

Giacomo De Nicola

Statistical approaches for modeling network and public health data

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 25.05.2024



Giacomo De Nicola

Statistical approaches for modeling network and public health data

Dissertation an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

Eingereicht am 25.05.2024

Erster Berichterstatter: Prof. Dr. Göran Kauermann (LMU München)
Zweiter Berichterstatter: Prof. Dr. Helmut Küchenhoff (LMU München)
Dritter Berichterstatter: Prof. Dr. Ernst-Jan Camiel Wit (Università della Svizzera italiana)

Tag der Disputation: 26.07.2024

Acknowledgments

I extend my deepest gratitude to my wonderful doctoral advisor, Göran Kauermann. You showed full trust in me from the beginning of our journey, and always provided me with the perfect balance of freedom and guidance. Your leadership and overall approach to statistics and science continue to inspire me every day, and for that, I thank you.

Next, I'd like to thank Helmut Küchenhoff for serving as the internal reviewer for this thesis, as well as for all of our invaluable discussions with the rest of the CODAG-group. These conversations were not just very instructive and productive, but also a lot of fun. I am also very grateful to Ernst Wit for acting as the external reviewer for this dissertation, and to Christian Heumann and Thomas Nagler for agreeing to be part of the doctoral committee.

This thesis is largely a product of collaboration and teamwork, and I therefore want to thank all my co-authors for their help in making it happen. A huge thank you goes to my former office mate, Cornelius Fritz, for being a constant source of inspiration and motivation, as well as for our regular afternoon walks in the English Garden, which sparked countless ideas. I am also grateful to my current office mate, Martje Rave, for her unwavering support in both good and bad times, her resourcefulness, and the endless supply of snacks. My gratitude extends to Benjamin Sischka for our many hours of brainstorming and engaging conversations, and to Victor Tuekam for always pushing me to follow my dreams. I further thank Marc Schneble for our incredible synergy in working together during the COVID-19 lockdowns, Marius Mehrl for his formidable substantive input, and Ursula Berger for her infectious enthusiasm and encouragement. A special thank you goes to Maximilian Weigert, for not only being an amazing colleague, but also a true friend.

The collaborative, stimulating, and overall fun atmosphere at the Department of Statistics played no small part in the genesis of this thesis, and many of its current or former members have supported my work without directly being co-authors. I'd especially like to thank Alexander Bauer for his friendship and his inspiring approach to life. I am grateful to Cornelia Gruber for her positive energy as well as her invaluable advice on the practical aspects of life in Munich, and to Michael Lebacher for his always entertaining pessimistic optimism. I further thank Ivan Melev for our stimulating discussions on every imaginable topic, Nurzhan Sapargali for his never dull remarks, Michael Windmann for being an immovable pillar of the canteen group, and Constanze Schmaling for her singing guidance and overall dependability.

Finally, I am grateful to those outside the department who supported me throughout this journey. I thank my former flatmate, Nikolai Hörmann, for motivating and inspiring me with his passion for science, as well as for keeping me sane throughout the various stages of the pandemic. I am grateful to all my friends from the Pisa group, still holding strong since high school times. A particular thank you goes to Sergio Buttazzo, who I didn't know where to place as he is now also a partner in crime at the department, and Valerio Orsatti, for our enduring friendship of 25 years. I'd further like to thank my mother Susanna, my father Rocco, and my sister Beatrice, for their unwavering support and encouragement. Simply put, I would not have made it without them. Last but not least, an enormous thank you goes to my life partner Serena, who has been by my side each step of the way, offering me her unconditional support every single day.

Summary

The rapid technological advancement that characterized the past few decades has brought about an increasingly large amount and variety of data. This wealth of data naturally comes with further complexity, thus requiring increasingly sophisticated and efficient methodologies to extract valuable information from it. In this context, statistical models can serve as effective tools to obtain interpretable insight from the data while adequately quantifying and accounting for the underlying uncertainty. This thesis deals with the statistical modeling of two broad data categories that are prominent in modern times: network data and public health data. After an introductory Part I, the thesis comprises a total of eleven contributions, which can be divided into three further parts.

Part II, composed of four contributions, deals with the statistical analysis of network data. Networks can broadly be defined as groups of interconnected people or things. This thesis focuses mostly on social and economic networks, and on statistical models aimed at capturing and explaining the mechanisms leading to the formation of ties between actors within the network. We specifically concern ourselves with two broad model families, namely latent variable models and exponential random graph models. The first two contributions in this section introduce and compare several models from these classes, and showcase them by applying them to real-world network data. The following two contributions extend and apply these models to answer substantive questions in the social sciences. More specifically, the third contribution extends exponential random graph models to deal with the modeling of a massive dynamic bipartite network of patents and inventors to explore the drivers of innovation, while the fourth one uses latent distance models to map the network of popular Twitter users discussing the COVID-19 pandemic, with the goal of investigating polarization on the platform.

Part III, which also comprises four contributions, addresses statistical challenges related to the real-time monitoring and modeling of public health data. More specifically, the chapter tackles questions that emerged during the early stages of the COVID-19 pandemic, mainly by adapting and extending the class of generalized additive mixed models (GAMMs). The fifth contribution develops a statistical model using reported fatal infections data to predict how many of the registered infections will turn out to be lethal in the near future, thereby enabling to effectively monitor the current state of the pandemic. The sixth contribution instead focuses on all reported infections, and proposes a model to nowcast locally detected (but not yet centrally reported) cases by accounting for expected reporting delays, as well as to forecast infections at the regional level in the near future. The seventh contribution proposes a statistical tool to study the dynamics of the case-detection ratio over time, allowing for comparisons of infection figures between different pandemic phases. The chapter is concluded by the eighth contribution, which further demonstrates the effectiveness of GAMMs by applying them to three relevant pandemic-related issues, i.e. the interdependence among infections in different age groups among school children, the nowcasting of COVID-19 related hospitalizations, and the modeling of the weekly occupancy of intensive care units.

Finally, Part IV, composed of three contributions, focuses on the principled estimation of excess mortality, which can generally be defined as the number of deaths from all causes during a crisis beyond what would have been expected had the crisis not occurred. More specifically, the ninth contribution develops a point-estimation method by deploying a corrected version of classical life tables to calculate age-adjusted excess mortality, and applies it to obtain estimates the first year of

the COVID-19 pandemic (i.e. 2020) in Germany. The tenth contribution applies the same method to provide updated age-specific estimates for 2021. Finally, the eleventh contribution extends the method to incorporate uncertainty quantification, and deploys it at a broader scale to obtain estimates for 30 developed countries in the first two years of the COVID-19 crisis. The results are further compared with existing estimates published in other major scientific outlets, highlighting the importance of proper age adjustment to obtain unbiased figures.

Zusammenfassung

Der rasche technologische Fortschritt, der die letzten Jahrzehnte geprägt hat, hat eine zunehmend große Menge und Vielfalt an Daten mit sich gebracht. Diese Fülle an Daten geht natürlich mit einer erhöhten Komplexität einher und erfordert daher immer ausgefeiltere und effizientere Methoden, um wertvolle Informationen daraus zu extrahieren. In diesem Kontext können statistische Modelle als effektive Werkzeuge dienen, um interpretierbare Einblicke aus den Daten zu gewinnen und gleichzeitig die zugrunde liegende Unsicherheit angemessen zu quantifizieren und zu berücksichtigen. Diese Dissertation befasst sich mit der statistischen Modellierung von zwei wichtigen Arten von Daten der modernen Zeit: Netzwerkdaten und Gesundheitsdaten. Nach einem einführenden Teil I umfasst die Dissertation insgesamt elf Beiträge, die in drei weitere Teile unterteilt werden können.

Teil II besteht aus vier Beiträgen und behandelt die statistische Analyse von Netzwerkdaten. Netzwerke können allgemein als Gruppen von miteinander verbundenen Personen oder Objekten definiert werden. Diese Dissertation konzentriert sich hauptsächlich auf soziale und wirtschaftliche Netzwerke und auf statistische Modelle, die darauf abzielen, Mechanismen zu erfassen und zu erklären, die zur Entstehung von Verbindungen zwischen Akteuren im Netzwerk führen. Wir befassen uns insbesondere mit zwei großen Modellfamilien, nämlich latenten Variablenmodellen und exponentiellen Random-Graph-Modellen. Die ersten beiden Beiträge in diesem Abschnitt stellen mehrere Modelle aus diesen Klassen vor und vergleichen sie, indem sie auf reale Netzwerkdaten angewendet werden. Die folgenden zwei Beiträge erweitern und wenden diese Modelle an, um substantielle Fragen in den Sozialwissenschaften zu beantworten. Genauer gesagt, erweitert der dritte Beitrag exponentielle Random-Graph-Modelle, um die Modellierung eines massiven dynamischen bipartiten Netzwerks von Patenten und Erfindern zu ermöglichen und die Treiber der Innovation zu erforschen, während der vierte Beitrag latente Distanzmodelle verwendet, um das Netzwerk populärer Twitter-Nutzer zu kartieren, die über die COVID-19-Pandemie diskutieren, mit dem Ziel, die Polarisierung auf der Plattform zu untersuchen.

Teil III, der ebenfalls aus vier Beiträgen besteht, befasst sich mit statistischen Herausforderungen im Zusammenhang mit der Echtzeitüberwachung und Modellierung von Gesundheitsdaten. Genauer gesagt, behandelt das Kapitel Fragen, die in den frühen Stadien der COVID-19-Pandemie aufkamen, hauptsächlich durch Anpassung und Erweiterung der Klasse der generalisierten additiven gemischten Modelle (GAMMs). Der fünfte Beitrag entwickelt ein statistisches Modell unter Verwendung von gemeldeten tödlichen Infektionsdaten, um vorherzusagen, wie viele der registrierten Infektionen in naher Zukunft tödlich verlaufen werden, wodurch eine effektive Überwachung des aktuellen Pandemiestands ermöglicht wird. Der sechste Beitrag konzentriert sich auf alle gemeldeten Infektionen und stellt ein Modell vor, um lokal erkannte (aber noch nicht zentral gemeldete) Fälle unter Berücksichtigung erwarteter Meldeverzögerungen nowzucasten sowie Infektionen auf regionaler Ebene in naher Zukunft vorherzusagen. Der siebte Beitrag schlägt ein statistisches Werkzeug vor, um die Dynamik des “case-detection Ratios” im Laufe der Zeit zu untersuchen, was Vergleiche der Infektionszahlen zwischen verschiedenen Pandemiephasen ermöglicht. Das Kapitel wird durch den achten Beitrag abgeschlossen, der die Effektivität von GAMMs weiter demonstriert, indem sie auf drei relevante pandemiebezogene Themen angewendet werden, nämlich die Interdependenz von Infektionen in verschiedenen Altersgruppen bei Schulkindern, das Nowcasting von COVID-19-bedingten Hospitalisierungen und die Modellierung der wöchentlichen Belegung von Intensivstationen.

Teil IV, der aus drei Beiträgen besteht, konzentriert sich schließlich auf die grundlegenden Schätzung der Übersterblichkeit. Diese wird im Allgemeinen definiert, als die Anzahl der Todesfälle auf Grund von allen Ursachen während einer Krise, die über das hinausgeht, was ohne die Krise erwartet worden wäre. Genauer gesagt, entwickelt der neunte Beitrag eine Punkteschätzungsmethode durch die Verwendung einer korrigierten Version klassischer Sterbetafeln, um altersbereinigte Übersterblichkeit zu berechnen, und wendet diese Methode an, um Schätzungen für das erste Jahr der COVID-19-Pandemie (d.h. 2020) in Deutschland zu erhalten. Der zehnte Beitrag wendet dieselbe Methode an, um aktualisierte altersspezifische Schätzungen für 2021 zu liefern. Abschließend erweitert der elfte Beitrag die Methode, um die Unsicherheitsquantifizierung zu integrieren, und setzt sie in größerem Maßstab ein, um Schätzungen für 30 Industriestaaten in den ersten beiden Jahren der COVID-19-Krise zu erhalten. Die Ergebnisse werden weiter mit bestehenden Schätzungen verglichen, die in anderen bedeutenden wissenschaftlichen Publikationen veröffentlicht wurden, und heben die Bedeutung einer korrekten Altersanpassung hervor, um unverzerrte Zahlen zu erhalten.

Contents

I. Introduction and background	1
1. Introduction	3
2. Models for network data	5
2.1. Networks as random variables	6
2.1.1. Setting and notation	6
2.1.2. Random Graphs	6
2.2. Exponential random graph models	9
2.3. Latent variable network models	10
2.3.1. Stochastic blockmodels	11
2.3.2. Latent distance models	12
2.3.3. Additive and Multiplicative Effects Model	14
2.4. Contributions and discussion	15
3. Models for monitoring epidemics	19
3.1. From linear regression to GLMs	20
3.1.1. The linear regression model	20
3.1.2. Generalized linear models	21
3.2. Generalized additive mixed models	22
3.3. Contributions and discussion	23
4. Models for assessing excess mortality	27
4.1. Age-adjusted estimation	28
4.2. Uncertainty quantification	30
4.3. Contributions and discussion	30
References	33
II. Statistical network analysis	41
5. Mixture models and networks: The stochastic blockmodel	43
6. Dependence matters: Statistical models to identify the drivers of tie formation in economic networks	73
7. Modelling the large and dynamically growing bipartite network of German patents and inventors	87
8. COVID-19 and social media: Beyond polarization	109
III. Modeling and monitoring epidemics	119
9. Nowcasting fatal COVID-19 infections on a regional level in Germany	121
10. Regional now- and forecasting for data reported with delay: toward surveillance of COVID-19 infections	141

11. A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020	163
12. Statistical modelling of COVID-19 data: Putting generalized additive models to work	175
IV. Estimating excess mortality during crises	201
13. On assessing excess mortality in Germany during the COVID-19 pandemic	203
14. An update on excess mortality in the second year of the COVID-19 pandemic in Germany	221
15. Estimating excess mortality in high-income countries during the COVID-19 pandemic	227
Contributing Publications	247
Eidesstattliche Versicherung (Affidavit)	249

Part I.

Introduction and background

1. Introduction

“It is easy to lie with statistics. It is hard to tell the truth without it.”

- A. Dunkels, *Swedish mathematician (1939-1998)*

It’s hard to overstate the centrality of data in society. Data, in the broad sense of the word, shaped human behavior for millennia. Consider, for instance, a scenario in which a caveman repeatedly consumes a poisonous berry, experiencing adverse effects each time. Through observation, the caveman associates the ingestion of this berry with negative outcomes, prompting him to stop eating it and likely signaling its peers to avoid it as well. This is a primitive example of data-driven decision-making. With time, civilizations evolved and expanded, and thus got better at collecting, storing and sharing information. This upward trajectory took a decisive acceleration in the last decades, which, together with the advent of the computer, brought upon what has been described as the information age, a time characterized by its wealth of available data (Kline, 2015). But the truth is that reality is complex, often much more so than it may look at a first glance: as data becomes more and more ubiquitous, it also becomes increasingly clear that raw information is nothing without proper tools to extract real insight from it. What good is a raw figure without understanding the context surrounding it? In fact, improper use of data can even be worse than having no data at all, as it can be misleading or worse, aid in coating lies with a veil of “factual evidence”. In this environment, statistical models have emerged as principled tools to obtain interpretable insight from different types of data, while at the same time adequately quantifying and accounting for the underlying uncertainty. In fact, past decades have been defined as “a golden age of statistical methodology” (Efron and Hastie, 2021) for the central role that such tools have played in recent times. But, crucially, each data constellation is unique in its own way, and thus requires tailored modeling approaches to reveal its secrets. In this context, statistics, and particularly applied statistics, must rise up to the challenge to address the growing need of society to answer substantive questions in diverse domains. More in particular, applied data scientists are tasked with the triple role of *designing*, *implementing*, and *applying* sensible statistical methods to solve different classes of problems in the empirical sciences by means of data analysis. All three aspects are necessary: *designing*, as datasets need appropriate theories on how the data-generating process functions; *implementing*, as the best theories can often remain in the shadows if they are not accompanied by functioning tools to put them to work; and *applying*, as there is no better way to facilitate the diffusion of a new technology than demonstrating its effectiveness in the empirical realm. The present work is set in this context, and is concerned with all three of these dimensions. More specifically, the thesis is motivated by data-driven challenges posed by modern society, and deals with designing, extending, and leveraging modern statistical tools to answer classes of questions posed within the economic, social, and public health sciences, with particular focus on applications with tangible real-world impact. To be more precise, the scope of this work can broadly be categorized in three parts, namely (a) social and economic network modeling, (b) models for monitoring epidemics, and (c) methods for excess mortality

estimation. These three topics respectively make up Parts [II](#), [III](#), and [IV](#) of the thesis, where each of these parts is composed by multiple contributions, for a total of eleven published journal articles. The current section, i.e. Part [I](#), instead serves as an introduction to the three topics just outlined. More specifically, the goal of this introductory part is to frame each of the parts in its broader statistical context, summarize the statistical tools upon which the contributions build, and give a brief overview of the research questions tackled and the original contributions provided. In essence, this introduction aims at equipping the reader with the necessary knowledge and tools to navigate the remainder of the thesis, which as a whole can be viewed as a demonstration of how statistics can be put to work to break down complexities present in modern society.

2. Models for network data

Networks can broadly be defined as systems of interconnected people or things. Within the context of this thesis, we can more specifically define networks as data structures that can be represented through use of graphs, where nodes represent actors in the network, and edges represent ties within them. Such structures are ubiquitous in diverse empirical settings, stemming from, e.g., social, political, economic, health, and natural sciences. As the data class is so general, networks can extensively be studied from various different perspectives. This thesis will focus on analyzing and modeling networks from a statistical view. Specifically, given a network, we here focus on understanding the mechanisms contributing to its generation, thereby investigating the drivers behind the formation of ties between actors. The special characteristic of network data, which makes this feat particularly tricky, is that ties between actors are generally dependent on one another. This means that standard regression techniques assuming conditional independence between observations are usually not applicable. Instead, modeling approaches tailored to networks need to be employed. This chapter is dedicated to introducing the different families of network models that are discussed, extended and applied in the contributions included in Part II of this thesis.

Note that, of course, networks can be very different from one another, and different networks can have very different generative mechanisms. For example, a friendship network will evolve fundamentally differently from a network of neurons, a power grid, or an ecological web. As a result, there is no single network model capable of capturing the generative mechanisms of all networks. However, some network models are flexible enough to be adapted to different dependence structures. Moreover, networks pertaining to similar phenomena tend to follow similar patterns. The contributions included in this thesis predominantly focus on networks driven by interactions between humans or groups of humans, such as e.g. networks of friendship, collaborations, or other types of active relationships linking together people or other human-driven entities, such as companies or countries. The models that we will introduce here will therefore mainly be discussed using this type of network structures as their target. Note that this section does not aim to be an exhaustive treatment of statistical network analysis, but merely aims at introducing the general framework and providing the reader with the necessary tools to more easily dive into the related contributions. For an overview of the field, we instead refer to [Goldenberg et al. \(2010\)](#) and [Kolaczyk \(2009\)](#).

The remainder of this chapter is structured as follows. Section 2.1 introduces the required notation and presents the general framework of random graph models. Section 2.2 discusses the class of exponential random graph models, while Section 2.3 focuses on latent variable models for network data. Section 2.4 concludes the chapter by discussing the contributions contained in Part II, and gives some remarks on promising future developments in the field.

2.1. Networks as random variables

2.1.1. Setting and notation

Before digging into network models, we briefly introduce the mathematical framework for networks, as well as the necessary notation. As anticipated in the previous section, in this thesis we formalize networks as graphs. Formally, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is composed by a set of N vertices $\mathcal{V} = \{1, \dots, N\}$ and a set of M edges $\mathcal{E} = \{(i_1, j_1), \dots, (i_M, j_M)\} \subseteq \mathcal{P} = \{(i, j); i, j \in \mathcal{V}\}$, with $|\mathcal{E}| = M$. In this notation, \mathcal{P} is the set of all possible pairs of nodes, and the set of edges \mathcal{E} includes all pairs where the actors are actually connected in the graph. Vertices can also interchangeably be referred to as actors or nodes, while edges are also known as links or ties. To make them easier to handle, graph can also be expressed in matrix form through their so-called “adjacency matrix”. Given a binary graph, its adjacency matrix $\mathbf{y} = (y_{ij})_{i,j=1,\dots,n}$ is an $N \times N$ matrix where each entry y_{ij} is given by:

$$y_{ij} = \begin{cases} 1, & \text{if } (i, j) \in \mathcal{E} \\ 0, & \text{otherwise.} \end{cases}$$

In other words, $y_{ij} = 1$ indicates the presence of an edge from node i to node j , while $y_{ij} = 0$ translates to no edge between the two. Note that a graph can either be “directed” or “undirected”. In a directed graph, the direction of an edge carries additional information, while if the network is undirected $y_{ij} = y_{ji}$ holds $\forall i, j \in \mathcal{V}$. Various examples of networks of both types will be illustrated over the course of this thesis. Furthermore, since self-ties are not admitted for most studied networks, the diagonal of \mathbf{y} is usually left unspecified or set to zero.

In addition to the edges, we often observe additional information on the nodes composing the network, as well as on the relationships between them. We denote covariate information at the nodal level, such as e.g. the gender or political affiliation of a person, by $x_{\text{node}} = (x_1, \dots, x_N)$. Similarly, we indicate dyadic covariates, such as for example an indicator of whether or not two people live in the same area, or the age differential between them, with $x_{\text{dyad}} = (x_{12}, \dots, x_{N(N-1)})$. In theory, covariates at higher levels (such as e.g. triadic covariates) would also be admissible; however, these are not present in the cases studied in this thesis. Also note that, in this introductory section, we start by discussing the simplest type of networks, that is static (i.e. non time dependent), unimodal (i.e. with only one node type), and binary (i.e. where each edge can only be present or absent). However, the thesis also considers extensions and applications to dynamic, bipartite, and weighted networks. Those more complex cases will be gradually introduced as they come up in the contributions, building on the foundational framework given in this introduction.

2.1.2. Random Graphs

To study the generative mechanisms behind networks, we first need to make a key abstraction, that is to consider networks not as constant, static entities, but rather as random variables. More specifically, we consider random graphs as entities where the set of nodes is assumed to be fixed, while the presence (or absence) of each edge is random. Our goal is then to specify a statistical model for the random graph \mathbf{Y} , that is the random variable corresponding to the observed network \mathbf{y} . A natural way to do this is to specify a probability distribution over the space of all possible

2.1 Networks as random variables

networks, which we define by the set \mathcal{V} . The simplest possible model for this task is the so-called Bernoulli graph, also termed Erdős-Rényi-Gilbert model, where all edges are assumed to be independent and to have the same probability of being observed (Erdős and Rényi, 1959; Gilbert, 1959). In stochastic terms, each observed tie is then a realization of a Bernoulli random variable with success probability π , i.e.

$$\mathbb{P}_\pi(Y_{ij} = 1) = \pi \quad (2.1)$$

for any pair of nodes $i, j \in \mathcal{V}$. The above, given the assumed edge independence, leads to the following probability distribution over the whole network:

$$\mathbb{P}_\pi(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^n \prod_{j \neq i} \pi^{y_{ij}} (1 - \pi)^{1 - y_{ij}}. \quad (2.2)$$

Evidently, model (2.2), which implies equal probability for all possible ties, is too restrictive to be applied to real world problems. It is unrealistic to assume edges to be completely random and independent from any other factor, including nodal and pairwise characteristics. The natural next step is, therefore, to let the edge probability π vary depending on the features of the nodes involved in each tie, leading to edge-specific probabilities π_{ij} . To do so, we can follow the commonly used logistic regression framework, and parameterize the log-odds of an edge by $\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \theta^\top x_{ij}$, where x_{ij} is a vector of covariates with the first entry set to 1 to incorporate an intercept. For the probability of the whole network, we then get

$$\mathbb{P}_\theta(\mathbf{Y} = \mathbf{y} | x) = \prod_{i=1}^n \prod_{j \neq i} \left(\frac{\exp\{\theta^\top x_{ij}\}}{1 + \exp\{\theta^\top x_{ij}\}} \right)^{y_{ij}} \left(\frac{1}{1 + \exp\{\theta^\top x_{ij}\}} \right)^{1 - y_{ij}}. \quad (2.3)$$

Note that the logistic structural assumption is just one of the possible different model formulations. In fact, looking at (2.3), it is apparent that this model is just a special case of a generalized linear model, which allows for flexible model choice within the exponential family (Nelder and Wedderburn, 1972). But while this formulation is already a lot more flexible than the Bernoulli graph we started with, a key ingredient for properly modeling real-world networks is still missing. While we are allowing the probability of an edge to depend on the nodal and pairwise characteristics of the nodes involved in a pair, model (2.3) still assumes ties to be independent conditionally on such covariates. This assumption is often unreasonable in practice: In the context of an international relations network, it would, for example, imply that Germany imposing economic sanctions on Russia is independent of Italy imposing sanctions on Russia, and, in the directed case, even of Russia imposing them on Germany itself. In general, real world networks, and particularly networks driven by human interaction, are often heavily shaped by dependence between edges. In the following, some of the most common endogenous network mechanisms present in human-driven networks are described.

Reciprocity: In a directed network, reciprocity describes the tendency of directed ties to be reciprocated. This tendency was first documented by Newcomb (1979). For example, given that person i likes person j , it is often more likely for person j to also like person i .

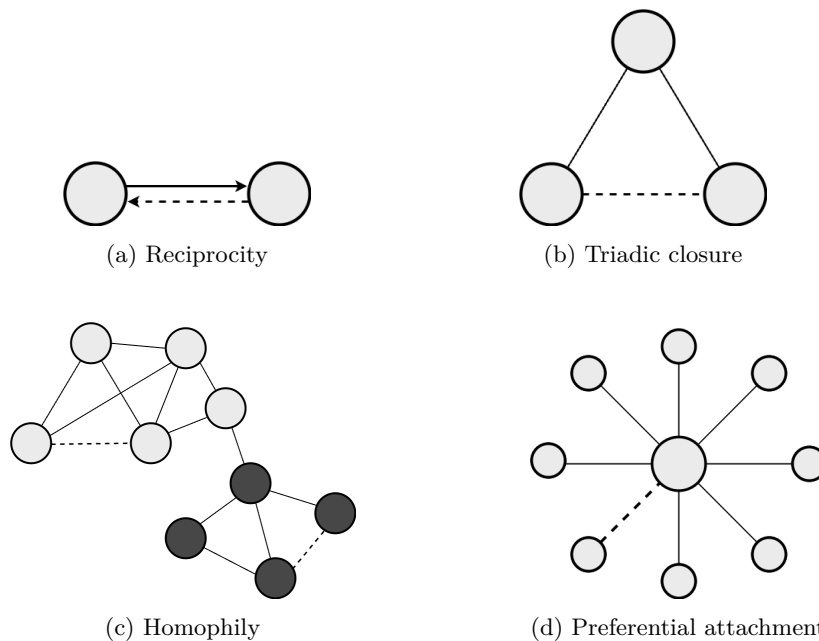


Figure 2.1.: Typical patterns observed in networks driven by human interaction

Transitivity: Also known as triadic closure, transitivity describes the importance of shared ties in forming new ones (Davis, 1970). This mechanism can also be summarized with the common saying “a friend of my friend is (likely to be) my friend”. For example, if country i has a good relationship with country j , and j has a good relationship with country z , i is likely to also be on good terms with z .

Homophily: This term describes the tendency of similar nodes to form connections with each other (Rivera et al., 2010). Human-driven networks are often homophilic: for example, people of similar age, or sharing similar political opinions, are often more likely to form social connections with each other. Note, however, that these characteristics may either be observed or unobserved. This distinction gives rise to two different types of homophily, which can be defined as “observed homophily” and “latent homophily”, respectively. In the former case, the strength of the mechanism can simply be estimated as a function of exogenous covariates. In contrast, the latter case can be viewed as an endogenous network mechanism, as the latent similarity can only be tracked down through the connectivity behavior of the nodes.

Preferential attachment: This phenomenon, also known as “rich get richer”, refers to the tendency of nodes with many ties to attract even more. For example, accounts with lots of friends or followers on social media are generally more likely than others to attract new connections.

These four mechanisms, for which a visual representation is given in Figure 2.1, are just examples of the possible network-driven dependencies that can arise between ties. Such dependencies are the reason why we need to go beyond standard regression models, and instead use special techniques to model network data. We are not in a classical $X \rightarrow Y$ setting, but rather in an $X, Y \rightarrow Y$ one, i.e. there is simultaneous dependency within the response variable (the edges of the network itself). To account for this endogenous dependence, several extensions to regression models have been

2.2 Exponential random graph models

introduced over the years. In this thesis, we will specifically focus on two of the most prominent model families that have been proposed for this task, namely exponential random graph models (Robins et al., 2007a) and latent variable network models (Matias and Robin, 2014). Albeit from very different angles, both of these model classes are suitable to simultaneously capture the mechanisms leading to network formation, i.e. to measure how the probability of forming a tie is influenced by (a) nodal characteristics, (b) pairwise covariates, and (c) the rest of the network. The following two sections motivate and introduce the two model families, with the goal of enabling the reader to quickly dive into the contributions relating to this chapter.

2.2. Exponential random graph models

The exponential random graph model (in short ERGM) is one of the most popular models for analyzing network data. First introduced by Holland and Leinhardt (1981) as a model class that builds on the platform of exponential families, it was later extended with respect to fitting algorithms and more complex dependence structures (Lusher et al., 2012; Robins et al., 2007b). The general idea of this model class is to generalize the simplistic Erdős-Rényi-Gilbert random graph model and its version with covariates, given in (2.2) and (2.3), respectively, to account for specific dependence structures between the network’s edges. To do so, the ERGM exploits the properties of the exponential family of distributions, building on the framework of generalized linear models while incorporating additional statistics into the model equation. More specifically, while in a traditional regression model the set of sufficient statistics is only a function of covariates, in the realm of ERGMs it can also include statistics related to the network itself, such as for example the count of reciprocated ties, or the number of triangles in the network.

An exponential random graph model can thus be seen as an extension of a generalized regression model for the joint distribution of the edges of the network \mathbf{Y} . But this generalization did not come all at once: a first extension in this direction was performed by Holland and Leinhardt (1981), who extended model (2.2) to such settings with the so-called p_1 model. To represent reciprocity, the authors assume dyads, defined by (Y_{ij}, Y_{ji}) , to be independent of one another, which again yields an exponential family distribution similar to (2.3) with structural statistics counting the number of mutual ties ($s_{\text{Mut}}(\mathbf{y}) = \sum_{i < j} y_{ij}y_{ji}$), of edges ($s_{\text{Edges}}(\mathbf{y}) = \sum_{i=1}^n \sum_{j \neq i} y_{ij}$), and the in- and out-degree statistics for all degrees observed in the networks. Note that the actors’ in- and out-degrees are their number of incoming and outgoing edges, and relate to their relative position in the network (Wasserman and Faust, 1994).

Next to reciprocity, another important endogenous network mechanism is transitivity, originating in the structural balance theory of Heider (1946) and adapted to binary networks by Davis (1970). Transitivity affects the clustering in the network, implying that a two-path between actors i and j , i.e. $y_{ih} = y_{hj} = 1$ for some other actor h , affects the edge probability of Y_{ij} . Put differently, Y_{ij} and Y_{kh} are assumed to be independent iff $i, j \neq k$ and $i, j \neq h$. Frank and Strauss (1986) proposed the Markov model to capture such dependencies. This model incorporates counts of triangular structures as well as star-statistics, which are counts of sub-structures in the network where one actor has edges to between 0 and $n - 1$ other actors.

After allowing for specific type of edge dependence, the next step was to allow for more general dependence structures, and specify a probabilistic model for \mathbf{Y} directly through the sufficient statistics. Wasserman and Pattison (1996) introduced this model as

$$\mathbb{P}_\theta(\mathbf{Y} = \mathbf{y}) = \frac{\exp\{\theta^\top \mathbf{s}(\mathbf{y})\}}{\kappa(\theta)}, \quad (2.4)$$

where θ is a p -dimensional vector of parameters to be estimated, $\mathbf{s}(\mathbf{y})$ is a function calculating the vector of p sufficient statistics for network \mathbf{y} , and $\kappa(\theta) = \sum_{\tilde{\mathbf{y}} \in \mathcal{Y}} \exp\{\theta^\top \mathbf{s}(\tilde{\mathbf{y}})\}$ is a normalizing constant to ensure that (2.4) sums up to one over the set of all possible networks. To estimate θ , Handcock (2003) adapted the Monte Carlo Maximum Likelihood technique of Geyer and Thompson (1992), approximating the logarithmic likelihood ratio of θ and a fixed θ_0 via Monte Carlo quadrature (see Hunter et al., 2012, for an in-depth discussion).

Within the general framework of model (2.4), the possible range of network-based sufficient statistics is virtually endless, allowing to test for the presence of endogenous mechanisms such as reciprocity and transitivity, and to measure the tendency of patterns such as cycles and stars to form (see Morris et al., 2008 for a survey on possible model configurations). However, it is important to keep in mind that the estimation of ERGMs is prone to degeneracy (Handcock, 2003; Schweinberger, 2011). Degenerate models are characterized by probability distributions that put most probability mass either on the empty or on the full network, i.e., where either all or no ties are observed (Hunter et al., 2008). Such models are, of course, not at all appropriate to represent real data. To avoid degeneracy, it is important to keep the model specification concise, and to choose the sufficient statistics sensibly, while monitoring goodness of fit (see Hunter et al., 2012). For more details on ERGM specification and estimation we refer to Lusher et al. (2012), who also discuss the model class from a broader perspective.

We further note that, along with capturing endogenous network statistics, it is also possible to extend the ERGM framework to include the temporal dimension, that is, to model longitudinal network data. This can be done quite naturally through use of a Markov assumption on the temporal dependence of subsequently observed networks, giving rise to the Temporal Exponential Random Graph Model (TERGM). We refer to Hanneke et al. (2010) for an introduction to this temporal variant (see also Krivitsky and Handcock, 2014 for further developments). In summary, the ERGM allows to account for network dependencies via explicitly specifying them in $\mathbf{s}(\mathbf{y})$. A large variety of potential network statistics can be included in $\mathbf{s}(\mathbf{y})$, enabling to test for their influence in the formation of the observed network. However, this wide range of possibilities does not come without caveats, including the previously mentioned issue of degeneracy, thus requiring the user to at least have an implicit theory regarding what types of network dependence should exist in the studied network before fitting the model.

2.3. Latent variable network models

Another prominent option for modeling network data is offered by latent variable network models. Models within this broad class start from the key assumption that endogenous dependence between edges can be explained by unobserved nodal characteristic. These models thus postulate the existence of latent variables Z_i in association with each node i , and that, crucially, all edges Y_{ij} are independent conditionally on these latent variables. This implies that the probability of the whole network \mathbf{Y} can factorize as:

2.3 Latent variable network models

$$\mathbb{P}_\theta(\mathbf{Y} = \mathbf{y} | \mathbf{z}) = \prod_{i=1}^N \prod_{j \neq i} \mathbb{P}_\theta(Y_{ij} | z_i, z_j). \quad (2.5)$$

Moreover, the conditional distribution of each tie Y_{ij} is assumed to only depend on Z , that is

$$\mathbb{P}_\theta(Y_{ij} | z_i, z_j) = f(z_i, z_j). \quad (2.6)$$

While these assumption may seem strict, this general definition is quite broad, as the latent variables can take many different forms, leading to flexible model specifications. Clearly, the main difficulty lies in estimating the latent structure, which will allow us to gain an understanding of the underlying network mechanisms at play, or to at least account for them. One main distinction within the class of latent variable models can be drawn between discrete latent variables (e.g. indicating group memberships for each node) and continuous ones (Matias and Robin, 2014). In this section, we will briefly go over the most prominent of these models, which will then be covered more in depth, applied, and extended in the contributions relating to this chapter.

2.3.1. Stochastic blockmodels

One of the simplest, and yet perhaps still the most widely popular latent variable model for network data, is the so-called stochastic blockmodel (SBM). The model assumes that each node belongs to a latent, categorical class (also known as group, or block). Nodes within each class are assumed to be stochastically equivalent in their connectivity behavior, meaning that the probability of two nodes to connect depends solely on their group memberships. From a statistical perspective, the SBM can be viewed as a mixture model in which each mixture component is given by group membership. More formally, we assume the independent discrete group indicator coefficients $Z_i \in \{1, \dots, K\}$ for $i = 1, \dots, N$, with

$$\mathbb{P}(Z_i = k) = \pi_k \quad \text{for } k = 1, \dots, K,$$

and $\sum_{k=1}^K \pi_k = 1$. For a binary, undirected network, each tie probability between nodes i and j is then a simple Bernoulli random variable governed solely by the connection probability between the two groups to which i and j belong, i.e.

$$\mathbb{P}(Y_{ij} = y_{ij} | z_i, z_j) = (p_{z_i z_j})^{y_{ij}} (1 - p_{z_i z_j})^{(1 - y_{ij})}, \quad (2.7)$$

where $\mathbf{P} = (p_{h,l})_{h,l=1,\dots,K}$ is a $K \times K$ block-probability matrix, in which each entry p_{hl} is given by the connection probability between groups h and l . Note that the number of blocks K is generally assumed to be known, and has to be appropriately chosen by the user (see Lee and Wilkinson, 2019 for a list of possible approaches to this task).

The initial versions of the SBM, as introduced by Holland et al. (1983), assumed group memberships for each node to be known, rendering the model fairly trivial. The first steps towards “a posteriori” blockmodeling, that is modeling with initially unknown group structure, were taken

by [Snijders and Nowicki \(1997\)](#) and [Nowicki and Snijders \(2001\)](#), who proposed estimation routines for, respectively, two groups and any known number of groups. From there, the model class gained traction, mostly as a principled way to classify nodes into groups based on their connectivity behavior. In a sense, SBMs can be seen as a tool for performing community detection, which can generally be understood as clustering nodes into densely connected communities, as depicted in [Figure 2.1c](#) (see e.g. [Clauset et al., 2004](#), [Fortunato, 2010](#), [Fortunato and Hric, 2016](#)). However, for community detection one typically assumes that $p_{hh} > p_{hl}$ for all $h \neq l$, which is not a requirement for SBMs. In fact, the block-structure may describe clusters of nodes that behave similarly from a connectivity standpoint without necessarily being more densely connected, thus allowing for other types of structures, such as disassortative communities and core-periphery (see [Fortunato and Hric, 2016](#) for more details).

Following their initial formulation, stochastic blockmodels have been extended in various ways for different purposes. A well known extension of the classical SBM is the degree-corrected stochastic blockmodel, introduced by [Karrer and Newman \(2011\)](#). In their work, the authors show how the standard SBM implicitly assumes the degree structure within communities to be relatively homogeneous. This, combined with the fact that many real-world networks exhibit extremely skewed degree distributions ([Simon, 1955](#)), leads the model to often only be able to find core-periphery type block structures, where node grouping is predominantly driven by degree similarity. To bypass this issue, [Karrer and Newman \(2011\)](#) introduced the idea of degree correction, making the probability of an edge depend not only on group membership, but also on node-specific heterogeneity parameters. This leads the model to find traditionally “dense” groups, more in line with the conventional concept of community detection. Besides the degree-corrected SBM, other notable extensions include the mixed membership model ([Airoldi et al., 2008](#)), in which nodes can belong to multiple communities simultaneously, and the hierarchical stochastic blockmodel ([Peixoto, 2017](#)), in which communities are comprised of meta-communities, leading to a hierarchical block structure. It is also possible to add covariates to the analysis, as initially proposed by [Tallberg \(2005\)](#) and further elaborated by [Choi et al. \(2012\)](#), [Sweet \(2015\)](#), and [Huang et al. \(2023\)](#). A further extension is the mixture of experts SBM (see [Gormley and Murphy, 2010](#) and [White and Murphy, 2016](#)), which allows covariates to enter the latent position cluster model in a number of ways, yielding different model interpretations. Surveys on recent developments in this field have been published by [Abbe \(2018\)](#) and [Lee and Wilkinson \(2019\)](#).

The stochastic blockmodel owes its success largely to its simplicity and flexibility in uncovering and describing subgroups of nodes within networks. Being able to classify nodes in different categories based on their connectivity behavior is indeed attractive for diverse applications. The simplicity of the SBM, however, can also be viewed as its major shortcoming, as in several domains discrete groupings fail to adequately represent the observed data. For settings where agents behave more heterogeneously than can be described via simple groupings, it can thus be useful to replace the discrete random variables with continuous ones, as showcased in the upcoming sections.

2.3.2. Latent distance models

A prominent approach in the realm of continuous latent variable models for network data is offered by the latent distance model (LDM). Initially proposed by [Hoff et al. \(2002\)](#), the model postulates that agents are positioned in a latent Euclidean “social space”, and that the closer they are within

2.3 Latent variable network models

it, the more likely they are to form ties. More precisely, the classical latent distance model specifies the probability of observing an edge between nodes i and j through

$$\mathbb{P}_\theta(Y_{ij} = 1 | \mathbf{Z}) = \frac{\exp\{\theta^\top x_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\|\}}{1 + \exp\{\theta^\top x_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\|\}}, \quad (2.8)$$

where $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ denotes the latent positions of the nodes in the d -dimensional latent space, where d is assumed to be known, and θ is the coefficient vector for the covariates x_{ij} . The latent positions \mathbf{Z} are assumed to originate independently from a spherical Gaussian distribution, i.e. $Z \sim N_d(0, \tau^2 \mathbf{I}_d)$, where \mathbf{I}_d indicates a d -dimensional identity matrix.

Latent distance models are useful tools to graphically represent network structures, since low-dimensional Euclidean spaces lend themselves well to visualization. Unlike standard graph visualization algorithms, which are based on heuristics (see e.g. Kamada et al., 1989 and Fruchterman and Reingold, 1991), the resulting plot will have a probabilistic interpretation, enhancing its interpretability. Indeed, visually inspecting the estimated latent space can provide very useful in understanding the overall structure of the network, as well as the dynamics at play. Moreover, it is possible to use LDMs for clustering purposes: Handcock et al. (2007) extended the class by including model-based clustering to the original formulation of the latent distance model, allowing for the actors' positions in the latent space to come from a mixture of G normal distributions, where each mixture component represents a cluster. More formally, for the latent positions Z_i holds

$$Z_i \stackrel{iid}{\sim} \sum_{g=1}^G \lambda_g \text{MVN}_d(\mu_g, \tau_g^2 \mathbf{I}_d). \quad (2.9)$$

Furthermore, it's possible to add nodal random effects to the model to control for agent-specific heterogeneity in the propensity to form edges (Krivitsky et al., 2009). The model then becomes

$$\mathbb{P}_\theta(Y_{ij} = 1 | \mathbf{Z}, a, b) = \frac{\exp\{\theta^\top x_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\| + a_i + b_j\}}{1 + \exp\{\theta^\top x_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\| + a_i + b_j\}}, \quad (2.10)$$

where $a = (a_1, \dots, a_n)$ and $b = (b_1, \dots, b_n)$ are node-specific sender and receiver effects that account for the individual agents' propensity to form ties, with $a \sim N_n(0, \tau_a^2 \mathbf{I}_n)$ and $b \sim N_n(0, \tau_b^2 \mathbf{I}_n)$. Several other extensions to the model class have been proposed: we refer to Kaur et al. (2023) for an overview.

Latent distance models are particularly attractive for social networks in which both homophily (see Figure 2.1c) and triadic closure (see Figure 2.1b) play a major role. This is because the euclidean distance naturally lends itself to representing those mechanisms: if two points are both close to a third point in the space, they will also automatically be close to each other. Similarly, if two nodes are similar in terms of connectivity behavior, they will tend to be close to the same nodes, which in turn will also lead them to be close to one another. However, despite its advantages and fairly simple interpretation, a Euclidean latent space is unable to effectively approximate the behavior of networks where nodes that are similar in terms of connectivity behavior are not necessarily more likely to form ties (Hoff, 2008), such as, e.g., many networks of amorous relationships (Ghani et al., 1997; Bearman et al., 2004). Indeed, if nodes who have similar connections tend not to connect with one another, the latent space will not be able to properly place them. More generally, the latent distance model tends to perform poorly for networks in which stochastic equivalence does not imply homophily, i.e., when nodes which behave similarly in terms of connectivity patterns

towards the rest of the network do not necessarily have a higher probability of being connected among themselves. A better option for such cases is offered by multiplicative latent positions, as discussed in the next section.

2.3.3. Additive and Multiplicative Effects Model

Many real world networks exhibit varying degrees and combinations of stochastic equivalence, triadic closure and homophily. Moreover, it is often *a priori* unclear which of these mechanisms are at play in a given network. In this context, node-specific multiplicative random effects (in place of the Euclidean latent positions) allow for simultaneously representing all these patterns (Hoff, 2005). Further developments of this innovation have led to the modern specification of the additive and multiplicative effects network model (AME, Hoff, 2011), which, from a matrix representation perspective, generalizes both the stochastic blockmodel and the latent distance model (Hoff, 2021).

