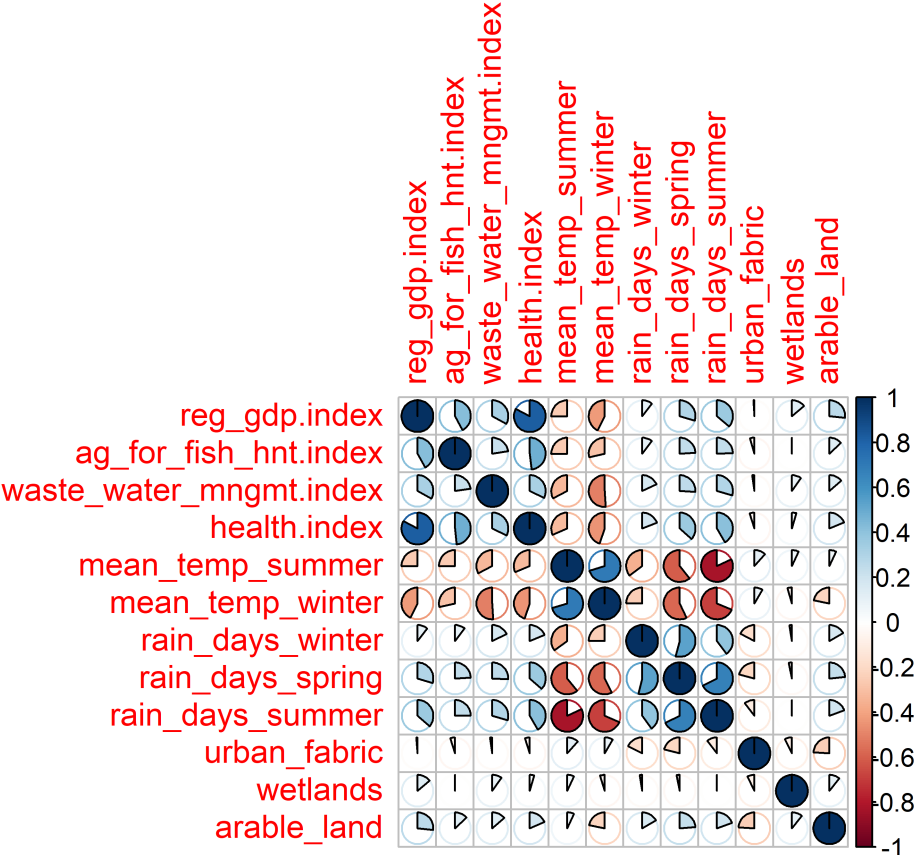


Figure B9: Spatial Residuals Tweedie model.

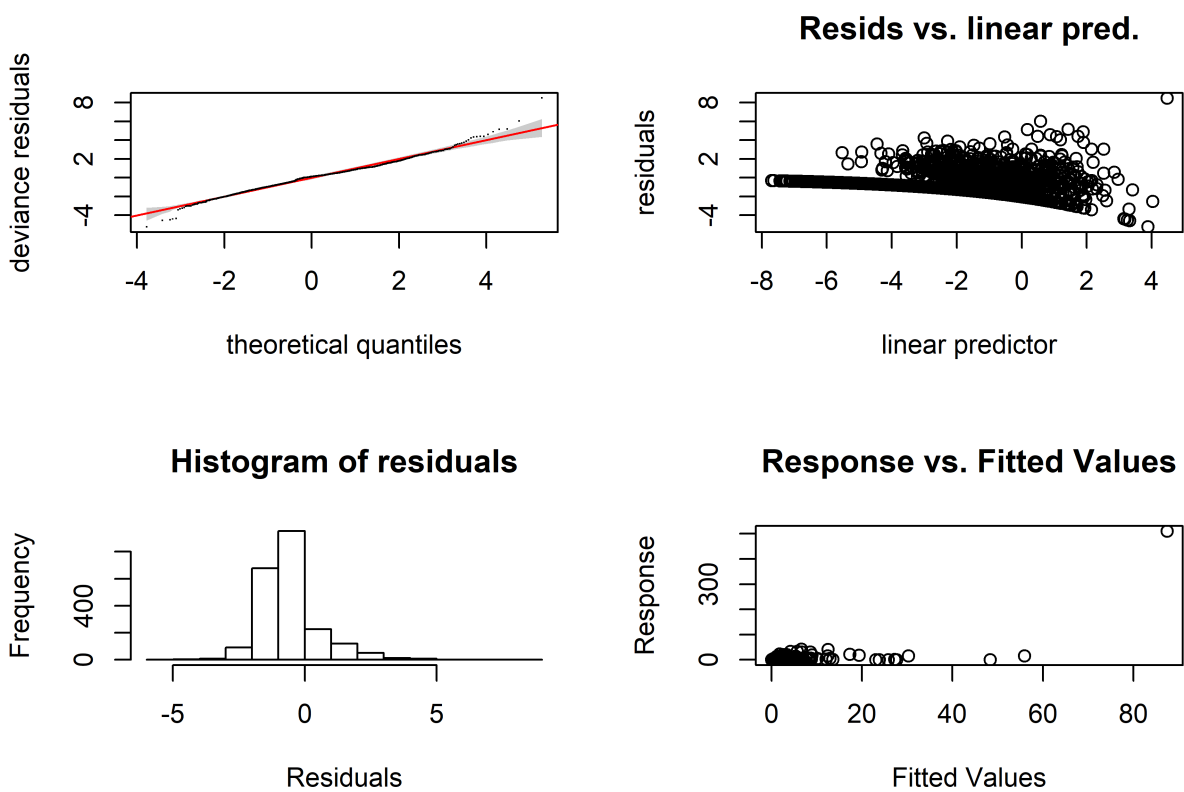


Figure B10: Diagnostics Tweedie model.

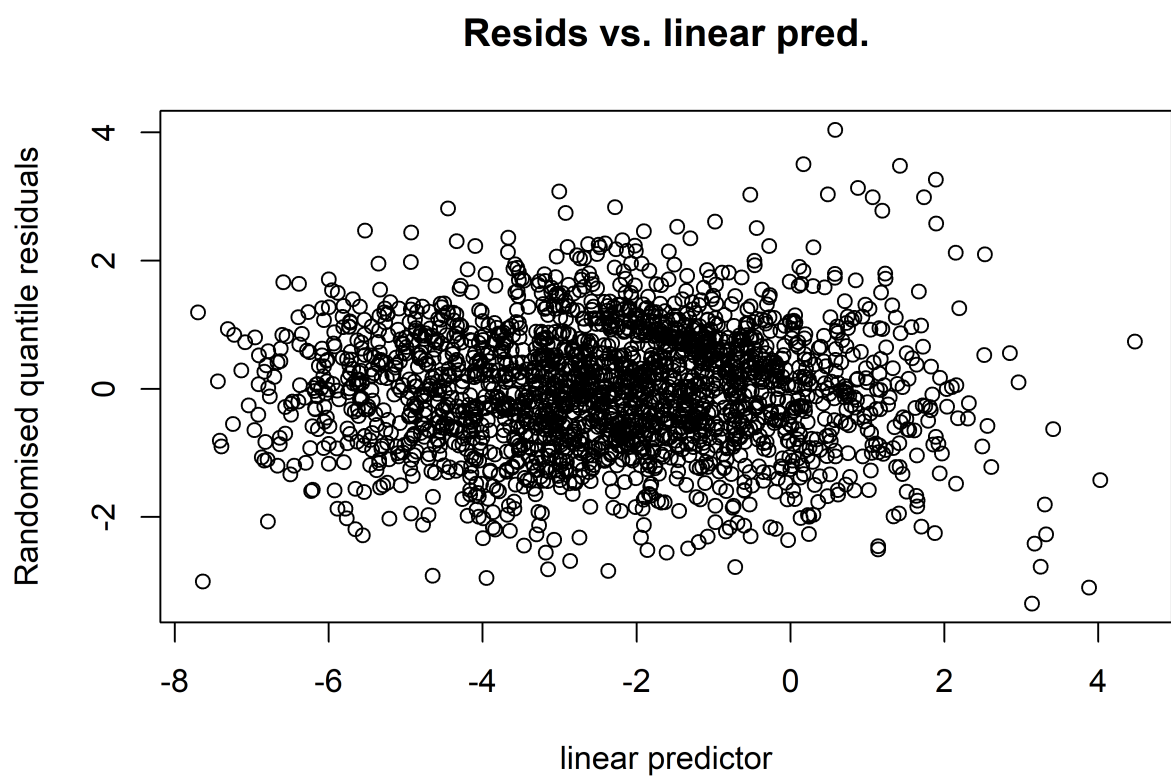


Figure B11: Diagnostics 2 Tweedie model.