The AME approach can be motivated by considering that network data often exhibit first-, second- and third-order dependencies. *First-order effects* capture agent-specific heterogeneity in sending (or receiving) ties within a network. For example, in the case of companies and legal disputes, first-order effects can be viewed as the propensity of each firm to initiate (or be hit by) legal disputes. *Second-order effects*, i.e. reciprocity, describe the statistical dependence of the directed relationship between two agents in the network, as illustrated in Figure 2.1a. In the previous example, this effect can be described as the correlation between (a) company i initiating a legal dispute against company j and (b) j doing the same towards i . Of course, second-order effects can only occur in directed networks. *Third-order effects* can instead be described as the dependency within triads, defined as the connections between three agents, and relate to the triangular statistics previously depicted in Figure 2.1b. How likely is it that “a friend of a friend is also my friend”? Or, returning to the previous example: given that i has legal disputes with j and k , how likely are disputes to occur between j and k ?

The AME network model is designed to simultaneously capture these three orders of dependencies. More specifically, it extends the classical (generalized) linear modeling framework by incorporating extra terms into the systematic component to account for them. In the case of binary network data, we can make use of the Probit AME model. As is well known, the classical Probit regression model can be motivated through a latent variable representation in which y_{ij} is the binary indicator that some latent normal random variable, say $L_{ij} \sim \mathcal{N}(\theta^\top \mathbf{x}_{ij}, \sigma^2)$, is greater than zero (Albert and Chib, 1993). But an ordinary Probit regression model assumes that L_{ij} , and thus the binary indicators (edges) y_{ij} , are independent, which is generally inappropriate for network data. In contrast, the AME Probit model specifies the probability of a tie y_{ij} from agent i to agent j , conditional on a set of latent variables W , as

$$\mathbb{P}(Y_{ij} = 1|W) = \Phi(\theta^\top \mathbf{x}_{ij} + e_{ij}), \quad (2.11)$$

where Φ is the cumulative distribution function of the standard normal distribution, $\theta^\top \mathbf{x}_{ij}$ accommodates the inclusion of dyadic, sender, and receiver covariates, and e_{ij} can be viewed as a structured residual, containing the latent terms in W to account for the network dependencies described above. In the directed case, e_{ij} is composed as

$$e_{ij} = a_i + b_j + u_i v_j + \epsilon_{ij}. \quad (2.12)$$

2.4 Contributions and discussion

In this context, a_i and b_j are zero-mean additive effects for sender i and receiver j accounting for first-order dependencies, jointly specified as

$$(a_1, b_1), \dots, (a_n, b_n) \stackrel{\text{i.i.d.}}{\sim} N_2(0, \Sigma_1), \quad \text{with} \quad \Sigma_1 = \begin{pmatrix} \sigma_a & \sigma_{ab} \\ \sigma_{ab} & \sigma_b \end{pmatrix}. \quad (2.13)$$

The parameters σ_a and σ_b measure the variance of the additive sender and receiver effects, respectively, while σ_{ab} relates to the covariance between sender and receiver effects for the same node. Going back to (2.12), ϵ_{ij} is a zero-mean residual term which accounts for second order dependencies, i.e. reciprocity. More specifically, it holds that

$$\{(\epsilon_{ij}, \epsilon_{ji}) : i < j\} \stackrel{\text{i.i.d.}}{\sim} N_2(0, \Sigma_2), \quad \text{with} \quad \Sigma_2 = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad (2.14)$$

where σ^2 denotes the error variance and ρ determines the correlation between ϵ_{ij} and ϵ_{ji} , thus quantifying the tendency towards reciprocity. Finally, \mathbf{u}_i and \mathbf{v}_j in (2.12) are d -dimensional multiplicative sender and receiver effect vectors that account for third-order dependencies, and for which $(u_1, v_1), \dots, (u_n, v_n) \sim \mathcal{N}_{2d}(0, \Sigma_3)$ holds. Note that the dimensionality d is generally assumed to be known. Also note that we here limited ourselves to the directed case: For undirected networks, the model is defined in a slightly different way (see Hoff, 2021).

As noted above, AME is able to represent a wide variety of network structures, generalizing several other latent variable model classes. Moreover, the model class is suitable for modeling networks where edges represent “negative” ties, such as animosity or acts of violence (see e.g. Dorff et al., 2020). This generality comes at the price of a high level of complexity for the estimated latent structure. This can make the model class a sub-optimal choice if one wants to interpret the latent structure with respect to, e.g., clustering. On the other hand, its flexibility makes it an ideal fit when the underlying network dependencies are unknown, and the researchers’ interest mainly lies in evaluating and interpreting the effect of dyadic and nodal covariates on tie formation while controlling for network effects to avoid confounding and bias in the estimates (see Lee and Ogburn, 2021).

2.4. Contributions and discussion

The field of network analysis has experienced significant development in recent years, extending its roots in various disciplines, from sociology to physics, to economics and computer science. This chapter specifically focused on networks as seen from a statistical perspective. More in particular, some of the most prominent techniques for understanding and modeling the process leading to tie formation were covered. These models include the exponential random graph model as well as several types of latent variable models, such as stochastic blockmodels, latent distance models, and additive and multiplicative effects models. The focus on those is also motivated by the fact that these techniques are discussed, extended, and applied in the contributions included in Part II, which relates to this chapter. In the following, a short summary of these contributions is provided.

Chapter 5, [De Nicola, Sischka, and Kauermann \(2022\)](#): In this contribution, we consider stochastic blockmodels and some of their variants and extensions by framing them in the general context of mixture models. We also explore some of the main classes of estimation methods available, and propose an alternative approach based on the reformulation of the blockmodel as a graphon. In addition to the discussion of inferential properties and estimating procedures, we focus on the application of the models to several real-world network datasets, showcasing the advantages and pitfalls of different approaches.

Chapter 6, [De Nicola, Fritz, Mehrl, and Kauermann \(2023a\)](#): Networks are often the subject of economic research on organizations, trade, and many other areas. Despite this, empirical research in the field often relies on outdated statistical methods which implicitly assume conditional edge independence. The goal of this chapter is to introduce modern statistical modeling tools for empirical research on economic networks. More specifically, we focus on ERGMs and AME models, contrasting their different uses. The ERGM allows one to explicitly specify and test the influence of particular network structures, making it a natural choice if one is substantively interested in estimating endogenous network effects. In contrast, AME captures these effects by introducing actor-specific latent variables, making it a good choice if the interest mainly lies in capturing the effect of exogenous covariates on tie formation while controlling for network effects. After introducing the two model classes, we showcase them by applying them to substantively relevant real-world networks of international arms trade and foreign exchange activity.

Chapter 7, [Fritz, De Nicola, Kevork, Harhoff, and Kauermann \(2023\)](#): This chapter proposes an extension of the temporal ERGM for dynamic bipartite networks. The model is designed and tailored to analyze the dynamic bipartite network of all inventors and patents registered within the field of electrical engineering in Germany in the past two decades, with the goal of exploring the drivers of innovation. To deal with the sheer size of the data, we decompose the network by exploiting the fact that most inventors tend to only stay active for a relatively short period. We thus allow the node set to vary over time, and introduce network statistics to capture covariate effects specific to bipartite data. Our results corroborate that inventor characteristics and team formation play a key role in the dynamics of invention.

Chapter 8, [De Nicola, Tuekam Mambou, and Kauermann \(2023b\)](#): Popular social media users play a major role during crises, as they are able to influence public opinion through their massive reach. In this contribution, we consider the network of influential Twitter users discussing the COVID-19 pandemic, and model it through use of a latent distance model incorporating model-based clustering. This effectively produces an interpretable map of the COVID-19 social media universe. The results suggest the existence of two distinct communities, which respectively favor and oppose vaccine mandates, thus corroborating the presence of echo chamber effects on the platform. We further show that the two groups are not entirely homogeneous: instead, the social map describes an entire spectrum of beliefs between the two extremes, demonstrating that polarization is not the only relevant factor at play, and that moderate users are central to the discussion.

While the field of statistical network analysis has seen major evolution in recent years, it is clear that there is still much room for further development. All model classes considered in this chapter suffer from diverse issues, from slow and unstable estimation routines, to limits in applicability and

2.4 Contributions and discussion

interpretability (see [Crane, 2018](#)). The current limitations, however, make future developments even more exciting.

One area in particular in which I believe there to be great potential is the field of latent variable models. Since its inception, this model family has been extensively developed to represent network structure in different ways without the use of covariates, or by relegating them to a secondary role. On the other hand, covariate information on nodes (such as geographic location or socio-economic information) and edges (such as similarity measures or shared traits between actors) can be extremely valuable in explaining and predicting the formation of ties. In this context, there is great unexplored potential for latent variable approaches to be extended in the direction of simultaneously modeling network and covariate effects, to help clearly distinguish between the two, and to measure the influence each of them has on network formation. For example, I believe there is potential to make use of latent variable approaches for developing a statistical test to detect (endogenous) edge dependence in network data. Such a test would provide empirical researchers with a clear answer to the question: “Why should I use a complicated network model instead of a simpler regression approach?”. A tool of this type would provide a solid basis for selecting an appropriate statistical model for the application at hand. Furthermore, a test for endogeneity would likely increase the uptake of available state-of-the-art tools for network analysis, thus helping bridge the gap between theory and applications.

The natural next step after establishing the presence (or absence) of dependence between edges would be to measure how much this dependence matters in the network. It would thus also be interesting to leverage latent variable approaches for quantifying how much of the overall variance is attributable to network effects, how much of it is driven by exogenous covariates, and how much of it is due to noise. The eventual goal of this endeavor would be to develop a standard set of procedures that empirical researchers could use to analyze and model graphs at various resolutions, and to answer empirical research questions related to tie formation in networks.

3. Models for monitoring epidemics

Infectious diseases have posed a major challenge to humanity since the dawn of time. Epidemics, defined as outbreaks of disease that spread quickly and affect many individuals at the same time, are especially dangerous in the increasingly interconnected landscape of modern society. Seasonal influenza alone is estimated to kill an average of around 700,000 people each year (Dattani and Spooner, 2022). Pandemics, i.e. epidemics that escalated to the point of spreading widely across multiple countries or continents, represent a particularly pernicious threat. As an example, the so-called “Spanish Flu”, quite possibly the deadliest pandemic in human history, is estimated to have killed between 50 and 100 million people globally between the years 1918 and 1920 (Barry, 2020). Much more recently, the COVID-19 pandemic also claimed the lives of millions of people worldwide.

In this context, society has devoted increasingly more resources in efforts to prevent, control, and extinguish epidemics over the years. The biggest positive change factor was played, of course, by developments in the medical field, with the emergence of tools and techniques to better treat and prevent disease (through e.g. antiseptics and vaccines). But another aspect of pandemic management which gained more relevance in recent years is the collection of quality data. Indeed, the increased capacity of society to gather and store information has led to big improvements in this area. There is more available data on cases, deaths, mobility, and other factors related to infectious disease today than ever before, giving rise to the field of public health data science (Goldsmith et al., 2021). However, the available data is far from perfect, and more efforts in this direction are certainly needed (Chiolero et al., 2023). The COVID-19 pandemic taught us several lessons in this regard (Fritz et al., 2022a). Available data has proven to be insufficient, and, in several cases, contradictory, enabling widespread misinterpretation and fueling misinformation. In general, public health data is often incomplete and prone to several kinds of biases. These widespread issues make the methods that are used to analyze the data even more important. In general, trying to draw conclusions from raw data is not a good idea, however “big” the data may be; this is especially true if the data at hand is structurally flawed. In some cases, however, statistical methodology can help address some of the deficiencies in the data, and aid in extracting reliable information from it. This part of the thesis is concerned with addressing this issue, that is, with extracting information from incomplete or delayed public health records. More specifically, the contributions contained in Part III are motivated by real-world questions that emerged from the COVID-19 pandemic. Most of the addressed questions arose in the context of the very early stages of the crisis, when governments around the world were scrambling to interpret available information in order to better manage public health interventions. The focus is on data from Germany, and the addressed questions relate to very central aspects of the pandemic, such as “how many of the infected people are going to die?”, “how many are presently infected?”, or, further, “which proportion of the total infections were authorities able to detect over time”? We address most of these issues by utilizing and extending the broad model class known as generalized additive mixed models (GAMMs). This framework combines the flexibility of generalized additive models (GAMs, Hastie and Tibshirani, 1987) with the ability to account for random effects in mixed

models. To this day, GAMMs are still one of the most used tools for empirical statistical modeling, as their flexibility allows for their extension and tailoring for diverse applications. Because of that, we were able to put them to work to address some of the most pressing data-related questions relating to the pandemic in Germany. Moreover, note that, despite being motivated by substantive research questions, most of the proposed models are suitable for use in different applications which present similar data constellations, such as, e.g., data reported with delay.

As was the case for the previous one, this chapter serves as an introduction to the related contributions, which are collected in Part III of the thesis. More specifically, Section 3.1 provides a step-by-step introduction to the GLM framework, while Section 3.2 discusses its extension towards GAMM. Section 3.3 concludes the chapter by summarizing and discussing the original contributions provided.

3.1. From linear regression to GLMs

This section will introduce generalized linear models by progressively building upon the linear regression framework. Note that the model class has been widely explored, and goes far beyond what is discussed here. In the interest of conciseness, this section is limited to introducing the basics of the framework as utilized in the related contributions. Instead, we refer to Wood (2017) and Fahrmeir et al. (2013) for a more in-depth overview of the field.

3.1.1. The linear regression model

The simple linear model aims at quantifying how the mean of a variable of interest Y , often termed “target”, “response” or “dependent” variable, depends on a single (independent) variable x . As implied by the name, the relationship is assumed to be linear, and takes the following form:

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{3.1}$$

Here, β_0 is the intercept, while β_1 is the slope of the regression line. The term ϵ is called the error term and consists of random deviations with mean 0. Because of that, the expected value of Y is assumed to be a linear function of x , that is

$$E(Y|x) = \beta_0 + \beta_1 x.$$

In other words, we assume (a) a linear relationship between x and Y , and (b) that the relationship is only disturbed by the random perturbation ϵ . Note that the relationship does not necessarily need to be causal in nature, but can also be a simple (linear) association. In addition, the error term is often assumed to be normally distributed and independent of the covariate, that is

$$\epsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n$$

for each subject i , where σ^2 is known as the error variance.

The parameters of the model, i.e. β_0 and β_1 , can be estimated through maximum likelihood. The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are also known as “least squares estimates”, as they can equivalently be obtained through the method of least squares (Charnes et al., 1976).

3.1 From linear regression to GLMs

The case of a single covariate x having an effect on response variable Y can be extended to multiple covariates in straightforward fashion. This results in the *multiple linear regression model*:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon, \quad (3.2)$$

where x_1, x_2, \dots, x_k are k different covariates assumed to influence the response variable Y , $\beta_1, \beta_2, \dots, \beta_k$ are the coefficients associated with each covariate, and ϵ is the random error term. As before, the individual errors ϵ_i are often assumed to follow a normal distribution with mean 0 and variance σ^2 .

Note that the linear regression model does not make any assumption on the nature of the covariates x_1, \dots, x_k . In fact, the independent variables can be transformed in many ways, therefore increasing the range of alternative dependence structures at our disposal far beyond simple linear associations. Possibilities include discrete indicator variables, variable transformations, and polynomials. It is also common to use products of variables, e.g. including variables x_1, x_2 and their product $x_3 = x_1 \cdot x_2$ together in the model equation. In this case, x_3 is named interaction term, as it quantifies how x_1 and x_2 interact with each other relatively to their effect on Y . The fact that non-linearity can be captured may seem conflicting with the model's name: after all, the method is called "*linear regression*". This confusion can easily be solved by noting that the model class is always linear with respect to the parameters, but not necessarily in the covariates, i.e. the covariates can be transformed non-linearly. However, such linearity assumption can also be relaxed, as demonstrated in the next subsection.

3.1.2. Generalized linear models

The linear regression model is suitable for response variables which approximately follow a normal distribution conditionally on the covariates. In the real world, however, there are many cases of non-Gaussian variables of interest. Some variables are not even continuous, but rather categorical, i.e. taking a finite number of non-ordinal values. The classical linear model is clearly not suitable for modeling such variables, but it can be extended to accommodate them. More specifically, we can limit ourselves to modeling the conditional expectation $E(Y_i | x_{i1}, x_{i2}, \dots, x_{ik})$ instead of predicting observations Y_i , and by assuming the model:

$$E(Y_i | x_{i1}, x_{i2}, \dots, x_{ik}) = h(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}), \quad (3.3)$$

where $h(\cdot)$ is any function compatible with the distribution of the response variable Y . By modeling the expectation we do not need to make any assumption on the errors; further, by using a suitable function h to transform the output, we can ensure that the predicted expected values will only take acceptable values. The argument of the h function is termed *linear predictor*, and usually indicated with the letter η , i.e.

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}. \quad (3.4)$$

These models can accommodate a wide variety of different response variable types, following e.g. Poisson, Exponential, Gamma, and many more distributions. This extension to the linear model gives rise to the class of generalized linear models (GLMs), introduced by [Nelder and Wedderburn \(1972\)](#). More precisely, GLMs possess the following properties:

1. The (conditional) distribution of the response variable Y belongs to the broad class of exponential family distributions. This class contains most commonly used distributions (see [Fahrmeir et al., 2013](#) for more details).
2. The expected value of the response value, $E(Y|x)$, is connected to the linear predictor $\eta = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k$ through a response function $h(\cdot)$. In other words, we have:

$$E(Y|x) = \mu = h(\eta)$$

The GLM framework is apparently quite general and encompasses many different regression scenarios. A very important role is played by the response function h , which needs to be strictly monotonic and invertible, and should be chosen to be compatible with the type of response variable Y that is modeled. Examples include the logistic regression model, in which the response function has the main job of transforming the linear predictor, which is generally unbounded, into a probability, which only takes real values between 0 and 1. Another common type of GLM is Poisson regression, which typically employs an exponential response function to map real values into positive ones. A very special case is given by a normally distributed response variable: In this scenario, given that normal distributions can take all values in the real numbers, there is no need to transform the linear predictor η . In other words, the response function h is given by the identity function, i.e. $h(\eta) = \eta$, reverting to the case of the classical linear model. Note that, as in linear models, parameter estimation in GLMs can typically be carried out via maximum likelihood. We refer to [Wood \(2017\)](#) for more detail.

3.2. Generalized additive mixed models

Generalized linear models are already remarkably flexible in comparison to their linear ancestor. The wide range of distributional families available, together with the choice of the response function, provides the user with great flexibility. However, they can further be extended to accommodate for more scenarios. More specifically, the models we work with in the contributions included in Part III generalize the GLM framework in two main ways. Firstly, it is possible to model data that includes repeated measurements over the same individuals over time (i.e. longitudinal data) or data for units which belong to groups or spatial units (i.e. clustered data). For brevity, from now on we will refer to such constellations simply as clustered data. Observations within the same cluster are usually correlated with each other, rendering the standard “iid” assumption of GLMs invalid. One possibility is to account for clusters by including a fixed effect for each of them; However, this is often impractical, since the number of parameters to be estimated becomes quite large relative to the sample size, especially if the number of clusters is high. A better alternative is to instead account for within-cluster correlation by including so-called “random effects” into the model equation. The simplest example of a model of this type is the linear random intercept model, which is a linear model with the addition of a random intercept for each cluster:

$$y_{ij} = \beta_0 + \beta_1x_{ij} + \gamma_i + \epsilon_{ij}, \tag{3.5}$$

noting that the index i indicates the cluster, and j the unit within each cluster. This formulation is equivalent to (3.1), but with the addition of the random term γ_i . This additional intercept accounts for cluster-specific heterogeneity, and is therefore equal for each unit within the same

3.3 Contributions and discussion

cluster i . The key difference from a fixed effect is that this quantity is assumed to be a (typically Gaussian) random variable, i.e. we have

$$\gamma_i \stackrel{iid}{\sim} N(0, \tau^2).$$

This randomness assumption imposes regularization properties on the parameter, and saves degrees of freedom in the estimation (see [Fahrmeir et al., 2013](#) for more details). Adding random effects to GLMs results in the class of generalized linear mixed models (GLMMs), where the word “mixed” indicates that the model includes both fixed and random effects ([Clayton, 1996](#)).

The second major extension to GLMs, attributable to [Hastie and Tibshirani \(1987\)](#), is the inclusion of non-parametric, unknown smooth functions inside the linear predictor η , giving rise to the class of generalized additive models (GAMs). More specifically, GAMs can be defined by

$$E(Y_i) = h(\eta_i) = h(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + f_1(w_{i1}) + \dots + f_q(w_{iq})). \quad (3.6)$$

This specification is equivalent to (3.3), with the addition of the functions $f_1(w_1), \dots, f_q(w_q)$, which are nonlinear smooth effects of the covariates w_1, \dots, w_q , and are modeled and estimated in a non-parametric way ([Wood, 2011](#)). These smooth functions can replace traditional linear parameter specifications within the predictor η , and allow for highly flexible and agnostic estimation of the covariate effects.

It is further possible to combine GLMMs and GAMs, bringing us to the destination of our modeling journey, that is generalized additive mixed models (GAMMs). GAMMs combine the flexibility of GAMs with the ability to account for random effects available from mixed effect models. Building on (3.6), GAMMs can be defined by

$$E(Y_{ij}) = h(\eta_{ij}) = h(\beta_0 + \beta_1 x_{ij1} + \dots + \beta_k x_{ijk} + f_1(w_{ij1}) + \dots + f_q(w_{ijq}) + \gamma_{i0} + \gamma_{i1} u_{i1} + \dots + \gamma_{il} u_{il}), \quad (3.7)$$

where the newly added $\gamma_{i0} + \gamma_{i1} u_{i1} + \dots + \gamma_{il} u_{il}$ constitutes the random part of the model. More specifically, u_{i1}, \dots, u_{il} are the variables included in the random effects design, and $\gamma_{i0}, \dots, \gamma_{il}$ are the corresponding random coefficients. Note that the random effects design is usually fairly simple, with the simplest and most frequently used case being $u_i \equiv 1$, resulting in a random intercept model. Similarly to GAMs, GAMMs enable the representation of non-linear relationships between the response variable and covariates by modeling them through smooth functions. However, GAMMs additionally allow for the inclusion of random effects, which are needed to account for cluster-specific variability.

Given the general formulation of the smooth functions in (3.7), many possible estimation methods are available. In the context of this thesis, such functions are consistently fit through use of polynomial splines, and more specifically P-splines. Such splines prevent overfitting by penalizing overly flexible effects, thus ensuring a smooth fit. We refer to [Wood \(2011\)](#) and [Wood \(2017\)](#) for more details on the estimation of smooth effects, and, more in general, of GAMMs.

3.3. Contributions and discussion

Generalized additive mixed models are powerful tools to measure associations between variables under very general conditions. The wide range of distributional families available, the option of

modeling effects semi-parametrically, and the possibility of including random effects provide the user with remarkable flexibility. In the context of this thesis, GAMMs are put to work to answer complex public health questions revolving around measuring and monitoring key quantities related to the COVID-19 pandemic. Several characteristics make this type of models particularly useful for public health data, such as the possibility of using smooth effects for modeling varying spatial and time trends, random effects to account for region-specific heterogeneity, offsets for modeling rates and account for reporting delays, and flexible autoregressive components in parallel to traditional covariates. The specific contributions contained in Part III of this thesis can be summarized as follows:

Chapter 9, [Schneble, De Nicola, Kauermann, and Berger \(2021a\)](#): This chapter analyzes the temporal and regional structure in mortality rates related to COVID-19 infections, making use of openly available data on registered cases in Germany published by the Robert Koch Institute on a daily basis. Estimates for the number of present-day infections that will, at a later date, prove to be fatal, are derived through a nowcasting model which relates the day of death of each deceased patient to the corresponding day of registration of the infection. This allows to obtain real-time estimates on the severity of the pandemic without waiting for fatalities to happen. Further, our district-level modeling approach for fatal infections disentangles spatial variation into a global pattern for Germany, district-specific long-term effects, and short-term dynamics, while also taking the age and gender structure of the regional population into account. This enables to highlight areas with unexpectedly high disease activity. The analysis of death counts contributes to a better understanding of the spread of the disease while being less dependent on testing strategy and capacity in comparison to infection counts. The proposed approach and the presented results thus provide reliable insight into the state and the dynamics of the pandemic during the early phases of the infection wave in spring 2020 in Germany, when little was known about the disease and limited data were available.

Chapter 10, [De Nicola, Schneble, Kauermann, and Berger \(2022b\)](#): COVID-19 cases in Germany enter the central register days after they are first detected, with this delay deferring an up-to-date view of the state of the pandemic. This contribution provides a stable tool for monitoring current infection levels as well as predicting infection numbers in the immediate future at the regional level. We accomplish this by nowcasting cases that have not yet been reported, as well as through short term predictions of future infections. We apply our model to German data, for which our focus lies in predicting and explain infectious behavior by district.

Chapter 11, [Schneble, De Nicola, Kauermann, and Berger \(2021b\)](#): The case detection ratio of COVID-19 infections varies over time due to changing testing capacities, differing testing strategies, and the evolving underlying number of infections itself. This chapter shows a way of quantifying these dynamics by jointly modeling the reported number of detected COVID-19 infections with nonfatal and fatal outcomes. The proposed methodology also allows to explore the temporal development of the actual number of infections, both detected and undetected, thereby shedding light on the infection dynamics. We exemplify our approach by analyzing German data from 2020, making only use of data available since the beginning of the pandemic. Our modeling approach can be used to quantify the effect of different testing strategies, visualize the dynamics of the case detection ratio over time, and obtain information about the underlying true infection

3.3 Contributions and discussion

numbers, thus enabling us to get a clearer picture of the course of the COVID-19 pandemic in 2020.

Chapter 12, [Fritz, De Nicola, Rave, Weigert, Khazaei, Berger, Küchenhoff, and Kauermann \(2022b\)](#): This article showcases the potential of GAMMs for public health applications, demonstrating their flexibility by focusing on three relevant pandemic-related issues. First, we examine the interdependency among infections in different age groups, concentrating on school children. In this context, we derive the setting under which parameter estimates are independent of the (unknown) case-detection ratio, which plays an important role in COVID-19 surveillance data. Second, we model the incidence of hospitalizations, for which data is only available with a temporal delay. We illustrate how correcting for this reporting delay through a nowcasting procedure can be naturally incorporated into the GAMM framework as an offset term. Third, we propose a multinomial model for the weekly occupancy of intensive care units (ICU), where we distinguish between the number of COVID-19 patients, other patients and vacant beds. With these three examples, we aim to showcase the practical and “off-the-shelf” applicability of GAMMs to gain new insights from real-world data.

As a final consideration, I note that, while these contributions were motivated by specific problems relating to the recent pandemic, several of the methods are applicable to situations presenting similar data constellations. In particular, the use of GAMMs for nowcasting public health data has vast potential, as data from official sources tends to inevitably enter central databases some time after it is first recorded at the local level. In such contexts, nowcasting techniques can aid in obtaining up-to-date estimates for key quantities to monitor. One particularly promising application is in the area of all-cause mortality, where fatalities also typically enters central databases with significant delay. Incorporating nowcasting ideas in this context would enable real-time assessment of excess mortality during crises, which would be particularly useful for managing emergencies. More of this will be discussed in the next chapter.

4. Models for assessing excess mortality

Excess mortality can generally be defined as the number of deaths from all causes during a crisis beyond what would have been expected under normal conditions. In this context, the word “crisis” can encompass any situation causing significant disruption or upheaval, such as natural disasters, epidemics, wars, or other emergencies. All-cause excess mortality is generally considered to be a more reliable way of assessing death tolls extracted by a crisis in comparison to directly considering fatalities officially related to the crisis itself. This is due to the fact that all-cause mortality data is typically more robust and less subject to problems such as, e.g., under-counting and regional variations in reliability (Leon et al., 2020; Beaney et al., 2020). Accurately quantifying excess mortality is thus crucial for understanding the factors driving a crisis, to be able to evaluate the effectiveness of government responses, and to accordingly inform decision makers in modulating policy responses to ongoing and future emergencies.

The concept of excess mortality is well established, and has long been utilized for analyzing the impact of wars, natural disasters, and pandemics (Johnson and Mueller, 2002; Simonsen et al., 2013), with its application dating as far back as the Great Plague of London in 1665 (see Boka and Wainer, 2020). Today, the concept is routinely employed by governments around the world. Despite this long tradition, however, estimating excess mortality remains a challenge, and no single, unified method for doing so has yet been established (Nepomuceno et al., 2022; Acosta, 2023). The main difficulty lies in estimating the “counterfactual” expected mortality, i.e. the number of deaths that would have been expected had the crisis not occurred. This is challenging as mortality rates, trends, and data availability vary greatly across different regions and periods of time. Estimating expected mortality requires (i) choosing a reference period, and (ii) using some model to project mortality rates from the reference period to the period of interest. When choosing a model, it is particularly important to consider factors exogenous to the crisis which may influence mortality, such as varying life expectancy, due to e.g. changes in living conditions, and shifts in the age structure of the population over time. Basic approaches, such as simply using the mean number of deaths during the reference period as the expected mortality for the crisis period, implicitly assumes the total (expected) number of deaths to be constant over both the reference and the crisis period, thereby disregarding factors that may influence mortality, such as shifts in the age structure of the population over time. Ignoring the role of age can be particularly damning, as the age structure within a population can change considerably over short periods of time. Moreover, countries can show large variation in how their populations evolve over time, even when their income levels are comparable. It is thus crucial to take age into account to avoid systematic bias in the estimates.

Several high profile studies tackling the estimation of excess mortality during the COVID-19 pandemic in multiple countries try to capture trends in mortality by fitting various regression models to the data in the reference period, and then extrapolating the trend to the period of the crisis to obtain expected mortality figures for that period (Karlinsky and Kobak, 2021; The Economist, 2023; Wang et al., 2022; Knutson et al., 2023). While incorporating a trend can

account for some of the variation in expected mortality rates over the years, the approach is still not free of problems, as not explicitly accounting for age in the model implicitly assumes the age pyramid to be smooth. This is often not the case, e.g., for many modern-day European countries, where demographic traces of World War II are still visible. For this reason, a trend alone is often not capable of capturing the effect of age. Furthermore, incorporating country-specific trends in the estimation has the effect of projecting any evolution in the overall death rate observed during the reference period on the period of interest, including those due to factors other than age. While, on the one hand, capturing true long-term trends in mortality would be desirable, mortality rates have been shown to exhibit large variance across short periods even in the same region (Bergeron-Boucher and Kjærgaard, 2022). This large variation can lead to predicting large decreases or increases in expected mortality based on variance alone, especially if the trend estimate is based on a period of only 3-5 years, ultimately resulting in overly sensitive and less stable estimates (see Levitt et al., 2023 and Ioannidis et al., 2023 for more detail).

Given the major role that age plays in mortality, many have argued explicit age adjustment to be a sensible way forward (Levitt et al., 2022; Nepomuceno et al., 2022; Stang et al., 2020; Gianicolo et al., 2021). Several prominent multi-nation studies also do take age into account in their analysis (see e.g. Islam et al., 2021; Konstantinou et al., 2022; Levitt et al., 2022). However, these methods simply divide the population into a small number of broad age strata. Doing so reduces the magnitude of the age-induced bias, and is thus certainly better than not accounting for age at all; on the other hand, simply partitioning the population into age groups is equivalent to assuming age structure to be homogeneous within those age groups. To eradicate bias from the estimates, it is thus necessary to perform age adjustment at a finer level, when appropriate data is available. One way to do this is by making use of population standardization approaches. Such approaches have a long tradition in demography when comparing mortality across different regions with different age structure (Keiding and Clayton, 2014; Kitagawa, 1964). The contributions included in Part IV of this thesis introduce novel methods for excess mortality estimation using age standardization, and apply them to calculate excess mortality in high income countries during the COVID-19 pandemic. The papers also introduce an empirical approach for quantifying uncertainty in the estimation, an open issue in the field of excess mortality modeling. The present chapter is meant to introduce the general problem and put those contributions into context. More specifically, Section 4.1 introduces our point-estimation method for calculating yearly excess mortality figures with fine-grained age adjustment. Section 4.2 discusses the issue of uncertainty quantification, and introduces our approach to tackle it. Finally, Section 4.3 provides a brief summary of the related contributions, and discusses potential future avenues for research in this field.

4.1. Age-adjusted estimation

The main challenge in estimating excess mortality during any crisis lies in estimating the “counterfactual” expected mortality, i.e. the number of deaths that would have been expected had the crisis not occurred. One way to do this is to consider mortality rates observed shortly before the crisis and project them onto the period of interest. To do so, however, one needs to decide for a specific method of projection, as well as on what “shortly before” exactly means. In other words, we need to choose a model and a reference period, as mentioned in the previous section. Let’s start with the first point, i.e. model choice. A natural approach is to consider age-specific mortality

4.1 Age-adjusted estimation

data contained in official life tables, which give the probability q_x of a person who has completed x years of age to die before completing their next life-year, i.e. before their $x + 1^{\text{th}}$ birthday. Note that the calculation of a life table, as simple as it sounds, is not straightforward, and is an age-old actuarial problem. First references date far back, to [Price \(1771\)](#) and [Dale \(1772\)](#). A historical digest of the topic is provided by [Keiding \(1987\)](#). Despite this, calculation methods for life tables provided by most countries are fairly consistent with one another. Furthermore, there are entities which provide life tables for many different countries using the same method for all of them, which greatly facilitates comparisons. For example, [HMD \(2023\)](#) provides life tables for tens of different countries, all calculated with the same method ([Wilmoth et al., 2021](#)). As demonstrated in [De Nicola et al. \(2022a\)](#), however, further adjustments to the tables are recommendable to relate the expected number of deaths to recently observed ones. In particular, population data and life tables need to be appropriately matched, since life tables count the number of deaths of x -year-old people over the course of a year, while population data typically gives the number of x -year-old people at a fixed time point (typically the beginning of the year). This requires correction (4.1), which accounts for the fact that a person that dies at x years of age in a given year t was either x years old or $x - 1$ years old at the beginning of the year, i.e. the time point used for the population data. We therefore apply this additional correction, which consists of calculating the adjusted age-specific death probabilities \tilde{q}_x at age x as

$$\tilde{q}_x = \frac{1}{2}q_x + \frac{1}{2}q_{x+1}, \quad (4.1)$$

where q_x are the death probabilities for age x contained in the life tables before the adjustment. We can then compute the overall expected number of deaths in year t as

$$E_t = \sum_{x=1}^{x_{\max}} \tilde{q}_x P_{x,t}, \quad (4.2)$$

where $P_{x,t}$ is the population aged x at the beginning of year t , and x_{\max} is the maximum possible age assumed in the life tables (usually set between 100 and 120 years). We can then obtain the excess mortality estimate Δ_t for a given year t by simply subtracting the expected mortality estimate E_t from the observed death toll O_t in the same year:

$$\Delta_t = O_t - E_t.$$

Let us now focus on the second key modeling decision, that is the choice of a reference period, i.e. the period that will be used to define “normal” mortality levels. In general, it is desirable to have a reference period that is (a) long enough to provide robust data evidence and not fall prey of variance, and (b) short enough to be as similar to the period of interest as possible. Using too long of a reference period, such as, e.g., a ten years window, is problematic due to the fact that baseline mortality levels can change heavily over such a wide time window. Using data from ten years ago to calculate expected mortality today would, e.g., downweigh potential changes in life gains in life expectancy due to changes in living condition and advances in medical technology over time. On the other hand, using a very short reference period, such as e.g. a single year, would also be problematic, as yearly death rates exhibit considerable variation, well beyond what can be explained by underlying changes in life expectancy over time. Mortality in a given year can be heavily skewed by a single factor, such as for example a strong or weak seasonal influenza. As such, we recommend and use reference periods in the order of 3 to 5 years, to strike a balance between bias and variance. Note the choice of reference period, while crucial, will see its influence mitigated when shifting focus from point estimates to range estimates, as described in the upcoming section.

4.2. Uncertainty quantification

In the previous section, we introduced our method to obtain point estimates for excess mortality in a given period. However, it would be very useful to have interval estimates available, especially given the many sources of variance associated with the mortality process. Uncertainty quantification is generally an open issue when it comes to expected and excess mortality estimation. While many existing approaches propose standard confidence intervals based on distributional assumptions, classical probability models do not seem very realistic here, as variation in mortality is in large part driven by external factors, such as, e.g., the strength of an influenza wave and other exogenous shocks. For this reason, our method explicitly refrains from pursuing such model-based approaches, and instead proposes a data-driven empirical assessment of variability. Specifically, we make use of age-specific single-year mortality rates to provide what we can call a “plausible range” for expected mortality. To do so, we can consider the single-year life tables for each year of the reference period, and use them to calculate expected mortality for the years of interest in the same way as above, i.e. using (4.1) and (4.2). Assuming the reference period to contain a total of K years, we will obtain K different excess mortality estimates. We can then take the lowest and highest resulting estimates as the upper and lower bound of our plausible expected mortality range. In other words, we use mortality rates from the “worst” and “best” years of the reference period to obtain a plausible range for expected mortality in the years of interest. To be more precise, the upper mortality bound for year t can be written as:

$$E_t^{\text{upper}} = \max(\tilde{E}_{t,1}, \tilde{E}_{t,2}, \dots, \tilde{E}_{t,K}), \quad (4.3)$$

where $\tilde{E}_{t,k}$ represents expected mortality for year t calculated using the (corrected) single-year life tables from year k . Analogously, the lower bound can be defined as

$$E_t^{\text{lower}} = \min(\tilde{E}_{t,1}, \tilde{E}_{t,2}, \dots, \tilde{E}_{t,K}). \quad (4.4)$$

These expected mortality bounds can then be used to obtain excess mortality intervals in a straightforward manner. Note that these bounds do not give a probabilistic measure of uncertainty, as no distributional model is used. Instead, they provide us with plausible high-mortality and low-mortality scenarios for expected mortality in the years of interest based on levels observed during reference years. In a sense, this is akin to the multiverse approach proposed by [Levitt et al. \(2023\)](#), whereas instead of presenting all possible universes we only present the average one, the best one and the worst one.

4.3. Contributions and discussion

The recent COVID-19 pandemic has led to a spike in interest in the area of excess mortality modeling. The topic has sparked broad debate, which led to the development of new techniques for estimating excess deaths over different time horizons and at different resolutions. This thesis contributes to the field by focusing on a specific task, that is estimating yearly excess mortality in countries where high-resolution population and all-cause deaths data is available. The specific contributions offered can be summarized as follows:

4.3 Contributions and discussion

Chapter 13, De Nicola, Kauermann, and Höhle (2022a): This chapter introduces the method for computing excess mortality presented in Section 4.1. The paper puts particular focus on the role of age, and demonstrates the importance of correctly accounting for it by comparing the performance of different methods on past data. After validating our model, we apply it to age-stratified mortality data from Germany to compute age group-specific yearly excess mortality during the COVID-19 pandemic in 2020. To zoom in on the different pandemic phases, we also provide estimates on the weekly level.

Chapter 14, De Nicola and Kauermann (2022): In this short note, we apply the yearly method introduced in Chapter 13 to newly observed data from 2021 to provide estimates of all-cause excess mortality in Germany for that year. The analysis reveals a preliminary excess mortality of approximately 2.3%, mainly driven by significantly higher excess mortality in the 60-79 age group.

Chapter 15, De Nicola and Kauermann (2024): This chapter extends the method proposed in Chapter 13 by introducing the uncertainty quantification method presented in Section 4.2. We then apply our method to 30 countries with publicly available data, and obtain estimates and uncertainty bounds for each of them. The results uncover considerable variation in pandemic outcomes across different countries. We finally compare our findings with existing estimates published in other major scientific outlets, highlighting how much the method matters, and how important it is to properly account for age to obtain unbiased figures.

The proposed methods for excess mortality estimation have the potential to be further improved and extended to cover a broader set of empirical settings. More specifically, while the existing method is effective in what it sets out to do, i.e. estimating yearly excess mortality for geographical units in which complete and high-resolution data is available, several challenges to make the methodology suitable for broader settings and under more general conditions remain. Specifically, it would be of interest to extend the method in the following directions:

a) Incorporating changes in life expectancy over time: The current version of our method does not adjust for changes in the life expectancy of individuals over time, thereby implicitly assuming constant age-specific hazards over the considered years. This is reasonable for applications in which both the period of interest and the reference period are relatively short. However, if the aim is to monitor excess mortality over longer time frames, accounting for (expected) changes in life expectancy becomes necessary. Examples of such applications would be evaluating the long-term consequences of a pandemic, or the exposure to polluting agents over multi-year periods. An extension in this direction could be pursued by adapting existing projection techniques, such as, e.g., the Lee-Carter model (Lee, 2000).

b) Extensions to varying time and space windows: Existing estimation methods typically make use of country-level data computed on yearly or multi-year windows. To effectively monitor mortality on different spatial units and over shorter time windows (i.e. days, weeks or months), obtaining flexible, time- and region-specific hazard rates is required. This is simple to achieve if data at the desired temporal and spatial resolution is present, but challenging when only spatially and temporally aggregated death counts are available, as is common. In such cases, flexible estimation techniques accounting for historical seasonality and spatial heterogeneity are needed.

These extensions are essential for applications to shorter and region-specific crises, such as floods, wildfires, or the initial stages of a war.

c) Extensions for incomplete or deficient data: Chapter 15 estimates excess mortality in regions for which high-resolution age-specific data is available. This includes many high-income countries, but largely excludes the majority of the low-income and developing world. It would therefore be of great interest to extend the method for cases in which official data is lacking or unavailable. While certainly not straightforward, this feat can be attempted through use of multiple imputation and regression techniques.

d) Real-time mortality monitoring: Retrospective excess mortality estimation using historical data is effective to understand the impact of past crises, and to learn policy lessons in preparation for future ones. An even bigger challenge is posed by the estimation of excess mortality levels in real time. Such real-time monitoring would be vital to inform policy making, particularly with respect to handling an ongoing crisis to minimize its impact. This could be attempted through use of nowcasting techniques similar to those introduced in Chapters 9 and 10, adapting such approaches for all-cause mortality data.

Each of these four extensions could be pursued on its own, to unlock relevant substantive applications to different empirical settings. Taken together, however, they could be functional to the broader goal of building a comprehensive tool capable of providing real-time monitoring of excess mortality levels over flexible spatial and temporal dimensions. Such a tool would enable governments and institutions around the world to make more informed policy decisions in situations of uncertainty during crises induced by, e.g., climate change-related adverse events, pollution, epidemics, conflicts, and economic downturns.

References

- Abbe, E. (2018). Community detection and stochastic block models. *Foundations and Trends in Communications and Information Theory*, 14(1-2): 1–162.
- Acosta, E. (2023). Global estimates of excess deaths from COVID-19. *Nature*, 613: 31–33.
- Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed Membership Stochastic Blockmodels. *Journal of machine learning research*, 9: 1981–2014.
- Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422): 669–679.
- Barry, J. M. (2020). *The great influenza: The story of the deadliest pandemic in history*. Penguin UK.
- Beaney, T., Clarke, J. M., Jain, V., Golestaneh, A. K., Lyons, G., Salman, D., and Majeed, A. (2020). Excess mortality: the gold standard in measuring the impact of COVID-19 worldwide? *Journal of the Royal Society of Medicine*, 113(9): 329–334.
- Bearman, P. S., Moody, J., and Stovel, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *American journal of sociology*, 110(1): 44–91.
- Bergeron-Boucher, M.-P. and Kjærgaard, S. (2022). Mortality forecasting at age 65 and above: an age-specific evaluation of the lee-carter model. *Scandinavian Actuarial Journal*, 2022(1): 64–79.
- Boka, D. M. and Wainer, H. (2020). How can we estimate the death toll from COVID-19? *Chance*, 33(3): 67–72.
- Charnes, A., Frome, E. L., and Yu, P.-L. (1976). The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. *Journal of the American Statistical Association*, 71(353): 169–171.
- Chiolero, A., Tancredi, S., and Ioannidis, J. P. (2023). Slow data public health. *European journal of epidemiology*, 38: 1219–1225.
- Choi, D. S., Wolfe, P. J., and Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika*, 99(2): 273–284.
- Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6): 066111.
- Clayton, D. G. (1996). Generalized linear mixed models. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov chain Monte Carlo in practice*, pages 275–302. Chapman & Hall/CRC.