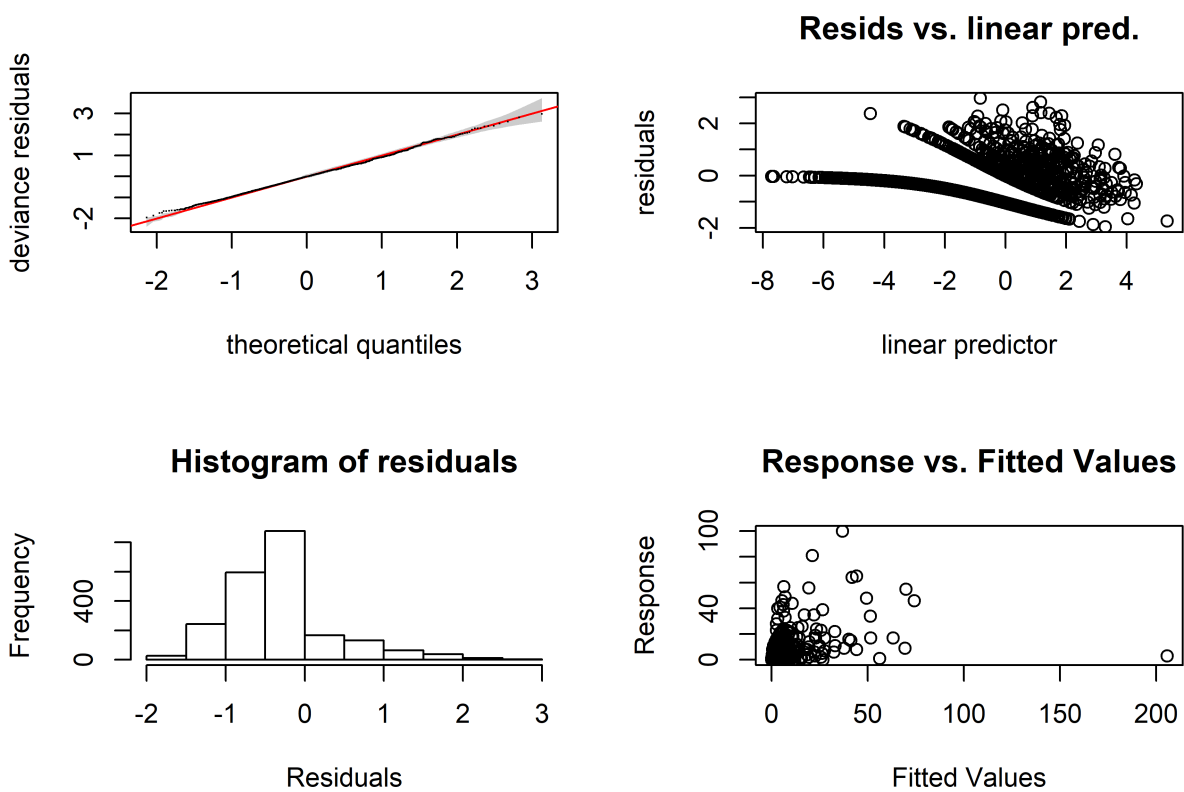


Figure B12: Diagnostics negbin model.

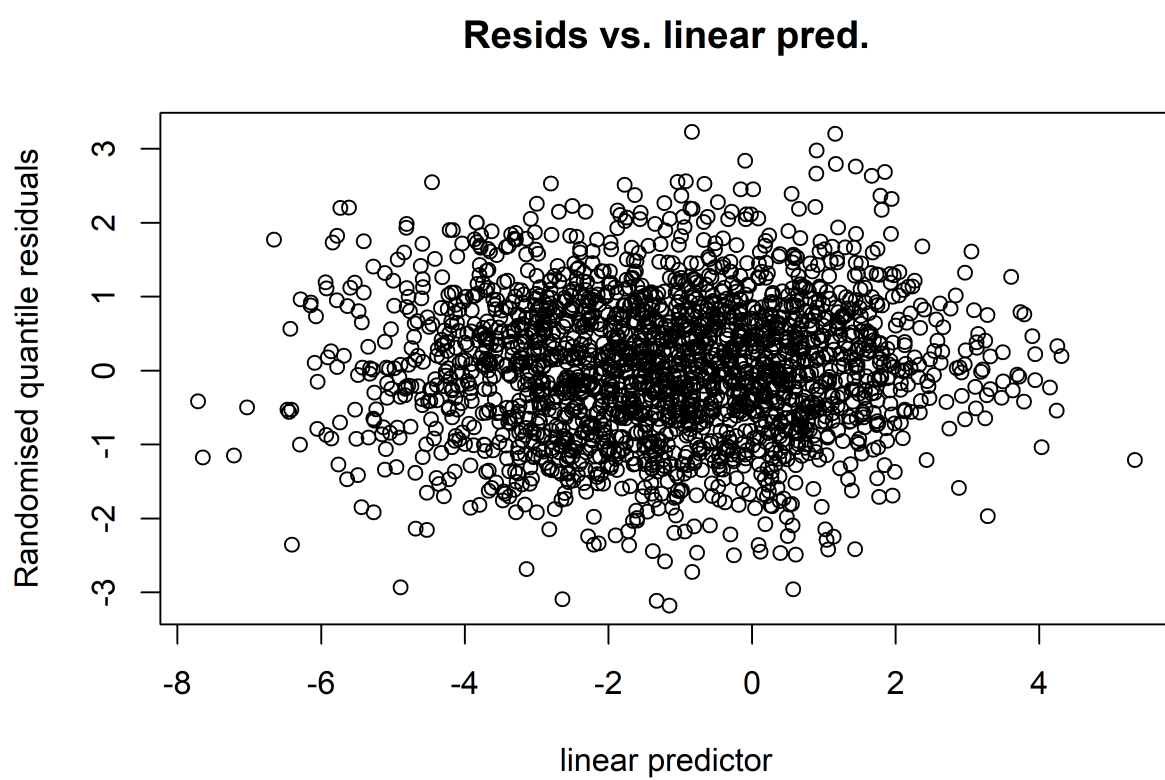


Figure B13: Diagnostics 2 negbin model.

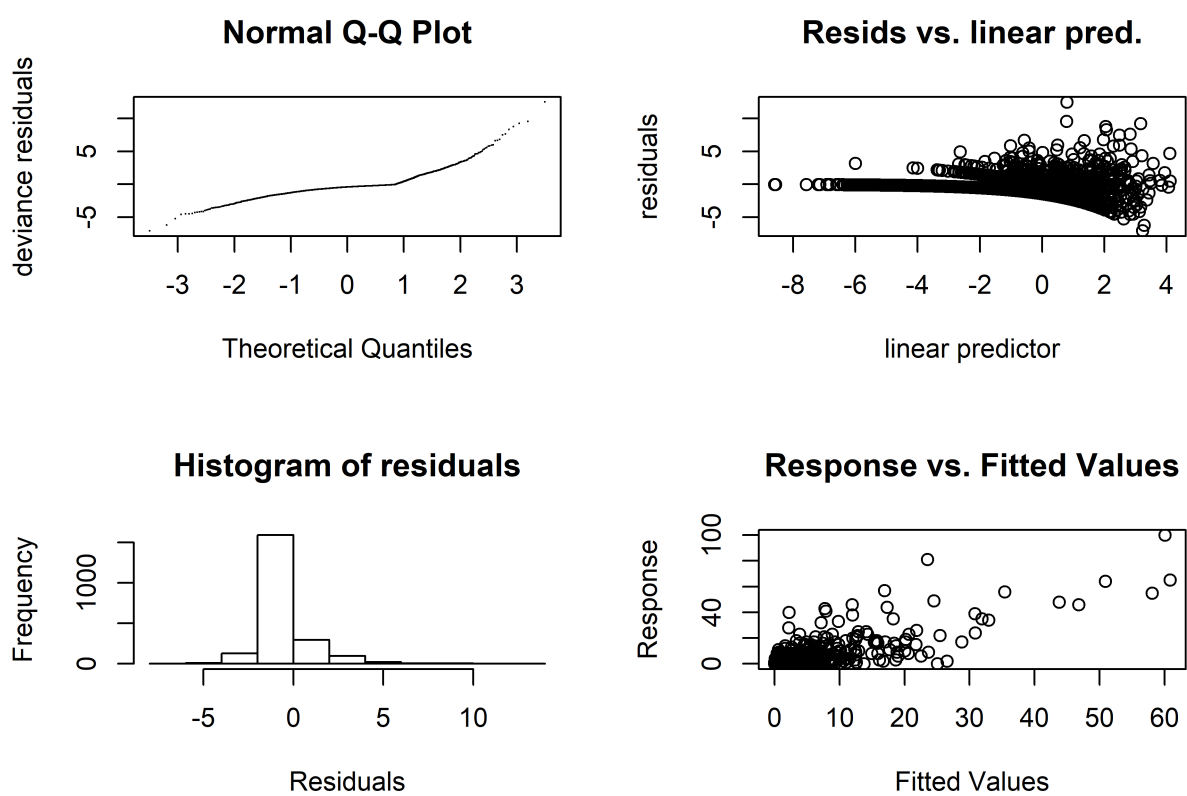


Figure B14: Diagnostics Quasipoisson model.

Table B1: WNF Cases Per Country 2006-2019

Country	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Austria	0	0	2	1	0	0	0	2	6	5	6	20	4
Bulgaria	0	0	0	0	0	2	0	1	2	1	1	15	6
Croatia	0	0	0	0	0	6	20	1	1	2	5	57	1
Cyprus	0	0	0	0	0	0	0	0	0	1	0	1	24
Czechia	0	0	0	0	0	0	1	0	0	0	0	5	1
France	0	0	0	0	0	0	0	0	1	0	2	27	1
Germany	0	0	0	0	0	0	0	0	0	0	0	1	4
Greece	0	0	0	262	100	157	85	15	0	0	48	312	228
Hungary	4	19	7	18	4	17	35	10	18	44	20	216	72
Italy	0	0	0	4	18	45	80	24	61	76	53	610	54
Portugal	0	0	0	0	0	0	0	0	1	0	0	0	0
Romania	4	2	2	57	11	15	24	23	32	93	66	279	68
Slovakia	0	0	0	0	0	0	0	0	0	0	0	0	1
Slovenia	0	0	0	0	0	0	1	0	0	0	0	4	0
Spain	0	0	0	2	0	0	0	0	0	4	0	0	0
Turkey	0	0	0	47	5	0	0	0	0	1	7	26	10