- Crane, H. (2018). *Probabilistic foundations of statistical network analysis*. Chapman and Hall/CRC.
- Dale, W. (1772). *Calculations deduced from first principles, in the most familiar manner, by plain arithmetic, for the use of the societies instituted for the benefit of old age: intended as an introduction to the study of the doctrine of annuities. By a member of one of the societies*. London: J. Ridley.
- Dattani, S. and Spooner, F. (2022). How many people die from the flu? *Our World in Data*. <https://ourworldindata.org/influenza-deaths>.
- Davis, J. A. (1970). Clustering and hierarchy in interpersonal relations: Testing two graph theoretical models on 742 sociomatrices. *American Sociological Review*, 35(5): 843–851.
- De Nicola, G., Fritz, C., Mehrl, M., and Kauermann, G. (2023a). Dependence matters: Statistical models to identify the drivers of tie formation in economic networks. *Journal of Economic Behavior & Organization*, 215: 351–363.
- De Nicola, G. and Kauermann, G. (2022). An update on excess mortality in the second year of the COVID-19 pandemic in Germany. *ASTA Wirtschafts-und Sozialstatistisches Archiv*, 16: 21–24.
- De Nicola, G. and Kauermann, G. (2024). Estimating excess mortality in high-income countries during the COVID-19 pandemic. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, qnae031.
- De Nicola, G., Kauermann, G., and Höhle, M. (2022a). On assessing excess mortality in Germany during the COVID-19 pandemic. *ASTA Wirtschafts-und Sozialstatistisches Archiv*, 16: 5–20.
- De Nicola, G., Schneble, M., Kauermann, G., and Berger, U. (2022b). Regional now- and forecasting for data reported with delay: Towards surveillance of COVID-19 infections. *ASTA Advances in Statistical Analysis*, 106: 407–426.
- De Nicola, G., Sischka, B., and Kauermann, G. (2022). Mixture models and networks: The stochastic blockmodel. *Statistical Modelling*, 22(1-2): 67–94.
- De Nicola, G., Tuekam Mambou, V. H., and Kauermann, G. (2023b). COVID-19 and social media: Beyond polarization. *PNAS Nexus*, 2(8): pgad246.
- Dorff, C., Gallop, M., and Minhas, S. (2020). Networks of violence: Predicting conflict in Nigeria. *Journal of Politics*, 82(2): 476–493.
- Efron, B. and Hastie, T. (2021). *Computer age statistical inference, student edition: algorithms, evidence, and data science*. Cambridge University Press.
- Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen*, 6: 290.
- Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression - Models, Methods and Applications*. Springer.
- Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5): 75–174.

References

- Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659: 1–44.
- Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395): 832–842.
- Fritz, C., De Nicola, G., Günther, F., Rügamer, D., Rave, M., Schneble, M., Bender, A., Weigert, M., Brinks, R., Hoyer, A., Berger, U., Küchenhoff, H., and Kauermann, G. (2022a). Challenges in interpreting epidemiological surveillance data – experiences from Germany. *Journal of Computational and Graphical Statistics*, 32(3): 765–766.
- Fritz, C., De Nicola, G., Kevork, S., Harhoff, D., and Kauermann, G. (2023). Modelling the large and dynamically growing bipartite network of German patents and inventors. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 186(3): 557–576.
- Fritz, C., De Nicola, G., Rave, M., Weigert, M., Khazaei, Y., Berger, U., Küchenhoff, H., and Kauermann, G. (2022b). Statistical modelling of COVID-19 data: Putting generalized additive models to work. *Statistical Modelling (OnlineFirst)*. <https://doi.org/10.1177/1471082X221124628>.
- Fruchterman, T. M. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11): 1129–1164.
- Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3): 657–683.
- Ghani, A. C., Swinton, J., and Garnett, G. P. (1997). The role of sexual partnership networks in the epidemiology of gonorrhoea. *Sexually Transmitted Diseases*, 24(1): 45–56.
- Gianicolo, E. A., Russo, A., Büchler, B., Taylor, K., Stang, A., and Blettner, M. (2021). Gender specific excess mortality in Italy during the COVID-19 pandemic accounting for age. *European Journal of Epidemiology*, 36(2): 213–218.
- Gilbert, E. N. (1959). Random graphs. *Annals of Mathematical Statistics*, 30(4): 1141–1144.
- Goldenberg, A., Zheng, A. X., Fienberg, S. E., Airoldi, E. M., et al. (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2): 129–233.
- Goldsmith, J., Sun, Y., Fried, L., Wing, J., Miller, G. W., and Berhane, K. (2021). The emergence and future of public health data science. *Public Health Reviews*, 42: 1604023.
- Gormley, I. C. and Murphy, T. B. (2010). A mixture of experts latent position cluster model for social network data. *Statistical Methodology*, 7(3): 385–405.
- Handcock, M. (2003). Assessing degeneracy in statistical models of social networks. *Working Paper no. 39, University of Washington*.
- Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2): 301–354.

- Hanneke, S., Fu, W., and Xing, E. P. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4: 585–605.
- Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398): 371–386.
- Heider, F. (1946). Attitudes and cognitive organization. *Journal of Psychology*, 21(1): 107–112.
- HMD. (2023). Human Mortality Database. Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France). Available at www.mortality.org (accessed 14/3/2023).
- Hoff, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. *Advances in Neural Information Processing Systems*, 20: 657–664.
- Hoff, P. (2021). Additive and multiplicative effects network models. *Statistical Science*, 36(1): 34–50.
- Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the American Statistical Association*, 100(469): 286–295.
- Hoff, P. D. (2011). Hierarchical multilinear models for multiway data. *Computational Statistics & Data Analysis*, 55(1): 530–543.
- Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460): 1090–1098.
- Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social networks*, 5(2): 109–137.
- Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373): 33–50.
- Huang, S., Sun, J., and Feng, Y. (2023). PCABM: Pairwise covariates-adjusted block model for community detection. *Journal of the American Statistical Association*. <https://doi.org/10.1080/01621459.2023.2244731>.
- Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481): 248–258.
- Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012). Computational statistical methods for social network models. *Journal of Computational and Graphical Statistics*, 21(4): 856–882.
- Ioannidis, J. P., Zonta, F., and Levitt, M. (2023). Flaws and uncertainties in pandemic global excess death calculations. *European Journal of Clinical Investigation*, 23(8): e14008.
- Islam, N., Shkolnikov, V. M., Acosta, R. J., Klimkin, I., Kawachi, I., Irizarry, R. A., Alicandro, G., Khunti, K., Yates, T., Jdanov, D. A., White, M., Lewington, S., and Lacey, B. (2021). Excess deaths associated with COVID-19 pandemic in 2020: age and sex disaggregated time series analysis in 29 high income countries. *BMJ*, 373:n1137.
- Johnson, N. P. and Mueller, J. (2002). Updating the accounts: global mortality of the 1918-1920 “Spanish” influenza pandemic. *Bulletin of the History of Medicine*, pages 105–115.

References

- Kamada, T., Kawai, S., et al. (1989). An algorithm for drawing general undirected graphs. *Information processing letters*, 31(1): 7–15.
- Karlinsky, A. and Kobak, D. (2021). Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset. *eLife*, 10: e69336.
- Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1): 016107.
- Kaur, H., Rastelli, R., Friel, N., and Raftery, A. (2023). Latent position network models. In P. Carrington J. McLevey and J. Scott, editors, *Sage Handbook of Social Network Analysis (2nd Edition)*, chapter 36, pages 526–541. Sage.
- Keiding, N. and Clayton, D. (2014). Standardization and control for confounding in observational studies: a historical perspective. *Statistical Science*, pages 529–558.
- Keiding, N. (1987). The method of expected number of deaths, 1786-1886-1986. *International Statistical Review*, 55(1): 1–20.
- Kitagawa, E. M. (1964). Standardized comparisons in population research. *Demography*, 1(1): 296–315.
- Kline, R. R. (2015). *The cybernetics moment: Or why we call our age the information age*. JHU Press.
- Knutson, V., Aleshin-Guendel, S., Karlinsky, A., Msemburi, W., and Wakefield, J. (2023). Estimating global and country-specific excess mortality during the COVID-19 pandemic. *The Annals of Applied Statistics*, 17(2): 1353–1374.
- Kolaczyk, E. (2009). *Statistical analysis of network data: Methods and models*. Springer.
- Konstantinoudis, G., Cameletti, M., Gómez-Rubio, V., Gómez, I. L., Pirani, M., Baio, G., Larrauri, A., Riou, J., Egger, M., Vineis, P., et al. (2022). Regional excess mortality during the 2020 COVID-19 pandemic in five European countries. *Nature Communications*, 13(1): 482.
- Krivitsky, P. N. and Handcock, M. S. (2014). A separable model for dynamic networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1): 29–46.
- Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social networks*, 31(3): 204–213.
- Lee, C. and Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1): 1–50.
- Lee, R. (2000). The Lee-Carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal*, 4(1): 80–91.
- Lee, Y. and Ogburn, E. L. (2021). Network dependence can lead to spurious associations and invalid inference. *Journal of the American Statistical Association*, 116(535): 1060–1074.
- Leon, D. A., Shkolnikov, V. M., Smeeth, L., Magnus, P., Pechholdová, M., and Jarvis, C. I. (2020). COVID-19: a need for real-time monitoring of weekly excess deaths. *The Lancet*, 395(10234): e81.

- Levitt, M., Zonta, F., and Ioannidis, J. P. (2022). Comparison of pandemic excess mortality in 2020–2021 across different empirical calculations. *Environmental Research*, 213: 113754.
- Levitt, M., Zonta, F., and Ioannidis, J. P. (2023). Excess death estimates from multiverse analysis in 2009–2021. *European Journal of Epidemiology*, 38: 1129–1139.
- Lusher, D., Koskinen, J., and Robins, G. (2012). *Exponential Random Graph Models for Social Networks*. Cambridge University Press.
- Matias, C. and Robin, S. (2014). Modeling heterogeneity in random graphs through latent space models: A selective review. *ESAIM: Proceedings and Surveys*, 47: 55–74.
- Morris, M., Handcock, M. S., and Hunter, D. R. (2008). Specification of exponential-family random graph models: Terms and computational aspects. *Journal of Statistical Software*, 24(4): 1548.
- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3): 370–384.
- Nepomuceno, M. R., Klimkin, I., Jdanov, D. A., Alustiza-Galarza, A., and Shkolnikov, V. M. (2022). Sensitivity analysis of excess mortality due to the COVID-19 pandemic. *Population and Development Review*, 48(2): 279–302.
- Newcomb, T. M. (1979). Reciprocity of interpersonal attraction: A nonconfirmation of a plausible hypothesis. *Social Psychology Quarterly*, 42(4): 299–306.
- Nowicki, K. and Snijders, T. A. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455): 1077–1087.
- Peixoto, T. P. (2017). Nonparametric Bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95(1): 012317.
- Price, R. (1771). *Observations on reversionary payments; on schemes for providing annuities for widows, and for persons in old age; on the method of calculating the values of assurances on lives, and on the national debt*. London: Cadell.
- Rivera, M. T., Soderstrom, S. B., and Uzzi, B. (2010). Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *Annual Review of Sociology*, 36: 91–115.
- Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007a). An introduction to exponential random graph (p^*) models for social networks. *Social Networks*, 29(2): 173–191.
- Robins, G., Snijders, T., Wang, P., Handcock, M., and Pattison, P. (2007b). Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29(2): 192–215.
- Schneble, M., De Nicola, G., Kauermann, G., and Berger, U. (2021a). Nowcasting fatal COVID-19 infections on a regional level in Germany. *Biometrical Journal*, 63(3): 471–489.
- Schneble, M., De Nicola, G., Kauermann, G., and Berger, U. (2021b). A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020. *Biometrical Journal*, 63(8): 1623–1632.
- Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association*, 106(496): 1361–1370.

References

- Simon, H. A. (1955). On a Class of Skew Distribution Functions. *Biometrika*, 42(3-4): 425–440.
- Simonsen, L., Spreeuwenberg, P., Lustig, R., Taylor, R. J., Fleming, D. M., Kroneman, M., Van Kerkhove, M. D., Mounts, A. W., Paget, W. J., and Teams, G. C. (2013). Global mortality estimates for the 2009 Influenza Pandemic from the GLaMOR project: a modeling study. *PLoS medicine*, 10(11): e1001558.
- Snijders, T. A. and Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1): 75–100.
- Stang, A., Standl, F., Kowall, B., Brune, B., Böttcher, J., Brinkmann, M., Dittmer, U., and Jöckel, K.-H. (2020). Excess mortality due to COVID-19 in Germany. *Journal of Infection*, 81(5): 797–801.
- Sweet, T. M. (2015). Incorporating covariates into stochastic blockmodels. *Journal of Educational and Behavioral Statistics*, 40(6): 635–664.
- Tallberg, C. (2005). A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, 29(1): 1–23.
- The Economist. (2023). Tracking COVID-19 excess deaths. <https://www.economist.com/graphic-detail/coronavirus-excess-deaths-tracker> (accessed 14/3/2023).
- Wang, H., Paulson, K. R., Pease, S. A., Watson, S., Comfort, H., Zheng, P., Aravkin, A. Y., Bisignano, C., Barber, R. M., Alam, T., et al. (2022). Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21. *The Lancet*, 399(10334): 1513–1536.
- Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.
- Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^* . *Psychometrika*, 61(3): 401–425.
- White, A. and Murphy, T. B. (2016). Mixed-membership of experts stochastic blockmodel. *Network Science*, 4(1): 48–80.
- Wilmoth, J., Andreev, K., Jdanov, D., Gleit, D., and Riffe, T. (2021). Methods Protocol for the Human Mortality Database (Version 6). Available at www.mortality.org/File/GetDocument/Public/Docs/MethodsProtocolV6.pdf (accessed 14/3/2023).
- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1): 3–36.
- Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Part II.

Statistical network analysis

5. Mixture models and networks: The stochastic blockmodel

Contributing article

De Nicola, G., Sischka, B., and Kauermann, G. (2022). Mixture models and networks: The stochastic blockmodel. *Statistical Modelling*, 22(1-2):67–94. <https://doi.org/10.1177/1471082X211033169>.

Data and code

Available at <http://www.statmod.org/smij/Vol22/Iss1-2/DeNicola/Abstract.html>.

Copyright information

This article is distributed under a Creative Commons 4.0 International license (CC BY 4.0).

Author contributions

The idea of framing the stochastic blockmodel as a mixture modeling approach for network data can be attributed to Göran Kauermann. The literature review for this survey article was mainly done by Giacomo De Nicola, who also took care of categorizing the papers and putting them into context. Moreover, the writing was done for the most part by Giacomo De Nicola, where the methodological formulation of the reviewed estimation techniques (Section 4) was strongly supported by Göran Kauermann. Benjamin Sischka was mainly responsible for formulating the graphon representation (Section 3.3) as well as for developing and implementing the corresponding estimation routine (Section 4.3). Giacomo De Nicola further designed and implemented the application studies in Section 5, for which Benjamin Sischka provided the analysis results for the military alliance network. All authors contributed through fruitful comments and extensive proofreading of the manuscript.



Statistical Modelling 2022; **22**(1-2): 67–94

Mixture models and networks: The stochastic blockmodel

Giacomo De Nicola¹, Benjamin Sischka¹ and Göran Kauermann¹

¹Department of Statistics, Faculty of Mathematics, Informatics and Statistics,
Ludwig-Maximilians-Universität München, Munich, Germany

Abstract: Mixture models are probabilistic models aimed at uncovering and representing latent subgroups within a population. In the realm of network data analysis, the latent subgroups of nodes are typically identified by their connectivity behaviour, with nodes behaving similarly belonging to the same community. In this context, mixture modelling is pursued through stochastic blockmodelling. We consider stochastic blockmodels and some of their variants and extensions from a mixture modelling perspective. We also explore some of the main classes of estimation methods available and propose an alternative approach based on the reformulation of the blockmodel as a graphon. In addition to the discussion of inferential properties and estimating procedures, we focus on the application of the models to several real-world network datasets, showcasing the advantages and pitfalls of different approaches.

Key words: community detection, mixture models, statistical network analysis, stochastic blockmodels

1 Introduction

The underlying idea of a mixture model is rather simple. Instead of assuming that the target variable follows a plain distribution, one considers a mixture of multiple distributions. Specifically, for a random variable Y , one assumes

$$Y \sim \sum_{k=1}^K \pi_k f_k(y), \quad (1.1)$$

where π_k is a weighting coefficient, with $\sum_{k=1}^K \pi_k = 1$, and $f_k(\cdot)$ is the k th mixture distribution. Commonly, the mixture components come from the same distributional family but differ in their parameters, that is, $f_k(\cdot) = f(\cdot | \theta_k)$, where θ_k parametrizes the k th mixture component. An early (maybe the first) reference in this direction

Address for correspondence: Göran Kauermann, Department of Statistics, Faculty of Mathematics, Informatics and Statistics, Ludwig-Maximilians-Universität München, Ludwigstr. 33, 80539 Munich, Germany.

E-mail: goeran.kauermann@stat.uni-muenchen.de



© 2021 The Author(s)

10.1177/1471082X211033169

dates back to Pearson (1894) and focuses on the estimation of a mixture of two normal distributions. An early mathematical treatment of the topic, more in the style of convolution, is provided in Robbins (1948). In a series of papers, Teicher (1960) discusses identifiability issues, where the cited work puts the focus on finite mixtures in the style of (1.1). A first survey on mixture models is provided by Gupta and Huang (1981), presenting the different estimation routines that had been developed and used by that time. A central algorithm in this respect, which is not included in the above survey article (certainly because of simultaneous time of publication), is the work of Aitkin and Wilson (1980; see also Aitkin, 1980) who propose the use of the at the time recently developed Expectation–Maximization (EM) algorithm (see Dempster et al., 1977) to estimate the finite mixture distribution. Though the focus of their paper lies in the modelling of outliers, the authors make use of the idea that a finite mixture model can be comprehended as a missing data problem. Under this modelling framework, one assumes that the discrete valued random variable Z takes values $\{1, \dots, K\}$ with

$$P(Z = k) = \pi_k, \quad (1.2)$$

where again $\sum_{k=1}^K \pi_k = 1$. Conditional on $Z = k$, one then observes Y from the k th mixture component, that is,

$$Y|(Z = k) \sim f_k(y) \quad \text{for } k = 1, \dots, K.$$

Treating Z as unobserved (or unobservable) enables the framing of estimation in a missing data situation, where the considered likelihood (1.1) can be maximized with the EM algorithm. The results are generalized and extended towards hypothesis tests in Aitkin and Rubin (1985). A comprehensive overview on finite mixture models is given in the early book of Everitt and Hand (1980), followed by the monographs of Titterton et al. (1985), Lindsay (1995), Böhning (1999), McLachlan and Peel (2000) and Frühwirth-Schnatter (2006). We also refer to the recent *Handbook of Mixture Analysis* (Frühwirth-Schnatter et al., 2019). For software implementations of mixture models, Leisch (2004) is a central reference (see also Benaglia et al., 2009). Allowing the mixture components and/or the mixing proportions π_k to depend on additional covariates extends mixture models towards regression models. The resulting model class is also known as mixture of experts, tracing back to Jacobs et al. (1991). A survey from the perspective of machine learning can be found in Masoudnia and Ebrahimpour (2014, see also Gormley and Frühwirth-Schnatter, 2019).

While most of the literature cited above deals with a univariate response variable Y , in this article we aim to look at multivariate data with Y expressing a network. Network data have a simple binary structure resulting from a network as follows. Assume a set of N actors, where we define with $V = \{v_1, \dots, v_N\}$ the set of nodes in a network. We call $E \subset V \times V$ the edge set, and the resulting network can be

represented with an adjacency matrix Y such that $Y \in \{0, 1\}^{N \times N}$ and

$$Y_{ij} = \begin{cases} 1, & \text{if } (v_i, v_j) \in E \\ 0, & \text{otherwise.} \end{cases}$$

If the network is undirected, $Y_{ij} = Y_{ji}$ holds. Furthermore, the diagonal of Y often remains undefined, meaning that self-loops are not contemplated. The statistical analysis of network data has achieved increasing interest in the last two decades: We refer to Kolaczyk (2009) and Kolaczyk and Csárdi (2014) for a general introduction to the topic (see also Goldenberg et al., 2009; Hunter et al., 2008; Fienberg, 2012; Lusher et al., 2013; Salter-Townshend et al., 2012; Biagini et al., 2019).

If we consider Y as set of random variables $\{Y_{ij}; 1 \leq i, j \leq N, i \neq j\}$, we can transfer the mixture model setting (1.1) towards network data. This leads to what is commonly referred to as (*a posteriori*) stochastic blockmodelling. A survey on the latest theoretical developments in this field has recently been published by Abbe (2018; see also Lee and Wilkinson, 2019 for a comprehensive review). Stochastic blockmodels can be seen as a tool for performing community detection (see, e.g., Clauset et al., 2004; Fortunato, 2010; Fortunato and Hric, 2016). While community detection and stochastic blockmodels have a lot in common, the latter specifically focuses on the modelling aspect and will therefore be considered here. A stochastic blockmodel (SBM) is in fact a mixture model where each mixture component is specified by the group or community membership. The latent subgroups of nodes are typically identified by their connectivity behaviour, with nodes behaving similarly belonging to the same community. The class of stochastic blockmodels evolved from its deterministic counterpart, which dates back to White et al. (1976). The stochastic version of the blockmodel was introduced by Holland et al. (1983) in the statistical literature. Similar modelling proposals, developed independently, trace back to the computer science literature (see, for example, Bui et al., 1987). Wang and Wong (1987) were the first to apply the stochastic blockmodel to directed graphs, even though they still assumed the block structure to be known. The first steps towards *a posteriori* blockmodelling, that is modelling with initially unknown group structure, were taken by Snijders and Nowicki (1997) and Nowicki and Snijders (2001), who proposed estimation routines for, respectively, two groups and any known number of groups. From there, the model class gained traction. Recent literature on the classical version of stochastic blockmodels includes Daudin et al. (2008), Gormley and Murphy (2010) and Aitkin et al. (2014), using Bayesian approaches (see also Vu and Aitkin, 2015). Following their initial formulation, stochastic blockmodels have been extended in various ways. Some of such variants and extensions will be reviewed and treated in Section 3, and some of those will be put to practice later on. The aim of this article is to illuminate on the connection between mixture models and stochastic blockmodels, exploring some of the different approaches within the model class and demonstrating their applicability by making use of real data. The article also introduces a different formulation of the stochastic blockmodel through the

graphon framework, using this reformulation to propose an alternative estimation routine.

The rest of the article is organized as follows: Section 2 presents some real-world network datasets together with the potential questions that we face in analysing them. Those datasets will be later used to demonstrate the capabilities of stochastic blockmodels. Section 3 describes the blockmodelling framework in more detail, and introduces some of its most prominent variants and extensions. Section 4 compares the different estimation routines that are available, and introduces a Monte Carlo-based EM estimation routine under graphon representation. The empirical analysis of the previously introduced datasets is then carried out in Section 5, making use of the previously described models to tackle the questions posed in Section 2. Finally, Section 6 ends the article with some comments and conclusive remarks.

2 Data description

In order to demonstrate the capabilities of stochastic blockmodels, we have chosen network datasets pertaining to three different domains, namely political science, biology and sociology. Despite the different domains, the networks share the presence of some form of underlying community structure, or at least the appearance thereof. They all therefore lend themselves to be modelled through the use of mixture components. General descriptive measures of the data examples, which consist of undirected graphs, are given in Table 1, which shows that all three networks are of medium size and range from very dense to relatively sparse.

2.1 International alliances network

The first network that we introduce is constructed using data from the Alliance Treaty Obligations and Provisions project (Leeds et al., 2002). The dataset provides information on military alliance agreements pertaining to all countries of the world. For the analysis we consider alliances that were in force in the year 2016. The countries are taken as nodes, and an edge between two countries is present if the two countries take part in a ‘strong’ military alliance treaty. More specifically, the alliances that we consider strong are defensive and offensive ones. This means, respectively, ‘alliances in which the members promise to provide active military support in the event of attack on the sovereignty or territorial integrity of one or more alliance partners’ and ‘alliances in which the members promise to provide active military support under any conditions not precipitated by attack on the

Table 1 Descriptive statistics for the studied networks

	Alliances	Butterflies	E-mails
Nodes	141	832	548
Edges	1703	86528	5433
Density	0.173	0.250	0.036

Mixture models and networks: The stochastic blockmodel 71

sovereignty or territorial integrity of an alliance partner, regardless of whether the goals of the action are to maintain the status quo' (see Leeds et al., 2002). Note that, in general, an alliance can involve more than two nodes: Representing the network using dyadic edges only thus leads to the loss of some information. For example, pairwise edges between countries i , j and k could mean three pairwise treaties, or a treaty that involves all three of them. While using pairwise edges as we do here is standard in network modelling, hypergraph representations (Berge, 1984) offer a viable alternative, and models representing this kind of data in a more natural way have been explored (see, e.g., Chodrow, 2020). Looking at the network from a blockmodelling perspective, there are several questions that we can pose. First of all, do alliances between countries induce a partition of the network that is meaningful from a geopolitical perspective? Moreover, will the blocks found be in line with geographic proximity and political affinity, or will there be some other characteristics driving the grouping? And finally, what can the resulting block structure tell us about the global system of alliances?

2.2 Butterfly similarity network

The second real-world instance is a butterfly similarity weighted network, constructed using the data presented by Wang et al. (2009) and available from Zitnik et al. (2018). Each node represents a butterfly, and valued edges depict visual similarities between them. More specifically, pairs of butterflies with some positive degree of similarity between them are connected by a weighted edge, while no edge is present if the similarity score between the two is zero. The absence of an edge is thus equivalent to the presence of an edge with weight zero. The similarity scores lie in the interval $[0, 1.55]$, with a higher value implying a higher degree of similarity. Scores are computed using butterfly images, as described in Wang et al. (2009). Information on the species to which each butterfly belongs is also available, with each unit belonging to a single species. A total of ten species are present, implying a 'natural' partition of the network in ten blocks. In this case, there is one clear question that emerges: Are the communities found using visual similarity scores in agreement with how biologists categorized butterfly species? In other words, are we able to recover the 'ground truth' communities of the network via stochastic blockmodelling?

2.3 Email exchange network

The last network considered consists of anonymized email data from a large European research institution, collected between October 2003 and May 2005 (Leskovec and Krevl, 2014). Each node in the network represents a person, and an edge between nodes i and j is present if person i sent person j at least one email in the examined period. The nodes featured in this network are all members of the institution, meaning that only emails within the institution itself are considered. Moreover, only nodes belonging to the largest ten departments are included. Note that, similarly as for the previously described alliances data, this

binary representation disregards the multi-dimensional nature of the edges (as an email can have multiple recipients). Since department memberships are known and individuals from the same department are expected to behave similarly, we can consider the departments as ‘ground truth’ communities for the network. Given that, the questions that we pose are straightforward: Are we able to find some form of meaningful community structure in the network considering emails alone? And if so, will the structure recovered be similar to the partition induced by department memberships? And finally, what can email exchanges tell us about the structure of the institution and the relationships between departments? To analyse this and the other previously introduced networks and to investigate the correspondingly raised questions, we will introduce the appropriate model variants and related estimation procedures in the following sections.

3 Stochastic blockmodels: formulations and variants

3.1 The standard stochastic blockmodel

As anticipated in the introduction, if we consider the network \mathbf{Y} as set of random variables $\{Y_{ij}; 1 \leq i, j \leq N, i \neq j\}$, we can transfer the mixture model setting (1.1) towards network data. This leads to the stochastic blockmodel, that is a mixture model for which each mixture component is specified by the group or community membership. More specifically, we assume the independent discrete group indicator coefficients $Z_i \in \{1, \dots, K\}$ for $i = 1, \dots, N$ with

$$\mathbb{P}(Z_i = k) = \pi_k \quad \text{for } k = 1, \dots, K$$

and, as above, $\sum_{k=1}^K \pi_k = 1$. An edge between node i and j then exists with probability

$$Y_{ij} | (\mathbf{Z} = \mathbf{z}) \sim \text{Bernoulli}(p_{z_i z_j}), \quad (3.1)$$

where $\mathbf{P} = [p_{kl}]_{k,l=1,\dots,K}$ is the $K \times K$ dimensional block-probability matrix. For community detection one typically assumes that $p_{kk} > p_{kl}$ for all $l \neq k$, but this is not a requirement for stochastic blockmodels in general. In fact, the block structure may describe clusters of nodes that behave similarly from a connectivity standpoint without necessarily being more densely connected, thus allowing for other types of structures, such as disassortative communities and core-periphery.

For estimation, a numerically simpler setting can result by approximating the binomial distribution through a Poisson distribution. This approximation is justified since the network density is usually low, implying that p_{kl} is typically small. In this case, (3.1) is replaced by

$$Y_{ij} | (\mathbf{Z} = \mathbf{z}) \sim \text{Poisson}(\lambda_{ij}), \quad (3.2)$$

where $\lambda_{ij} = \exp\{\omega_{z_i z_j}\}$, with $\mathbf{\Omega} = [\omega_{kl}]_{k,l=1,\dots,K}$ as block-connectivity parameter matrix. One of the main allures of the Poisson model variant lies in the fact that there is a

closed form for integrating out parameters, as seen in, for example, McDaid et al. (2013; see also Lee and Wilkinson, 2019 for an illustration of this).

3.2 Degree correction

A well-known extension of the classical SBM is the degree-corrected stochastic blockmodel, introduced by Karrer and Newman (2011). In their work, the authors show how the standard stochastic blockmodel implicitly assumes the degree structure within communities to be relatively homogeneous. This, combined with the fact that many real-world networks exhibit extremely skewed degree distributions (Simon, 1955; Barabási and Albert, 1999), leads the model to often only be able to find core-periphery type block structures, where node grouping is predominantly driven by degree similarity. To bypass this issue, Karrer and Newman (2011) introduced the idea of degree correction, making the probability of an edge depend not only on group membership, but also on node-specific heterogeneity parameters. More precisely, the original version of the degree-corrected SBM can be written in the same way as (3.2), but in this case

$$\lambda_{ij} = \exp\{\gamma_i + \gamma_j + \omega_{z_i z_j}\}. \quad (3.3)$$

In this notation $\exp\{\gamma_i\}$ quantifies the heterogeneity specific of node i , and $\exp\{\omega_{z_i z_j}\}$ can be viewed as a measure of the propensity to form ties between the groups to which nodes i and j belong. Note that the degree-corrected SBM is not, in general, strictly better than the standard one, as the two models imply different underlying structures of the network (see, e.g., Yan et al., 2014; Yan, 2016; Wang and Bickel, 2017). The choice of one over the other simply depends on what is the kind of structure one wishes to find. It is also possible to combine the two approaches, as done in Aicher et al. (2015) and Lu and Szymanski (2019). All three versions of the model, namely (3.1), (3.2) and (3.3), will be applied to the previously introduced data examples.

3.3 Graphon representation

The stochastic blockmodel can also be formulated through the graphon model class, which recently received a lot of attention concerning the modelling of complex networks. Although the scope of the structures representable as a graphon is quite large, its formulation is rather simple. Let us therefore introduce U_i , for $i = 1, \dots, N$, as node-specific continuous random variables which can be described as

$$U_i \stackrel{i.i.d.}{\sim} \text{Uniform}[0, 1]. \quad (3.4)$$

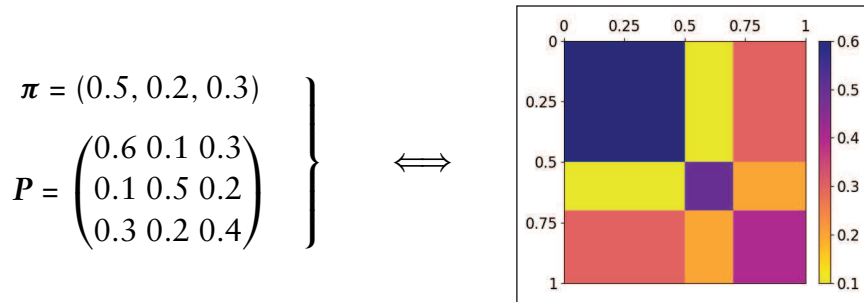
The network entries are then assumed, conditionally and independently from one another, to follow

$$Y_{ij} | (U = \mathbf{u}) \sim \text{Bernoulli}(p(\mathbf{u}_i, \mathbf{u}_j)),$$

where $p : [0, 1] \times [0, 1] \rightarrow [0, 1]$ is a function (sometimes called graphon). This function $p(\cdot, \cdot)$ is commonly assumed to be at least piecewise continuous, meaning to fulfill some Lipschitz or Hölder condition in segments. A representation of the SBM can then be generated by restricting the graphon function to be locally constant in a rectangular pattern. More precisely we define, for a SBM with K groups,

$$p(u_i, u_j) = \sum_{k=1}^K \sum_{l=1}^K \mathbb{1}_{\{\tau_{k-1} \leq u_i < \tau_k\}} \mathbb{1}_{\{\tau_{l-1} \leq u_j < \tau_l\}} p_{kl} \tag{3.5}$$

with $\mathbb{1}_{\{\cdot\}}$ as indicator function, $0 = \tau_0 < \tau_1 < \dots < \tau_K = 1$ as boundaries, and p_{kl} representing the edge probability between and within groups, as defined above. The group memberships Z_i are here substituted by the node-specific quantities U_i , which are also latent. This additionally implies that the community proportions are now represented by the boundaries τ_k , $k = 1, \dots, K - 1$. Note that from the uniform distribution of U_i specified in (3.4) follows that $\tau_k = \sum_{l=1}^k \pi_l$. An instance of such relationship can be given through the following illustration:



It is thus not difficult to see how this formulation of graphon models is equivalent to SBMs.

In this context, it should be noted that the graphon model suffers from major identifiability issues, which, with regard to the SBM representation, also include the label switching problem (we refer to the Appendix for more details and illustrations). This non-identifiability arises from the fact that any permutation of $p(\cdot, \cdot)$ represents the same network-generating model as $p(\cdot, \cdot)$ itself. Even more generally, two graphon functions $p(\cdot, \cdot)$ and $\tilde{p}(\cdot, \cdot)$ represent the same network-generating model if and only if there exist two measure preserving functions $\varphi, \tilde{\varphi} : [0, 1] \rightarrow [0, 1]$ such that $p(\varphi(u), \varphi(v)) = \tilde{p}(\tilde{\varphi}(u), \tilde{\varphi}(v))$ for almost every $(u, v) \in [0, 1]^2$ (Diaconis and Janson, 2008). A common approach to resolve this issue is the postulation of a monotonically non-decreasing marginal function $g(u) = \int_0^1 p(u, v) dv$ (see, e.g., Bickel and Chen, 2009 or Chan and Airolidi, 2014). With regard to SBMs, that means ordering the communities, $k = 1, \dots, K$, by $\sum_{l=1}^K p_{kl} \Delta \tau_l$ with $\Delta \tau_l = \tau_l - \tau_{l-1}$, inducing the additional constraint of $\sum_{l=1}^K p_{kl} \Delta \tau_l \neq \sum_{l=1}^K p_{jl} \Delta \tau_l$ for all $k \neq j$. This assumption,

however, might yield only an imperfect identification, especially when the marginal functions $\sum_{l=1}^K p_{kl} \Delta \tau_l$ are similar (see Nowicki and Snijders, 2001). Moreover, this is a strong restriction to the generality of graphon models. We therefore aim to circumvent this issue by formulating an adequate estimation procedure (see Section 4.3).

3.4 Further variants and extensions

Many other variants and extensions of SBMs exist. These include the mixed membership model (Airoldi et al., 2008), in which nodes can belong to multiple communities simultaneously, and the hierarchical stochastic blockmodel (Peixoto, 2017), in which communities are comprised of meta-communities, leading to a hierarchical block structure. A matter of simplifying the model representation is what motivates the microcanonical variant of the SBM (see, e.g., Peixoto, 2012), where the structural pattern is strictly fixed in absolute values. This, in turn, allows for fitting more elaborate generative models, which usually require Markov Chain Monte Carlo (MCMC) techniques for evaluation, to larger networks and to an increased number of groups, as demonstrated by Peixoto (2017). The same author also proposed a nested hierarchical variant of the SBM (Peixoto, 2014a) in which the generative model inferred at an upper level serves as prior information to the one at a lower level, thus also providing an increased resolution when performing model selection. Despite its more elaborate formulation, this hierarchical model remains tractable, and it is feasible to apply it to very large networks. It is also possible to add covariates to the analysis, as initially proposed by Tallberg (2005) and further elaborated by Choi et al. (2012), Sweet (2015) and Huang and Feng (2018). A further extension is the mixture of experts SBM (see Gormley and Murphy, 2010; White and Murphy, 2016), which allows covariates to enter the latent position cluster model in a number of ways, yielding different model interpretations. Extensions for more specific purposes have also been developed: Bouveyron et al. (2018) introduced the stochastic topic blockmodel, a probabilistic model for networks with textual edges. Their model addresses the problem of discovering meaningful clusters of vertices that are coherent with regards to both the network interactions and the text contents. Finally, another relevant approach that can be seen as a generalization of the SBM is the latent position cluster model proposed by Handcock et al. (2007) (originating from Hoff et al., 2002, see also Krivitsky et al., 2009). It is worth noting that most of the mentioned specifications can be applied to binary data as well as to valued and count data (see, e.g., Nowicki and Snijders, 2001). In this article, we do not concentrate on these extensions, but focus on the more ‘classical’ SBMs.

4 Estimation techniques

4.1 Variational methods

The EM algorithm proved to be a powerful and numerically efficient way for estimating parameters in mixture models (see Aitkin, 1980 or Friedl and

Kauermann, 2000). Unfortunately, this does not extend to the estimation of stochastic blockmodels. The complete data log-likelihood resulting from (3.2) in the case of an undirected network equals

$$l_C(\boldsymbol{\Omega}, \boldsymbol{\pi}) = \sum_{i=1}^N \sum_{j=i}^N \sum_{k,l=1}^K \mathbb{1}_{\{z_i=k\}} \mathbb{1}_{\{z_j=l\}} (y_{ij} \omega_{kl} - \exp\{\omega_{kl}\}) + \sum_{i=1}^N \sum_{k=1}^K \mathbb{1}_{\{z_i=k\}} \log(\pi_k) \quad (4.1)$$

with the side constraint $\sum_{k=1}^K \pi_k = 1$. Applying the EM algorithm would in this case mean calculating the posterior distribution

$$P(Z_i = k, Z_j = l | Y = \mathbf{y})$$

with \mathbf{y} being the observed adjacency matrix. This posterior, due to the resulting dependence structure of Z_i and Z_j , is numerically intractable (Mariadassou et al., 2010). To circumvent such numerical hurdles, Jordan et al. (1999) proposed variational methods, which are based on an approximation of the likelihood. Let $P(\mathbf{y}; \boldsymbol{\Omega}, \boldsymbol{\pi})$ be the probability of the data, resulting through

$$P(\mathbf{y}; \boldsymbol{\Omega}, \boldsymbol{\pi}) = \sum_{k_1=1}^K \dots \sum_{k_N=1}^K \pi_{k_1} \dots \pi_{k_N} \prod_{i=1}^N \prod_{j>i}^N \lambda_{k_i k_j}^{y_{ij}} \exp\{-\lambda_{k_i k_j}\},$$

which is apparently too complex from a numerical perspective. We define the lower bound function

$$J(\tilde{P}(\mathbf{z}; \boldsymbol{\xi}); \boldsymbol{\Omega}, \boldsymbol{\pi}) = \log P(\mathbf{y}; \boldsymbol{\Omega}, \boldsymbol{\pi}) - \text{KL}(\tilde{P}(\mathbf{z}; \boldsymbol{\xi}), P(\mathbf{z} | \mathbf{y}; \boldsymbol{\Omega}, \boldsymbol{\pi})),$$

where $\text{KL}(\cdot, \cdot)$ defines the Kullback–Leibler divergence. If we choose $\tilde{P}(\mathbf{z}; \boldsymbol{\xi})$ to be the posterior distribution of \mathbf{Z} given $\boldsymbol{\xi}$, we obtain $J(\cdot)$ to be equal to the log-likelihood of the observed data. Since this is numerically problematic, we compute the posterior distribution of \mathbf{Z} given $\boldsymbol{\xi}$ through independence:

$$\tilde{P}(\mathbf{z}; \boldsymbol{\xi}) = \prod_{i=1}^N \prod_{k=1}^K \xi_k^{\mathbb{1}_{\{z_i=k\}}},$$

where $\boldsymbol{\xi}_k = (\xi_{k1}, \dots, \xi_{kN})$ is a vector containing the probabilities for each of the N nodes to be in group k , with $\sum_{k=1}^K \xi_{ki} = 1$ needing to hold for every $i \in 1, \dots, N$. $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_K)$ is known as variational parameter, and needs to be chosen such that $J(\tilde{P}(\mathbf{z}; \boldsymbol{\xi}); \boldsymbol{\Omega}, \boldsymbol{\pi})$ is maximized with respect to all parameters. It can be shown that $J(\cdot)$ can, up to an intractable constant, be written in a simple numerical form which allows for fast and numerically feasible estimation. The remaining unknown component expresses the approximation error which is typically difficult to quantify (see Lee et al., 2020).

4.2 Vertex switching algorithms

Another possibility for the estimation of stochastic blockmodels is to maximize the likelihood through vertex switching routines. The basic idea of this type of algorithms is the following: starting from an initial, possibly random group assignment, a starting value of the likelihood is computed. From there, one or more vertices are moved from one group into another, and the likelihood is computed again. The new allocation is then accepted or rejected based on a function of the previous and the subsequent likelihood, and such procedure runs iteratively until convergence is reached, that is until a maximum is found. Algorithms of this type include single-vertex Monte Carlo (see, e.g., Peixoto, 2013, 2014b) and a local heuristic routine inspired by the Kernighan–Lin algorithm used in minimum-cut graph partitioning (Kernighan and Lin, 1970; Karrer and Newman, 2011). In principle, computing the likelihood that many times may seem quite expensive. On the other hand, it is not always necessary to calculate the complete likelihood at each step. Depending on the model specification, it is often possible to write the change in the likelihood in a computationally efficient way, so that the algorithm becomes quite competitive in terms of speed. The chief issue with this type of algorithm is that, given the heuristic maximization routine, it is not possible to obtain a measure of uncertainty for group assignments. The procedure will only produce the graph partitioning that (locally) maximizes the likelihood, without any additional information. This is fine if the problem at hand is one of pure community detection, but can become problematic if the goal is proper mixture modelling, as the stochastic component of the mixture is lost. Another potential issue is the possibility to get stuck at local maxima, which usually is tackled by running the procedure several times with different (random) starting points.