Table B2: Final model specification comparisons by distribution

	Tweedie model	Negbin model	Quasi Poisson model
Intercept	−2.35*** (0.40)	−13.82*** (0.36)	−13.86*** (0.45)
Mean temp summer (C)	1.00* (1.00)	1.50* (1.70)	2.00** (2.00)
Mean temp winter (C)	1.94*** (1.99)	1.96*** (1.99)	2.00*** (2.00)
Days of rain in summer	1.00** (1.00)	1.00** (1.00)	1.00 (1.00)
Summer surface water extent (30m2)	1.02*** (1.03)	1.40* (1.64)	1.62*** (1.85)
Regional GDP index (2007=100%)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
Agri, forest + fish spending (2007=100%)	1.93*** (1.99)	1.94*** (1.99)	2.00*** (2.00)
Waste water managment spending (2007=100%)	1.10*** (1.19)	1.45*** (1.69)	1.75*** (1.93)
Continuous urban fabric %	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
Discontinuous urban fabric %	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
Wetlands %	1.00 (1.00)	1.00 (1.00)	1.01 (1.01)
Arable land %	1.74** (1.84)	1.83*** (1.90)	1.00** (1.00)
Year	11.56*** (12.00)	11.55*** (12.00)	11.62*** (12.00)
Spatial lag	76.19*** (106.52)	83.79*** (115.75)	117.20*** (141.05)
AIC	3907.56	4659.28	-
BIC	4538.85	5335.48	-
Log Likelihood	−1842.57	−2210.53	-
Deviance	3520.85	1207.27	4607.49
Deviance explained	0.65	0.64	0.69
Dispersion	2.73	1.00	3.22
R <sup>2</sup>	0.25	−0.32	0.60
GCV score	1871.90	2367.22	2.45
Num. obs.	2158	2158	2158
Num. smooth terms	13	13	13

\*\*\* $p < 0.01$ ; \*\* $p < 0.05$ ; \* $p < 0.1$

## Bibliography

- [1] A. Aswi, S. M. Cramb, P. Moraga, and K. Mengersen. Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review. *Epidemiology and infection*, 147:1–14, 2018.
- [2] Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013.
- [3] Richard C. Cornes, Gerard van der Schrier, Else J. M. van den Besselaar, and Philip D. Jones. An ensemble version of the e-obs temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, 123(17):9391–9409, 2018.
- [4] EU. Copernicus land monitoring service 2018, 2018.
- [5] Eurostat. Agriculture, forestry and fishery statistics. Report, Eurostat, 2019.
- [6] Eurostat. Nuts - nomenclature of territorial units for statistics, 2020.
- [7] Robert J. Hijmans and Jacob van Etten. *Geographic analysis and modeling with raster data*, 2012. R package version 2.0-12.
- [8] Christoph F. Kurz. Tweedie distributions for fitting semicontinuous health care utilization cost data. *BMC Medical Research Methodology*, 17(1), 12 2017.
- [9] Leo Lahti, Janne Huovari, Markus Kainu, and Przemyslaw Biecek. eurostat r package, 2017. Version 3.7.5.
- [10] Edzer Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018.
- [11] Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S. Belward. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633):418–422, 2016.
- [12] S N Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, 2011.
- [13] Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.
- [14] Achim Zeileis and Gabor Grothendieck. zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27, 2005.

## Appendix C

**Macro-Level drivers of  
SARS-CoV-2 transmission: A  
data-driven analysis of factors  
contributing to epidemic growth  
during the first wave of outbreaks  
in the United States**

## **C.1 Covid policy tracker**

See file

## **C.2 Data availability**

## **C.3 Diagnostics: Infection mode**

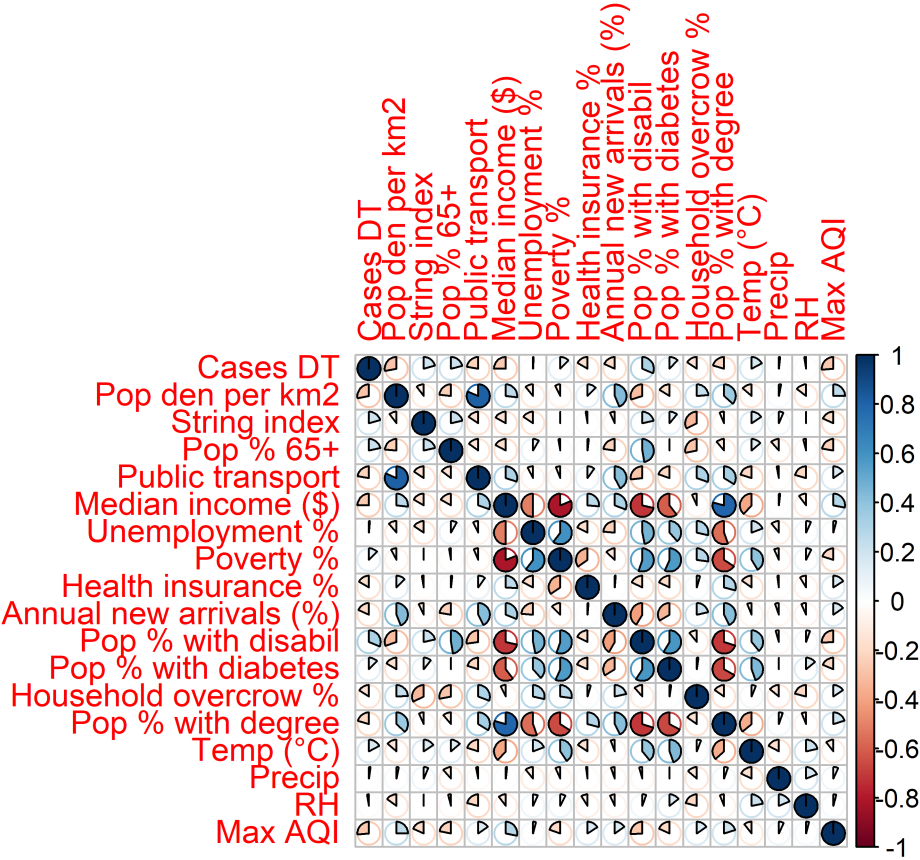


Figure C1: Correlation plot - infection model data.