4.3 Monte–Carlo-based EM estimation under graphon representation

A third and to some extent novel estimation routine is to estimate the block structure using its graphon representation. Although it is not quite clear how this approach competes with already existing methods, our ambition here is to demonstrate a further possible form of representing and estimating mixture models in networks. Such a model can be fitted appropriately by applying an EM-type algorithm including Gibbs sampling in the E-step. As mentioned above, EM-based algorithms are a common approach to estimate mixture models as well as other models involving latent variables, although in the case of networks the task becomes analytically intractable and numerically demanding. We therefore make use of MCMC techniques to approximate the complex posterior distribution of the latent quantities, which here reflect the group assignments. In this approach, we thus slightly reformulate the stochastic blockmodelling procedure, relating it to graphon estimation (see, e.g., Latouche and Robin, 2016 or, for the reverse link, Olhede and Wolfe, 2014 and Airolidi et al., 2013). We here want to follow the estimation approach of Sischka and Kauermann (2019), applying it to SBMs. The idea is to make use of model (3.5) and estimate, in the M-step, the parameters of $p(\cdot, \cdot)$, namely the interval boundaries τ_k and the blockwise heights p_{kl} , $k, l = 1, \dots, K$,

directly yielding estimates for the SBM quantities $\boldsymbol{\pi}$ and \mathbf{P} . The group assignments Z_1, \dots, Z_N can be determined by considering the positions U_1, \dots, U_N in relation to $\boldsymbol{\tau} = (\tau_0, \dots, \tau_K)$. To carry out the E-step, we assume the function $p(\cdot, \cdot)$ to be given (or to be set to the current estimate). In this regard, as follows from (3.5), the full conditional posterior can be formulated as

$$g_j(\mathbf{u}_j | \mathbf{u}_1, \dots, \mathbf{u}_{j-1}, \mathbf{u}_{j+1}, \dots, \mathbf{u}_N, \mathbf{y}) \propto \prod_{\substack{i=1 \\ i \neq j}}^N p(\mathbf{u}_j, \mathbf{u}_i)^{y_{ji}} (1 - p(\mathbf{u}_j, \mathbf{u}_i))^{1-y_{ji}}.$$

This allows for applying Gibbs sampling in a straightforward manner. Details on this sampling scheme, as well as remarks on the associated potential issues of label switching and non-identifiability, are given in the Appendix. In this context, we underline how the issue of label switching is prevented through the EM algorithm on the primary level of the estimation procedure (apart from the exceptional case of complete symmetry, as described in the Appendix). In comparison, label switching is a common problem when making use of an overall Bayesian estimation procedure (if the MCMC scheme is run for sufficiently long, see Stephens, 2000), where one randomly draws quantities from the corresponding posterior distributions in alternating fashion. This, in contrast, is circumvented in the EM framework, since in the E- and M-steps the results of the respective other step are kept fixed and, based on that, the ‘optimal’ solution is carried out to be used for the next iteration. Parameter estimates are thus not achieved by averaging over several iterations but are given for each iteration separately. Therefore, with regard to our estimation routine, no post-hoc relabelling is required, and assignments can be adopted as deduced from the subordinate Gibbs sampling scheme. Making use of the sampling sequence, we specify the posterior mode in the m th iteration using $\hat{u}_j^{(m)} = (\tau_{k'-1} + \tau_{k'})/2$ for $j = 1, \dots, N$, where the index k' is defined as $\arg \max_k \sum_{t=1}^n \mathbb{1}_{\{\tau_{k-1} \leq u_j^{<t>} < \tau_k\}}$. In that regard, $u_j^{<t>}$ is the value of the j th element in the Markov chain at time t , and $n \in \mathbb{N}$ is the number of considered states of the MCMC sequence extracted by thinning factor $r \in \mathbb{N}$. To take into account that the U_i are uniformly distributed and therefore expected to spread proportionally to interval size, we additionally apply a subsequent adjustment. This concerns both the latent quantities U_i and the interval boundaries $\tau_1, \dots, \tau_{K-1}$. Assuming that $\hat{\tau}_k^{(m)}$ represents the current estimate of τ_k , we then set

$$\hat{\tau}_k^{(m+1)} = \delta^{(m+1)} \frac{\sum_{i=1}^N \mathbb{1}_{\{\hat{u}_i^{(m)} < \hat{\tau}_k^{(m)}\}}}{N} + (1 - \delta^{(m+1)}) \frac{k}{K}$$

and accordingly adjust the estimates of U_j in the form of $\tilde{u}_j^{(m)} = (\hat{\tau}_{k'-1}^{(m+1)} + \hat{\tau}_{k'}^{(m+1)})/2$, with index k' defined through the previous assignment in the form of $\hat{u}_j^{(m)} \in [\hat{\tau}_{k'-1}^{(m)} + \hat{\tau}_{k'}^{(m)})$. Regarding the specification of $\hat{\tau}_k^{(m+1)}$, the weighting $\delta^{(m+1)} \in [0, 1]$ with $\delta^{(m+1)} \geq \delta^{(m)}$ induces a step-size adaptation from *a priori* equidistant boundaries to observed

boundaries implied by frequencies. Such step-size adaptation is recommendable to prevent the community size to shrink too substantially before the structure of the community has been evolved properly. In general, $\delta^{(m)}$ is chosen to be one in the last iteration.

The M-step is then carried out by maximizing the likelihood conditionally on $U = \tilde{\mathbf{u}}^{(m)}$ and for $\tau_1, \dots, \tau_{K-1}$, taking the estimates adjusted as above. This is easily done by setting

$$\hat{p}^{(m+1)}(u, v) = \frac{\sum_{i=1}^N \sum_{j=1}^N \mathbb{1}_{\{\tau_{k-1} \leq u_i < \tau_k\}} \mathbb{1}_{\{\tau_{l-1} \leq u_j < \tau_l\}} y_{ij}}{\sum_{i=1}^N \sum_{j=1}^N \mathbb{1}_{\{\tau_{k-1} \leq u_i < \tau_k\}} \mathbb{1}_{\{\tau_{l-1} \leq u_j < \tau_l\}}}$$

for all $u \in [\tau_{k-1}, \tau_k)$ and $v \in [\tau_{l-1}, \tau_l)$. As it is done for the previously mentioned vertex switching algorithms, we run this MCEM algorithm several times with varying initialization of U , and then choose the outcome with the highest likelihood, which should here also prevent us from getting stuck at a local maximum. To determine the optimal number of blocks, typical model selection criteria can be applied. We here make use of the AIC, for which both quantities required for the computation, namely the likelihood and the number of parameters, can easily be determined. The major advantage of the reformulation of model (3.1) to model (3.5) is that the graphon function $p(\cdot, \cdot)$ could be also formulated in more complex fashion, that is, instead of just being local constant one could allow for more complex structures within each segment. This is not pursued in this article, but we refer to this new research strand discussed, among others, in Vu et al. (2013).

In contrast to non-stochastic estimation routines, such as the vertex switching algorithm discussed in Section 4.2, this modelling approach naturally yields information about the inherent uncertainty of the proposed group allocation. In order to achieve this, we run the E-step one more time after the algorithm has converged. The resulting Gibbs sampling sequence of this last iteration then reveals the distribution of the node allocation with respect to the model estimate $(\hat{p}(\cdot, \cdot), \hat{\tau} = (0, \hat{\tau}_1, \dots, \hat{\tau}_{K-1}, 1))$. A normalized Gini coefficient calculated over the assignment frequencies of a single vertex can then be used as a measure of uncertainty, where a value near one (zero) implies a low (high) level of uncertainty.

4.4 Choosing the number of blocks

A general big challenge in mixture models (and hence also in stochastic blockmodels) lies in the choice of the number of mixture components (blocks). In fact, most of the variants presented so far require that number to be known *a priori*. This is typically not true in real-world applications. In mixture models the question of choosing the number of mixture components is tackled, for instance, in Aitkin (2011). In the field of stochastic blockmodels, a comprehensive list of different approaches is provided by Lee and Wilkinson (2019). Approaches based on penalized likelihood criteria have emerged. In particular, Wang and Bickel (2017) consider an approach based on the log-likelihood ratio statistic, enabling the use of a likelihood-based

model selection criterion that is asymptotically consistent. Other techniques are also available: Chen and Lei (2018) develop a network cross-validation approach which is based on a block-wise node-pair splitting technique, combined with an integrated step of community recovery using sub-blocks of the adjacency matrix. Mariadassou et al. (2010) base the choice on an Integrated Classification Likelihood criterion. The number of blocks can also be estimated using ‘collapsed’ approaches, where the model parameters are integrated out in a Bayesian formulation of the model. The model space and cluster allocations can then be estimated using a greedy search routine (Côme and Latouche, 2015) or using MCMC (McDaid et al., 2013). Another possible approach is that of Peixoto (2013), who uses the Minimum Description Length principle, which seeks to minimize the total amount of information required to describe the network and avoid overfitting. This also allows to deduce general bounds on the detectability of any prescribed block structure, given the number of nodes and edges in the sampled network. Finally, Riolo et al. (2017) (see also Newman and Reinert, 2016) present a method for estimating the number of communities in a network using a combination of Bayesian inference and an efficient Monte Carlo sampling scheme. While other approaches have been proposed, we will not go into further detail here. For modelling the previously described networks, when possible we select K such that the resulting number of blocks coincides with the ground truth. If such ground truth is not available, we make use of the Akaike Information Criterion (AIC), which can be easily calculated when using the graphon representation-driven algorithm.

5 Application to real world networks

5.1 International alliances network

To model the network we use the standard version of the stochastic blockmodel, as in (3.1). Estimation was performed using the Monte Carlo-based EM routine under graphon representation. Applying the AIC yields seven communities as the optimal dimensionality of the blockmodel. The resulting fitted block decomposition is given in Figure 1. Network visualization, as for the rest of the examples in this section, is carried out through use of the open-source tool Gephi (Bastian et al., 2009). The associated world map is shown in Figure 2, where countries are coloured by block. States coloured in grey on the map are isolates in the network, meaning that they were not involved in any strong military alliance in 2016. Moreover, China, Cuba and North Korea are only connected to each other, and are thus isolated from the rest of the network. Those countries have therefore been excluded from the model fitting. The plots show how the blocks recovered by the stochastic blockmodel are very much related and in accordance with the geopolitical structure of the modern world, while also revealing some interesting patterns. The network representation can be visually split into two large components. In the first component, on the left side of the plot in Figure 1, the central block contains most European countries together with Canada. This block is very densely linked, as most of the countries inside it

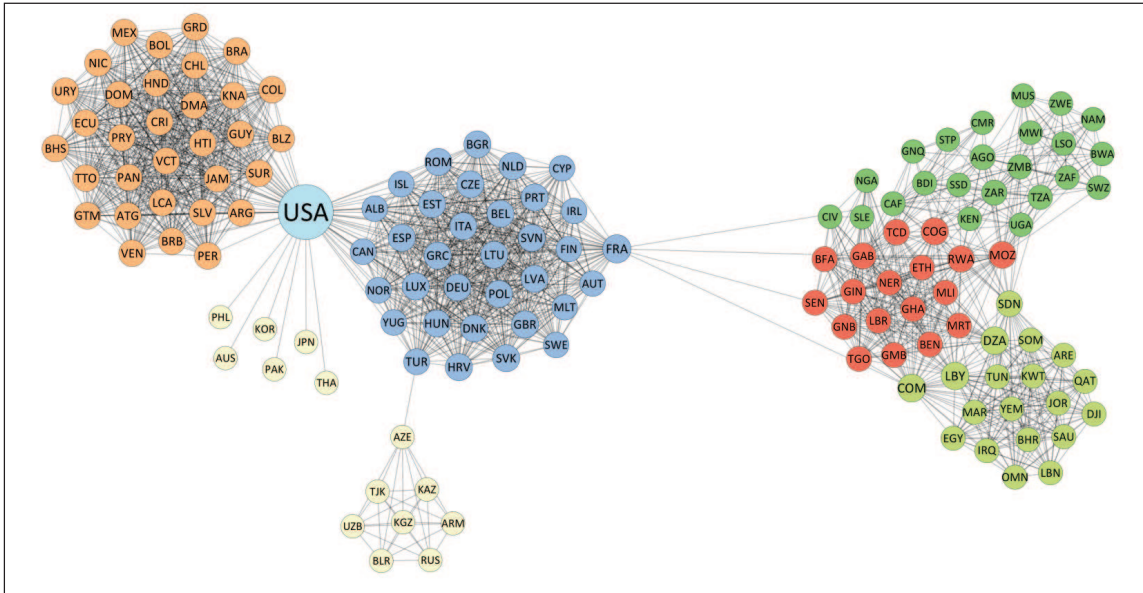


Figure 1 Global network of political alliances in 2016. Two countries are connected if they have taken part in a strong alliance treaty. Labels indicate country codes, while nodes are coloured by block memberships found through the standard stochastic blockmodel

belong to NATO and other major alliances. The block on the very left pretty much coincides with Central and South America, and it is also quite dense. The European and the American block are linked by the USA, which, given its unique connectivity behaviour, constitutes a block on its own. The bottom block includes mostly Asiatic countries as well as some Pacific states, which share a very low edge density. The other component of the network, on the right-hand side of Figure 1, is made out of three blocks. The block on the bottom contains all countries from the Middle East together with Northern African countries such as Libya, Tunisia, Egypt and Morocco. The middle block includes countries from Central and Western Africa. Finally, the upper block is composed of Southern African countries. The central block is well connected with both the northern and the southern blocks, mostly through countries that share borders, while the latter two blocks are instead only directly bridged by Sudan. As an additional note, we can observe that the two major components of the network are linked exclusively through France, that, while belonging to the European block, acts as a bridge between Africa and Europe itself. Finally, it is evident how transferring the group assignments to the world map in Figure 2 clearly reveals a general geographic proximity of countries belonging to the same community.

In addition to the detected block structure, we also investigate the uncertainty of the node allocation, using the Monte Carlo-based posterior samples. We therefore consider the last Gibbs sampling sequence after the algorithm has converged. More specifically, we take a look at the three countries with the lowest values of the normalized Gini coefficient calculated over the allocation frequencies, which in turn

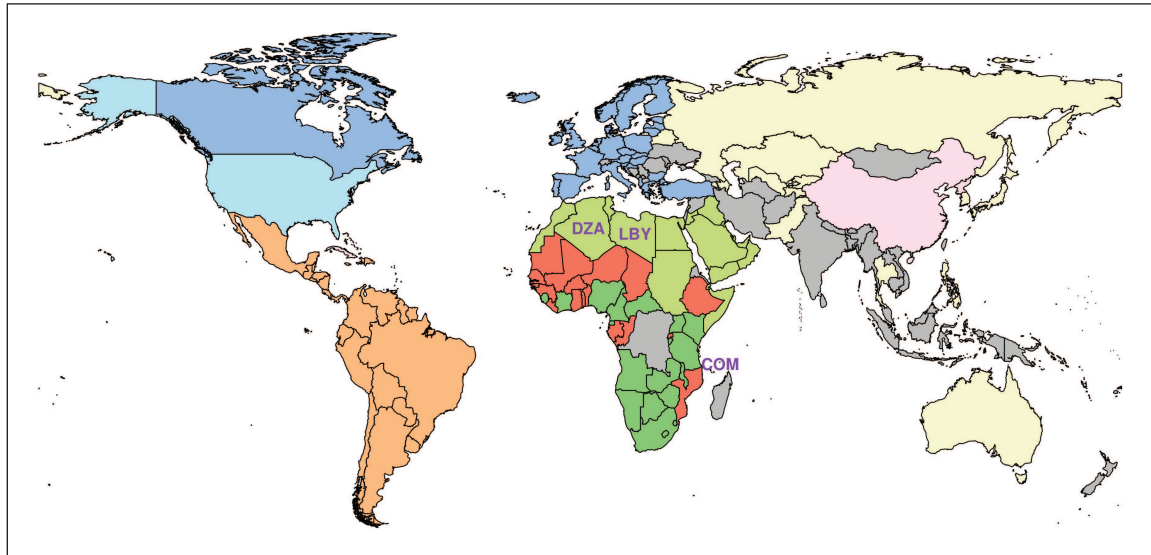


Figure 2 World map with countries coloured by block membership. Colours are kept consistent with Figure 1. Countries not included in Figure 1 are isolates, meaning that they were not part of any strong military alliance as of 2016, with the exception of China, Cuba and North Korea, which are only connected among themselves. Labels indicate the three countries with the highest uncertainty in block membership

imply the highest uncertainty. These countries are Libya (LIBY), Algeria (DZA) and Comoros (COM), which all belong to the Arabic block. The switching of communities exhibited by Libya throughout the posterior sampling is illustrated as an example in Figure 3. It shows how the sample for U_{Libya} mostly appears within $[0.87, 1]$ (the interval of the Arabic block) while also exhibiting some states where it is within $[0, 0.13]$ (the interval of the Western African block). The posterior frequencies for Libya as well as for Comoros and Algeria with respect to the different groups are shown in Table 2, which also comprises the corresponding Gini coefficients. The table shows how all three countries have a substantial tendency to move to the Western African block. According to the fitted blockmodel, in 15% to 18% of the MCMC sample states the three countries are assigned to this block. Turning our attention to all other countries, we observe Gini coefficients which are close to one and thus exhibit only very little uncertainty in block membership. Altogether, this reveals how the estimated community structure appears to be quite strong.

5.2 Butterfly similarity network

The standard SBM as showcased in the previous section is suitable for modelling binary networks. As described in Section 2, this dataset is, however, comprised of similarity scores which lie in the interval $[0, 1.55]$. While it would be possible to binarize the data, for example defining a threshold within the domain as cut-off, this would lead to considerable information loss. We therefore use the Poisson

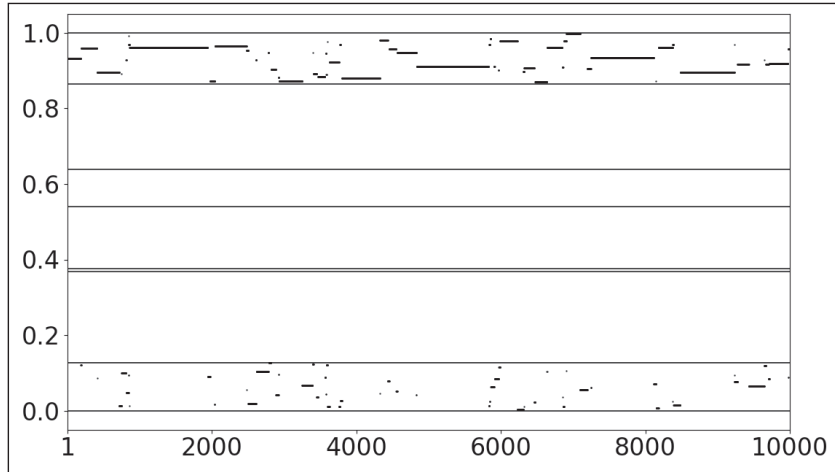


Figure 3 Posterior sample of the latent quantity U for Libya plotted against the MCMC states. Horizontal lines represent community boundaries

Table 2 Posterior frequencies for the three countries with the highest uncertainty in their community memberships. The corresponding normalized Gini coefficient is depicted in the rightmost column

Country	Community							Gini coefficient
	Western African	European	USA	Southern African	Asian/Pacific	South American	Arabic	
Comoros	0.1748	0	0	0	0	0	0.8252	0.9417
Libya	0.1598	0	0	0	0	0	0.8402	0.9467
Algeria	0.1558	0	0	0	0	0	0.8442	0.9481

version of the stochastic blockmodel as defined in (3.2), taking advantage of the fact that this variant is suitable to treat multi-edged networks as well as binary ones. To fit this model, we discretized underlying similarity measures into count data through binning. More specifically, each similarity measure was multiplied by 100 and rounded to the nearest integer, resulting in natural values between 0 and 155. Estimation on the resulting multi-edged network was performed using the Variational EM approach developed by Mariadassou et al. (2010; see also Daudin et al., 2008) and implemented in the R software by Leger (2016). In this case, since we know that the real number of species is ten, we can simply use the same number of communities for the estimation. Figure 4 shows the results of the model fit compared with the partition of butterflies into species. At a first glance, we can see that the communities recovered mirror the real species relatively well. The most evident difference lies in the fact that two of the species (located towards the centre of the plot) are apparently really similar according to the utilized measure of visual similarity, and are therefore split up by the blockmodel. It is also interesting to note how communities found by the stochastic blockmodel seem to be visually clearer than ground truth ones. This

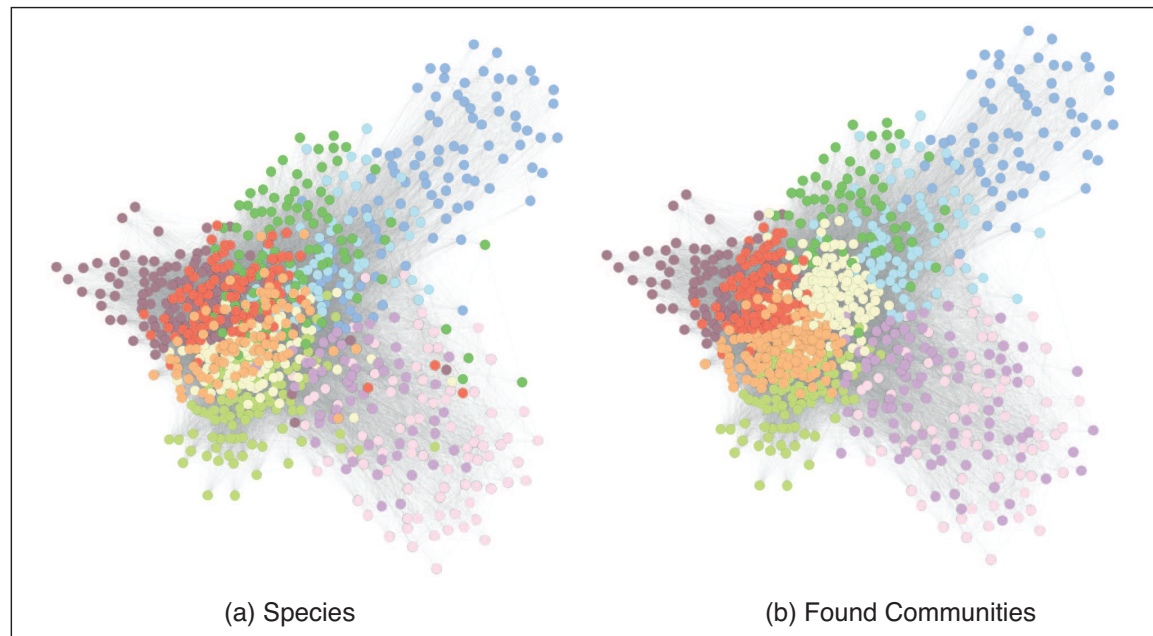


Figure 4 Comparison between ‘ground truth’ communities (species) and groups found by the Poisson stochastic blockmodel in a network of butterflies, with weighted edges representing the degree of visual similarity between them

is attributable to the fact that network visualization techniques and the clustering algorithm utilized are both based solely on the ties between the nodes, and thus tend to be more in accordance. The ground truth, on the other hand, is always given *a priori*, and can easily have outliers in terms of connectivity behaviour. In this specific case, visualization and blockmodelling are both based on the aforementioned measure of visual similarity between butterflies, while the ground truth communities are given by the classification of butterflies into species by biologists. This at least partially explains the discrepancy between the ground truth communities and the positioning of the nodes in the visualized graph. Despite this discrepancy, in general, the structure that was found does not appear to present major differences from the biological classification of the species. To quantify the goodness of the recovered block structure compared to the ‘ground truth’ communities, several measures are available (we refer to Jebabli et al., 2018 for a comprehensive survey). Here we opted for the Rand index, a measure of similarity between two data clusterings that can simply be described as the number of agreements in classifying pairs divided by the total number of pairs (Rand, 1971). The index takes values between 0 and 1, and in this case it is equal to 0.91, indicating that, given two Butterflies chosen at random, the blockmodel is able to correctly identify if they belong to the same species or not 91% of times.

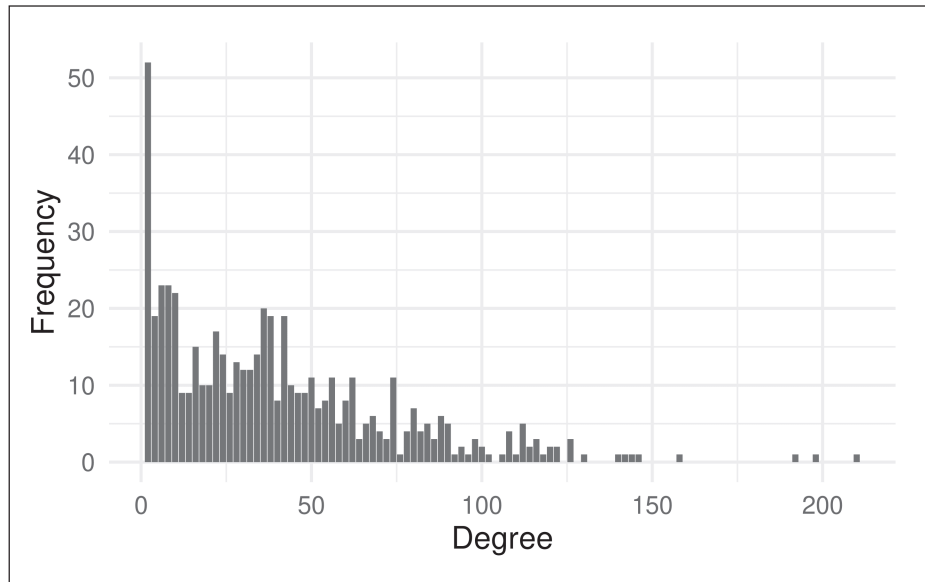


Figure 5 Empirical degree distribution of the email exchange network

5.3 Email exchange network

This network of emails within a research institution exhibits a skewed degree distribution, as shown in Figure 5. This type of degree distribution is typical of real-world social networks, and leads the classical SBM to often only be able to find core-periphery type block structures, with nodes grouped mostly on the basis of degree similarity. As explained above, one way to circumvent this issue is to use degree correction. For this application, we therefore made use of the original version of the degree-corrected stochastic blockmodel as in (3.3) (Karrer and Newman, 2011). The results of the model fitting, together with the partitioning of the network into real departments, are visualized in Figure 6. Looking at the plots, it is evident how the model with degree correction is able to recover the communities quite accurately. Comparing the partition discovered by the degree-corrected SBM with the actual departments, one small department (depicted towards the upper-centre of the figure) merges into another one close to it, and an additional block is therefore found at the bottom-centre of the plot, splitting the larger bottom department into two. Other than that, the structure found is remarkably similar to the partition induced by the departments, with some exceptions due to the existence of disconnected components within departments. In this case the Rand index is equal to 0.95, indicating a very high level of agreement among the partitions. For comparison purposes, we also fit a standard SBM to the same data, and computed the Rand index for the partition found with that as well. The resulting value of the index only amounts to 0.86, underlining the importance of applying degree correction in this case.

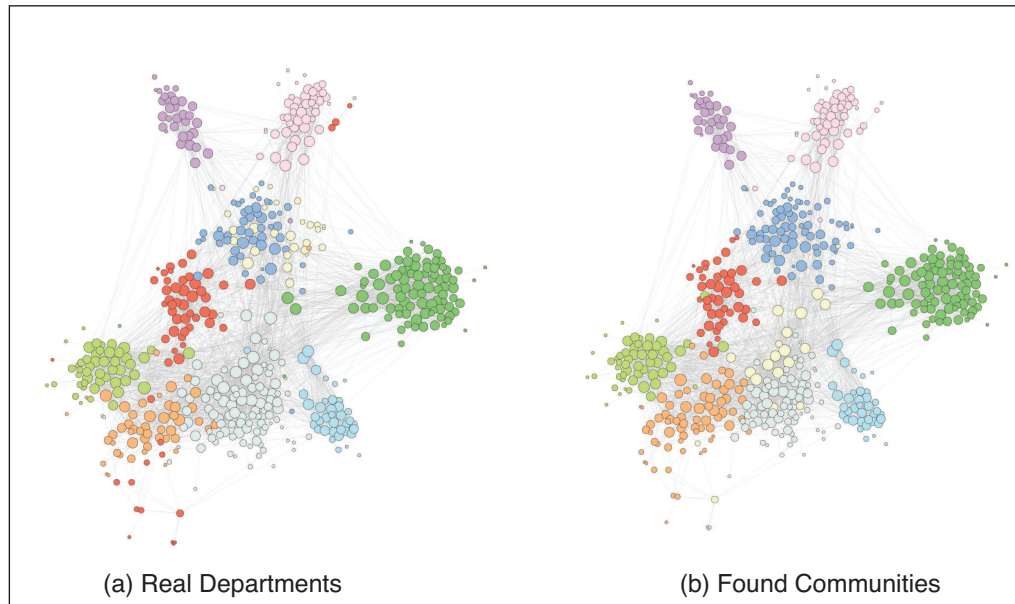


Figure 6 Comparison between ‘ground truth’ communities (departments) and groups found by the degree-corrected stochastic blockmodel in a network of email exchanges within a large European research institution

6 Conclusions

Mixture modelling can be extended to network data through stochastic blockmodels. Networks are rather complex structures, leading to computationally demanding estimation routines. Several algorithms specific for this class of problems have emerged over time, some of which are discussed in this article. We also provided an overview of different types of blockmodels by applying them to real-world network datasets. Among others, one of the models that we showcased is the degree-corrected stochastic blockmodel, which is particularly well suited for networks with a highly skewed degree structure.

Considering stochastic blockmodels (and community detection problems) as mixture models opens up a new avenue of extensions and novel models. Looking at the many model proposals in the field of mixture, ranging from mixing different distributions towards the mixture of experts, it is evident that these extensions can be brought forward in network modelling with mixtures as well. In fact, block-wise constant connectivity probabilities could be extended towards non-constant ones. Moreover, covariates could also be included. These extensions lie well beyond the scope of this article, but it is evident how the long history of mixture models, which started with Pearson (1894), has not come to an end, and extends promisingly in the realm of networks.

Acknowledgements

We would like to thank the European Cooperation in Science and Technology [COST Action CA15109 (COSTNET)]. The authors of this work take full responsibility for its content. The first author would also like to thank Cornelius Fritz for invaluable comments and discussions. Finally, the last author would like to thank Murray Aitkin for his enthusiasm, ingenuity and open-mindedness with respect to statistics, and, last but not least, for his friendship.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This work was also partly supported by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A as well as the Elite Network of Bavaria (ESG Data Science).

Appendix

Details on MCMC sampling scheme

Assuming $\mathbf{u}^{<t>} = (u_1^{<t>}, \dots, u_N^{<t>})$ to be the current state of the Markov chain, we can update the j th component as follows. At first, we set $u_l^{<t+1>} = u_l^{<t>}$ for all $l \neq j$, while for component u_j we draw a new potential state u_j^* from a uniform proposal with regard to the domain $[0, 1] \setminus [\tau_{k(j, <t>)-1}, \tau_{k(j, <t>)})$ and with $[\tau_{k(j, <t>)-1}, \tau_{k(j, <t>)})$ being the subinterval that includes $u_j^{<t>}$. This leads to the acceptance probability

$$\min \left\{ 1, \prod_{\substack{l=1 \\ l \neq j}}^N \left[\left(\frac{p(u_j^*, u_l^{<t>})}{p(u_j^{<t>}, u_l^{<t>})} \right)^{y_{jl}} \left(\frac{1 - p(u_j^*, u_l^{<t>})}{1 - p(u_j^{<t>}, u_l^{<t>})} \right)^{1 - y_{jl}} \right] \cdot \frac{1 - (\tau_{k(j, <t>)} - \tau_{k(j, <t>)-1})}{1 - (\tau_{k(j, *)} - \tau_{k(j, *)-1})} \right\}, \quad (\text{A.1})$$

where $[\tau_{k(j,*)-1}, \tau_{k(j,*)})$ represents the subinterval which includes u_j^* . If we accept the alteration, we set $u_j^{<t+1>}$ to the value u_j^* , while in the event of rejection, we remain with the previous value $u_j^{<t>}$. Running the Markov chain, we get a sampling sequence from which we derive a simulation-based estimate of the group mode, which concludes the E-step. It should be mentioned that, in the beginning, the number of Gibbs sampling states taken into account for approximating the posterior mode can be rather small, since the early model configurations are potentially far from the truth and thus already imply a deviating reallocation.

Label switching and non-identifiability

As has been extensively discussed in other works, approaching the conceptual formulation of mixture models by MCMC methods induces the label switching problem (see, for example, Stephens, 2000). This issue describes the invariance of the likelihood under relabelling of the mixture components. However, since in the proposed MCEM algorithm the model parameters are not part of the MCMC scheme but rather given as fixed based on the M-step, the label switching problem reduces to the exceptional case of symmetric parametrization. The two different situations can be exemplified by the configurations shown in Figure A.7. In both of the depicted cases (a) and (b), the two respective models describe and capture the exact same structure of the respective given network. That means none of them is preferable, and it is thus unclear beforehand to which label ordering the algorithm will tend. Nevertheless, regarding the non-symmetric configuration in (a), we point out that our MCEM algorithm will remain in either one of the partitions once that has been reached. At the stage of convergence, fixing the model parameters based on the M-step will leave the partition unchanged in the MCMC-based E-step (if the Gibbs sampling sequence is chosen to be sufficiently large). This is because the posterior distribution of the allocations is not invariant to label switching when $p(\cdot, \cdot)$ is fixed. Only in the symmetric case (b) a label switching might occur, which here exclusively refers to the node assignments, since now not only the likelihood but also $p(\cdot, \cdot)$ is invariant to label switching. In the worst case, this might lead to a fuzzy estimate in the M-step representing an in-between state of the different partitions. However, we argue that the case of two (or more) groups exhibiting a very similar connectivity behaviour in regard to all other groups (and among themselves) is an extraordinary one, that is unlikely to occur in real-world applications.

Another issue similar to the label switching problem which is inherent in graphon models is that of non-identifiability. This issue consists in the fact that different arrangements of the function $p(\cdot, \cdot)$ represent the same model. More precisely, as has been shown by Diaconis and Janson (2008), two graphon functions $p(\cdot, \cdot)$ and $\tilde{p}(\cdot, \cdot)$ represent the same network-generating model if and only if there exist two measure preserving functions $\varphi, \tilde{\varphi} : [0, 1] \rightarrow [0, 1]$ such that $p(\varphi(u), \varphi(v)) = \tilde{p}(\tilde{\varphi}(u), \tilde{\varphi}(v))$ for almost every $(u, v) \in [0, 1]^2$. Accordingly, this also includes the label switching problem, although it only refers to the model specification (and not to the likewise

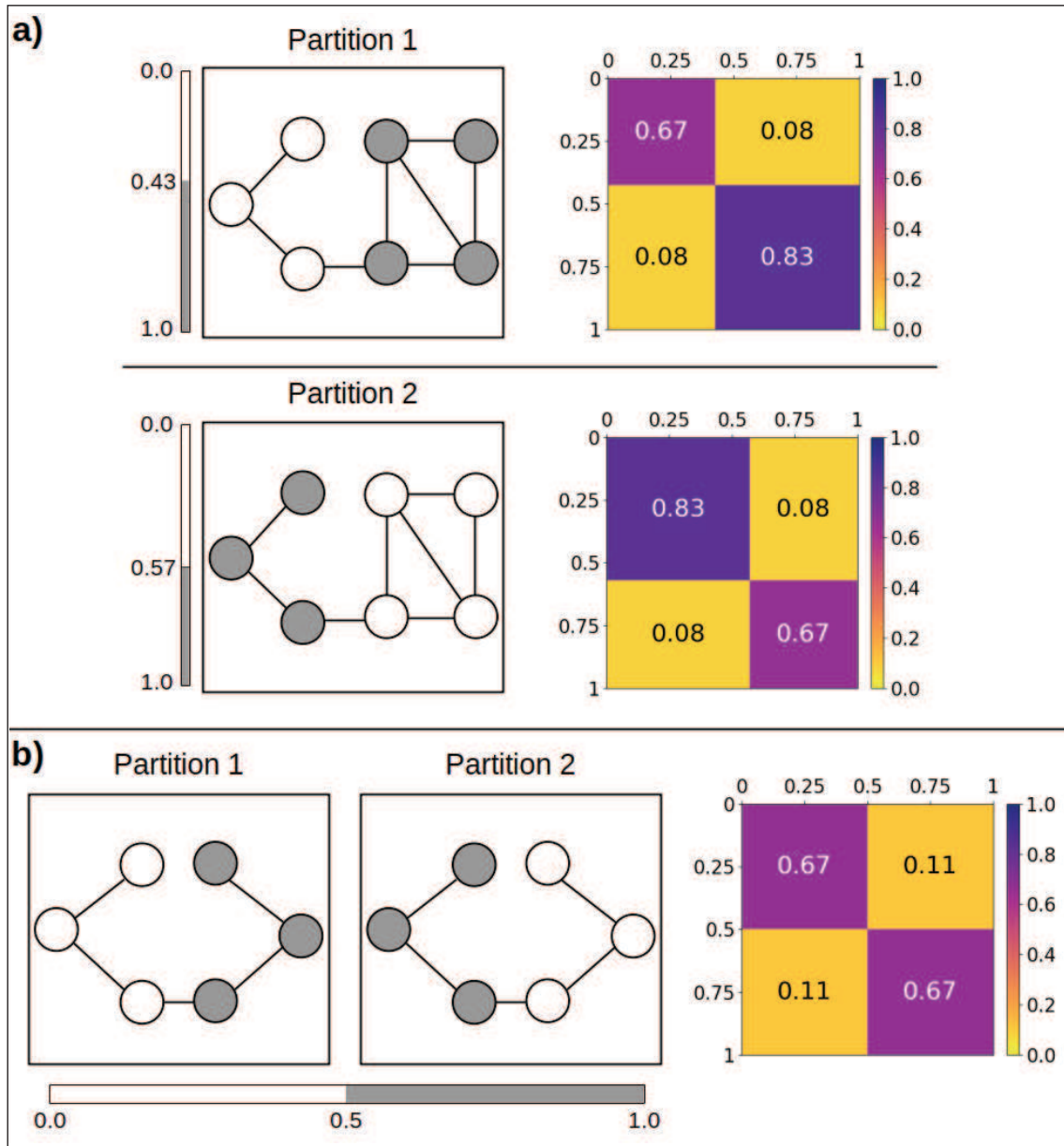


Figure A.7 Simple examples to illustrate the label switching problem in SBMs expressed through graphon representation. **a)** Two plausible blockmodels for the same seven-node network which are specified by $U_i \leq 3/7$ and $U_i \leq 4/7$, respectively (illustrated by node colors), and a corresponding step function $p(\cdot, \cdot)$ (depicted in the right column). Both models yield the same value for the likelihood and can be transferred into one another through label switching. **b)** Two potential partitions of a six-node network, each forming a blockmodel. The partitions can again be transferred into one another through label switching, but in this case they both refer to the same function $p(\cdot, \cdot)$. Only b) poses a label switching problem for the proposed MCEM algorithm

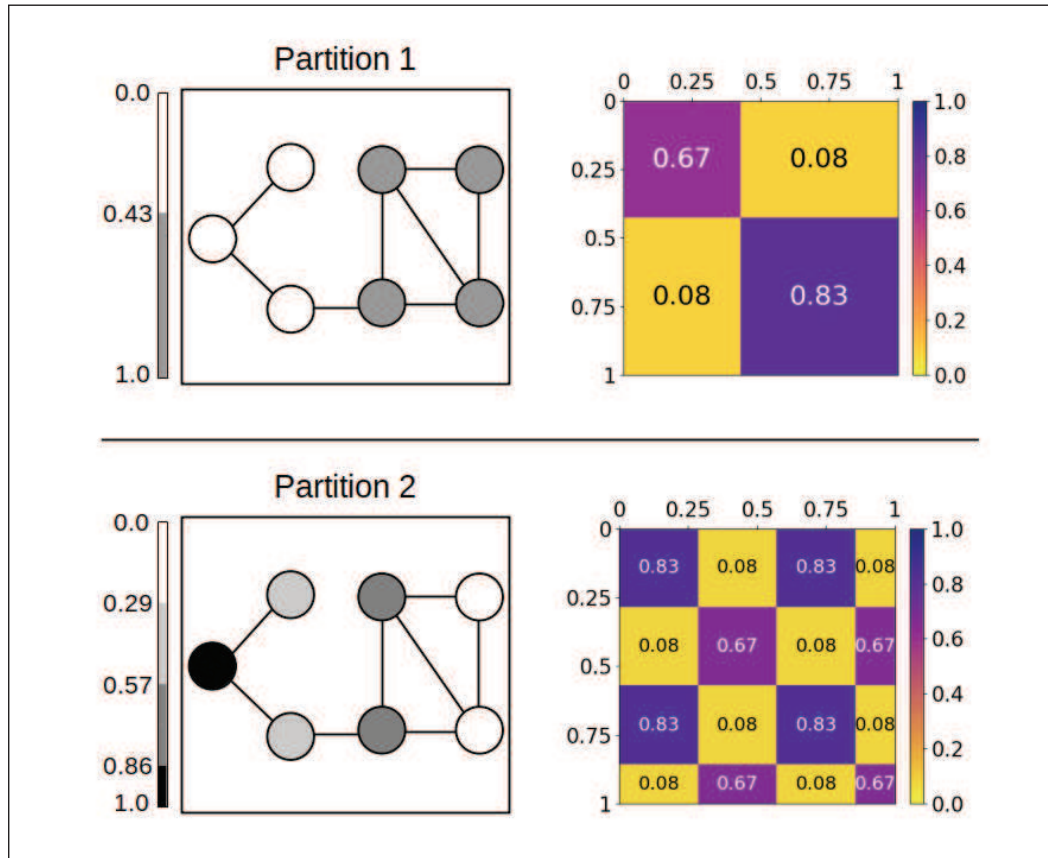


Figure A.8 Simple example to illustrate the splitting of groups in SBMs expressed through graphon representation. The node colouring exhibits node assignments, with the corresponding graphon functions depicted on the right. Both of the models describe and capture the same block structure in the given network. Nevertheless, the upper model is preferable to the lower model with respect to the following three criteria: a monotonically non-decreasing marginal function, merging similarly behaving nodes, and parsimony in terms of the number of communities

affected node assignments). Another potential instance of the identifiability issue in SBMs represented as graphon models lies in the splitting of groups. To illustrate that, we consider the two blockmodels in Figure A.8, which both capture the same structure in the given network. As has been mentioned in Section 3.3, the identifiability issue can be resolved by assuming a monotonically non-decreasing marginal function. This condition only applies to the upper model representation. However, considering our MCEM algorithm, the E-step aims to merge nodes with similar connectivity behaviour and therefore naturally prevents the splitting of groups. In addition, the lower representation is that of a blockmodel with four groups, a number which, compared to the upper representation, appears to be unnecessarily inflated. We hence argue that the identifiability issue in regards to the splitting of groups is a matter of the applied blockmodel dimensionality and can be also prevented

through an appropriate choice of the number of mixture components. We therefore avoid the additional constraint of a monotonically non-decreasing marginal function.

References

- Abbe E (2018) Community detection and stochastic block models. *Foundations and Trends in Communications and Information Theory*, **14**, 1–162.
- Aicher C, Jacobs AZ and Clauset A (2015) Learning latent block structure in weighted networks. *Journal of Complex Networks*, **3**, 221–48.
- Airoldi EM, Blei DM, Fienberg SE and Xing EP (2008) Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, **9**, 1981–2014.
- Airoldi EM, Costa TB and Chan SH (2013) Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems* **26**, pages 692–700.
- Aitkin M (1980) Mixture applications of the EM algorithm in GLIM. In *Proceedings of COMPSTAT 1980*, pages 537–41.
- Aitkin M (2011) How many components in a finite mixture? In *Mixture Estimation and Applications*, pages 277–92. Hoboken, NJ: Wiley.
- Aitkin M and Rubin DB (1985) Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society: Series B*, **47**, 67–75.
- Aitkin M, Vu D and Francis B (2014) Statistical modelling of the group structure of social networks. *Social Networks*, **38**, 74–87.
- Aitkin M and Wilson TG (1980) Mixture models, outliers and the EM algorithm. *Technometrics*, **22**, 325–31.
- Barabasi AL and Albert R (1999) Emergence of scaling in random networks. *Science*, **286**, 509–12.
- Bastian M, Heymann S and Jacomy M (2009) Gephi: An open source software for exploring and manipulating networks. *Proceedings of the International AAAI Conference on Web and Social Media*, **3**, 361–62.
- Benaglia T, Chauveau D, Hunter DR and Young DS (2009) mixtools: An R package for analyzing mixture models. *Journal of Statistical Software*, **32**, 1–29.
- Berge C (1984) *Hypergraphs: Combinatorics of Finite Sets* (Vol. 45). Amsterdam: Elsevier.
- Biagini F, Kauermann G and Meyer-Brandis T (2019) *Network Science*. Berlin: Springer-Verlag.
- Bickel PJ and Chen A (2009) A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, **106**, 21068–73.
- Böhning, D. (1999) *Computer Assisted Analysis of Mixtures and Applications: Meta Analysis, Disease Mapping and Others*. Boca Raton, FL: CRC Press.
- Bouveyron C, Latouche P and Zreik R (2018) The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*, **28**, 11–31.
- Bui TN, Chaudhuri S, Leighton FT and Sipser M (1987) Graph bisection algorithms with good average case behavior. *Combinatorica*, **7**, 171–91.
- Chan SH and Airoldi EM (2014) A consistent histogram estimator for exchangeable graph models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 208–16.
- Chen K and Lei J (2018) Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, **113**, 241–51.
- Chodrow PS (2020) Configuration models of random hypergraphs. *Journal of Complex Networks*, **8**.
- Choi DS, Wolfe PJ and Airoldi EM (2012) Stochastic blockmodels with a growing number of classes. *Biometrika*, **99**, 273–84.