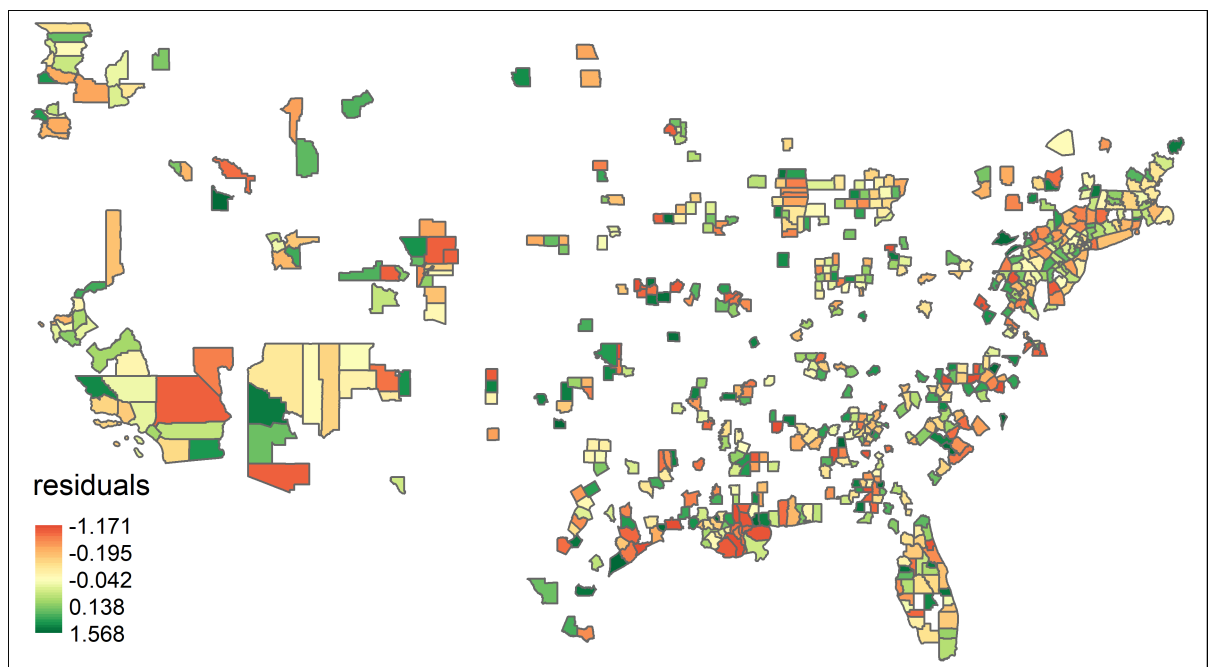


Figure C2: Spatial residuals - infection model.