- Clauset A, Newman ME and Moore C (2004) Finding community structure in very large networks. *Physical Review E*, **70**, 066111.
- Côme E and Latouche P (2015) Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. *Statistical Modelling: An International Journal*, **15**, 564–89.
- Daudin JJ, Picard F and Robin S (2008) A mixture model for random graphs. *Statistics and Computing*, **18**, 173–83.
- Dempster A, Laird N and Rubin D (1977) Maximum likelihood from incomplete observations. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.
- Diaconis P and Janson S (2008) Graph limits and exchangeable random graphs. *Rendiconti di Matematica*, **28**, 33–61.
- Everitt BS and Hand DJ (1980) *Finite Mixture Distributions*. Chapman & Hall.
- Fienberg SE (2012) A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, **21**, 825–39.
- Fortunato S (2010) Community detection in graphs. *Physics Reports*, **486**, 75–174.
- Fortunato S and Hric D (2016) Community detection in networks: A user guide. *Physics Reports*, **659**, 1–44.
- Friedl H and Kauermann G (2000) Standard errors for EM estimates in generalized linear models with random effects. *Biometrics*, **56**, 761–67.
- Frühwirth-Schnatter S (2006) *Finite Mixture and Markov Switching Models*. Berlin: Springer-Verlag.
- Frühwirth-Schnatter S, Celeux G and Robert CP (2019) *Handbook of Mixture Analysis*. London: Chapman & Hall.
- Goldenberg A, Zheng AX, Fienberg SE and Airoldi EM (2009) A survey of statistical network models. *Foundations and Trends in Machine Learning*, **2**, 129–233.
- Gormley IC and Frühwirth-Schnatter S (2019) Mixture of experts models. In *Handbook of Mixture Analysis*, pages 271–307. Boca Raton, FL: CRC Press.
- Gormley IC and Murphy TB (2010) A mixture of experts latent position cluster model for social network data. *Statistical Methodology*, **7**, 385–405.
- Gupta SS and Huang WT (1981) On mixture of distributions: A survey and some new results on ranking and selection. *Sankhya: The Indian Journal of Statistics*, **43**, 45–290.
- Handcock MS, Raftery AE and Tantrum JM (2007) Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **170**, 301–54.
- Hoff PD, Raftery AE and Handcock MS (2002) Latent space approaches to social network analysis. *Journal of the American Statistical Association*, **97**, 1090–98.
- Holland PW, Laskey K and Leinhardt S (1983) Stochastic blockmodels: First steps. *Social Networks*, **5**, 109–37.
- Huang S and Feng Y (2018) Pairwise covariates-adjusted block model for community detection. arXiv:1807.03469.
- Hunter DR, Handcock MS, Butts CT, Goodreau SM and Morris M (2008) ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software*, **24**.
- Jacobs RA, Jordan MI, Nowlan SJ and Hinton GE (1991) Adaptive mixtures of local experts. *Neural Computation*, **3**, 79–87.
- Jebabli M, Cheri H, Cheri C and Hamouda A (2018) Community detection algorithm evaluation with ground-truth data. *Physica A: Statistical Mechanics and Its Applications*, **492**, 651–706.
- Jordan MI, Ghahramani Z, Jaakkola TS and Saul LK (1999) Introduction to variational methods for graphical models. *Machine Learning*, **37**, 183–233.
- Karrer B and Newman ME (2011) Stochastic blockmodels and community structure in networks. *Physical Review E*, **83**, 016107.
- Kernighan BW and Lin S (1970) An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, **49**, 291–307.
- Kolaczyk ED (2009) *Statistical Analysis of Network Data: Methods and Models*. Berlin: Springer.

Mixture models and networks: The stochastic blockmodel 93

- Kolaczyk ED and Csardi G (2014) *Statistical Analysis of Network Data with R*. Berlin: Springer.
- Krivitsky PN, Handcock MS, Raftery AE and Hoff PD (2009) Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, **31**, 204–13.
- Latouche P and Robin S (2016) Variational Bayes model averaging for graphon functions and motif frequencies inference in W-graph models. *Statistics and Computing*, **26**, 1173–85.
- Lee C and Wilkinson DJ (2019) A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, **4**, 1–50.
- Lee KH, Xue L and Hunter DR (2020) Model-based clustering of time-evolving networks through temporal exponential-family random graph models. *Journal of Multivariate Analysis*, **175**, 104540.
- Leeds BA, Ritter JM, Mitchell SML and Long AG (2002) Alliance treaty obligations and provisions, 1815–1944. *International Interactions*, **28**, 237–60.
- Leger J-B (2016) Blockmodels: A R-package for estimating in Latent Block Model and Stochastic Block Model, with various probability functions, with or without covariates. arXiv:1602.07587.
- Leisch F (2004) FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, **11**, 1–18.
- Leskovec J and Krevl A (2014) SNAP Datasets: Stanford Large Network Dataset Collection. URL <http://snap.stanford.edu/data>
- Lindsay BG (1995) Mixture models: Theory, geometry and applications. In NSF-CBMS regional conference series in *Probability and Statistics*. Institute of Mathematical Statistics and the American Statistical Association.
- Lu X and Szymanski BK (2019) A regularized Stochastic Block Model for the robust community detection in complex networks. *Scientific Reports*, **9**, 1–9.
- Lusher D, Koskinen J and Robins G (2013) *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge: Cambridge University Press.
- Mariadassou M, Robin S and Vacher C (2010) Uncovering latent structure in valued graphs: A variational approach. *Annals of Applied Statistics*, **4**, 715–42.
- Masoudnia S and Ebrahimpour R (2014) Mixture of experts: A literature survey. *Artificial Intelligence Review*, **42**, 275–93.
- McDaid AF, Murphy TB, Friel N and Hurley NJ (2013) Improved Bayesian inference for the stochastic block model with application to large networks. *Computational Statistics and Data Analysis*, **60**, 12–31.
- McLachlan GJ and Peel D (2000) *Finite Mixture Models*. Hoboken, NJ: Wiley.
- Newman MEJ and Reinert G (2016) Estimating the number of communities in a network. *Physical Review Letters*, **117**, 78301.
- Nowicki K and Snijders TA (2001) Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, **96**, 1077–87.
- Olhede SC and Wolfe PJ (2014) Network histogram and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, **111**, 14722–27.
- Pearson K (1894) Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society*, **185**, 71–110.
- Peixoto TP (2012) Entropy of stochastic blockmodel ensembles. *Physical Review E*, **85**, 056122.
- Peixoto TP (2013) Parsimonious module inference in large networks. *Physical Review Letters*, **110**, 148701.
- Peixoto TP (2014a) Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, **4**, 011047.
- Peixoto TP (2014b) Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E*, **89**, 012804.
- Peixoto TP (2017) Nonparametric Bayesian inference of the microcanonical stochastic

- block model. *Physical Review E*, **95**, 012317.
- Rand WM (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–50.
- Riolo MA, Cantwell GT, Reinert G and Newman MEJ (2017) Efficient method for estimating the number of communities in a network. *Physical Review E*, **96**, 32310.
- Robbins H (1948) Mixture of distributions. *The Annals of Mathematical Statistics*, **19**, 360–69.
- Salter-Townshend M, White A, Gollini I and Murphy TB (2012) Review of statistical network analysis: Models, algorithms, and software. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **5**, 243–64.
- Simon HA (1955) On a class of skew distribution functions. *Biometrika*, **42**, 425–40.
- Sischka B and Kauermann G (2019) EM based smooth Graphon estimation using Bayesian and Spline based Approaches. arXiv:1903.06936.
- Snijders TA and Nowicki K (1997) Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, **14**, 75–100.
- Stephens M (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62**, 795–809.
- Sweet TM (2015) Incorporating covariates into stochastic blockmodels. *Journal of Educational and Behavioral Statistics*, **40**, 635–64.
- Tallberg C (2005) A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, **29**, 1–23.
- Teicher H (1960) On the mixture of distributions. *The Annals of Mathematical Statistics*, **31**, 55–73.
- Titterton DM, Smith AF and Makov UE (1985) *Statistical analysis of finite mixture distributions*. Hoboken, NJ: Wiley.
- Vu DQ and Aitkin M (2015) Variational algorithms for biclustering models. *Computational Statistics and Data Analysis*, **89**, 12–24.
- Vu DQ, Hunter DR and Schweinberger M (2013) Model-based clustering of large networks. *The Annals of Applied Statistics*, **7**, 1010.
- Wang J, Markert K and Everingham M (2009) Learning models for object recognition from natural language descriptions. In *Proceedings of the 20th British Machine Vision Conference (BMVC)*, volume 1, page 2.
- Wang YX and Bickel PJ (2017) Likelihood-based model selection for stochastic block models. *Annals of Statistics*, **45**, 500–28.
- Wang YJ and Wong GY (1987) Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, **82**, 8–19.
- White A and Murphy TB (2016) Mixed-membership of experts stochastic blockmodel. *Network Science*, **4**, 48–80.
- White HC, Boorman SA and Breiger RL (1976) Social structure from multiple networks: I. Blockmodels of roles and positions. *American Journal of Sociology*, **81**, 730–80.
- Yan X (2016) Bayesian model selection of stochastic block models. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 323–28.
- Yan X, Shalizi C, Jensen JE, Krzakala F, Moore C, Zdeborova L, Zhang P and Zhu Y (2014) Model selection for degree-corrected block models. *Journal of Statistical Mechanics: Theory and Experiment*, **2014**, P05007.
- Zitnik M, Rok Susic S and Leskovec J (2018) BioSNAP Datasets: Stanford biomedical network dataset collection. URL <http://snap.stanford.edu/biodata>.

6. Dependence matters: Statistical models to identify the drivers of tie formation in economic networks

Contributing article

De Nicola, G., Fritz, C., Mehrl, M., and Kauermann, G. (2023). Dependence matters: Statistical models to identify the drivers of tie formation in economic networks. *Journal of Economic Behavior & Organization*, 215:351–363. <https://doi.org/10.1016/j.jebo.2023.09.021>.

Data and code

Available at <https://github.com/gdenicola/statistical-network-analysis-in-economics>.

Copyright information

© 2023 Elsevier B.V. The full article is included in accordance with Elsevier's terms on reuse in dissertations for non-commercial purposes.

Supplementary material

Supplementary material is available at Elsevier online.

Author contributions

The idea of writing an overview paper focused on applications to economic networks can be attributed to Göran Kauermann, Cornelius Fritz and Giacomo De Nicola. The manuscript, which juxtaposes two prominent network model classes and showcases them through independent applications, was jointly designed and drafted by Giacomo De Nicola, Cornelius Fritz and Marius Mehrl. Giacomo De Nicola was specifically responsible for writing Section 4, and for the modeling and visualization of the historical foreign exchange network. Giacomo De Nicola further contributed by writing major parts of Sections 1, 2 and 5. The analysis and the writing for Section 3 were mainly carried out by Cornelius Fritz. All authors contributed through fruitful comments and extensive proofreading of the manuscript.



Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo



Dependence matters: Statistical models to identify the drivers of tie formation in economic networks

Giacomo De Nicola ^{a,*}, Cornelius Fritz ^b, Marius Mehrl ^c, Göran Kauermann ^a

^a Department of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich, Germany

^b Department of Statistics, Pennsylvania State University, 100 Thomas Building, 16802 State College, PA, USA

^c School of Politics and International Studies, University of Leeds, LS2 9JT Leeds, United Kingdom

ARTICLE INFO

JEL classification:

C20
C49
F14
F31
L14

Keywords:

Inferential network analysis
Network data
Endogeneity
Arms trade
Foreign exchange networks
Statistical modeling

ABSTRACT

Networks are ubiquitous in economic research on organizations, trade, and many other areas. However, while economic theory extensively considers networks, no general framework for their empirical modeling has yet emerged. We thus introduce two different statistical models for this purpose – the Exponential Random Graph Model (ERGM) and the Additive and Multiplicative Effects network model (AME). Both model classes can account for network interdependencies between observations, but differ in how they do so. The ERGM allows one to explicitly specify and test the influence of particular network structures, making it a natural choice if one is substantively interested in estimating endogenous network effects. In contrast, AME captures these effects by introducing actor-specific latent variables affecting their propensity to form ties. This makes the latter a good choice if the researcher is interested in capturing the effect of exogenous covariates on tie formation without having a specific theory on the endogenous dependence structures at play. After introducing the two model classes, we showcase them through real-world applications to networks stemming from international arms trade and foreign exchange activity. We further provide full replication materials to facilitate the adoption of these methods in empirical economic research.

1. Introduction

The study of networks has established itself as a central topic in economic research (Jackson, 2008). Within the broader context of the study of complex and interdependent systems (see e.g. Flaschel et al., 1997, 2007, 2018), networks can be defined as interconnected structures which can naturally be represented through graphs. In the economic literature, networks have been extensively considered from a theoretical perspective, with the primary goal of understanding how economic behavior is shaped by interaction patterns (Jackson and Rogers, 2007). Indeed, the adequate modelling of such interactions has been described as one of the main empirical challenges in economic network analysis (Jackson et al., 2017). Research in this direction on, e.g., organizations as networks, diffusion in networks, network experiments, or network games, is surveyed in Bramoullé et al. (2016), Jackson (2014), and Jackson et al. (2017). These theoretical advances find application in many different fields in which network structures naturally arise, such as national and international trade, commercial agreements, firms' organization, and collaboration activity. However, such advances have not yet been accompanied by a corresponding shift in the standard methods used to empirically validate them.

* Corresponding author.

E-mail address: giacomo.denicola@stat.uni-muenchen.de (G. De Nicola).

<https://doi.org/10.1016/j.jebo.2023.09.021>

Received 28 October 2022; Received in revised form 7 July 2023; Accepted 18 September 2023

Available online 29 September 2023

0167-2681/© 2023 Elsevier B.V. All rights reserved.

Some recent contributions (see e.g. Atalay et al., 2011; Chaney, 2014; Morales et al., 2019) develop estimators tailored specifically to their network-based theoretical models, but more generally applicable modeling frameworks for the analysis of real-world network data have not yet emerged. Statistical methods specifically designed to empirically test theories where interdependencies arise from network structures, such as the Exponential Random Graph Model (ERGM), exist but are not yet widely used by economists. Jackson (2014), for instance, discusses ERGMs but argues that they “suffer from proven computational problems” (2014, p.76). Jackson et al. (2017) further explain that “it is practically impossible to estimate the likelihood of a given network at even a moderately large scale”, concluding that with ERGMs, “there is an important computational hurdle that must be overcome in working with data” (2017, p.85).

Contrasting this assessment, we argue that recent work in the realm of empirical network analysis provides robust and scalable methods with readily available implementations in the R statistical software (R Core Team, 2021). Computational issues thus do not represent an insurmountable barrier to employ robust inferential network methods anymore. In this paper, we demonstrate the effectiveness and usability of some of those methods by applying them to real economic data. We specifically focus on models which aim to capture the mechanisms leading to network formation, i.e. to measure how the probability of forming a tie is influenced by (a) nodal characteristics, (b) pairwise covariates, and (c) the rest of the network. In particular, our focus is on Exponential Random Graph Models (ERGM) (Robins et al., 2007a) and Additive and Multiplicative Effect (AME) network models (Hoff, 2021), respectively implemented in the R packages *statnet* (Handcock et al., 2008b) and *amen* (Hoff, 2015). We find these two model classes to be among the most promising ones for applications in the economic sciences, as they are well suited for answering two broad categories of research questions. The ERGM is an ideal fit if, based on economic theory, the researcher envisages a particular dependence structure for the existence of ties in the network at hand and wants to test whether their theory is corroborated by empirical data. On the other hand, AME, and more generally continuous latent variable models, are a good choice when the researcher is interested in capturing the effect of exogenous variables on tie formation without having prior knowledge on which endogenous network dependence mechanisms are at play. In this case, AME offers the possibility to estimate the effect of both nodal and pairwise covariates while simultaneously controlling for network effects, which may induce bias if ignored (see Lee and Ogburn, 2021). In addition, the estimated latent structure can provide insight on the underlying network mechanisms for which they are controlling.

The principal aim of this paper is to showcase ERGM and AME by focusing on their value for economic research. After introducing each model class, we demonstrate their empirical usage by respectively applying them to two relevant economic questions stemming from real-world networks. We first use the ERGM to model the international trade of major conventional weapons, where a directed tie exists if one country transfers arms to another. In line with Chaney (2014), network effects such as directed triadic closure (e.g. the positive impact of an increase in the volume of trade between countries A and B on the probability that country C, that already exports to A, starts exporting to B) are of explicit theoretical interest in this application, and the ERGM allows for their proper specification and testing. We then make use of the AME model to study a historical network of global foreign exchange activity, where a directed edge is present if one country's national currency is actively traded within the other country. AME allows us to estimate how relevant country features, such as per-capita gdp and the gold standard, and pairwise covariates, such as the distance between two countries and their reciprocal trade volume, influence tie formation, while controlling for network effects to provide unbiased estimates. We further compare the two model classes, weighing pros and cons of each approach and providing guidance on which tool is appropriate for applications to different empirical settings and research questions. Finally, in addition to a step-by-step analysis and interpretation of these application cases, we provide full replication code in our GitHub repository,¹ allowing for seamless reproducibility. We, therefore, demonstrate the “off-the-shelf” applicability of these methods, and offer applied researchers a head-start in employing them to study substantive economic problems.

Our contribution is related to various strands of the growing literature on economic networks (e.g. Jackson and Rogers, 2007; Jackson, 2008; Bramoullé et al., 2016). Due to its focus on economic questions, our work differs from surveys in physics (Newman, 2003), statistics (Goldenberg et al., 2010), or political science (Cranmer et al., 2017). Several articles provide overviews and surveys of existing economic network models from a theoretical perspective (Jackson, 2014; Graham, 2015; Jackson et al., 2017; De Paula, 2020). None of these articles concentrates on discussing broadly applicable statistical modeling frameworks, such as ERGM and AME, from an empirical perspective. In this sense our paper is similar in spirit to van der Pol (2019) who, however, only focuses on ERGM, without comparing alternative approaches. Indeed, one of the goals of this paper is to shed light on the emerging AME model class (and, more generally, on latent variable network models) for future applications in the economic literature.

The remainder of the paper is structured as follows. Section 2 discusses existing literature and presents the mathematical and notational framework used to define and discuss networks throughout the paper. Section 3 introduces the ERGM and applies it to the international arms trade network. Section 4 is dedicated to AME and its application to the global foreign exchange network. Section 5 concludes the paper with a brief discussion on the two model classes, contrasting their different uses and highlighting pros and cons of each approach.

¹ <https://github.com/gdenicola/statistical-network-analysis-in-economics>.

6. Dependence matters: Statistical models to identify the drivers of tie formation in economic networks

2. Economic networks

2.1. Related literature

Even though network structures naturally arise in many aspects of economics and are subject of prominent research in the field, much of the previous literature has ignored the implied interdependencies, instead opting for regression models assuming ties to be independent conditional on the covariates (e.g. Anderson and Van Wincoop, 2003, Rose, 2004, Lewer and Van den Berg, 2008). This assumption is often unreasonable in practice. It would, for example, imply that Germany imposing economic sanctions on Russia is independent of Italy imposing sanctions on Russia, and, in the directed case, even of Russia imposing them on Germany itself. While no standard framework for the modeling of empirical network data has emerged in economics so far, a number of contributions in – or adjacent to – the field do make use of statistical network models. We shortly survey these works here to show that the models we present are indeed suitable for the analysis of economic data. Possibly the most obvious kind of economic network is the international trade network (see Chaney, 2014) and many of these studies accordingly seek to model the formation of trade ties. In this vein, two early studies (Ward and Hoff, 2007; Ward et al., 2013) apply latent position models to show that trade exhibits a latent network structure beyond what a standard gravity model can capture (see also Fagiolo, 2010; Dueñas and Fagiolo, 2013). More recently, numerous contributions have used the ERGM to explicitly theorize and understand network interdependence in the general trade (Herman, 2022; Liu et al., 2022; Smith and Sarabi, 2022) as well as the trade in arms (Thurner et al., 2019; Lebacher et al., 2021), patents (He et al., 2019), and services (Feng et al., 2021).

That being said, empirical research on economic networks is not limited to trade. Smith et al. (2019) use multilevel ERGMs to study a production network consisting of ownership ties between firms at the micro-level and trade ties between countries at the macro-level, while Mundt (2021) explores the European Union's sector-level production network via ERGMs as well as an alternative methodology, the stochastic actor-oriented model (SAOM). The latter is another prominent tool in the realm of network analysis, which is suitable for modeling longitudinal network data. As we, in the interest of brevity, focus on models for static networks (i.e. networks that are observed only at one point in time), we do not treat the SAOM, and instead refer to Snijders (1996, 2017) for an introduction to the model class. Going back to empirical research on economic networks in the literature, Fritz et al. (2023) deploy ERGMs to investigate patent collaboration networks. Studies on foreign direct investments document network influences using latent position models (Cao and Ward, 2014), or seek to model them via extensions of the ERGM (Schoeneman et al., 2022). Finally, economists also study networks of interstate alliances and armed conflict (see e.g. Jackson and Nei, 2015; König et al., 2017), both of which have been modeled via ERGMs (Cranmer et al., 2012; Campbell et al., 2018) and AME (Dorff et al., 2020; Minhas et al., 2022). This short survey indicates that both ERGM and AME can be used to answer questions which are of substantive interest to economists.

2.2. Setup

Before introducing models for networks in which dependencies between ties are expected, we briefly introduce the mathematical framework for networks, as well as the necessary notation. Let $y = (y_{ij})_{i,j=1,\dots,n}$ be the adjacency matrix representing the observed binary network, comprising n fixed and known agents (nodes). In this context, $y_{ij} = 1$ indicates an edge from agent i to agent j , while $y_{ij} = 0$ translates to no edge between the two. Since self-loops are not admitted for most studied networks, the diagonal of y is left unspecified or set to zero. Depending on the application, the direction of an edge can carry additional information. If it does, we call the network directed. In this article, we mainly focus on this type of networks. Also note that all matrix-valued objects are written in bold font for consistency. In addition to the network connections, we often observe covariate information on the agents, which can be at the level of single agents (e.g. the gdp of a country) or at the pairwise level (e.g. the distance between two countries). We denote covariates by x_1, \dots, x_p , and our goal is to specify a statistical model for Y , that is the random variable corresponding to y , conditional on x_1, \dots, x_p . A natural way to do this is to specify a probability distribution over the space of all possible networks, which we define by the set \mathcal{Y} . Two main characteristics differentiate our modeling endeavor from classical regression techniques, such as Probit or logistic regression models. First, for most applications, we only observe one realization y from Y , rendering the estimation of the parameters to characterize this distribution particularly challenging. Second, the entries of Y are generally co-dependent; thus, most conditional dependence assumptions inherent to common regression models are violated. Generally, we term mechanisms that induce direct dependence between edges to be endogenous, while all effects external to the modeled network, such as covariates, are called exogenous.

3. The exponential random graph model

The ERGM is one of the most popular models for analyzing network data. First introduced by Holland and Leinhardt (1981) as a model class that builds on the platform of exponential families, it was later extended with respect to fitting algorithms and more complex dependence structures (Lusher et al., 2012; Robins et al., 2007b). We next introduce the model step-by-step to highlight its ability to progressively generalize by building on conditional dependence assumptions.

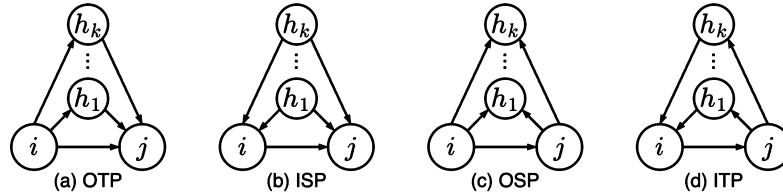


Fig. 1. Illustration of directed edgewise-shared partner statistics for k agents. Circles represent agents, and black lines represent edges between them. The names follow statenet nomenclature: OTP = “Outgoing Two-Path”, ISP = “Incoming Shared Partner”, OSP = “Outgoing Shared Partner”, and ITP = “Incoming Two-Path”.

3.1. Accounting for dependence in networks

We begin with the simplest possible stochastic network model, the Erdős-Rényi-Gilbert model (Erdős and Rényi, 1959; Gilbert, 1959), where all edges are assumed to be independent and to have the same probability of being observed. In stochastic terms, each observed tie is then a realization of a binomial random variable with success probability π , which yields

$$\mathbb{P}_\pi(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^n \prod_{j \neq i} \pi^{y_{ij}} (1 - \pi)^{1 - y_{ij}} \tag{1}$$

for the probability to observe \mathbf{y} . Evidently, model (1), which implies equal probability for all possible ties, is too restrictive to be applied to real world problems. In the next step, we, therefore, additionally incorporate covariates x_{ij} by letting π vary depending on those covariates, leading to edge-specific probabilities π_{ij} . Following the common practice in logistic regression, we parameterize the log-odds by $\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \theta^\top x_{ij}$, where x_{ij} is a vector of exogenous statistics with the first entry set to 1 to incorporate an intercept, and get

$$\mathbb{P}_\theta(\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^n \prod_{j \neq i} \left(\frac{\exp\{\theta^\top x_{ij}\}}{1 + \exp\{\theta^\top x_{ij}\}}\right)^{y_{ij}} \left(\frac{1}{1 + \exp\{\theta^\top x_{ij}\}}\right)^{1 - y_{ij}} \tag{2}$$

From (2), the analogy to standard logistic regression being a special case of generalized linear models (Nelder and Wedderburn, 1972) becomes apparent. The joint distribution of \mathbf{Y} can be formulated in exponential family form, yielding

$$\mathbb{P}_\theta(\mathbf{Y} = \mathbf{y} | \mathbf{x}) = \frac{\exp\{\theta^\top s(\mathbf{y})\}}{\kappa(\theta)}, \tag{3}$$

where $s(\mathbf{y}) = (s_1(\mathbf{y}), \dots, s_p(\mathbf{y}))$, $s_q(\mathbf{y}) = \sum_{i=1}^n \sum_{j \neq i} y_{ij} x_{ij,q} \forall q = 1, \dots, p$, with $x_{ij,q}$ as q -th entry in x_{ij} and $\kappa(\theta) = \prod_{i=1}^n \prod_{j \neq i} (1 + \exp\{\theta^\top x_{ij}\})$. In the jargon of exponential families, we term $s(\mathbf{y})$ sufficient statistics.

Newcomb (1979) observed that many observed networks exhibit complicated relational mechanisms, including reciprocity, which we can account for by extending the set of sufficient statistics. Under reciprocity, an edge Y_{ji} influences the probability of its reciprocal edge Y_{ij} to occur. Analyzing social networks, we would expect that the probability of agent i nominating agent j to be a friend is higher if agent j has nominated agent i as a friend. Holland and Leinhardt (1981) extended model (1) to such settings with the so-called p_1 model. To represent reciprocity, we assume dyads, each of them defined by (Y_{ij}, Y_{ji}) , to be independent of one another, which again yields an exponential family distribution similar to (3) with sufficient statistics that count the number of mutual ties ($s_{\text{Mut}}(\mathbf{y}) = \sum_{i < j} y_{ij} y_{ji}$), of edges ($s_{\text{Edges}}(\mathbf{y}) = \sum_{i=1}^n \sum_{j \neq i} y_{ij}$), and the in- and out-degree statistics for all degrees observed in the networks.² Agents’ in- and out-degrees are their number of incoming and outgoing edges, and relate to their relative position in the network (Wasserman and Faust, 1994).

Next to reciprocity, another important endogenous network mechanism is transitivity, originating in the structural balance theory of Heider (1946) and adapted to binary networks by Davis (1970). Transitivity affects the clustering in the network, implying that a two-path between agents i and j , i.e. $y_{ih} = y_{hj} = 1$ for some other agent h , affects the edge probability of Y_{ij} . Put differently, Y_{ij} and Y_{kh} are assumed to be independent iff $i, j \neq k$ and $i, j \neq h$. Frank and Strauss (1986) proposed the Markov model to capture such dependencies. For this model, the sufficient statistics are star-statistics, which are counts of sub-structures in the network where one agent has (incoming and outgoing) edges to between 0 and $n - 1$ other agents, and counts of triangular structures. If the network is directed it is possible to define different types of triangular structures, as depicted in Fig. 1.

3.2. Extension to general dependencies

Starting from the Erdős-Rényi-Gilbert model, which is a special case of a generalized linear model, we have consecutively allowed for more complicated dependencies between edges, resulting in the Markov graphs of Frank and Strauss (1986). Over this course, we showed that each model can be stated in exponential family form, characterized by a particular set of sufficient statistics. We now

² We provide more details on this derivation in the Supplementary Material.

6. Dependence matters: Statistical models to identify the drivers of tie formation in economic networks

make this more explicit to allow for more general dependence structures, and specify a probabilistic model for Y directly through the sufficient statistics.³ Wasserman and Pattison (1996) introduced this model as

$$\mathbb{P}_\theta(Y = y) = \frac{\exp\{\theta^\top s(y)\}}{\kappa(\theta)}, \tag{4}$$

where θ is a p -dimensional vector of parameters to be estimated, $s(y)$ is a function calculating the vector of p sufficient statistics for network y , and $\kappa(\theta) = \sum_{\bar{y} \in \mathcal{Y}} \exp\{\theta^\top s(\bar{y})\}$ is a normalizing constant to ensure that (4) sums up to one over all $y \in \mathcal{Y}$. To estimate θ , Handcock (2003) adapted the Monte Carlo Maximum Likelihood technique of Geyer and Thompson (1992), approximating the logarithmic likelihood ratio of θ and a fixed θ_0 via Monte Carlo quadrature (see Hunter et al., 2012, for an in-depth discussion).

A problem often encountered when fitting model (4) to networks is degeneracy (Handcock, 2003; Schweinberger, 2011). Degenerate models are characterized by probability distributions that put most probability mass either on the empty or on the full network, i.e., where either all or no ties are observed. To detect this behavior, one can use a goodness-of-fit procedure where observed network statistics are compared to statistics of networks simulated under the estimated model (Hunter et al., 2008). To address it, Snijders et al. (2006) and Hunter and Handcock (2006) propose weighted statistics that, in many cases, have better empirical behavior. Degeneracy commonly affects model specifications encompassing statistics for triad counts and multiple degree statistics. For in-degree statistics, we would thus incorporate the geometrically-weighted in-degree,

$$GWIDEG(y, \alpha) = \exp\{\alpha\} \sum_{k=1}^{n-1} (1 - \exp\{-\alpha\})^k IDEG_k(y), \tag{5}$$

where $IDEG_k(y)$ is the number of agents in the studied network with in-degree k and α is a fixed decay parameter. One can substitute $IDEG_k(y)$ in (5) with the number of agents with a specific out-degree, $ODEG_k(y)$, to capture the out-degree distribution. We term these statistics geometrically weighted since the weights in (5) are a geometric series.⁴ A positive estimate implies that an edge from a low-degree agent is more likely than an edge from a high-degree agent, resulting in a decentralized network. If, on the other hand, the corresponding coefficient is negative, one may interpret it as an indicator for a centralized network.

To capture clustering, we have to define the distribution of edgewise-shared partners (ESP). This distribution is defined as the relative frequency of edges in the network with a specific number of k shared partners, that we denote by $ESP(y)$ for $k \in \{1, \dots, n-2\}$. As shown in Fig. 1, various versions of edgewise-shared partner statistics can be found in directed networks, depending on the direction of the edges between the three agents involved. Geometrically weighted statistics can be stated for them in a similar manner as for degree statistics. For example, for the outgoing two-path (OTP, see Fig. 1a), this is

$$GWOTP(y, \alpha) = \exp\{\alpha\} \sum_{k=1}^{n-2} (1 - \exp\{-\alpha\})^k OTP_k(y). \tag{6}$$

In this case, a positive coefficient indicates that sharing ties with third actors increases the probability of observing an event between two agents.

Along with capturing endogenous network statistics, it is also possible to extend the ERGM framework to include the temporal dimension, that is, to model longitudinal network data. This is done quite naturally through use of a Markov assumption on the temporal dependence of subsequently observed networks, giving rise to the Temporal Exponential Random Graph Model (TERGM). As we here focus on static networks, we do not cover this in depth, and refer to Hanneke et al. (2010) for an introduction to the TERGM, and to Fritz et al. (2020) for a more general discussion on temporal extensions to the model class.

In summary, the ERGM allows to account for network dependencies via explicitly specifying them in $s(y)$. A large variety of potential network statistics, such as those given in (5) and (6), can be included in $s(y)$, enabling to test for their influence in the formation of the observed network. By allowing for this explicit inclusion and testing of network statistics, the ERGM requires researchers to at least have an implicit theory regarding what types of network dependence should exist in the network they study. Without such theory to guide the selection of network statistics, the range of potential network dependencies, and corresponding statistics, is virtually endless.⁵ As a result, the ERGM is best suited for research questions that explicitly concern interdependencies within the network. If these interdependencies are, instead, only a potential source of bias the researcher wants to control for, the AME model (introduced in Section 4) may be a better fit.

3.3. Application to the international arms trade network

We next make use of the ERGM to analyze the international arms transfer network. Recent studies on trade in Major Conventional Weapons (MCW), such as fighter aircraft or tanks, not only emphasize its networked nature, but also argue that this very nature is of substantive theoretical interest (Thurner et al., 2019; Fritz et al., 2021). In line with Chaney (2014), triadic trade structures are held to reveal information regarding the participants' economic and security interests. Explicitly modeling these structures allows us

³ Alternatively, (4) can also be derived as the equilibrium distribution of a strategic game where players myopically reassess and update their links to optimize their utility in the network (see Mele, 2017; Boucher and Mourifié, 2017).

⁴ Geometrically weighted statistics require setting the decay parameter α . We set $\alpha = \log(2)$, though it can also be estimated as an additional parameter given sufficient data (Hunter and Handcock, 2006).

⁵ For a survey of possible endogenous terms, see Morris et al. (2008).

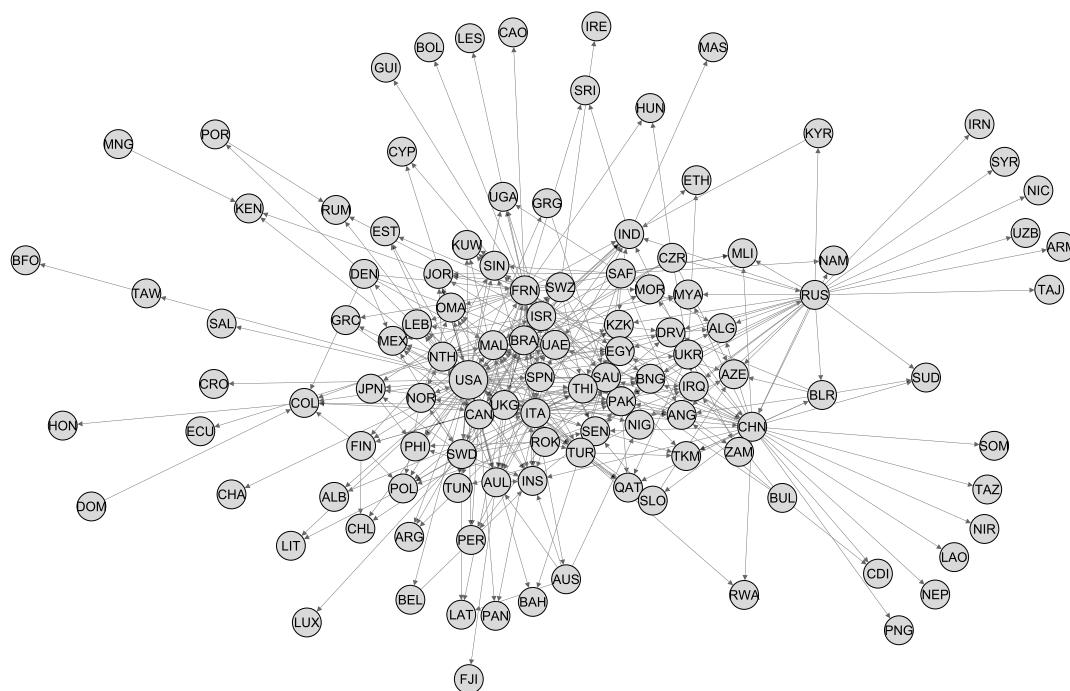


Fig. 2. Illustration of the international arms trade network in 2018. Countries are labeled by their ISO 3166-1 codes, and a directed edge from node i to node j indicates major conventional weapons being delivered from country i to country j .

to test hypotheses regarding their effects on further arms transfers. Accordingly, we seek to model the network of international arms transfers in the year 2018, where countries are nodes and a directed edge indicates MCW being delivered from country i to country j . Our interest here mainly lies in uncovering the network's endogenous mechanisms. MCW trade data come from SIPRI (2021), and the resulting network is depicted in Fig. 2, obtained using the Yifan Hu force-directed graph drawing algorithm (Hu, 2005) with the software Gephi (Bastian et al., 2009).

For estimating the parameters characterizing the ERGM, we use the R package *ergm* (Handcock et al., 2008a). Since evaluating $\kappa(\theta)$ from (4) necessitates calculating the sum of $|\mathcal{Y}| = 2^{n(n-1)}$ terms, we rely on MCMC approximations thereof to obtain the maximum likelihood estimates (see Handcock, 2003 and Hummel et al., 2012 for additional information on this topic). As discussed above, the ERGM allows us to use both exogenous (node-specific and pair-specific) attributes as well as endogenous structures to model the network of interest. Here, we select both types of covariates based on existing studies on the arms trade (Thurner et al., 2019; Fritz et al., 2021). In addition to an edges term, which corresponds to the intercept in standard regression models, we include importers' and exporters' logged GDP, whether they share a defense pact, their absolute difference in "polity" scores (a type of democracy index), and their geographical distance.⁶ We lag these covariates by three years, reflecting the median time between order and delivery for MCW delivered in 2018.⁷ More importantly, for the purpose of demonstrating how to model network data with the ERGM, we specify five endogenous network terms. In- and out-degree (IDEG and ODEG) measure, respectively, importers' and exporters' trade activity, and thus capture whether highly active importers and exporters are particularly attractive trading partners, or if they are instead less likely to form additional trade ties. Moreover, we specify a reciprocity term to capture whether countries tend to trade MCW uni- or bidirectionally. We further include two types of triadic structures, which represent transitivity and a shared supplier between countries i and j . The transitivity term counts how often country i exports arms to j while i exports to k , which in turn exports to j , thus capturing i 's tendency to directly trade with j if they engage in indirect trade (OTP, see Fig. 1a). In contrast, the shared supplier term counts how often country i sends arms to j while both import weapons from a shared supplier k (ISP, see Fig. 1b). Note that, given the issue of degeneracy discussed above, we use geometrically weighted versions of all endogenous statistics except reciprocity. Finally, we include a repetition term capturing whether arms transfer dyads observed in 2018 had already occurred in

⁶ Data for these covariates come from the *peacesciencer* package (Miller, 2022).

⁷ We use the median as the distribution of times between order and delivery is quite skewed. As shown in the Supplementary Materials, our substantive results remain unchanged when using 4- and 5-year lags instead, which reflect the average time between order and delivery. In particular, the ERGM outperforms the logistic regression model regardless of lag choice.

6. Dependence matters: Statistical models to identify the drivers of tie formation in economic networks

Table 1
Estimated coefficients and standard errors (in parentheses) of the ERGM and the logistic regression model for the international arms trade network in 2018.

	ERGM	Logit
Intercept	−15.356 (2.017)***	−28.197 (1.731)***
Repetition	3.254 (0.141)***	3.957 (0.141)***
Distance	−0.081 (0.087)	−0.239 (0.088)**
Abs. Diff. Polity	−0.001 (0.010)	−0.003 (0.012)
Alliance	0.350 (0.207)	0.209 (0.207)
log-GDP (Sender)	0.300 (0.050)***	0.588 (0.045)***
log-GDP (Receiver)	0.166 (0.049)***	0.355 (0.039)***
Mutual	−0.311 (0.438)	
GWIDEG	−1.478 (0.296)***	
GWODEG	−2.848 (0.296)***	
GWOTP	−0.146 (0.104)	
GWISP	0.210 (0.083)*	
AIC	1769.718	1891.984
BIC	1866.053	1948.179
Log Likelihood	−872.859	−938.992

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

any of the previous three years. Results of this ERGM, as well as, for comparison's sake, a logistic regression that includes the same exogenous covariates but does not capture any of the endogenous network structures, are presented in Table 1. These results can be compared directly, as, just like in a logistic regression model, coefficients in the ERGM indicate the additive change in the log odds of a tie occurring in association with a unit change in the respective variable. In this sense, the logistic regression model can be viewed as a special case of the ERGM in which the network effects are omitted. From the table, we can see how the two models differ both in their in-sample fit, as captured by AIC and BIC,⁸ as well as in the substantive effects they identify for the exogenous covariates. The repetition coefficient is positive and statistically significant in both models, but differs substantially in its size. An arms transfer edge having occurred at least once in 2015–17 increases the log odds of it occurring also in 2018 by 3.96 in the Logit, but only by 3.26 in the ERGM. Similarly, both models agree that the log odds of an arms transfer occurring increase with the economic size of the sender and receiver, as captured by their respective GDPs, but the coefficients retrieved by the Logit are approximately double the size of those in the ERGM, thus attributing more explanatory power to them. Also in this vein, the effect of the geographical distance between sender and receiver is three times as large in the logistic regression as in the ERGM and, while statistically significant in the former, indistinguishable from zero in the latter. Finally, both models report small and statistically insignificant effects for countries' polity difference and alliance ties. Taken together, however, there are clear, substantively meaningful differences in the effect sizes and, in the case of geographical distance, even statistical significance of the coefficients that the ERGM and Logit recover for the exogenous covariates.

Furthermore, three of the endogenous statistics included in the ERGM exhibit statistically significant effects on the probability of arms being traded. The results for in- and out-degree replicate the finding by Thurner et al. (2019), showing that highly active importers and exporters are less likely to form additional trade ties. In the ERGM, coefficients can also be interpreted at the global level, in addition to the edge-level interpretation given above. The shared supplier term having a (statistically significant) positive coefficient indicates, at the edge level, that an exporter is more likely to transfer weapons to a potential receiver if both of them import arms from the same source. Globally, on the other hand, the same coefficient means that the observed network exhibits more shared supplier configurations – where country i sends weapon to j while both receive arms from k – than would be expected in a random network of the same size. On the whole, the results presented in Table 1 offer an example for the striking differences that modeling network structures (instead of assuming them away) can make. The ERGM and Logit, while identical in their non-network covariates, report substantively different effects for these covariates, and, in the ERGM, network effects are also found to drive the formation of arms transfer edges.

4. The additive and multiplicative effect network model

4.1. Latent variable network models

Another way to account for network dependencies is by making use of latent variables. Models within this class assume that latent variables Z_i are associated with each node i . Depending on the type of model, these latent variables can either be discrete (e.g. indicating group memberships for each node) or continuous, and affect the connection probability in different ways (Matias and Robin, 2014). An early (but still popular) approach in this direction is the stochastic blockmodel, which assumes that each agent

⁸ As shown in the Supplementary Material, the ERGM also outperforms the Logit model when assessing their respective areas under the receiver-operator and precision-recall curves. In line with Hunter and Handcock (2006), one could also calculate the likelihood ratio test statistic from the log likelihoods reported in Table 1 for the same purpose.

possesses a latent, categorical class (or group membership). Nodes within each class are assumed to be stochastically equivalent in their connectivity behavior, meaning that the probability of two nodes to connect depends solely on their group memberships (Holland et al., 1983; De Nicola et al., 2022). This family of models is attractive due to its simplicity in detecting and describing subgroups of nodes in networks. In many applications, however, discrete groupings fail to adequately represent the observed data, as agents behave more heterogeneously. Moving from discrete to continuous latent variable network models, another prominent approach is the latent distance model. The latter postulates that agents are positioned in a latent Euclidean “social space”, and that the closer they are within it, the more likely they are to form ties (Hoff et al., 2002). More precisely, the classical latent distance model specifies the probability of observing an edge between nodes i and j , conditional on Z , through

$$\mathbb{P}_\theta(Y_{ij} = 1 | Z) = \frac{\exp\{\theta^\top x_{ij} - \|z_i - z_j\|\}}{1 + \exp\{\theta^\top x_{ij} - \|z_i - z_j\|\}}, \tag{7}$$

where $Z = (z_1, \dots, z_n)$ denotes the latent positions of the nodes in the d -dimensional latent space, and θ is the coefficient vector for the covariates x_{ij} . The latent positions Z are assumed to originate independently from a spherical Gaussian distribution, i.e. $Z \sim N_d(0, \tau^2 I_d)$, where I_d indicates a d -dimensional identity matrix.

Latent distance models are particularly attractive for social networks in which triadic closure plays a major role, and where nodes with similar characteristics tend to form connections with each other (i.e. homophilic networks, see Rivera et al., 2010). It is also possible to add nodal random effects to the model, to control for agent-specific heterogeneity in the propensity to form edges (Krivitsky et al., 2009). The model then becomes

$$\mathbb{P}_\theta(Y_{ij} = 1 | Z, a, b) = \frac{\exp\{\theta^\top x_{ij} - \|z_i - z_j\| + a_i + b_j\}}{1 + \exp\{\theta^\top x_{ij} - \|z_i - z_j\| + a_i + b_j\}}, \tag{8}$$

where $a = (a_1, \dots, a_n)$ and $b = (b_1, \dots, b_n)$ are node-specific sender and receiver effects that account for the individual agents' propensity to form ties, with $a \sim N_n(0, \tau_a^2 I_n)$ and $b \sim N_n(0, \tau_b^2 I_n)$.

Despite its advantages and its fairly simple interpretation, a Euclidean latent space is unable to effectively approximate the behavior of networks where nodes that are similar in terms of connectivity behavior are not necessarily more likely to form ties (Hoff, 2008), such as, e.g., many networks of amorous relationships (Ghani et al., 1997; Bearman et al., 2004). More generally, the latent distance model tends to perform poorly for networks in which stochastic equivalence does not imply homophily and triadic closure, i.e., when nodes which behave similarly in terms of connectivity patterns towards the rest of the network do not necessarily have a higher probability of being connected among themselves. This is often the case in economics, where real-world networks can exhibit varying degrees and combinations of stochastic equivalence, triadic closure and homophily. Moreover, it is often *a priori* unclear which of these mechanisms are at play in a given observed network. In this context, agent-specific multiplicative random effects instead of the additive latent positions allow for simultaneously representing all these patterns (Hoff, 2005). Further developments of this innovation have led to the modern specification of the Additive and Multiplicative Effects network model (AME, Hoff, 2011), which, from a matrix representation perspective, generalizes both the stochastic blockmodel and the latent distance model (Hoff, 2021).

4.2. AME: motivation and framework

The AME approach can be motivated by considering that network data often exhibit first-, second-, and third-order dependencies. *First-order effects* capture agent-specific heterogeneity in sending (or receiving) ties within a network. For example, in the case of companies and legal disputes, first-order effects can be viewed as the propensity of each firm to initiate (or be hit by) legal disputes. *Second-order effects*, i.e., reciprocity, describe the statistical dependency of the directed relationship between two agents in the network. In the previous example, this effect can be described as the correlation between (a) company i initiating a legal dispute against company j and (b) j doing the same towards i . Of course, second-order effects can only occur in directed networks. *Third-order effects* are described as the dependency within triads, defined as the connections between three agents, and relate to the triangular statistics previously illustrated in Fig. 1. How likely is it that “a friend of a friend is also my friend”? Or, returning to the previous example: given that i has legal disputes with j and k , how likely are disputes to occur between j and k ?

The AME network model is designed to simultaneously capture these three orders of dependencies. More specifically, it extends the classical (generalized) linear modeling framework by incorporating extra terms into the systematic component to account for them. In the case of binary network data, we can make use of the Probit AME model. As is well known, the classical Probit regression model can be motivated through a latent variable representation in which y_{ij} is the binary indicator that some latent normal random variable, say $L_{ij} \sim \mathcal{N}(\theta^\top x_{ij}, \sigma^2)$, is greater than zero (Albert and Chib, 1993). But an ordinary Probit regression model assumes that L_{ij} , and thus the binary indicators (edges) y_{ij} , are independent, which is generally inappropriate for network data. In contrast, the AME Probit model specifies the probability of a tie y_{ij} from agent i to agent j , conditional on a set of latent variables W , as

$$\mathbb{P}(Y_{ij} = 1 | W) = \Phi(\theta^\top x_{ij} + e_{ij}), \tag{9}$$

where Φ is the cumulative distribution function of the standard normal distribution, $\theta^\top x_{ij}$ accommodates the inclusion of dyadic, sender, and receiver covariates, and e_{ij} can be viewed as a structured residual, containing the latent terms in W to account for the network dependencies described above. In the directed case, e_{ij} is composed as

$$e_{ij} = a_i + b_j + u_i v_j + \varepsilon_{ij}. \tag{10}$$

6. Dependence matters: Statistical models to identify the drivers of tie formation in economic networks

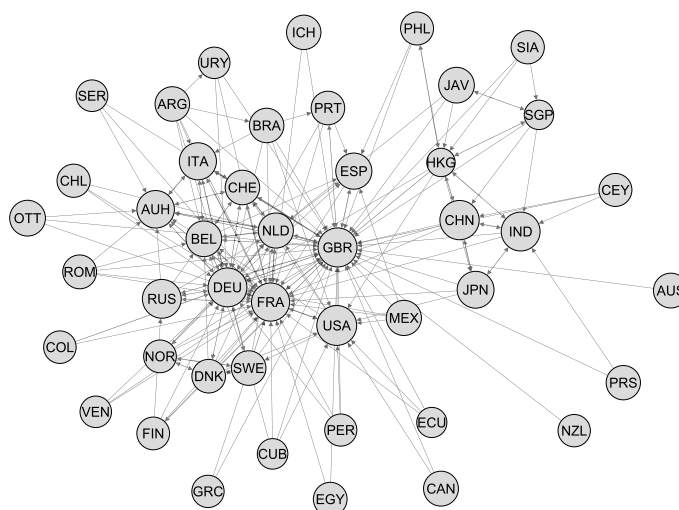


Fig. 3. Illustration of the global foreign exchange activity network in 1900. Countries are labeled by their ISO 3166-1 codes, and an edge from node i to node j indicates active trading of the currency from country j within a financial center of country i .

In this context, a_i and b_j are zero-mean additive effects for sender i and receiver j accounting for first-order dependencies, jointly specified as

$$(a_1, b_1), \dots, (a_n, b_n) \stackrel{\text{i.i.d.}}{\sim} N_2(0, \Sigma_1), \quad \text{with} \quad \Sigma_1 = \begin{pmatrix} \sigma_a & \sigma_{ab} \\ \sigma_{ab} & \sigma_b \end{pmatrix}. \quad (11)$$

The parameters σ_a and σ_b measure the variance of the additive sender and receiver effects, respectively, while σ_{ab} relates to the covariance between sender and receiver effects for the same node. Going back to (10), ε_{ij} is a zero-mean residual term which accounts for second order dependencies, i.e. reciprocity. More specifically, it holds that

$$\{(\varepsilon_{ij}, \varepsilon_{ji}) : i < j\} \stackrel{\text{i.i.d.}}{\sim} N_2(0, \Sigma_2), \quad \text{with} \quad \Sigma_2 = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad (12)$$

where σ^2 denotes the error variance and ρ determines the correlation between ε_{ij} and ε_{ji} , thus quantifying the tendency towards reciprocity. Finally, u_i and v_j in (10) are d -dimensional multiplicative sender and receiver effect vectors that account for third-order dependencies, and for which $(u_1, v_1), \dots, (u_n, v_n) \sim \mathcal{N}_{2d}(0, \Sigma_3)$ holds.

As noted above, AME is able to represent a wide variety of network structures, generalizing several other latent variable model classes. This generality comes at the price of a high level of complexity for the estimated latent structure. This can make the model class a sub-optimal choice if one wants to interpret the latent structure with respect to, e.g., clustering. On the other hand, its flexibility makes it an ideal fit when the underlying network dependencies are unknown, and the researchers' interest mainly lies in evaluating and interpreting the effect of dyadic and nodal covariates on tie formation while controlling for network effects. This strength has led to AME being used for several applications of this type (Koster, 2018; Minhas et al., 2019, 2022; Dorff et al., 2020). We next showcase the AME framework by applying it to the world foreign exchange activity network as of 1900, originally introduced and studied by Flandreau and Jobst (2005, 2009). This application highlights how using AME instead of classical regression can allow us to reconsider existing, influential answers to relevant questions via replication.

4.3. Application to the global foreign exchange activity network

In 1900, every financial center featured a foreign exchange market where bankers bought and sold foreign currency against the domestic one. Foreign exchange market activity was monitored in local bulletins, which allowed Flandreau and Jobst (2005) to collect a global dataset with all currencies used in the world at that time. In the resulting network structure, laid out in Fig. 3, countries are nodes, and a (directed) edge from country i to country j occurs if the currency of country j was actively traded in at least one financial center within country i . From the graph representation, laid out using a variant of the Yifan Hu force-directed graph drawing algorithm (Hu, 2005), we observe that the most actively traded currencies at the time belonged to large European economies, such as Great Britain, France and Germany. To determine the drivers of currency adoption, Flandreau and Jobst (2009) model this network as a function of several covariates by employing ordinary binary regression. As we show, it is possible to use AME to pursue the same goal while taking network dependencies into account.

Table 2
Estimated coefficients and related standard errors (in parentheses) for the AME model and the corresponding Probit model for the global foreign exchange activity network in 1900.

	AME	Classical Probit
Intercept	−4.845 (5.310)	−3.211 (1.580)*
Sender	Gold standard	−0.629 (0.397)
	log-GDP per-capita	−0.453 (0.419)
	Democracy index	−0.033 (0.064)
	Currency coverage	1.418 (0.405)***
Receiver	Gold standard	−0.599 (0.667)
	log-GDP per-capita	0.426 (0.703)
	Democracy index	0.121 (0.102)
	Currency coverage	2.734 (0.691)***
Dyadic	Distance	−1.019 (0.151)***
	log-trade volume	0.488 (0.081)***

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

We specify the AME model as in (9), using directed edges y_{ij} as response variable. The nodal covariates we use, sourced from and described in detail in the replication materials of Flandreau and Jobst (2009), are (log)per-capita GDP, democracy index score, coverage of foreign currencies traded in the country, and an indicator of whether the country's currency was on the gold standard. We also include, as dyadic covariates, the distance between two countries as well as their total trade volume. As specified in (10), the structured residual term e_{ij} comprises additive effects a_i and b_j for each node, which capture country-specific propensities to send and receive ties, respectively. Multiplicative effects u_i and v_j are included to account for third order dependencies. We here set the dimensionality of the multiplicative effects to two, which we assume to be sufficient given the relatively small size of the network.

To estimate the AME model, we make use of the R package `amen` (Hoff, 2015). As the likelihood involves intractable integrals arising from the combination of the transformation and dependencies induced by the model, closed form solutions are not available. The package thus uses reasonably standard Gibbs sampling algorithms to provide Bayesian inference on the model parameters. More details on the estimation routine can be found in Hoff (2021).

The results of the analysis, as well as, for comparison's sake, a Probit regression including the same covariates but ignoring network dependencies, are displayed in Table 2. Note that the classical Probit regression model can be seen as a special case of AME Probit in which both additive and multiplicative node-specific effects are omitted. Additional model diagnostics and goodness of fit measures, together with the estimated variance and covariance parameters, are provided in the Supplementary Material. The estimated coefficients (for both models) can be interpreted as in standard Probit regression: For the nodal covariate per-capita GDP, for example, a unit increase in the log-per-capita GDP for country i corresponds to a decrease of 0.453 in the linear predictor, therefore negatively influencing the expected probability of the country to *send* a tie. The same unit increase in the log-per-capita-gdp for country i corresponds to an increase of 0.426 in the linear predictor, and has therefore a positive impact on the expected probability of that country to *receive* a tie. In the case of a dyadic covariate, such as distance, a unit increase in distance between two countries leads to a decrease of 1.019 in the linear predictor, resulting in a decrease in the expected probability of the two countries to form a tie in either direction. Overall, we find that the principal drivers of the formation of a tie between i and j are the magnitude of the foreign exchange coverage of the two countries involved, the distance between them, and their reciprocal trade volume. These results correspond to the thesis of Kindleberger (1967) and to Flandreau and Jobst (2009), who suggest that the most important determinants of international adoption for a currency are size and convenience of use. At the same time, we note that, as for the ERGM in the arms trade example, the results of the Probit and AME model differ in several regards. In particular, several effects are statistically significant in the Probit but not significant in the AME model. Indeed, unacknowledged network dependence can cause downward bias in the estimation of standard errors, leading to spurious associations (Lee and Ogburn, 2021). This finding once again highlights how accounting for network dependencies can make a difference when it comes to the substantive results.

As a final note, we add that in this case we went with AME over ERGM as our interest lies in answering the research questions addressed by Flandreau and Jobst (2005, 2009), that is assessing the effect of the exogenous covariates in Table 2 on tie formation. AME allows us to do that without specifying the configuration of the endogenous network mechanisms at play, which are instead accounted for through the imposed latent structure. If, on the other hand, the researcher expects some specific network effects to play a role, and wishes to test for their presence and measure their influence on network formation, the ERGM may be a better tool. The latter model class can, for example, directly answer questions such as “Does the fact that both countries A and B trade the currency of country C influence the probability of A and B to be connected? And if so, to what extent?”. AME, on the other hand, is limited to accounting for those effects via the latent variables, without explicitly identifying them, to provide unbiased inference for the covariate effects. The choice between the two model classes is thus a matter of what assumptions can be made about the network and where the researcher's interest lies.

6. Dependence matters: Statistical models to identify the drivers of tie formation in economic networks

G. De Nicola, C. Fritz, M. Mehrl et al.

Journal of Economic Behavior and Organization 215 (2023) 351–363

5. Conclusion

Complex dependencies are ubiquitous in the economic sciences (Chiarella et al., 2005; Flaschel et al., 2008), and many economic interactions can be naturally perceived as networks. This area of research has thus received considerable interest in recent years. However, this attention has not yet been accompanied by a corresponding general take-up of empirical research methods tailored towards networks. Instead, researchers either develop their own estimators to reproduce the features of their theoretical network models, or use standard regression methods that assume conditional independence of the edges in the network. Against this background, this paper seeks to provide a hands-on introduction to two statistical models which account for network dependencies, namely the Exponential Random Graph Model (ERGM) and the Additive and Multiplicative Effects network model (AME). These two classes serve different purposes: While the ERGM is most appropriate when explicitly interested in testing the effects of endogenous network structure, the AME model allows one to control for network dependencies while substantively focusing on estimating the effects of exogenous covariates of interest. We present the statistical foundations of both models, and demonstrate their applicability to economic networks through examples in the international arms trade and foreign currency exchange, showing that modeling network dependencies can alter the substantive results of the analysis. We, moreover, provide the full data and code necessary to replicate these exemplary applications. We explicitly encourage readers to use these replication materials to get started with analyzing economic networks via ERGM and AME, beginning with the examples covered here to then transfer the code and methods to their own research.

We especially want to encourage the use of such methods as not accounting for interdependence between observations when it exists can lead to biased estimates and spurious findings. Our two applications demonstrate that this bias can result in very different empirical results, and thus affect substantive conclusions. It is therefore vital to account for network structure when studying interactions between economic agents such as individuals, firms, or countries, regardless of whether one is substantively interested in this structure. As shown by Lee and Ogburn (2021), our applications are just two examples of how unaccounted dependence in the observed data may lead to spurious findings.

At the same time, this paper can only serve as an introduction to statistical network data analysis in economics. We covered two general frameworks in this realm, but, in the interest of brevity, focused only on their simplest versions that apply to networks observed at only one time point and with binary edges. However, both frameworks have been extended to cover more general settings. For the ERGM, there are extensions for longitudinal data (Hanneke et al., 2010), distinguishing between edge formation and continuation (Krivitsky and Handcock, 2014), as well as to settings where edges are not binary but instead count-valued or signed (Krivitsky, 2012; Fritz et al., 2022). As for AME, approaches for longitudinal networks are described by Minhas et al. (2016), while versions for undirected networks as well as for non-binary network data are presented by Hoff (2021). Both the ERGM and the AME frameworks are thus flexible enough to cover a wide array of potential economic interactions. We believe that increasingly adopting these methods will, in turn, aid our understanding of these interactions.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The full data and code used for the analysis is publicly available on our GitHub repository, appropriately referenced in the paper and in the Supplementary Material.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.jebo.2023.09.021>.

References

- Albert, J.H., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.* 88 (422), 669–679.
- Anderson, J.E., Van Wincoop, E., 2003. Gravity with gravitas: a solution to the border puzzle. *Am. Econ. Rev.* 93 (1), 170–192.
- Atalay, E., Hortacsu, A., Roberts, J., Syverson, C., 2011. Network structure of production. *Proc. Natl. Acad. Sci.* 108 (13), 5199–5202.
- Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: an open source software for exploring and manipulating networks. In: *Third International AAAI Conference on Weblogs and Social Media*, pp. 361–362.
- Bearman, P.S., Moody, J., Stovel, K., 2004. Chains of affection: the structure of adolescent romantic and sexual networks. *Am. J. Sociol.* 110 (1), 44–91.
- Boucher, V., Mourifié, I., 2017. My friend far, far away: a random field approach to exponential random graph models. *Econom. J.* 20 (3), S14–S46.
- Bramoullé, Y., Galeotti, A., Rogers, B.W., 2016. *The Oxford Handbook of the Economics of Networks*. Oxford University Press.
- Campbell, B., Cranmer, S., Desmarais, B., 2018. Triangulating war: network structure and the democratic peace. arXiv:1809.04141.
- Cao, X., Ward, M.D., 2014. Do democracies attract portfolio investment? Transnational portfolio investments modeled as dynamic network. *Int. Interact.* 40 (2), 216–245.
- Chaney, T., 2014. The network structure of international trade. *Am. Econ. Rev.* 104 (11), 3600–3634.
- Chiarella, C., Flaschel, P., Franke, R., 2005. *Foundations for a Disequilibrium Theory of the Business Cycle: Qualitative Analysis and Quantitative Assessment*. Cambridge University Press.
- Cranmer, S.J., Desmarais, B.A., Kirkland, J.H., 2012. Toward a network theory of alliance formation. *Int. Interact.* 38 (3), 295–324.

- Cranmer, S.J., Leifeld, P., McClurg, S.D., Rolfe, M., 2017. Navigating the range of statistical tools for inferential network analysis. *Am. J. Polit. Sci.* 61 (1), 237–251.
- Davis, J.A., 1970. Clustering and hierarchy in interpersonal relations: testing two graph theoretical models on 742 sociomatrixes. *Am. Sociol. Rev.* 35 (5), 843–851.
- De Nicola, G., Sischka, B., Kauermann, G., 2022. Mixture models and networks: the stochastic blockmodel. *Stat. Model.* 22 (1–2), 67–94.
- De Paula, Á., 2020. Econometric models of network formation. *Annu. Rev. Econ.* 12, 775–799.
- Dorff, C., Gallop, M., Minhas, S., 2020. Networks of violence: predicting conflict in Nigeria. *J. Polit.* 82 (2), 476–493.
- Dueñas, M., Fagiolo, G., 2013. Modeling the international-trade network: a gravity approach. *J. Econ. Interact. Coord.* 8 (1), 155–178.
- Erdős, P., Rényi, A., 1959. On random graphs I. *Publ. Math. (Debr.)* 6, 290.
- Fagiolo, G., 2010. The international-trade network: gravity equations and topological properties. *J. Econ. Interact. Coord.* 5 (1), 1–25.
- Feng, L., Xu, H., Wu, G., Zhang, W., 2021. Service trade network structure and its determinants in the belt and road based on the temporal exponential random graph model. *Pac. Econ. Rev.* 26 (5), 617–650.
- Flandreau, M., Jobst, C., 2005. The ties that divide: a network analysis of the international monetary system, 1890–1910. *J. Econ. Hist.* 65 (4), 977–1007.
- Flandreau, M., Jobst, C., 2009. The empirics of international currencies: network externalities, history and persistence. *Econ. J.* 119 (537), 643–664.
- Flaschel, P., Charpe, M., Galanis, G., Proaño, C.R., Veneziani, R., 2018. Macroeconomic and stock market interactions with endogenous aggregate sentiment dynamics. *J. Econ. Dyn. Control* 91, 237–256.
- Flaschel, P., Franke, R., Semmler, W., 1997. *Dynamic Macroeconomics: Instability, Fluctuation, and Growth in Monetary Economies*. MIT Press.
- Flaschel, P., Groh, G., Proaño, C., Semmler, W., 2008. *Topics in Applied Macrodynamics Theory*, vol. 10. Springer Science & Business Media.
- Flaschel, P., Kauermann, G., Semmler, W., 2007. Testing wage and price Phillips curves for the United States. *Metroeconomica* 58 (4), 550–581.
- Frank, O., Strauss, D., 1986. Markov graphs. *J. Am. Stat. Assoc.* 81 (395), 832–842.
- Fritz, C., De Nicola, G., Kevork, S., Harhoff, D., Kauermann, G., 2023. Modelling the large and dynamically growing bipartite network of German patents and inventors. *J. R. Stat. Soc., Ser. A, Stat. Soc.* 186 (3), 557–576.
- Fritz, C., Lebacher, M., Kauermann, G., 2020. Tempus volat, hora fugit: a survey of tie-oriented dynamic network models in discrete and continuous time. *Stat. Neerl.* 74 (3), 275–299.
- Fritz, C., Mehrl, M., Thurner, P.W., Kauermann, G., 2022. Exponential random graph models for dynamic signed networks: an application to international relations. [arXiv:2205.13411](https://arxiv.org/abs/2205.13411).
- Fritz, C., Thurner, P.W., Kauermann, G., 2021. Separable and semiparametric network-based counting processes applied to the international combat aircraft trades. *Netw. Sci.* 9 (3), 291–311.
- Geyer, C.J., Thompson, E.A., 1992. Constrained Monte Carlo maximum likelihood for dependent data. *J. R. Stat. Soc., Ser. B, Methodol.* 54 (3), 657–683.
- Ghani, A.C., Swinton, J., Garnett, G.P., 1997. The role of sexual partnership networks in the epidemiology of gonorrhoea. *Sex. Transm. Dis.* 24 (1), 45–56.
- Gilbert, E.N., 1959. Random graphs. *Ann. Math. Stat.* 30 (4), 1141–1144.
- Goldenberg, A., Zheng, A.X., Fienberg, S.E., Airolidi, E.M., 2010. A survey of statistical network models. *Found. Trends Mach. Learn.* 2 (2), 129–233.
- Graham, B.S., 2015. Methods of identification in social networks. *Annu. Rev. Econ.* 7 (1), 465–485.
- Handcock, M., 2003. Assessing degeneracy in statistical models of social networks. Working Paper no. 39. University of Washington.
- Handcock, M.S., Hunter, D.R., Butts, C.T., Goodreau, S.M., Krivitsky, P.N., Morris, M., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M., Hunter, D.R., Butts, C.T., Goodreau, S.M., Krivitsky, P.N., Morris, M., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M., Martina, 2008a. *ergm: fit, simulate and diagnose exponential-family models for networks*. *J. Stat. Softw.* 24 (3), nihpa54860.
- Handcock, M.S., Hunter, D.R., Butts, C.T., Goodreau, S.M., Morris, M., 2008b. *statnet: software tools for the representation, visualization, analysis and simulation of network data*. *J. Stat. Softw.* 24 (1), 1548.
- Hanneke, S., Fu, W., Xing, E.P., 2010. Discrete temporal models of social networks. *Electron. J. Stat.* 4, 585–605.
- He, X., Dong, Y., Wu, Y., Jiang, G., Zheng, Y., 2019. Factors affecting evolution of the interprovincial technology patent trade networks in China based on exponential random graph models. *Phys. A, Stat. Mech. Appl.* 514, 443–457.
- Heider, F., 1946. Attitudes and cognitive organization. *J. Psychol.* 21 (1), 107–112.
- Herman, P.R., 2022. Modeling complex network patterns in international trade. *Rev. World Econ.* 158, 127–179.
- Hoff, P., 2008. Modeling homophily and stochastic equivalence in symmetric relational data. *Adv. Neural Inf. Process. Syst.* 20, 657–664.
- Hoff, P., 2021. Additive and multiplicative effects network models. *Stat. Sci.* 36 (1), 34–50.
- Hoff, P.D., 2005. Bilinear mixed-effects models for dyadic data. *J. Am. Stat. Assoc.* 100 (469), 286–295.
- Hoff, P.D., 2011. Hierarchical multilinear models for multiway data. *Comput. Stat. Data Anal.* 55 (1), 530–543.
- Hoff, P.D., 2015. Dyadic data analysis with *amen*. preprint. [arXiv:1506.08237](https://arxiv.org/abs/1506.08237).
- Hoff, P.D., Raftery, A.E., Handcock, M.S., 2002. Latent space approaches to social network analysis. *J. Am. Stat. Assoc.* 97 (460), 1090–1098.
- Holland, P.W., Laskey, K.B., Leinhardt, S., 1983. Stochastic blockmodels: first steps. *Soc. Netw.* 5 (2), 109–137.
- Holland, P.W., Leinhardt, S., 1981. An exponential family of probability distributions for directed graphs. *J. Am. Stat. Assoc.* 76 (373), 33–50.
- Hu, Y., 2005. Efficient, high-quality force-directed graph drawing. *Math. J.* 10 (1), 37–71.
- Hummel, R.M., Hunter, D.R., Handcock, M.S., 2012. Improving simulation-based algorithms for fitting ERGMs. *J. Comput. Graph. Stat.* 21 (4), 920–939.
- Hunter, D.R., Goodreau, S.M., Handcock, M.S., 2008. Goodness of fit of social network models. *J. Am. Stat. Assoc.* 103 (481), 248–258.
- Hunter, D.R., Handcock, M.S., 2006. Inference in curved exponential family models for networks. *J. Comput. Graph. Stat.* 15 (3), 565–583.
- Hunter, D.R., Krivitsky, P.N., Schweinberger, M., 2012. Computational statistical methods for social network models. *J. Comput. Graph. Stat.* 21 (4), 856–882.
- Jackson, M.O., 2008. *Social and Economic Networks*. Princeton University Press.
- Jackson, M.O., 2014. The past and future of network analysis in economics. In: *The Oxford Handbook of the Economics of Networks*. Oxford University Press.
- Jackson, M.O., Nei, S., 2015. Networks of military alliances, wars, and international trade. *Proc. Natl. Acad. Sci.* 112 (50), 15277–15284.
- Jackson, M.O., Rogers, B.W., 2007. Meeting strangers and friends of friends: how random are social networks? *Am. Econ. Rev.* 97 (3), 890–915.
- Jackson, M.O., Rogers, B.W., Zenou, Y., 2017. The economic consequences of social-network structure. *J. Econ. Lit.* 55 (1), 49–95.
- Kindleberger, C.P., 1967. The politics of international money and world language. International Finance Section, Department of Economics, Princeton University.
- Koster, J., 2018. Family ties: the multilevel effects of households and kinship on the networks of individuals. *R. Soc. Open Sci.* 5 (4), 172159.
- Krivitsky, P.N., 2012. Exponential-family random graph models for valued networks. *Electron. J. Stat.* 6, 1100–1128.
- Krivitsky, P.N., Handcock, M.S., 2014. A separable model for dynamic networks. *J. R. Stat. Soc., Ser. B, Stat. Methodol.* 76 (1), 29–46.
- Krivitsky, P.N., Handcock, M.S., Raftery, A.E., Hoff, P.D., 2009. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Soc. Netw.* 31 (3), 204–213.
- König, M.D., Rohner, D., Thoenig, M., Zilibotti, F., 2017. Networks in conflict: theory and evidence from the great war of Africa. *Econometrica* 85 (4), 1093–1132.
- Lebacher, M., Thurner, P.W., Kauermann, G., 2021. A dynamic separable network model with actor heterogeneity: an application to global weapons transfers. *J. R. Stat. Soc., Ser. A, Stat. Soc.* 184 (1), 201–226.
- Lee, Y., Ogburn, E.L., 2021. Network dependence can lead to spurious associations and invalid inference. *J. Am. Stat. Assoc.* 116 (535), 1060–1074.
- Lewer, J.J., Van den Berg, H., 2008. A gravity model of immigration. *Econ. Lett.* 99 (1), 164–167.
- Liu, L., Shen, M., Sun, D., Yan, X., Hu, S., 2022. Preferential attachment, R&D expenditure and the evolution of international trade networks from the perspective of complex networks. *Phys. A, Stat. Mech. Appl.* 603, 127579.
- Lusher, D., Koskinen, J., Robins, G., 2012. *Exponential Random Graph Models for Social Networks*. Cambridge University Press.

6. Dependence matters: Statistical models to identify the drivers of tie formation in economic networks

G. De Nicola, C. Fritz, M. Mehrl et al.

Journal of Economic Behavior and Organization 215 (2023) 351–363

- Matias, C., Robin, S., 2014. Modeling heterogeneity in random graphs through latent space models: a selective review. *ESAIM Proc. Surv.* 47, 55–74.
- Mele, A., 2017. A structural model of dense network formation. *Econometrica* 85 (3), 825–850.
- Miller, S.V., 2022. (peacescencer): an R package for quantitative peace science research. *Confl. Manage. Peace Sci.* 39 (6), 755–779.
- Minhas, S., Dorff, C., Gallop, M.B., Foster, M., Liu, H., Tellez, J., Ward, M.D., 2022. Taking dyads seriously. *Political Sci. Res. Methods* 10 (4), 703–721.
- Minhas, S., Hoff, P.D., Ward, M.D., 2016. A new approach to analyzing coevolving longitudinal networks in international relations. *J. Peace Res.* 53 (3), 491–505.
- Minhas, S., Hoff, P.D., Ward, M.D., 2019. Inferential approaches for network analysis: AMEN for latent factor models. *Polit. Anal.* 27 (2), 208–222.
- Morales, E., Sheu, G., Zahler, A., 2019. Extended gravity. *Rev. Econ. Stud.* 86 (6), 2668–2712.
- Morris, M., Handcock, M.S., Hunter, D.R., 2008. Specification of exponential-family random graph models: terms and computational aspects. *J. Stat. Softw.* 24 (4), 1548.
- Mundt, P., 2021. The formation of input–output architecture: evidence from the European Union. *J. Econ. Behav. Organ.* 183, 89–104.
- Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. *J. R. Stat. Soc. A, General* 135 (3), 370–384.
- Newcomb, T.M., 1979. Reciprocity of interpersonal attraction: a nonconfirmation of a plausible hypothesis. *Soc. Psychol. Q.* 42 (4), 299–306.
- Newman, M.E.J., 2003. The structure and function of complex networks. *SIAM Rev.* 45 (2), 167–256.
- R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing.
- Rivera, M.T., Soderstrom, S.B., Uzzi, B., 2010. Dynamics of dyads in social networks: assortative, relational, and proximity mechanisms. *Annu. Rev. Sociol.* 36, 91–115.
- Robins, G., Pattison, P., Kalish, Y., Lusher, D., 2007a. An introduction to exponential random graph (p^*) models for social networks. *Soc. Netw.* 29 (2), 173–191.
- Robins, G., Snijders, T., Wang, P., Handcock, M., Pattison, P., 2007b. Recent developments in exponential random graph (p^*) models for social networks. *Soc. Netw.* 29 (2), 192–215.
- Rose, A.K., 2004. Do we really know that the WTO increases trade? *Am. Econ. Rev.* 94 (1), 98–114.
- Schoeneman, J., Zhu, B., Desmarais, B.A., 2022. Complex dependence in foreign direct investment: network theory and empirical analysis. *Political Sci. Res. Methods* 10 (2), 243–259.
- Schweinberger, M., 2011. Instability, sensitivity, and degeneracy of discrete exponential families. *J. Am. Stat. Assoc.* 106 (496), 1361–1370.
- SIPRI, 2021. SIPRI arms transfers database. Stockholm international peace research institute, <https://www.sipri.org/databases/armstransfers>. (Accessed 20 October 2022).
- Smith, M., Gorgoni, S., Cronin, B., 2019. International production and trade in a high-tech industry: a multilevel network analysis. *Soc. Netw.* 59, 50–60.
- Smith, M., Sarabi, Y., 2022. How does the behaviour of the core differ from the periphery? – An international trade network analysis. *Soc. Netw.* 70, 1–15.
- Snijders, T.A., 1996. Stochastic actor-oriented models for network change. *J. Math. Sociol.* 21 (1–2), 149–172.
- Snijders, T.A., 2017. Stochastic actor-oriented models for network dynamics. *Annu. Rev. Stat. Appl.* 4, 343–363.
- Snijders, T.A.B., Pattison, P.E., Robins, G.L., Handcock, M.S., 2006. New specifications for exponential random graph models. *Sociol. Method.* 36 (1), 99–153.
- Thurner, P.W., Schmid, C.S., Cranmer, S.J., Kauermann, G., 2019. Network interdependencies and the evolution of the international arms trade. *J. Conf. Resolut.* 63 (7), 1736–1764.
- van der Pol, J., 2019. Introduction to network modeling using exponential random graph models (ergm): theory and an application using R-project. *Comput. Econ.* 54 (3), 845–875.
- Ward, M.D., Ahlquist, J.S., Rozenas, A., 2013. Gravity’s rainbow: a dynamic latent space model for the world trade network. *Netw. Sci.* 1 (1), 95–118.
- Ward, M.D., Hoff, P.D., 2007. Persistent patterns of international commerce. *J. Peace Res.* 44 (2), 157–175.
- Wasserman, S., Faust, K., 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.
- Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^* . *Psychometrika* 61 (3), 401–425.

7. Modelling the large and dynamically growing bipartite network of German patents and inventors

Contributing article

Fritz, C., De Nicola, G., Kevork, S., Harhoff, D., and Kauermann, G. (2023). Modelling the large and dynamically growing bipartite network of German patents and inventors. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(3):557–576. <https://doi.org/10.1093/jrsssa/qnad009>.

Data and code

Available at <https://github.com/corneliusfritz/Modelling-German-patents-and-inventors>.

Copyright information

© (RSS) Royal Statistical Society 2023. The full article is included, as a license to reuse it in this dissertation for non-commercial purposes has been obtained by the author.

Author contributions


The idea of modeling patent data as a bipartite network that grows over time can be attributed to Göran Kauermann. Cornelius Fritz and Giacomo De Nicola subsequently specified the model, and Cornelius Fritz implemented it. The manuscript was then mainly designed and drafted by Cornelius Fritz and Giacomo De Nicola, with contributions from the other authors. More specifically, the introduction was written jointly by Cornelius Fritz, Giacomo De Nicola, and Göran Kauermann. Section 2, on the other hand, was composed by Giacomo De Nicola, Cornelius Fritz, and Dietmar Harhoff. Section 3 was mainly written by Cornelius Fritz, while Giacomo De Nicola wrote Section 4, with contributions from Cornelius Fritz and Dietmar Harhoff. All authors contributed through fruitful comments and extensive proofreading of the manuscript.

Journal of the Royal Statistical Society Series A: Statistics in Society, 2023, **186**, 557–576
<https://doi.org/10.1093/jrsssa/qnad009>
Advance access publication 8 March 2023

Original Article



Modelling the large and dynamically growing bipartite network of German patents and inventors

Cornelius Fritz¹ , Giacomo De Nicola¹, Sevag Kevork¹, Dietmar Harhoff² and Göran Kauermann¹

¹Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstr. 33, Munich, 80539 Bavaria, Germany

²Max Planck Institute for Innovation and Competition, Marstallpl. 1, Munich, 80539 Bavaria, Germany

Address for correspondence: Cornelius Fritz, Department of Statistics, Ludwig-Maximilians-Universität München, Ludwigstr. 33, Munich, 80539 Bavaria, Germany. Email: cornelius.fritz@stat.uni-muenchen.de

Abstract

To explore the driving forces behind innovation, we analyse the dynamic bipartite network of all inventors and patents registered within the field of electrical engineering in Germany in the past two decades. To deal with the sheer size of the data, we decompose the network by exploiting the fact that most inventors tend to only stay active for a relatively short period. We thus propose a Temporal Exponential Random Graph Model with time-varying actor set and sufficient statistics mirroring substantial expectations for our analysis. Our results corroborate that inventor characteristics and team formation are essential to the dynamics of invention.

Keywords: bipartite networks, co-inventorship networks, inventors, knowledge flows, patent collaboration, temporal exponential random graph models

1 Introduction

In the social sciences, bipartite networks are often used to represent and study affiliation of the actors to some groups (such as directors on boards (Friel et al., 2016), or football players in teams (Onody & de Castro, 2004)) and participation of people to events (such as researchers citing papers (Small, 1973), or actors in movies (Ahmed et al., 2007)). Research on bipartite structures initially focused on unimodal projections of the networks (Breiger, 1974), where we consider two nodes of one type to be tied if they share at least one actor of the other kind. This practice forces the researcher to give priority to one type of node over another and thus comes with a loss of possibly relevant information (Koskinen & Edling, 2012). Direct bipartite network analysis has first been considered in Borgatti and Everett (1997), where traditional network analysis techniques are systematically discussed for bipartite networks. Latapy et al. (2008) further adjusted known concepts from unipartite networks, such as clustering and redundancy, to the bipartite case, with a focus on large networks.

For this paper, we consider high-dimensional bipartite networks where actors are related to one another through instantaneous events, which by definition only occur once. In particular, we focus on the network formed by inventors residing in Germany and patents submitted between 1995 and 2015, where a tie between an inventor and a patent is present if the individual is listed among the patent's inventors. The resulting data structure is visualised in Figure 1a, where we can assign each patent (or event, in the jargon of bipartite network analysis) to a time point and a set of co-inventors. For instance, inventors A and B filed the joint patent with ID 1. We may represent the bipartite network structure as an adjacency matrix with entries Y_{ij} , where

Received: January 21, 2022. Revised: January 4, 2023. Accepted: January 19, 2023

© (RSS) Royal Statistical Society 2023. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

$$Y_{ij} = \begin{cases} 1 & \text{if actor } i \text{ is on patent ID } j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

and $i \in \mathcal{I}$ and $j \in \mathcal{K}$, where we denote the complete set of inventors and patents by \mathcal{I} and \mathcal{K} , respectively. In our example, this bipartite network is of massive dimensions, with $|\mathcal{I}| = 78.412$ inventors on a total of $|\mathcal{K}| = 126.388$ filed patents.

The data allow us to gain insight into the dynamics and drivers of innovation, collaboration, and knowledge flows in the private sector. Moreover, inventorship status on a patent is more legally binding than authorship of academic papers, suggesting a greater degree of validity of the results of network analysis in this context. The data, however, present some obstacles to their study. First, the complete network is too massive, making analysis with most traditional network techniques prohibitive. Second, the data carry structural zero entries since not all inventors are active during the entire time period between 1995 and 2015. This phenomenon is partially due to the retirement of inventors, who hence suffer from natural ‘actor mortality’. Moreover, inventors may change their career track, e.g., by moving into managerial positions and ending their patenting activities, thus reinforcing the aforementioned actor mortality in our data. Vice versa, new inventors continuously enter the picture by producing their first patent, resulting in what we can call ‘actor natality’ in the network. These aspects imply that the bipartite network matrix at hand contains structural zeros for inventors which are not active at particular time points. To incorporate this feature into a statistical network model, we consider the network dynamically and discretise the time dimension by looking at yearly data, such that time takes values $t = 1, 2, \dots, T$, as sketched in Figure 1a. In this context, T denotes the number of observed time points. We then allow the actor set to change at each time point. For the adjacency matrix of Figure 1a, this leads to the matrix structure in Figure 1b, where e.g., inventor A retires after time point $t = 1$ and hence does not take part in the patent market at $t = 2$. To encode this information on the changing composition of actors, we define activity sets \mathcal{I}_t to include all actors that are active at time point t . Further, let \mathcal{K}_t denote the event set, containing all patents submitted in a particular time window. We assume that both sets are known for each time point $t = 1, \dots, T$. With this additional information, we decompose the observed massive bipartite network matrix into smaller dimensional bipartite submatrices denoted by

$$\mathbf{Y}_t = (Y_{t,ij} : i \in \mathcal{I}_t, j \in \mathcal{K}_t), \quad (2)$$

which are visualised for $t = 1$ and 2 by the grey-shaded areas in Figure 1b and where $Y_{t,ik}$ indicates whether inventor i is a co-owner of patent k at time point t . Instead of modelling the entire bipartite network, we break down our analysis to modelling \mathbf{Y}_t given the previous bipartite networks $\mathbf{Y}_1, \dots, \mathbf{Y}_{t-1}$. Incorporating the varying actor set as such in the analysis allows us to structurally account for the observed actor mortality and natality while also making the estimation problem more manageable, thus solving both issues simultaneously.

This change in perspective induces a structure that deviates from conventionally analysed networks. To accommodate for it in a probabilistic modelling framework, we extend the Temporal Exponential Random Graph Model (TERGM, Hanneke et al., 2010) towards dynamic bipartite networks with varying actor set. For TERGMs, we assume that a discrete Markov chain describes the generating process of the networks observed over time. The transition probabilities of jumping from one network to another one are determined by an Exponential Random Graph Model (ERGM, S. Wasserman & Pattison, 1996). ERGMs, on the other hand, were adapted to bipartite data by Faust and Skvoretz (1999), while adjustments to incorporate the model specifications of Snijders et al. (2006) were proposed in Wang, Robins, et al. (2013). These network models were already successfully applied to static (Metz et al., 2019) as well as dynamic networks (Broekel & Bednarz, 2018).

In addition to the dynamically varying actor set, the network at hand presents another particular feature for which we need to account in the modelling. Collaborations generally build up over time, rather than being confined to single time points. To adequately represent these mechanisms, we need to include covariate information from the past and on the pairwise level of one actor set in the model, which has not yet been implemented in the bipartite ERGM framework. We, therefore,

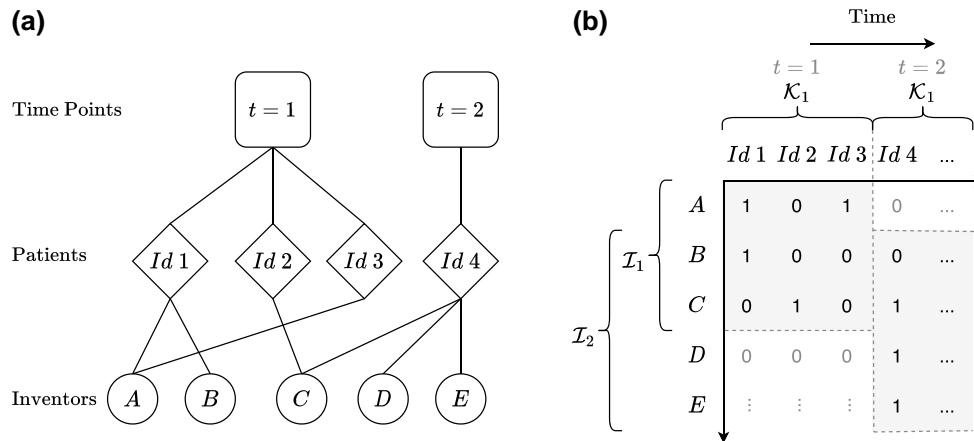


Figure 1. On the left side (a), the tripartite network structure of the patent data is illustrated with an example encompassing four patents (tilted squares) submitted at two different time points (squares with rounded edges) by five inventors (circles). The corresponding adjacency matrix is depicted on the right side (b). The two sub-matrices defined in (2) are shaded in grey are \mathcal{Y}_1 in the top left and \mathcal{Y}_2 on the bottom right. (a) Network structure and (b) adjacency matrix structure.

define and include sufficient network statistics in our model to account for this particular kind of dynamic interdependence.

Overall, the contributions of this paper are the following: We demonstrate how massive bipartite networks can be broken down in a way that allows their analysis, and propose sufficient statistics for modelling bipartite temporal networks with varying actor set. Using the proposed methodology, we then contribute to the growing literature on innovation by analysing a comprehensive patent and inventor population dataset. In particular, we study the network composed of all electrical engineering patents filed between 1995 and 2015 by inventors located in Germany. The unusually rich dataset allows us to study patterns of team formation in a more refined way than has been feasible to date. In particular, our modelling approach enables us to quantify how factors such as spatial proximity, teamwork, interlocking of collaborations, gender, and seniority affect the output of inventors. By answering these questions, our study contributes empirical findings to current discussions on the role of gender and seniority in innovation and, more generally, in the workplace.

The remainder of the paper is organised as follows: Section 2 gives a literature overview of the research in patent data. In this section, we also describe the data in detail. Section 3 motivates the model and introduces its novelties in more detail from a theoretical perspective. We present the results of our empirical analysis in Section 4, while Section 5 wraps up the paper with some concluding remarks.

2 Patent data

2.1 Research on patents and inventor teams

The analysis of patents and their impact and evolution over time is an important area of current economic research. Hall and Harhoff (2012) provide a general overview of the field and its recent developments. The holder of a patent receives a temporary right (typically for 20 years) to exclude others from using the patented technology. The patent right can be extremely valuable, e.g., when it becomes the foundation of an economic monopoly. Hence, patents can create powerful incentives and induce invention and innovation efforts. In addition, patents require disclosure of the patented invention and thus may invite others to build on the patented technology. These benefits have to be held against the welfare losses due to reduced competition. The study of patents in much of classic economic literature revolves around the trade-offs between these effects. Patent data are also often used in innovation research to explore how new technologies develop and spread, which innovation areas are the most active, how innovation areas and sub-areas are connected with one another, and how productive firms or nations are with regards to their patenting output. Patents

contain references to prior patents, so-called patent citations (Alstott et al., 2017). The study of patent citations and the network structures they form have become an important part of innovation economics, since citations can be interpreted as an indicator of knowledge flows. The study of citation networks has been an important area of research at least since the work of Garfield (1955) (see also de Solla Price, 1965; Egghe & Rousseau, 1990). Co-authorship networks have been extensively studied within the area of research publications (see, e.g., Leifeld, 2018; Melin & Persson, 1996; Newman, 2004). The techniques developed for general citation networks can naturally be applied to map patent citation networks as well (see, e.g., Li et al., 2007; Verspagen, 2012; von Wartburg et al., 2005). Moreover, since patent data always indicate the identity of the inventors contributing to the invention, they can be used to study the characteristics of inventor teams and inventor collaboration networks. The focus is then shifted from citations to co-inventorship of patents.