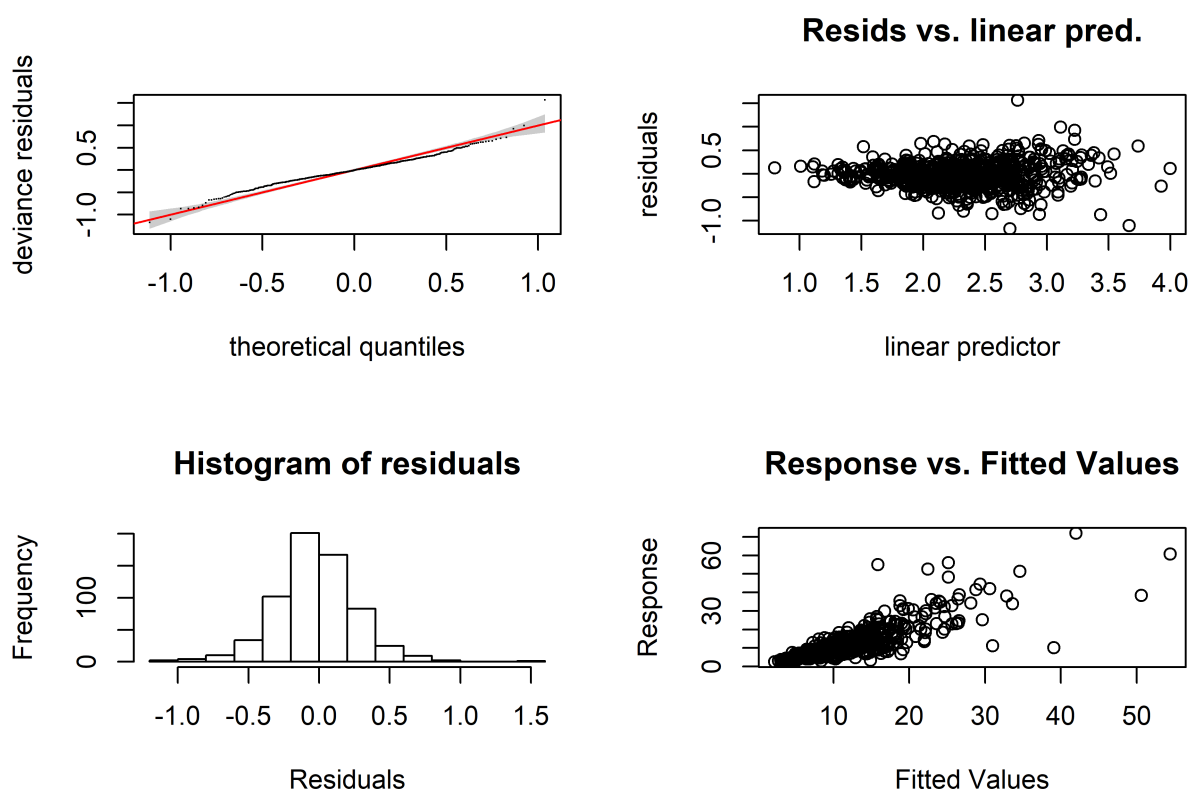


Figure C3: Model diagnostics - infection model.

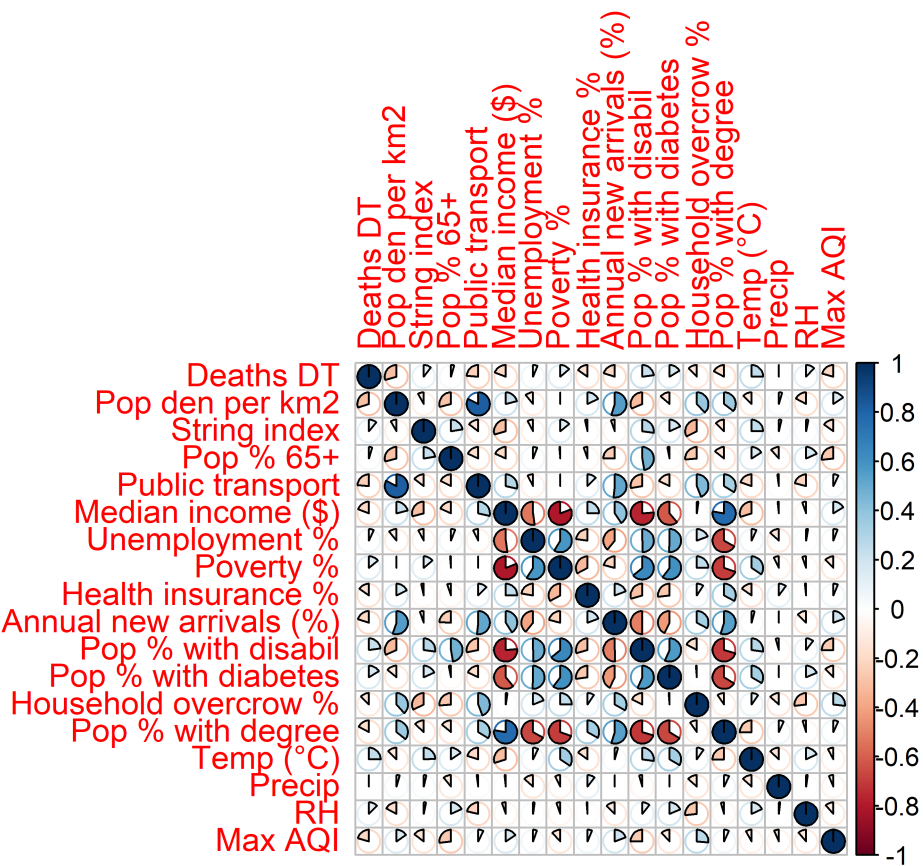


Figure C4: Correlation plot - mortality model data.