In both cases, i.e., patent citations and inventor teams, modern methods of network analysis can be applied to answer open research questions. In terms of the research questions tackled, our study differs substantially from patent citation studies, since we do not focus on knowledge flows, but rather on the logic of inventor team formation. We share this focus with studies of authorship teams in academia, but we note an important institutional difference: More than 93% of patents are filed by private enterprises (Giuri et al., 2007). Other than in scientific co-authoring, the composition of co-inventor teams does not just reflect the preferences of the authors (inventors), but it involves, in almost all cases, a managerial decision that is guided by profit concerns. Thus, the patterns we uncover in our analysis are not just a reflection of individual preferences, but also of the employer's productivity calculus. This feature of our setting will be particularly important when interpreting results and comparing them to results from other studies (e.g., for gender homophily).

For patent data, it is possible to construct the co-inventorship network in two main ways. One can directly analyse the bipartite network formed by the patents and their inventors (see, e.g., Balconi et al., 2004). Alternatively, one projects the bipartite structure on one of the two modes, which in the context of patent data is usually that of inventors. This entails a network composed only of inventors, in which two nodes (inventors) are connected if they have at least one patent in common (Bauer et al., 2022; Ejermo & Karlsson, 2006). Much of the literature in this area utilises such projections, since models for unimodal networks are developed to a greater extent. Several studies have used unipartite ERGMs to study knowledge diffusion networks in various domains (see e.g., Jiang et al., 2013; Keegan et al., 2012). As ERGMs allow for modelling networks with different types of ties (see Chen, 2021, for an overview), it is also possible to simultaneously model inventor-patent ties in a unipartite, multilayer network context, as done by Jiang et al. (2015). As explained in the introduction, however, projecting everything on one mode inevitably results in a loss of information on the mode that is excluded.

In the case of patents and inventors, the fact that two inventors collaborated on many patents together, together with the size of these patents, brings much information which is not available in the projection, where the inventors are simply linked together. This loss of information is made apparent by the fact that there are many bipartite graphs which lead to the same projection (Latapy et al., 2008). Preserving the original bipartite structure thus enables us to gain more detailed and accurate insight on the mechanisms at play by estimating effects which would not be visible by considering the projection, as will be shown in the application section.

2.2 Data description

We consider patent applications submitted to the European Patent Office or the German Patent and Trademark Office (Deutsches Patent- und Markenamt) between 1995 and 2015. More specifically, we look at patents filed within the main area of electrical engineering, for which at least one of the inventors listed on the patent has a residential address in Germany. For assigning each patent to a single time point, we use the priority date, i.e., the first-time filing date of a patent (which precedes the publication and the grant date). We focus on electrical engineering as it is one of the largest main areas and as it has seen particularly high growth rates since 2010. Moreover, collaborations between inventors are commonplace in this field. For our analyses, we focus on the data starting in 2000 and condition on the information from the first five years considered (i.e., from 1995 to 1999) to derive covariates from them. The dataset can be represented as a massive

7. Modelling the large and dynamically growing bipartite network of German patents and inventors

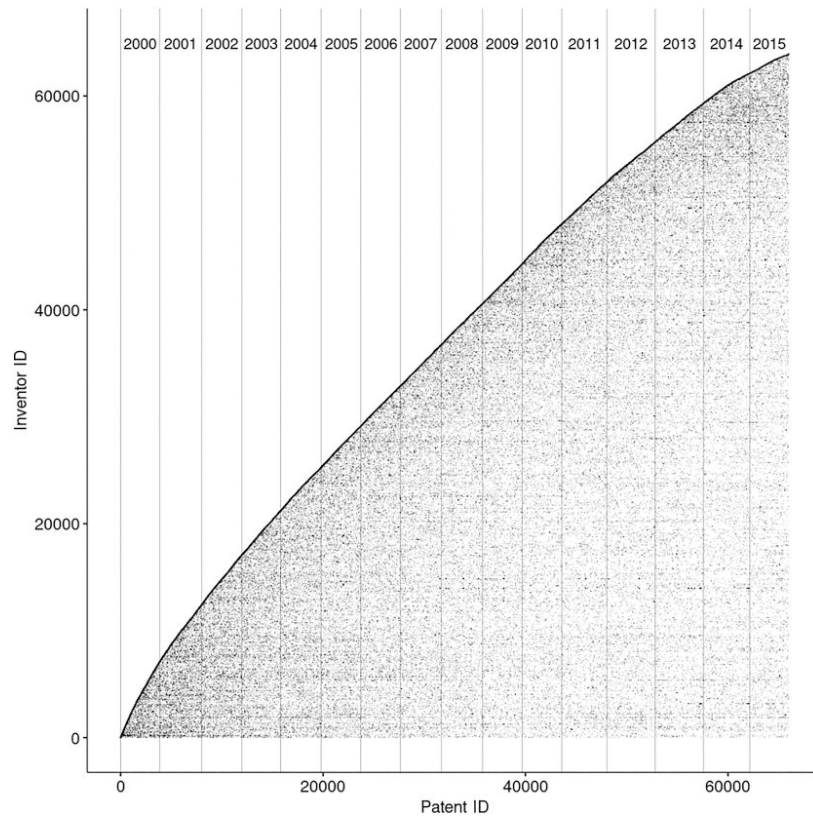


Figure 2. Graphical representation of the adjacency matrix of the patent–inventor network between 2000 and 2015. A dot in position (i, k) indicates that inventor i is a co-owner of patent k .

bipartite network, for which the observed adjacency matrix (1) is visualised in Figure 2. From the plot, we can get a clear sense of the previously described actor natality phenomenon, with new inventors becoming active at every time point. Moreover, the figure demonstrates the limits of descriptive analysis when dealing with such large networks, highlighting the need for adequate models to learn something from such data.

As described in Section 1, we instead consider this a dynamic bipartite network, discretising the time steps yearly such that time takes values $t = 1, 2, \dots, T$. In our notation, $t = 1$ translates to the year 2000. We also allow the actor set to change at each time point so that we end up with T bipartite networks in which the nodes are given by the active inventors at each time point. Resulting from this, we include new inventors that are active for the first time and remove inactive ones from the network at each time point t . The latter point is motivated by the empirical data, which suggests that if previously active inventors do not produce any patents for a long time, it is likely that they will not be active anymore. This phenomenon can stem from a change in career paths (moved up to a management position where writing patents is not among the work tasks) or retirement. To this point, we show the Kaplan–Meier estimate of the time passing between two consecutive patents by the same inventor in Figure 3. As indicated by the dashed grey lines, about 85% of patents by a specific inventor that already had at least one patent are submitted within two years from the previous one. Given this, we define an inventor as active at time t if they had at least one patent in the two years prior to t . Note that by doing so we do not disregard the remaining 15% of the data, but simply label these inventors as inactive for a specific period, i.e., until they appear on another patent.

As we are interested in investigating the drivers of patented innovation and inventor collaboration, we exclude patents developed by a single inventor from the modelled patent set. Moreover, we exclude inventors with no address in Germany from the actor set, as they make up less than 1%

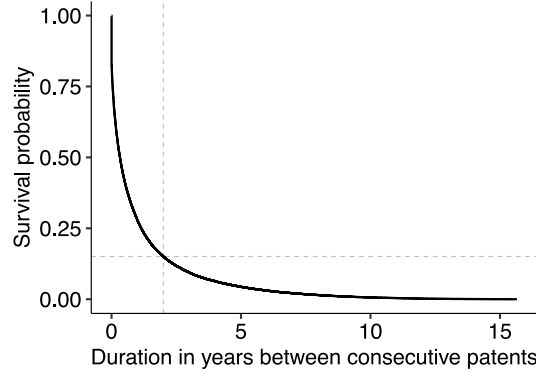


Figure 3. Kaplan–Meier estimate of the duration between consecutive patents submitted between 2000 and 2015.

of the population. In addition to the residence address of each inventor and the date of each patent, we also incorporate information on the gender of each inventor in our model. This set of exogenous covariates aligns with previous work on co-citation networks (Leifeld, 2018).

3 Modelling patent data as bipartite networks

3.1 Temporal exponential random graph models for bipartite networks

Having laid out the available data, we now formulate a generative network model for the bipartite networks at hand. This framework should allow us to differentiate between random and structural characteristics of the network to support or disregard our substantive expectations, such as, for example, whether or not two inventors that teamed up in the past are likely to produce another patent together in the future. To do so we first need to introduce some additional notation. As a general rule, we write \mathbf{Y}_t to denote the network when viewed as a random variable, and $\mathbf{y}_t = (y_{t,ik} : i \in \mathcal{I}_t, k \in \mathcal{K}_t)$ if we relate to the observed counterpart. In this context, $y_{t,ik} = 1$ translates to inventor i being a co-owner of patent k , while $y_{t,ik} = 0$ indicates the contrary. As a result, the observed networks are binary and undirected, i.e., $\mathbf{y}_t \in \{0, 1\}^{|\mathcal{I}_t| \times |\mathcal{K}_t|}$. We denote the space of all networks that could potentially be observed at time point t by \mathcal{Y}_t . For our application, as explained in the previous section, the latter is restricted to only allow for patents which have at least two inventors.

We specify the joint probability for the set of networks through

$$\mathbb{P}_\theta(\mathbf{Y}_1, \dots, \mathbf{Y}_T) = \prod_{t=1}^T \mathbb{P}_\theta(\mathbf{Y}_t | \mathcal{H}_t), \quad (3)$$

where \mathcal{H}_t defines the history, composed of the bipartite networks and covariates observed before t . The covariates can encompass dyadic and nodal information, but to make the notation less cumbersome we suppress the explicit inclusion of the covariates in the formulae. Following Hanneke et al. (2010), we simplify (3) by assuming that the temporal dependencies are constrained to a fixed time lag, i.e.,

$$\mathbb{P}_\theta(\mathbf{Y}_t | \mathcal{H}_t) = \mathbb{P}_\theta(\mathbf{Y}_t | \mathbf{Y}_{t-1} = \mathbf{y}_{t-1}, \dots, \mathbf{Y}_{t-s} = \mathbf{y}_{t-s}), \quad (4)$$

for $s \in \mathbb{N}$. The Markov property then allows us to postulate an ERGM for the transition probability (4) in the following form:

$$\mathbb{P}_\theta(\mathbf{Y}_t | \mathbf{Y}_{t-1} = \mathbf{y}_{t-1}, \dots, \mathbf{Y}_{t-s} = \mathbf{y}_{t-s}) = \frac{\exp\{\boldsymbol{\theta}^\top \mathbf{s}(\mathbf{y}_t, \dots, \mathbf{y}_{t-s})\}}{\kappa(\boldsymbol{\theta}, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-s})}, \quad (5)$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_q) \in \mathbb{R}^q$ is a q -dimensional vector of parameters, $\mathbf{s}: \mathcal{Y}_t \times \dots \times \mathcal{Y}_{t-s} \rightarrow \mathbb{R}^q$ is the function calculating the vector of sufficient statistics and $\kappa(\boldsymbol{\theta}, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-s}) := \sum_{\mathbf{y} \in \mathcal{Y}_t} \exp\{\boldsymbol{\theta}^\top \mathbf{s}(\mathbf{y}, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-s})\}$ is a normalising factor (see also

Cranmer et al., 2021, Chapter 6 and Leifeld et al., 2018). We obtain a canonical exponential family model with known characteristics (Barndorff-Nielsen, 1978), which come in handy when quantifying the uncertainty of the estimates of θ . Note that for the application to patent data, the coefficients governing the transition from one time point to another are not necessarily constant over time due to external shocks, such as, for example, the dot-com bubble and the 2008 financial crisis, which may affect the activity of inventors. For this reason, we let θ in (5) flexibly depend on time and estimate it separately for each time point t , but omit the subscript t from the formulae for notational simplicity. Thurner et al. (2018) and Cranmer et al. (2014) also opted for this parametrisation of dynamic coefficients, while smooth functions over time are employed in Lebacher et al. (2021).

Interpreting the coefficients θ can be done both at the global network level as well on the single tie level. We illustrate the interpretation for θ_p , defined as the coefficient corresponding to the p th sufficient statistic from (5). For the former, $\theta_p > 0$ implies that networks with higher values of the corresponding sufficient statistic become increasingly more likely, while $\theta_p < 0$ implies the converse. For the latter, we define so-called change statistics, which are the change in the sufficient statistics caused by switching the entry $y_{t,ik}$ from 0 to 1. Formally,

$$\Delta_{t,ik}(y_t, \dots, y_{t-s}) = s(y_{t,ik}^+, y_{t-1}, \dots, y_{t-s}) - s(y_{t,ik}^-, y_{t-1}, \dots, y_{t-s}), \tag{6}$$

where $y_{t,ik}^+$ is the network y_t with entry $y_{t,ik}$ fixed at 1, while the entry is set to 0 in $y_{t,ik}^-$. For each possible inventor-patent connection, we can then state the corresponding probability conditional on the remaining bipartite network denoted by $y_{t,ik}^C$, i.e., the complete network y_t excluding the single entry $y_{t,ik}$. This leads to

$$\mathbb{P}_\theta(Y_{t,ik} = 1 \mid Y_{t,ik}^C = y_{t,ik}^C) = \frac{\exp\{\theta^T \Delta_{t,ik}(y_t, \dots, y_{t-s})\}}{1 + \exp\{\theta^T \Delta_{t,ik}(y_t, \dots, y_{t-s})\}}. \tag{7}$$

Through this expression we can relate θ , the canonical parameter of (5), to the conditional probability of inventor i to be co-owner of patent k . We can thereby derive an interpretation of the coefficients reminiscent of common logistic regression: if adding the tie $y_{t,ik}$ to the network raises the p th entry of $\Delta_{t,ik}(y_t, \dots, y_{t-s})$ by one unit, the conditional log-odds of $Y_{t,ik}$ are, ceteris paribus, altered by the additive factor θ_p (Goodreau et al., 2009).

3.2 Sufficient statistics for bipartite patent data

The main ingredient of model (5) is the set of sufficient statistics, which translates to a particular dependence structure assumed for the edges in the observed bipartite network (Wang, Pattison, et al., 2013). A statistic that is typically included is the number of edges at time point t , i.e., $s_{\text{edges}}(y_t, \dots, y_{t-s}) = |y_t|$, which can be comprehended as the equivalent of an intercept term in standard regression models (Goodreau et al., 2009). As we are in a dynamic setting in which additional information on past networks is available, we can define statistics that depend on the past networks, such as the number of patents in the previous s years for each actor active at time point t :

$$s_{\text{pastpatent}}(y_t, \dots, y_{t-s}) = \sum_{i \in \mathcal{I}_t} \sum_{k \in \mathcal{K}_t} y_{t,ik} \sum_{u=t-s}^{t-1} \sum_{l \in \mathcal{K}_u} y_{u,il}. \tag{8}$$

As the patent network presents some particular dependence structures, more advanced types of statistics are needed, which we describe in the following.

3.2.1 Pairwise statistics of inventors

One drawback of representing our patent data as a bipartite adjacency matrix instead of the one-mode-projected version is that incorporating information on the pairwise inventor-to-inventor level is not straightforward. We therefore introduce assortative two-star statistics extending the work of Bomiriy (2014, Chapter 2) and Metz et al. (2019) on homophily, which is

defined as the mechanism driving ties between similar individuals (McPherson et al., 2001), for bipartite networks. In the context of relational event models for bipartite interactions, Malang et al. (2019) use tie-specific, as opposed to global, variants of these statistics based on exponentially decreasing temporal weights of past events. We take the patent-based two-star statistic as starting point, which for \mathbf{y}_t is defined by

$$s_{\text{two-star.patent}}(\mathbf{y}_t) = \frac{1}{2} \sum_{k \in \mathcal{K}_t} \sum_{i \in \mathcal{I}_t} y_{t,ik} \left(\sum_{j \neq i} y_{t,jk} \right). \quad (9)$$

The tendency to interact with one another is often based on the similarity of a factor variable $\mathbf{u}_t = (u_{t,i}; i \in \mathcal{I}_t)$. We therefore define the indicator matrix $\mathbf{x}_t \in \{0, 1\}^{|\mathcal{I}_t| \times |\mathcal{I}_t|}$ with entries $x_{t,ij} = \mathbb{1}(u_{t,i} = u_{t,j})$. In line with Bomiriyi (2014, Chapter 2), this allows to augment the two-star statistic (9) in the form

$$s_{\text{homophily.x}}(\mathbf{y}_t) = \frac{1}{2} \sum_{k \in \mathcal{K}_t} \sum_{i \in \mathcal{I}_t} y_{t,ik} \left(\sum_{j \neq i} y_{t,jk} x_{t,ij} \right). \quad (10)$$

Next, we follow Metz et al. (2019) and generalise (10) by not restricting ourselves to any particular definition of \mathbf{x}_t , but letting the matrix be an arbitrary function of the networks from the past s years and other exogenous information. To further correct for different sizes of patents, i.e., the number of inventors co-owning the patent, we normalise the statistic by the degree of each patent, whereby the resulting statistic is defined through:

$$s_{\text{assort.x}}(\mathbf{y}_t, \dots, \mathbf{y}_{t-s}) = \frac{1}{2} \sum_{k \in \mathcal{K}_t} \sum_{i \in \mathcal{I}_t} y_{t,ik} \left(100 \times \frac{\sum_{j \neq i} y_{t,jk} x_{t,ij}}{\sum_{j \neq i} y_{t,jk}} \right). \quad (11)$$

To obtain a less cluttered notation, we keep the dependence of $x_{t,ij}$ on $\mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-s}$ implicit. The corresponding change statistic for an edge between inventor i and patent k is then

$$\Delta_{t,ik,\text{assort.x}}(\mathbf{y}_t, \dots, \mathbf{y}_{t-s}) = 100 \times \frac{\sum_{j \neq i} y_{t,jk} x_{t,ij}}{\sum_{j \neq i} y_{t,jk}}, \quad (12)$$

which can be interpreted as the percentage of inventors on patent k that match with inventor i in matrix \mathbf{x} . We multiply the statistic by 100, which does not affect the model itself but eases interpretation (as a unit increase is now equivalent to a single percentage change). To give an example of a statistic of this type, we can combine (12) with matrix \mathbf{x}_t^P , for which entry $x_{t,ij}^P$ is 1 if inventor i and j already had a joint patent in the last s years and 0 otherwise. The resulting statistic measures how previous collaboration among inventors affects the propensity of future collaboration. Section 4 provides more examples of such statistics.

3.2.2 Node set statistics

As a result of the actor natality and mortality described in the Introduction, we can split the set of inventors \mathcal{I}_t at each time step $t = 1, \dots, T$ into new inventors with their first patent in t , $\mathcal{I}_t^+ = \{i \in \mathcal{I}_t; \sum_{u=t-s}^{t-1} \sum_{k \in \mathcal{K}_u} y_{u,ik} = 0\}$, and inventors that were already active prior to t , $\mathcal{I}_t^- = \{i \in \mathcal{I}_t; \sum_{u=t-s}^{t-1} \sum_{k \in \mathcal{K}_u} y_{u,ik} > 0\}$. We here use the term ‘new inventors’ for actors in \mathcal{I}_t^+ and ‘experienced inventors’ for those in \mathcal{I}_t^- . Given these sets, we define $\mathbf{y}_t^+ = (y_{t,ik})_{i \in \mathcal{I}_t^+, k \in \mathcal{K}_t}$ and $\mathbf{y}_t^- = (y_{t,ik})_{i \in \mathcal{I}_t^-, k \in \mathcal{K}_t}$ to be the sub-networks of \mathbf{y}_t made up of new and experienced inventors, respectively.

It is apparent that statistics on past behaviour, such as (8), are not meaningful for inventors from \mathcal{I}_t^+ , since no historical data is available for those inventors at time t . To account for this, we decompose the statistics $s(\mathbf{y}_t, \dots, \mathbf{y}_{t-s})$ into three types of terms, namely $s^+(\mathbf{y}_t^+)$, $s^-(\mathbf{y}_t^-, \dots, \mathbf{y}_{t-s})$, and $s^\pm(\mathbf{y}_t)$, which are defined as statistics that only relate to either \mathbf{y}_t^+ , \mathbf{y}_t^- and past networks or

the full set of inventors \mathbf{y}_t , respectively. Defining the corresponding coefficients $(\theta^+, \theta^-, \theta^\pm)$ and change statistics $(\Delta_{t,ik}^+, \Delta_{t,ik}^-, \Delta_{t,ik}^\pm)$ accordingly yields

$$\mathbb{P}_\theta(Y_{t,ik} = 1 | \mathbf{Y}_{t,ik}^C = \mathbf{y}_{t,ik}^C) = \begin{cases} \pi_{t,ik}^+(\mathbf{y}_t), & \text{if } i \in \mathcal{I}_t^+ \text{ (new inventor)} \\ \pi_{t,ik}^-(\mathbf{y}_t, \dots, \mathbf{y}_{t-s}), & \text{if } i \in \mathcal{I}_t^- \text{ (experienced inventor),} \end{cases} \quad (13)$$

where $\pi_{t,ik}^+(\mathbf{y}_t)$ and $\pi_{t,ik}^-(\mathbf{y}_t, \dots, \mathbf{y}_{t-s})$ are given by

$$\begin{aligned} \pi_{t,ik}^+(\mathbf{y}_t) &= \frac{\exp\{(\theta^+)^T \Delta_{t,ik}^+(\mathbf{y}_t^+) + (\theta^\pm)^T \Delta_{t,ik}^\pm(\mathbf{y}_t)\}}{1 + \exp\{(\theta^+)^T \Delta_{t,ik}^+(\mathbf{y}_t^+) + (\theta^\pm)^T \Delta_{t,ik}^\pm(\mathbf{y}_t)\}} \\ \pi_{t,ik}^-(\mathbf{y}_t, \dots, \mathbf{y}_{t-s}) &= \frac{\exp\{(\theta^-)^T \Delta_{t,ik}^-(\mathbf{y}_t^-, \dots, \mathbf{y}_{t-s}^-) + (\theta^\pm)^T \Delta_{t,ik}^\pm(\mathbf{y}_t)\}}{1 + \exp\{(\theta^-)^T \Delta_{t,ik}^-(\mathbf{y}_t^-, \dots, \mathbf{y}_{t-s}^-) + (\theta^\pm)^T \Delta_{t,ik}^\pm(\mathbf{y}_t)\}}. \end{aligned}$$

As an example, for the common edge statistic $s_{\text{edges}}(\mathbf{y}_t, \dots, \mathbf{y}_{t-s})$, the aforementioned decomposition means we can define $s_{\text{New}}(\mathbf{y}_t^+) = |\mathbf{y}_t^+|$ and $s_{\text{Experienced}}(\mathbf{y}_t^-, \dots, \mathbf{y}_{t-s}^-) = |\mathbf{y}_t^-|$, to allow for new and experienced inventors to generally have a different propensity to be part of a patent. Note that the splitting of the node set as in (13) does not assume any (in)dependence structure between \mathbf{Y}_t^+ and \mathbf{Y}_t^- , but rather serves as an aid to specify additional terms and interpret the coefficients at a finer level, as just exemplified for the edge statistic.

3.2.3 Adjustment for varying network size

As argued in Krivitsky et al. (2011), the task of comparing estimated coefficients of two models with identical specifications but different network sizes is non-trivial. This behaviour is due to the fact that including the edge count statistic from the previous paragraph in a TERGM assumes density invariance as the network grows. This characteristic seldom holds for real-world networks as it implies a linearly growing mean degree of all involved actors. In the case of our longitudinal patent network, the number and composition of inventors and patents change from year to year, thus correcting for this is of practical importance to be able to compare coefficient estimates at different time points. To solve the issue, we follow the suggestion of Krivitsky et al. (2011) and incorporate the offset term $1/(|\mathcal{I}_t| + |\mathcal{K}_t|)$ to achieve asymptotically constant mean-degree scaling as the composition of inventors and patents change over time.

3.3 Estimation and inference

We now seek to estimate the parameter θ by maximising the logarithmic likelihood constructed from (5) for the transition between time points $t-1$ and t . Analysing each transition one at a time enables the use of software for static networks, such as `ergm` (Hunter et al., 2013). If some of the coefficients are constant over some periods, one could apply the block-diagonal approach of Leifeld et al. (2018). We follow the Markov Chain Monte Carlo Maximum-Likelihood Estimation procedure introduced by Geyer and Thompson (1992) and adapted to ERGMs by Hunter and Handcock (2006). In our application, we repeat this for each available time step $t = 1, \dots, T$.

First, note that subtracting any constant from the logarithmic likelihood constructed from (5) does not change its maximum. We can therefore subtract the logarithmic likelihood evaluated at an arbitrary value of the parameter θ , i.e., θ_0 , which yields the equivalent objective function

$$\ell(\theta) - \ell(\theta_0) = (\theta - \theta_0)^T s(\mathbf{y}_t, \dots, \mathbf{y}_{t-s}) - \log(\mathbb{E}_{\theta_0}(\exp\{(\theta - \theta_0)^T s(\mathbf{Y}_t, \dots, \mathbf{y}_{t-s})\})), \quad (14)$$

where $\mathbb{E}_\theta(f(\mathbf{X}))$ is the expected value of random variable \mathbf{X} characterised by parameter θ and transformed through the arbitrary function $f(\cdot)$. As described in Hunter and Handcock (2006), one can evaluate this objective function by approximating the expected value by generating random networks $\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \dots, \mathbf{Y}^{(M)}$ from (5) under θ_0 . In particular, we approximate the expected value

in (14) through a Monte Carlo quadrature:

$$\mathbb{E}_{\theta_0}(\exp\{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{s}(\mathbf{Y}_t, \dots, \mathbf{y}_{t-s})\}) \approx \frac{1}{M} \sum_{m=1}^M \exp\{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathbf{s}(\mathbf{y}^{(m)}, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-s})\}. \quad (15)$$

For sufficiently large M , the convergence of this expectation is guaranteed, and we can plug (15) into (14) and apply Newton–Raphson-type methods to maximise it with respect to $\boldsymbol{\theta}$. Sampling from a probability distribution with intractable normalisation constant, such as (5), is achieved by a Metropolis–Hastings algorithm. In particular, we first sample an edge, defined as the tuple (i, k) , at random, and consecutively toggle the corresponding entry of \mathbf{Y}_t from 0 to 1 with probability equal to (7) (for more details see Hunter et al., 2013). Due to the large size of the patent networks, we start with the observed network, propose 15.000 of such changes and then stop the Markov chain. This procedure is hence equivalent to contrastive divergence as introduced by Hinton (2002) and adapted to ERGMs by Krivitsky (2017).

Inference on the estimates is drawn based on the Fisher matrix $\mathbf{I}(\boldsymbol{\theta})$, which equals the variance of the sufficient statistics for exponential family distributions (L. Wasserman, 2004). Thus, we can approximate the Fisher matrix through

$$\begin{aligned} \widehat{\mathbf{I}}(\boldsymbol{\theta}) = \text{Var}_{\boldsymbol{\theta}}(\mathbf{s}(\mathbf{Y}_t, \dots, \mathbf{y}_{t-s})) &\approx \frac{1}{M} \sum_{m=1}^M (\mathbf{s}(\mathbf{y}^{(m)}, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-s}) - \bar{\mathbf{s}}(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)})) \\ &\quad \times (\mathbf{s}(\mathbf{y}^{(m)}, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-s}) - \bar{\mathbf{s}}(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)}))^\top, \end{aligned}$$

where $\bar{\mathbf{s}}(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)}) = (1/M) \sum_{m=1}^M \mathbf{s}(\mathbf{y}^{(m)}, \mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-s})$ is the vector containing the averages of the sufficient statistics from the simulated networks $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)}$, which are, in turn, drawn from the fitted model, with the parameter $\boldsymbol{\theta}$ set to its maximum-likelihood estimate.

4 Application to inventor team formation

We now present the results of our application, in which we model inventor team formation using the patent data introduced in Section 2. For each statistic included in the model, we explain its meaning, interpret the corresponding estimated coefficient, and then discuss the relationship of our results to prior literature. Further details on the specification of each sufficient statistic can be found in Appendix A. We further provide MCMC diagnostics and goodness-of-fit assessments as proposed by Hunter et al. (2008) in the online Supplementary Material. Due to the slow inertia of patent submissions visible in Figure 3, we set $s = 5$, i.e., consider data from the last five years to be relevant for modelling the current network. This allows us to have enough information for capturing long-range dependence in the networks involving repeated patent submissions of single actors as well as groups of actors.

4.1 Network effects

4.1.1 Propensity to invent

To account for the changing activity levels over time, we incorporate a statistic that counts how many edges are in the network. Following Section 3.2, we split this term into separate statistics for experienced and new inventors. Heuristically, one can interpret the corresponding coefficients as the general propensity to form ties, i.e., participate in a patent, for the two inventor sets, respectively. Note that it would not be possible to estimate this effect by modelling a unipartite projection on inventors: in that case, the intercept term would only measure the propensity for inventors to collaborate, regardless of the number of patents produced. The plot of the estimates for the propensity to invent over time is shown in the upper left panel of Figure 4. It exhibits a different level of activity for new and experienced inventors. We expect this by design, as new inventors enter the network precisely because they are active at time t , while experienced ones might only have been active in the past. Overall, we observe a steady increase in activity in the network from 2008 onward for both sets of inventors.

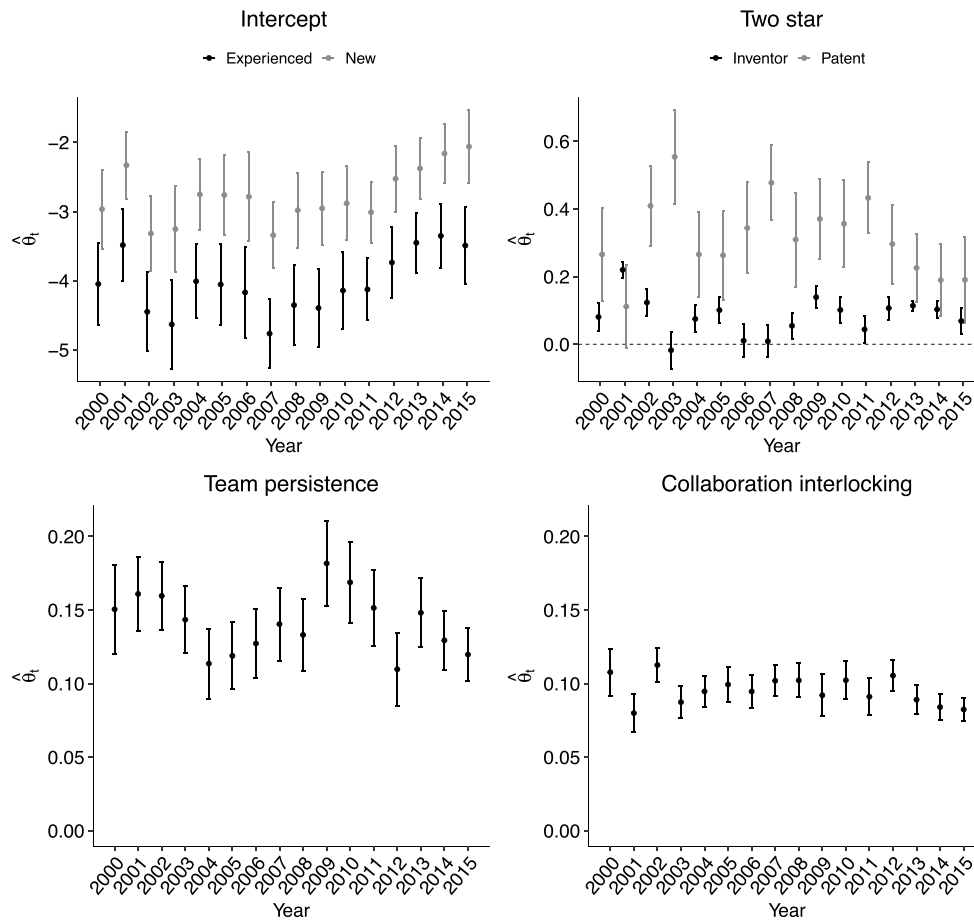


Figure 4. Estimated time-varying coefficients regarding the propensity to invent, two-star statistics, team persistence and collaboration interlocking.

4.1.2 Two-star statistics

Two-star statistics relate to the concept of centrality (S. Wasserman & Faust, 1994). For bipartite networks, they can be defined with respect to each of the two modes (inventors and patents, respectively). For inventors the statistic is given in Appendix A and expresses whether inventor i is more or less likely to invent an additional patent in year t , given that he/she is (co-)owner of at least another patent in that year. For patents, the statistic relates to the number of inventors per patent and is given in (9). These effects could not be estimated for a unipartite projection on inventors: in that case, the two-star statistic would simply relate to the propensity for inventors to have additional collaborators, with information on the number of patents and their size being lost. The top right panel of Figure 5 depicts both estimates for the two-star statistics. For inventors, the estimates take small positive values for most time points, without much temporal variation. This indicates a slight tendency towards centralisation for inventors, i.e., inventors aiming to submit multiple patents per year. For patents the corresponding two-star estimates are larger, i.e., patents tend to be owned by multiple inventors. The two-star effect slowly decreases since 2011, meaning that the number of owners per patent is getting smaller. The variance for the estimated two-star patent effect is generally larger than the estimate of the corresponding two-star inventor effect, which stems from the fact that there are fewer patents than inventors in a single year.

4.1.3 Team persistence

Most patented inventions are the result of team work (Giuri et al., 2007), which leads to the build-up of valuable team-specific capital (Jaravel et al., 2018). We therefore expect past collaboration

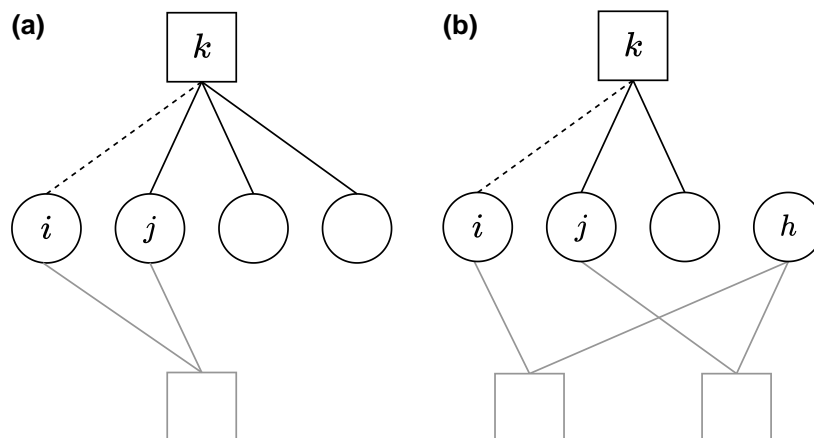


Figure 5. Illustration of the change statistics related to assortative network statistics for team persistence (a) and collaboration interlocking (b). Circles represent inventors, and squares are patents. The dashed line indicates a possible edge at time point t , while black lines represent edges given at time point t . Grey lines, on the other hand, display past connections, and grey squares stand for past patents. (a) Team persistence and (b) collaboration interlocking.

to positively affect the propensity for two inventors to collaborate again. To account for this effect, we include a team persistence statistic based on the pairwise statistics of inventors proposed in 3.2 in the model. The statistic, which could also be termed ‘repetition’ (or ‘reciprocity’, as defined in Leifeld & Brandenberger, 2019), is visually represented in Figure 5a, and rests on the definition of matrix \mathbf{x}_t^p , whose (i, j) th entry is 1 if inventors i and j have already co-invented a patent in the previous five years, and 0 otherwise. The bottom left panel of Figure 4 depicts the corresponding coefficient estimate, which is positive and significantly different from zero over time. This finding corroborates our anticipations that, controlling for the other factors, two inventors are more likely to jointly produce a patent if they already worked on an invention together in the past. Hence, teams of inventors play an important role in patent creation.

4.1.4 Collaboration interlocking

In addition to investigating the persistence of collaborations, it is of interest to understand how having had a common partner in the past influences the tendency to develop a joint patent in the present. We account for this by including the collaboration interlocking statistic in our model. By common partners we are referring to actors such as inventor h for inventors i and j in Figure 5b. We define the statistic again by pairwise statistics of inventors through the matrix \mathbf{x}_t^{ci} , where the binary information of whether or not inventors i and j have at least one common partner is encoded in the (i, j) th entry. The related coefficient estimates are shown in the bottom right panel of Figure 4, where we notice that the estimate attains significantly positive values throughout the observational period. This result suggests that if two inventors i and j both had a patent with the same inventor h , they are generally more likely to co-invent in the future. Our finding holds controlling for all other features in our model (including the previously described team persistence statistic). This effect can be considered similar to triadic closure in unimodal networks, i.e., ‘a collaborator of my collaborator is more likely to become my collaborator’. The result thus supports the idea that the creation of inventor teams is often promoted via common colleagues and that informal knowledge flows are key to the invention process (see Giuri & Mariani, 2013 and references cited therein).

4.2 Effects of inventor-specific covariates

4.2.1 Spatial proximity

Many patents are created in a workplace environment (Giuri et al., 2007). For this reason, we would expect inventors that live close to each other to be more likely to invent together. Moreover, there is empirical evidence that collaboration is more likely between inventors that

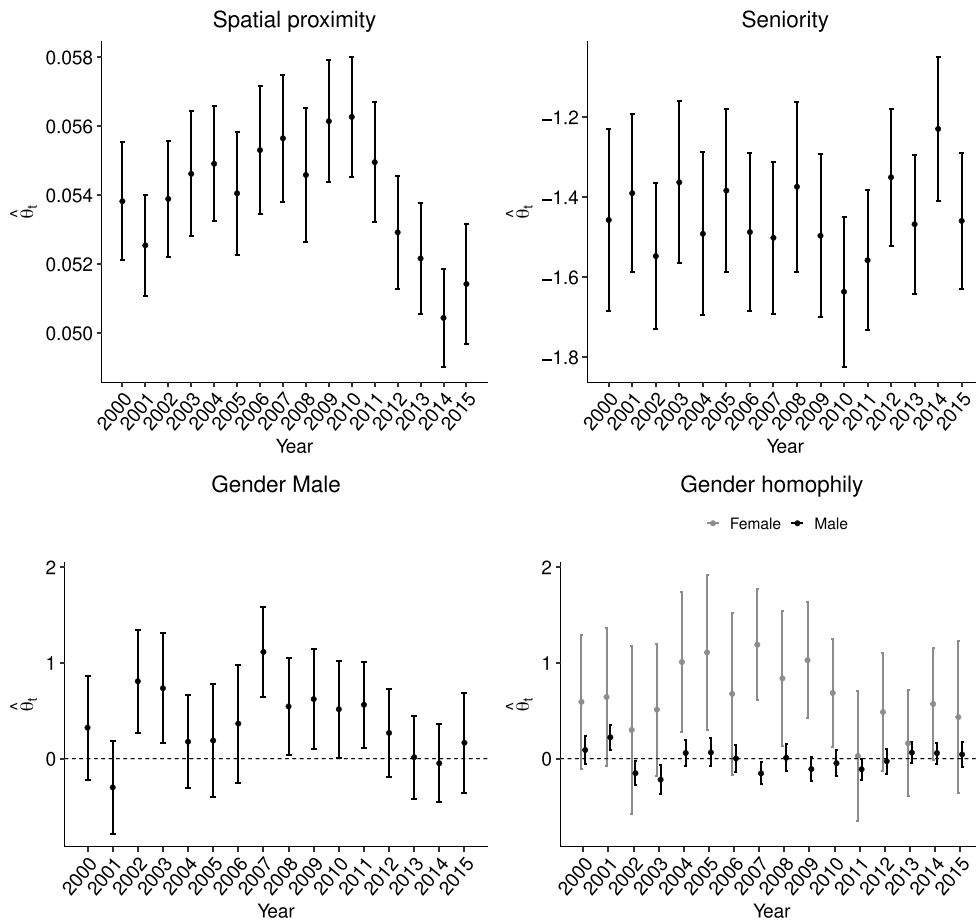


Figure 6. Estimated time-varying coefficients regarding the spatial proximity, seniority, and gender of inventors.

live close to one another even if they do not share the same employer (e.g., Crescenzi et al., 2016). For these reasons, we include a spatial proximity statistic in our model, where we define spatial proximity as living within a range of 50 km. We encode this proximity information in a binary matrix \mathbf{x}^{SP} and incorporate it in the model as a pairwise statistic of inventors. The top left panel of Figure 6 depicts the estimated coefficients for the statistic. The positive values attained over time confirm that inventors living near each other have a higher chance to collaborate. We can also see that the effect goes down over time from 2010 onward; this makes sense in an increasingly interconnected society, where more and more connections are formed through the web in addition to physical ones.

4.2.2 Seniority

The top right panel in Figure 6 depicts the effect of the number of previously owned patents by each inventor in the past five years. The corresponding statistic can be viewed as a measure of inventor seniority, where inventors with more patents in the past are considered to be senior. As this statistic would trivially be a structural zero for new inventors, it is only computed for the set of inventors which were previously active in the network (experienced inventors). This is another effect we would not be able to estimate if we only considered the unipartite projected network of inventors. The negative coefficient estimate here suggests that, conditional on all other statistics included in the model, senior inventors have a lower propensity to create new patents. Prior research has shown that career dynamics of inventors are complex as economic opportunities, productivity and personal preferences interact (see, e.g., Allen & Katz, 1992; Bell et al., 2019). But our

results are consistent with earlier results indicating that with greater seniority, inventors take over managerial responsibilities within the same firm, or that high visibility of their invention output also leads them to move to new employers and tasks, thus lowering (or halting) their invention output.

4.2.3 Gender and gender homophily

Another variable of interest in the realm of innovation research is gender. Many researchers have expressed concerns about the sparse representation of women among inventors (typically far less than 10%) and possible wage discrimination (see, e.g., [Hoisl & Mariani, 2017](#); [Jensen et al., 2018](#)). These studies established gender as an essential topic in innovation economics. We incorporate gender in our model in two ways, i.e., as a main effect and as a homophily effect (as introduced in (10)). The two plots at the bottom of [Figure 6](#) show the effects of gender on the propensity to create patents (left) and on homophily, i.e., the tendency of inventing together with people of the same gender (right). Note that both effects need to be interpreted keeping in mind that the vast majority of the actors in the network are male (96%). From the plot on the bottom left, we can see how, while male inventors seem to be slightly more active, all in all male and female inventors did not show significant differences in their propensity to invent. Note that this holds given the inclusion of those inventors in the network, i.e., given that they were already inventors. The gender homophily plot shows different results; here we see that, while male inventors seem to have the same likelihood to form patents with both genders, female inventors tend to have more collaborations with other females than with males. While the effect is quite sizeable in absolute value, the uncertainty here is considerable given the small number of female actors in the network. Still, we can see this as weak evidence for a gender homophily effect for female inventors. These results are consistent with earlier findings by [Whittington \(2018\)](#), who studies the role of gender in life science inventor teams.

5 Discussion

This paper analyses a massive bipartite network, consisting of all inventors and collaborative patents filed between 1995 and 2015 in electrical engineering. To account for the sheer size of the complete network and the structural zeros in the related bipartite adjacency matrix, we suggested a temporal decomposition of the data into multiple smaller networks. Guided by substantive questions posed by innovation research, we then proposed a set of bipartite network statistics focused on gender issues, team persistence, collaboration interlocking, and spatial proximity.

Time-varying actor sets due to actor mortality and natality are often observed in networks beyond the realm of patent data. For instance, scientific collaboration behaves similarly, as many PhD students do not pursue an academic career and hence have a short lifespan in the scientific collaboration network. At the same time, new PhD students continuously enter the scientific world. Therefore, the proposed temporal decomposition and the employed network terms exploiting pairwise information on either mode of actors can also be used in other settings.

In addition to the methodological contributions, our study offers several novel results concerning the substantive analysis. We utilise a population dataset spanning 20 years (1995–2015). The time span and the availability of population data are crucial to assess the team formation process reliably. Using a population dataset of this size is unique in the literature on inventor team formation. Moreover, while much of the literature has focused on the relationship between team characteristics and performance, there are very few studies on the actual process of inventor team formation. While some of the variables we are using have been discussed and utilised in other domains, we are unaware of inventor team studies employing data with a similar breadth of team and inventor descriptors. This breadth adds to the novelty of our study. We also note that our variable set reflects a number of meaningful concerns such as inclusiveness, gender equality, and seniority. The results should therefore be of considerable interest to policymakers.