C.4    Diagnostics: Mortality model

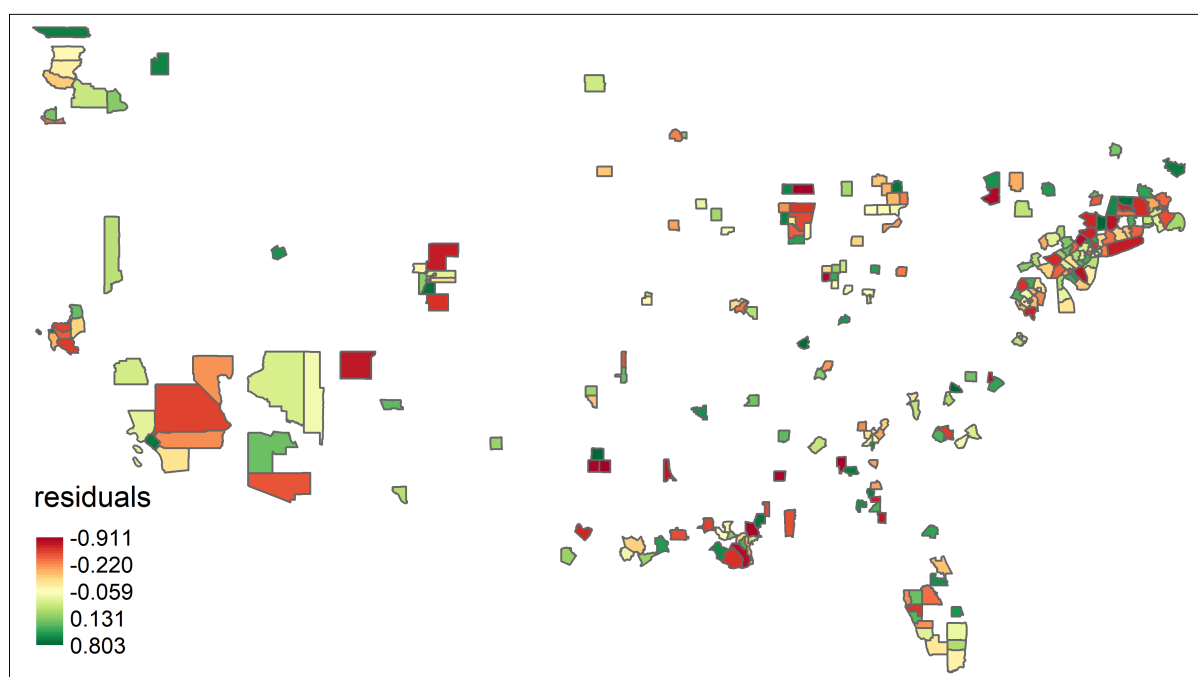


Figure C5: Spatial residuals - mortality model.

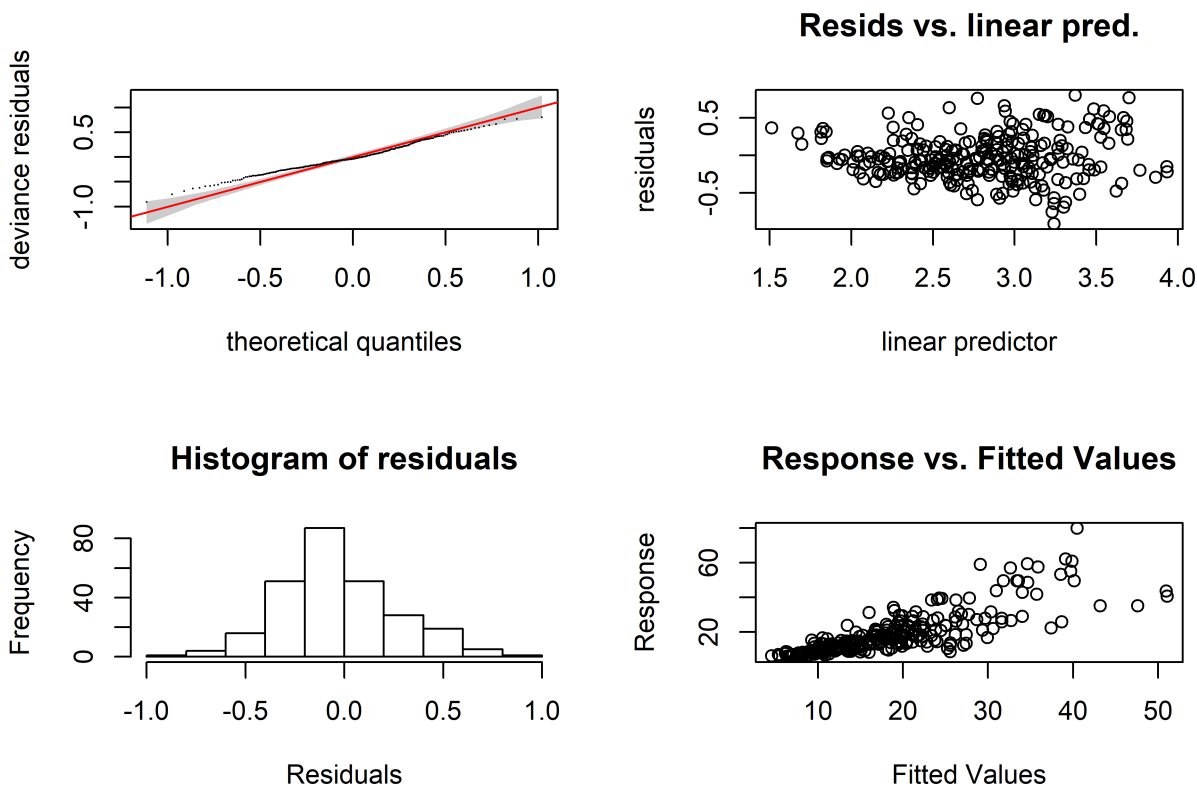


Figure C6: Model diagnostics - mortality model.