Still, we want to address some limitations in our analysis, which would benefit from further research. First, our definition of the actor sets is based on a simple heuristic we determined in a data-driven manner. However, this practice might bias our findings concerning degree-related statistics since the exact number of isolated inventors is not known but assumed. More complex methods to identify active inventors based on further exogenous data, such as job histories, might be a fruitful

future endeavour. Second, we assumed the parameters to be different each year. Extending the approach of Cranmer et al. (2014), one could incorporate a change-point detection directly into the TERGM framework to identify periods over which the coefficients are constant from the observed data. Note that, to facilitate building on our research, we make our implementation available through the R software package `patent.ergm`. Moreover, to guarantee the replicability of our results, we make the full data and code available online on a GitHub repository. This repository also includes the R package `patent.ergm`.

All in all, we show how spatial proximity, teamwork and interlocking of collaborations positively impact the output of inventors. Further, we demonstrate how inventors' characteristics, such as gender and seniority, play a significant role in the process, and identify gender homophily as a critical determinant of inventor team formation. Our application to inventor teams presents an alternative to classical forms of analysis of patenting and inventorship networks. While prior studies are almost exclusively focused on analysing the underlying mechanisms one at a time, we model them simultaneously in the framework of bipartite networks. Our study thus provides an effective alternative to classical forms of regression-based analysis of innovation and the mechanisms driving it.

Conflict of interest: No potential conflict of interest was reported by the author(s).

Funding

The work has been partially supported by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibility for its content.

Data availability

We provide full replication code and materials, including data, goodness-of-fit analysis and MCMC diagnostics, in our GitHub repository, available at <https://github.com/corneliusfritz/Modelling-German-patents-and-inventors>. Moreover, this repository includes the package `ergm.patent` that implements the pairwise statistics introduced in Section 3.2.1.

Appendix A. Sufficient statistics

In the following, we detail the mathematical definitions of all sufficient statistics incorporated in our model.

Propensity to invent: As already stated in Section 3.1, the standard term to incorporate in any ERGM specification is an edge statistic that counts how many edges are realised in the network. In accordance with Section 3.2, we split this term into the statistics $s_{\text{New}}(\mathbf{y}_t^+) = |\mathbf{y}_t^+| = \sum_{i \in \mathcal{I}_t^+} \sum_{k \in \mathcal{K}_t} y_{t,ik}$ and $s_{\text{Experienced}}(\mathbf{y}_t^-, \dots, \mathbf{y}_{t-s}^-) = |\mathbf{y}_t^-| = \sum_{i \in \mathcal{I}_t^-} \sum_{k \in \mathcal{K}_t} y_{t,ik}$. Figures A1a and A1b visualise the corresponding two network configurations.

Two-star statistics: Two-star statistics can be stated with regards to either set of actors in the case of bipartite networks. The definition of the two-star statistic for the patents is shown in Figure A1c and given by

$$s_{\text{twostar.patent}}(\mathbf{y}_t) = \frac{1}{2} \sum_{k \in \mathcal{K}_t} \sum_{i \in \mathcal{I}_t} y_{t,ik} \left(\sum_{j \neq i} y_{t,jk} \right),$$

while the version for the inventors is visualised in Figure A1d and defined as:

$$s_{\text{twostar.inventor}}(\mathbf{y}_t) = \frac{1}{2} \sum_{i \in \mathcal{I}_t} \sum_{k \in \mathcal{K}_t} y_{t,ik} \left(\sum_{l \neq k} y_{t,il} \right).$$

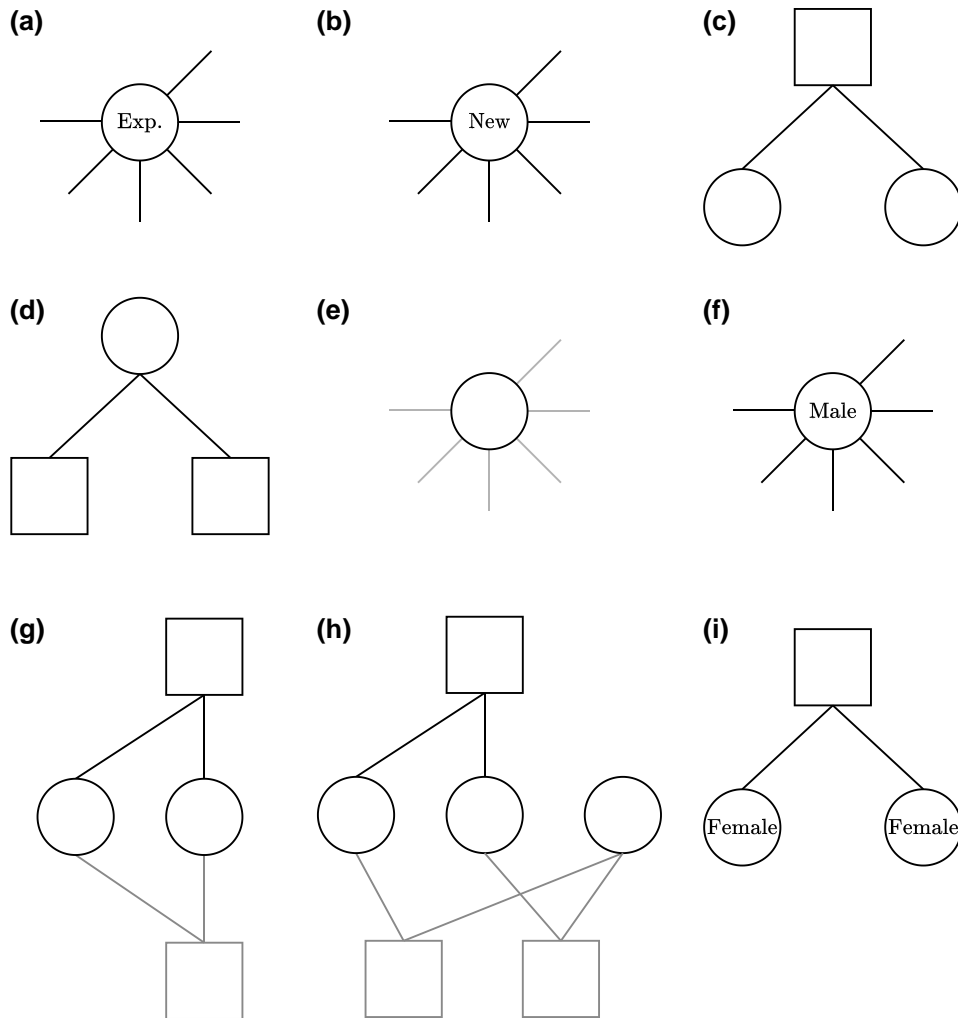


Figure A1. Network configurations for the general edge and two-star terms. Circles are inventors and squares patents and black lines are observed edges in the network at time point t , while grey lines are edges in the past. (a) Experienced inventors, (b) new inventors, (c) patent two-stars, (d) inventor two-stars, (e) seniority, (f) male inventors, (g) team persistence, (h) collaborative interlocking and (i) homophily of females.

Pairwise statistics of inventors: We include three versions of pairwise statistics of inventors introduced in Section 3.2. The statistics are given by

$$s_{\text{assort.x}}(\mathbf{y}_t, \dots, \mathbf{y}_{t-s}) = \frac{1}{2} \sum_{k \in \mathcal{K}_t} \sum_{i \in \mathcal{I}_t} y_{t,ik} \left(100 \times \frac{\sum_{j \neq i} y_{t,jk} x_{t,ij}}{\sum_{j \neq i} y_{t,jk}} \right).$$

Note that, in general, the matrix \mathbf{x} can be an arbitrary function of the past networks and nodal or dyadic exogenous information. Its definition differs between the three statistics of pairwise statistics of inventors:

1. Team persistence: For $i, j \in \mathcal{I}_t$ and $i \neq j$ the entries of \mathbf{x}_t^p are given by

$$x_{t,ij}^p = \begin{cases} 1, & \text{if } \sum_{u=t-s}^{t-1} \sum_{k \in \mathcal{K}_u} y_{u,ik} y_{u,jk} > 0 \\ 0, & \text{else} \end{cases}$$

and a graphical illustration of the statistic is provided in Figure A1g. Note that Leifeld and Brandenberger (2019) and Metz et al. (2019) describe a closely related mechanism as *reciprocity* and *collaboration*, respectively. One can comprehend this statistic as a particular type of the four-cycle statistic (Wang, Pattison, et al., 2013) where one half already occurred in the past, and the other half might occur in the present.

2. Collaboration interlocking: For $i, j \in \mathcal{I}_t$ and $i \neq j$, the entries of \mathbf{x}_t^{CI} are defined by

$$x_{t,ij}^{\text{CI}} = \begin{cases} 1, & \text{if } \sum_{u=t-1}^{t-1} \sum_{b \in \mathcal{I}_t} \sum_{k,l \in \mathcal{K}_u} y_{u,ik} y_{u,bk} y_{u,jl} y_{u,bl} > 0 \\ 0, & \text{else} \end{cases}$$

and a graphical illustration of the statistic is provided in Figure A1h. Coming back to the representation as cycle-statistics, this term is a six-cycle statistic in which four of the six edges happened in the time frame from $t-5$ to $t-1$ and two in year t .

3. Spatial proximity: For $i, j \in \mathcal{I}_t$ and $i \neq j$ the entries of \mathbf{x}_t^{SP} are defined as

$$x_{t,ij}^{\text{SP}} = \begin{cases} 1, & \text{if } \text{dist}(x_{\text{coord},i}, x_{\text{coord},j}) > 50 \text{ km} \\ 0, & \text{else} \end{cases}$$

where $x_{\text{coord},i}$ and $x_{\text{coord},j}$ define the longitude and latitude of inventors i and j , respectively, and the function $\text{dist}(x_{\text{coord},i}, x_{\text{coord},j})$ computes the distance in kilometres between them via the haversine formula. A continuous form of this statistic based on the Euclidean distance itself was employed in Metz et al. (2019).

Seniority: The respective binary indicator is based on the past patent statistic given in (8), but in this case we define it on the inventor level:

$$s_{\text{seniority},i}(y_t, \dots, y_{t-s}) = \sum_{u=t-s}^{t-1} \sum_{k \in \mathcal{K}_u} y_{u,ik}$$

We binarise this inventor-specific covariate by first computing the median of $s_{\text{seniority},i}(y_t, \dots, y_{t-s})$ over all inventors and then using this value to split the inventors into two groups (i.e., seniors and juniors). The resulting categorical covariate relates to the number of patents in the past and is represented in Figure A1e.

Gender and gender homophily: The main effect of gender is depicted in Figure A1f and defined by:

$$s_{\text{gender}}(y_t) = \sum_{i \in \mathcal{I}_t} \sum_{k \in \mathcal{K}_t} y_{t,ik} \mathbb{1}(x_{\text{gender},i} = \text{'male'}),$$

where $x_{\text{gender},i} \in \{\text{'male'}, \text{'female'}\}$ indicates the gender of inventor i . The homophily effect, on the other hand, is for males defined by:

$$s_{\text{homophily,male}}(y_t) = \frac{1}{2} \sum_{k \in \mathcal{K}_t} \sum_{i \in \mathcal{I}_t} y_{t,ik} \left(\sum_{j \neq i} y_{t,jk} \mathbb{1}(x_{\text{gender},i} = \text{'male'}) \mathbb{1}(x_{\text{gender},j} = \text{'male'}) \right).$$

and for females the formula reads:

$$s_{\text{homophily,female}}(y_t) = \frac{1}{2} \sum_{k \in \mathcal{K}_t} \sum_{i \in \mathcal{I}_t} y_{t,ik} \left(\sum_{j \neq i} y_{t,jk} \mathbb{1}(x_{\text{gender},i} = \text{'female'}) \mathbb{1}(x_{\text{gender},j} = \text{'female'}) \right).$$

Figure A1i visualises the homophily statistic for females. The equivalent statistic for males can be defined in the same manner.

References

- Ahmed A., Batagelj V., Fu X., Hong S.-H., Merrick D., & Mrvar A. (2007). Visualisation and analysis of the internet movie database. In *6th International Asia-Pacific Symposium on Visualization* (pp. 17–24). IEEE.
- Allen T. J., & Katz R. (1992). Age, education and the technical ladder. *IEEE Transactions on Engineering Management*, 39(3), 237–245. <https://doi.org/10.1109/17.156557>
- Alstott J., Triulzi G., Yan B., & Luo J. (2017). Mapping technology space by normalizing patent networks. *Scientometrics*, 110(1), 443–479. <https://doi.org/10.1007/s11192-016-2107-y>
- Balconi M., Breschi S., & Lissoni F. (2004). Networks of inventors and the role of academia: An exploration of Italian patent data. *Research Policy*, 33(1), 127–145. [https://doi.org/10.1016/S0048-7333\(03\)00108-2](https://doi.org/10.1016/S0048-7333(03)00108-2)
- Barndorff-Nielsen O. (1978). *Information and exponential families in statistical theory*. Wiley.
- Bauer V., Harhoff D., & Kauermann G. (2022). A smooth dynamic network model for patent collaboration data. *AStA Advances in Statistical Analysis*, 106(1), 97–116. <https://doi.org/10.1007/s10182-021-00393-w>
- Bell A., Chetty R., Jaravel X., Petkova N., & Van Reenen J. (2019). Do tax cuts produce more einsteins? The impacts of financial incentives versus exposure to innovation on the supply of inventors. *Journal of the European Economic Association*, 17(3), 651–677. <https://doi.org/10.1093/jeaa/jvz013>
- Bomiriya R. P. (2014). *Topics in exponential random graph modeling* [Phd thesis]. Pennsylvania State University.
- Borgatti S. P., & Everett M. G. (1997). Network analysis of 2-mode data. *Social Networks*, 19(3), 243–269. [https://doi.org/10.1016/S0378-8733\(96\)00301-2](https://doi.org/10.1016/S0378-8733(96)00301-2)
- Breiger R. L. (1974). The duality of persons and groups. *Social Forces*, 53(2), 181–190. <https://doi.org/10.2307/2576011>
- Broekel T., & Bednarz M. (2018). Disentangling link formation and dissolution in spatial networks: An application of a two-mode STERGM to a project-based R&D network in the German biotechnology industry. *Networks and Spatial Economics*, 18(3), 677–704. <https://doi.org/10.1007/s11067-018-9430-1>
- Chen T. H. Y. (2021). Statistical inference for multilayer networks in political science. *Political Science Research and Methods*, 9(2), 380–397. <https://doi.org/10.1017/psrm.2019.49>
- Cranmer S. J., Desmarais B. A., & Morgan J. W. (2021). *Inferential network analysis*. Cambridge University Press.
- Cranmer S. J., Heinrich T., & Desmarais B. A. (2014). Reciprocity and the structural determinants of the international sanctions network. *Social Networks*, 36(1), 5–22. <https://doi.org/10.1016/j.socnet.2013.01.001>
- Crescenzi R., Nathan M., & Rodríguez-Pose A. (2016). Do inventors talk to strangers? on proximity and collaborative knowledge creation. *Research Policy*, 45(1), 177–194. <https://doi.org/10.1016/j.respol.2015.07.003>
- de Solla Price D. J. (1965). The science of science. *Bulletin of the Atomic Scientists*, 21(8), 2–8. <https://doi.org/10.1080/00963402.1965.11454842>
- Egghe L., & Rousseau R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Elsevier.
- Ejermo O., & Karlsson C. (2006). Interregional inventor networks as studied by patent coinventorships. *Research Policy*, 35(3), 412–430. <https://doi.org/10.1016/j.respol.2006.01.001>
- Faust K., & Skvoretz J. (1999). Logit models for affiliation networks. *Sociological Methodology*, 31(1), 253–280.
- Friel N., Rastelli R., Wyse J., & Raftery A. E. (2016). Interlocking directorates in Irish companies using a latent space model for bipartite networks. *Proceedings of the National Academy of Sciences*, 113(24), 6629–6634. <https://doi.org/10.1073/pnas.1606295113>
- Garfield E. (1955). Citation indexes for science. *Science*, 122(3159), 108–111. <https://doi.org/10.1126/science.122.3159.108>
- Geyer C. J., & Thompson E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 54(3), 657–699. <https://doi.org/10.1111/j.2517-6161.1992.tb01443.x>
- Giuri P., et al. (2007). Inventors and invention processes in Europe: Results from the PatVal-EU survey. *Research Policy*, 36(8), 1107–1127. <https://doi.org/10.1016/j.respol.2007.07.008>
- Giuri P., & Mariani M. (2013). When distance disappears: Inventors, education, and the locus of knowledge spillovers. *Review of Economics and Statistics*, 95(2), 449–463. https://doi.org/10.1162/REST_a_00259
- Goodreau S., Kitts J. A., & Morris M. (2009). Birds of a feather, or friend of a friend?: Using exponential random graph models to investigate adolescent social networks. *Demography*, 46(1), 103–125. <https://doi.org/10.1353/dem.0.0045>
- Hall B. H., & Harhoff D. (2012). Recent research on the economics of patents. *Annual Review of Economics*, 4(1), 541–565. <https://doi.org/10.1146/annurev-economics-080511-111008>
- Hanneke S., Fu W., & Xing E. P. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4, 585–605. <https://doi.org/10.1214/09-EJS548>
- Hinton G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8), 1771–1800. <https://doi.org/10.1162/089976602760128018>

7. Modelling the large and dynamically growing bipartite network of German patents and inventors

- Hoisl K., & Mariani M. (2017). It's a man's job: Income and the gender gap in industrial research. *Management Science*, 63(3), 766–790. <https://doi.org/10.1287/mnsc.2015.2357>
- Hunter D. R., Goodreau S. M., & Handcock M. S. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481), 248–258. <https://doi.org/10.1198/016214507000000446>
- Hunter D. R., Goodreau S. M., & Handcock M. S. (2013). Ergm.userterms: A template package for extending statnet. *Journal of Statistical Software*, 52(2), 1–25. <https://doi.org/10.18637/jss.v052.i02>
- Hunter D. R., & Handcock M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3), 565–583. <https://doi.org/10.1198/106186006X133069>
- Jaravel X., Petkova N., & Bell A. (2018). Team-specific capital and innovation. *American Economic Review*, 108(4–5), 1034–1073. <https://doi.org/10.1257/aer.20151184>
- Jensen K., Kovács B., & Sorenson O. (2018). Gender differences in obtaining and maintaining patent rights. *Nature Biotechnology*, 36(4), 307–309. <https://doi.org/10.1038/nbt.4120>
- Jiang S., Gao Q., & Chen H. (2013). Statistical modeling of nanotechnology knowledge diffusion networks. *ICIS 2013 Proceedings*. Atlanta, GA, United States: Association for Information Systems.
- Jiang S., Gao Q., Chen H., & Roco M. C. (2015). The roles of sharing, transfer, and public funding in nanotechnology knowledge-diffusion networks. *Journal of the Association for Information Science and Technology*, 66(5), 1017–1029. <https://doi.org/10.1002/asi.23223>
- Keegan B., Gergle D., & Contractor N. (2012). Do editors or articles drive collaboration? Multilevel statistical network analysis of wikipedia coauthorship. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 427–436). New York, NY, United States: Association for Computing Machinery.
- Koskinen J., & Edling C. (2012). Modelling the evolution of a bipartite network – peer referral in interlocking directorates. *Social Networks*, 34(3), 309–322. <https://doi.org/10.1016/j.socnet.2010.03.001>
- Krivitsky P. N. (2017). Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models. *Computational Statistics and Data Analysis*, 107, 149–161. <https://doi.org/10.1016/j.csda.2016.10.015>
- Krivitsky P. N., Handcock M. S., & Morris M. (2011). Adjusting for network size and composition effects in exponential-family random graph models. *Statistical Methodology*, 8(4), 319–339. <https://doi.org/10.1016/j.stamet.2011.01.005>
- Latapy M., Magnien C., & Del Vecchio N. (2008). Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1), 31–48. <https://doi.org/10.1016/j.socnet.2007.04.006>
- Lebacher M., Thurner P. W., & Kauermann G. (2021). A dynamic separable network model with actor heterogeneity: An application to global weapons transfers. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(1), 201–226. <https://doi.org/10.1111/rssa.12620>
- Leifeld P. (2018). Polarization in the social sciences: Assortative mixing in social science collaboration networks is resilient to interventions. *Physica A: Statistical Mechanics and its Applications*, 507, 510–523. <https://doi.org/10.1016/j.physa.2018.05.109>
- Leifeld P., & Brandenberger L. (2019). Endogenous coalition formation in policy debates. Preprint, arXiv:1904.05327.
- Leifeld P., Cranmer S. J., & Desmarais B. A. (2018). Temporal exponential random graph models with btergm: Estimation and bootstrap confidence intervals. *Journal of Statistical Software*, 83(6), 1–36. <https://doi.org/10.18637/jss.v083.i06>
- Li X., Chen H., Huang Z., & Roco M. C. (2007). Patent citation network in nanotechnology (1976–2004). *Journal of Nanoparticle Research*, 9(3), 337–352. <https://doi.org/10.1007/s11051-006-9194-2>
- Malang T., Brandenberger L., & Leifeld P. (2019). Networks and social influence in European legislative politics. *British Journal of Political Science*, 49(4), 1475–1498. <https://doi.org/10.1017/S0007123417000217>
- McPherson M., Smith-Lovin L., & Cook J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1), 415–444. <https://doi.org/10.1146/annurev.soc.27.1.415>
- Melin G., & Persson O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, 36(3), 363–377. <https://doi.org/10.1007/BF02129600>
- Metz F., Leifeld P., & Ingold K. (2019). Interdependent policy instrument preferences: A two-mode network approach. *Journal of Public Policy*, 39(4), 609–636. <https://doi.org/10.1017/S0143814X18000181>
- Newman M. E. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(suppl_1), 5200–5205. <https://doi.org/10.1073/pnas.0307545100>
- Onody R. N., & de Castro P. A. (2004). Complex network study of Brazilian soccer players. *Physical Review E*, 70(3), 037103. <https://doi.org/10.1103/PhysRevE.70.037103>
- Small H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269. <https://doi.org/10.1002/asi.4630240406>
- Snijders T. A., Pattison P. E., Robins G. L., & Handcock M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology*, 36(1), 99–153. <https://doi.org/10.1111/j.1467-9531.2006.00176.x>

- Thurner P. W., Schmid C. S., Cranmer S. J., & Kauermann G. (2018). Network interdependencies and the evolution of the international arms trade. *Journal of Conflict Resolution*, 63(7), 1736–1764. <https://doi.org/10.1177/0022002718801965>
- Verspagen B. (2012). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in Complex Systems*, 10(1), 93–115. <https://doi.org/10.1142/S0219525907000945>
- von Wartburg I., Teichert T., & Rost K. (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy*, 34(10), 1591–1607. <https://doi.org/10.1016/j.respol.2005.08.001>
- Wang P., Pattison P., & Robins G. (2013). Exponential random graph model specifications for bipartite networks: A dependence hierarchy. *Social Networks*, 35(2), 211–222. <https://doi.org/10.1016/j.socnet.2011.12.004>
- Wang P., Robins G., Pattison P., & Lazega E. (2013). Exponential random graph models for multilevel networks. *Social Networks*, 35(1), 96–115. <https://doi.org/10.1016/j.socnet.2013.01.004>
- Wasserman L. (2004). *All of statistics*. Springer.
- Wasserman S., & Faust K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.
- Wasserman S., & Pattison P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^* . *Psychometrika*, 61(3), 401–425. <https://doi.org/10.1007/BF02294547>
- Whittington K. B. (2018). A tie is a tie? Gender and network positioning in life science inventor collaboration. *Research Policy*, 47(2), 511–526. <https://doi.org/10.1016/j.respol.2017.12.006>

8. COVID-19 and social media: Beyond polarization

Contributing article

De Nicola, G., Tuekam Mambou, V.H., and Kauermann, G. (2023). COVID-19 and social media: Beyond polarization. *PNAS Nexus*, 2(8):pgad246. <https://doi.org/10.1093/pnasnexus/pgad246>.

Data and code

Available at <https://github.com/gdenicola/latent-space-covid-twitter-elites>.

Copyright information

This article is distributed under a Creative Commons 4.0 International license (CC BY 4.0).

Supplementary material

[Supplementary material](#) is available at PNAS Nexus online.

Author contributions

The idea of analyzing the network of popular users tweeting about COVID-19 to investigate polarization can be attributed to Giacomo De Nicola. The latter further contributed by supervising and supporting data management, analysis and visualization, which were carried out for the most part by Victor H. Tuekam Mambou. Giacomo De Nicola was responsible for conceptualizing, designing and writing the entire paper. All authors contributed through fruitful comments and extensive proofreading of the manuscript.

COVID-19 and social media: Beyond polarization

Giacomo De Nicola ^{a,*}, Victor H. Tuekam Mambou ^{a,b} and Göran Kauermann ^a

^aDepartment of Statistics, Ludwig Maximilian University of Munich, 80539 Munich, Germany

^bIfo Institute – Leibniz Institute for Economic Research at the University of Munich, 81679 Munich, Germany

*To whom correspondence should be addressed: Email: giacomo.denicola@stat.uni-muenchen.de

Edited By: N. Contractor

Abstract

The COVID-19 pandemic brought upon a massive wave of disinformation, exacerbating polarization in the increasingly divided landscape of online discourse. In this context, popular social media users play a major role, as they have the ability to broadcast messages to large audiences and influence public opinion. In this article, we make use of openly available data to study the behavior of popular users discussing the pandemic on Twitter. We tackle the issue from a network perspective, considering users as nodes and following relationships as directed edges. The resulting network structure is modeled by embedding the actors in a latent social space, where users closer to one another have a higher probability of following each other. The results suggest the existence of two distinct communities, which can be interpreted as “generally pro” and “generally against” vaccine mandates, corroborating existing evidence on the pervasiveness of echo chambers on the platform. By focusing on a number of notable users, such as politicians, activists, and news outlets, we further show that the two groups are not entirely homogeneous, and that not just the two poles are represented. To the contrary, the latent space captures an entire spectrum of beliefs between the two extremes, demonstrating that polarization, while present, is not the only driver of the network, and that more moderate, “central” users are key players in the discussion.

Keywords: polarization, COVID-19, network analysis, Twitter, latent space models

Significance Statement

Popular social media users play a major role in the COVID-19 infodemic, as they can influence public opinion through their massive reach. Using state-of-the-art statistical network modeling techniques, we embed popular Twitter users discussing the pandemic in a latent social space, producing a map of the COVID-19 social media universe. The results suggest the existence of two distinct communities, which respectively favor and oppose vaccine mandates, thus corroborating the presence of echo chamber effects on the platform. We further show that the two groups are not entirely homogeneous: instead, the social map describes an entire spectrum of beliefs between the two extremes, demonstrating that polarization is not the only relevant factor, and that moderate users are central to the discussion.

Introduction

COVID-19 dramatically affected the lives of billions of people around the globe. Given its massive impact, the pandemic naturally assumed a central role in both private and public discourse, dominating the discussion on- and offline. Social media, in particular, has been extensively used to exchange pandemic-related information as well as disinformation, leading to what has been defined as an “infodemic” alongside the pandemic (1–3). This context saw the emergence of pandemic-related social media elites, accounts with a large number of followers that regularly discuss the pandemic and the issues surrounding it (4, 5). These actors play a central role in public communication, as they can shape popular sentiment and public discourse and thus potentially influence political decision-making (6). This is especially true in a setting characterized by increasing polarization and historically low trust in mainstream news, which allows politically and financially motivated actors to emerge (7–10). Because of this,

understanding the role that popular social media users play and the ways in which they operate is crucial for tackling arising challenges in public communication (11). In this article, we tackle this issue with the aim of drawing an explanatory map of the network of COVID-19 Twitter elites. We first identify users that are popular in the discussion related to the pandemic on Twitter^a, and go on to study their (directed) network, where an edge between two actors is present if one follows the other. To analyze the resulting network structure we make use of latent space models, which postulate that nodes in the network are embedded in a latent social space, where the probability for two actors to connect is inversely related to their distance within the space (12). We, in particular, make use of the latent cluster random effects model, which incorporates model-based clustering, allowing it to identify cohesive communities in the network, as well as additional nodal parameters to account for actor-specific heterogeneity in the propensity to form edges (13). The results suggest that the network can be

Competing Interest: The authors declare no competing interest.

Received: October 16, 2022. **Revised:** May 15, 2023. **Accepted:** July 21, 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of National Academy of Sciences. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

partitioned into two macro-communities. By focusing on a number of notable users, such as politicians, activists, and news outlets, we show how the two communities can be interpreted as “generally pro” and “generally against” pandemic containment measures and vaccine mandates. This finding supports the extensive body of literature that demonstrates the existence of significant polarization on social media (14–18). The central role of polarization has also been demonstrated for the specific case of pandemic-related online conversations, especially with respect to opinions on vaccination (19–23). However, our results also demonstrate how polarization, while prevalent, is far from being the only driver in the network. The continuous latent space enables us to see that substantial within-cluster heterogeneity is present: not all users in the two communities have the same opinions, and not just the two polar opposites are represented. On the contrary, a full spectrum of beliefs between the two poles is found. In particular, more radical users are found to be positioned towards the extremes of the latent space, while more moderate and neutral actors, such as health ministers and news outlets, are closer to the center. These central users thus occupy a uniquely powerful position, as they can act as a bridge between the two communities and thereby mitigate polarization. In addition to these results, our analysis demonstrates how, by making use of latent space models, it is possible to accurately map the COVID-19 Twitter landscape by only modeling information on who follows whom within the elite network. This finding highlights the strength and the pervasiveness of echo chamber effects on the platform, and showcases the power of latent space network models for studying communication on social media.

Data and methods

Identifying the network of COVID-19 Twitter elites

Social media elites can be broadly understood as users with the ability to influence (24). The term typically refers to a group of highly influential and popular users with considerable reach who significantly impact conversations, trends, and narratives circulating on social media. These users often include celebrities, politicians, journalists, thought leaders and influencers, who have a large and engaged audience and are frequently retweeted, quoted, and mentioned by others. While informative, this characterization is quite broad and does not indicate a unique way of identifying elites in practice. Operational definitions for empirical applications are often based on engagement metrics, such as the number of followers of each user, and engagement metrics, such as likes, shares, quotes, and replies. As our focus lies on analyzing the behavior of actors who actively engage in the discussion of the pandemic and that exert significant influence on the conversation, we here choose to identify elites as those who authored the most popular tweets, where the popularity of a tweet is given by the sum of its likes, replies, and retweets (including quotes).

Table 1. Structure of the analyzed dataset. Only columns relevant to our study are displayed.

Tweet ID	Author	Likes	Replies	Retweets
138712...	AnikaBlub	1,162	61	53
135224...	goetageblatt	1	2	1
140697...	galottom	1	0	0
146632...	1_FCM	171	26	35
135269...	covid_watch	0	0	0
...

Based on this characterization, we will therefore first need to identify popular tweets discussing COVID-19 and then relate those tweets to their authors. More motivation and details on this choice, as well as robustness checks, are included in the [supplementary material](#). For our study, we make use of the COVID-19 Twitter dataset published by Banda et al. (25), which comprises IDs of tweets containing pandemic-related keywords from January 1st, 2020 onward. These keywords were handpicked and continuously tracked to provide a global and real-time overview of the chatter related to the COVID-19 pandemic. The dataset was collected using Twitter’s streaming API, which allows free access to a random 1% sample of publicly available tweets in real time (26). At the time of the analysis, the entire dataset contains about 1.32 billion tweet IDs, representing both tweets and retweets in all languages, 340 million of which are unique (without retweets). Each tweet’s creation time and language are also provided. Using the tweet IDs, we are then able to recover additional information on the tweets, such as the text, the author, and metrics such as likes and retweets counts.

As a global platform, Twitter is host to speakers of many different languages, which induce the formation of largely separate communities. Since our goal is to map the latent space of COVID-19 elites, we choose to limit our analysis to a single language, as doing otherwise would return a fragmented map shaped mainly by language. In principle, it is possible to work with any single language, and we here opt for using tweets in German. The choice is motivated by the combination of two facts: Firstly, German is predominantly spoken by people from Germany, and to a smaller extent from Austria and parts of Switzerland, thereby guaranteeing a reasonable degree of geographical homogeneity. This prevents the estimated latent positions of the actors (and the resulting clusters) from being predominantly driven by their geographical locations. Secondly, German is used by a relevant proportion of the Twitter user base, allowing for a more than sufficient sample size. As the first COVID-19 vaccines started to be available to the public towards the very end of 2020, and given that one of the points we are most interested in investigating is attitude towards vaccination, we limit our sample to 2021 only, spanning from January 1st to December 31st. Considering all tweets in German from 2021 results in a total of 1.51 million unique tweets from 184,406 accounts. The data, sketched in Table 1, allow us to pinpoint popular users by looking at the authors of tweets with the highest interaction metrics. More

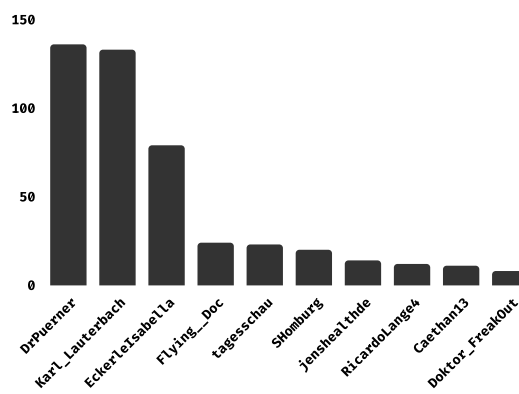


Fig. 1. Number of tweets authored by the 10 most popular users in our sample.

specifically, we classify a user as elite if they authored a tweet that achieved a popularity score of at least 2000, where we define popularity as the sum of likes, replies, and retweets (including quotes) gathered. This threshold results in 1024 popular tweets spanning all months of 2021, with each month represented by 53–156 tweets. Those 1024 tweets were produced by 372 users, 31.7% of which were granted verified status by Twitter, meaning that the platform deemed them both authentic and of public interest (27). In contrast, only 2.4% of the user base in the initial sample was verified. This confirms that more notable accounts and public figures are, on average, more central to the discussion, as we would expect. The bar plot in Fig. 1 depicts the number of tweets authored by the top 10 most popular users in our final sample, displayed by their Twitter usernames. From it, it is apparent how certain actors play a very prominent role in the conversation, with some accounts having authored more than 100 popular tweets in our 1% sample, meaning that one can expect them to have as much as 100 times more than that overall. This tells us how truly influential elites can be on Twitter, and also indicates that, given the sheer amount of popular tweets by the most prominent accounts, it is quite likely that they will be captured in our 1% sample.

After pinpointing these accounts as COVID-19 elites, we are able to define their following network in a natural way. Specifically, we consider the users as the nodes, and establish that a (directed) edge from actor i to actor j is present if, at the time of the analysis, i follows j . After removing the only nine users with no connections, the resulting network is composed of 363 nodes connected by a total of 12,182 edges, and is visualized in Fig. 2. From the plot, it is immediately apparent that the network is quite dense: in fact, 9.2% of all possible edges are observed. Given that the network is composed of users who produced popular tweets about the same topic, the fact that many of them follow each other makes intuitive sense. Moreover, from the graph representation, laid out using a variant of the Yifan Hu force-directed graph drawing algorithm (28), the network seems to be approximately split into two main groups of different sizes. This already gives a first impression of the two main poles in the network, which will be investigated in more detail in the Results section.

Latent space models for social network data

To model the network data, we make use of the latent cluster random effects model for social networks (13). This model is part of

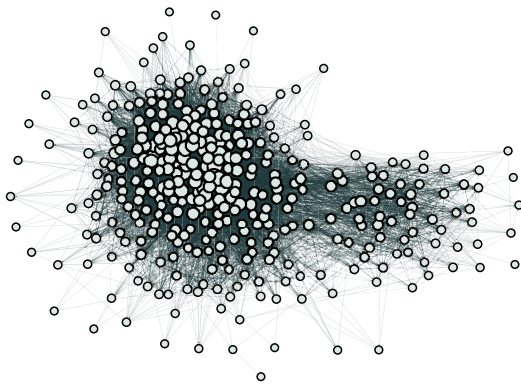


Fig. 2. Graphical representation of the network of COVID-19 elites on German-speaking Twitter.

the general family of latent space models, originating from the latent distance model proposed by Hoff et al. (12). Latent space network models postulate that each actor has an unobserved position in a d -dimensional Euclidean latent social space, and that the probability for two actors to form an edge is inversely related to their distance in the space. This family of models is particularly suitable for social networks, in which mechanisms such as homophily and triadic closure often play a major role (29). Handcock et al. (30) added the idea of model-based clustering to the original latent distance model, allowing for the actors' positions in the latent space to come from a mixture of normal distributions, where each mixture component represents a cluster. Krivitsky et al. (13) further extend this by adding nodal random effects to control for actor-specific heterogeneity in the propensity to form edges. More precisely, without the inclusion of nodal or edgewise covariates, the model specifies the probability of an edge y_{ij} between nodes i and j through:

$$\begin{aligned} \text{logit}(\mathbb{P}(y_{ij} = 1 | \beta_0, \mathbf{Z}, \boldsymbol{\delta}, \boldsymbol{\gamma})) \\ = \beta_0 - \|\mathbf{z}_i - \mathbf{z}_j\| + \delta_i + \gamma_j, \end{aligned} \quad (1)$$

where $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ are the latent positions of the nodes in the d -dimensional latent space, β_0 is an intercept, and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)$ and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n)$ are node-specific sender and receiver effects that account for the individual users' propensity of following or being followed, respectively. Here, the latent positions \mathbf{Z} are assumed to originate from a finite spherical multivariate mixture of independent normal distributions, and the random effects $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ are assumed to be drawn independently from normal distributions with mean 0 and variances σ_δ^2 and σ_γ^2 , respectively. The model is estimated through the R package `latentnet`, which implements a Bayesian routine based on the use of a Markov chain Monte Carlo algorithm (31). It is interesting to note that this model can be viewed as a generalization of the (latent) fitness model for networks (32, 33), as the node-specific random effects δ_i and γ_i can be seen as measuring the intrinsic fitness of node i to send and receive ties, while its latent position \mathbf{z}_i affects its probability of forming ties differently for each (potential) connection.

Homophily and triadic closure are generally prevalent in social media, particularly on Twitter and between popular accounts (34, 35). Those mechanisms often lead to the formation of subgroups of actors based on shared beliefs or other characteristics. Identifying such clusters can be helpful in understanding the drivers of polarization and, more in general, grouping behavior. The general task of identifying assortative, tightly knit groups in networks is a large area of research, known under the umbrella term of "community detection" (36). Notable examples of such methods include modularity maximization algorithms (37) and stochastic blockmodels (38). Classical community detection techniques are well suited for finding group structures, but they have the drawback of only returning a discrete partition of the network into clusters, where the connectivity behavior of each actor is fully described by its group label. In other words, two nodes in the same group are considered identical in all aspects. This is generally quite simplistic for social networks, in which cohesive groups often do exist, but where members of each group can also be very different from one another. Within a single group, for example, some nodes might be more "extreme" and isolated from all other communities. In contrast, others might be more central to the network and have many connections to other groups. We expect this to be the case in our network of COVID-19 Twitter elites: While we can assume polarization and grouping behavior to be present, we also expect the social positioning and political

beliefs of the actors to be more accurately described through a continuous, multidimensional spectrum rather than with discrete labels. Because of that, we are not only interested in the clear-cut grouping of nodes but also in uncovering the (continuous) social positioning of the users relative to one another. The chosen latent cluster random effects model is, therefore, particularly well suited for our application, as it combines clustering and latent position modeling, thereby enabling us to simultaneously capture polarization and grouping behavior as well as the positioning of the actors relative to each other in the socio-political spectrum.

Results

We fit the latent cluster random effects model to our data, setting both the number of clusters k and the number of dimensions d to 2. The choice of two clusters is backed by the approximated Bayesian Information Criterion for data-driven model selection proposed by Handcock et al. (30). Moreover, since much of the literature concerns itself with investigating polarization in the online discussion revolving around the COVID-19 pandemic, and given that polarization suggests the existence of two subgroups (39), setting $k = 2$ appears to be the natural choice from a substantive perspective. With regards to the choice of d , while dimensionality for latent space network models is generally an open question, setting $d = 2$ is considered to be the standard for applications in which interpretability of the positions is central, as it simplifies the visualization and description of social relationships (40). We also experimented with different values of d and observed that using higher dimensionality did not greatly impact the cluster assignments.

The results of the model fitting are visualized in Fig. 3. The axes correspond to the two latent dimensions Z_1 and Z_2 , respectively, and the nodes' colors indicate the estimated community memberships. More specifically, the node-specific pie charts represent the posterior probabilities for each user to belong to the one or the other cluster. Node sizes are scaled by each actor's total degree within the network. Note that, as defined by the model, two nodes that are closer to one another have a higher probability of forming an edge, i.e. of following each other. Also note that estimates of the node-specific random effects γ and δ , incorporating information on how active specific nodes are with respect to following or being followed, are made available in the [supplementary materials](#). At first glance, we see that the two communities are distributed along the horizontal axis Z_1 , with the more numerous blue community occupying the left and center parts of the figure, and the orange one being located towards the right-hand side. Moreover, from the posterior membership probabilities we can see that group memberships are fairly clear for most nodes. Nonetheless, significant uncertainty can be observed for a non-negligible proportion of the actors, which lie in between the two clear communities in the space.

As our task is of unsupervised nature, we do not have a set-in-stone "ground truth" with which to compare the model-based labeling and the estimated positions of the actors. To interpret the results, we therefore need to dig into the data and consider the emerging patterns. As the network is limited in size, and thanks to the naturally high propensity of elite users to voice their opinions, it is relatively straightforward to identify some of the more prominent actors and gauge their views on pandemic-related governmental interventions based on public information. Through this process, we can appreciate how the latent position of each actor in the network is strongly associated with their public stances on government mandates. More

specifically, despite substantial within-cluster heterogeneity in stances (and their intensity) on several issues, users in the blue community tend to hold views that can be summarized as "generally for" interventions and vaccine mandates. The opposite is true for actors in the orange community, which can be described as "generally against" such measures. Moreover, the positioning of nodes within communities is also informative on the actors' beliefs, capturing the within-cluster heterogeneity mentioned. Specifically, more central (external) positions in the overall latent space are associated with more moderate (extreme) stances. To showcase these patterns, we highlighted and labeled some notable users in Fig. 3, where each user is indicated with their Twitter username. The very center of the space is occupied by the most popular actors, most of whom, despite having connections to both groups thanks to their "elite among elites" status, reside firmly in the blue camp: A prime example is Karl_Lauterbach, an exponent of the Social Democratic Party who, at the time of writing, has been serving as the health minister of Germany since December 8th, 2021. He is known to be a strong proponent of vaccination and mandatory vaccination for all (41). Two other notable members of this group are Christian Drosten (c_drosten), a prominent virologist who has been described by major media outlets as "the country's real face of the coronavirus crisis" and "the nation's corona-explainer-in-chief" (42), and Melanie Brinkmann (BrinkmannLab), another well-known virologist who was among the proponents of the No-COVID strategy (43). Moving a bit further left in the space, another very popular user in the network is Flying_Doc, a medical doctor who has been outspoken in his support for policy proposals such as a vaccine mandate for all adults, and allowing access to events only to people who are both fully vaccinated and tested ("1G+" in the German political jargon). Looking even more toward the left on the Z_1 dimension, we encounter positions that are increasingly more in the direction of decisive government interventions. Examples of this are dr_heartbreaker, a medical professional who has expressed his support for hard lockdowns and the aforementioned No-COVID strategy, and NavomDienst and Doktor_Freakout, two anonymous medical doctors who also vehemently voiced their dissent for what they deemed to be bland policy making, and vouched their support for stronger restrictions. To conclude our outlook on the blue community, we also labeled two more peripheral, less Twitter-popular nodes. On the bottom-left of the plot we find Danzickler, an intensive care doctor who also expressed his support for more decisive action by the government, while on the top left we find MuttivsFaschos, who tweeted at the hashtags #ZeroCovid and #harderLockdownJetzt ("harder lockdown now"). All in all, our analysis highlights how users categorized in the blue group generally tend to openly support governmental efforts to contain the pandemic, and that the estimated dimension Z_1 is associated with the intensity of the actors' voiced stances on policy.

We now shift our focus to the orange community, composed of actors who have, on average, significantly fewer followers within this elite network, and tend to more or less strongly oppose pandemic-related government mandates. We start our overview with DrPuermer, the user with the highest number of popular tweets in our dataset. A medical doctor, Puermer rose to prominence during the pandemic for his stark criticism of COVID measures and opposition to government mandates. While not downplaying the dangers posed by COVID-19, he attracted following and praise from conspiracy theorists and the populist right-wing party "Alternative for Germany" ("AfD"), notorious for its antisystem beliefs (44). Closer to DrPuermer in the latent space

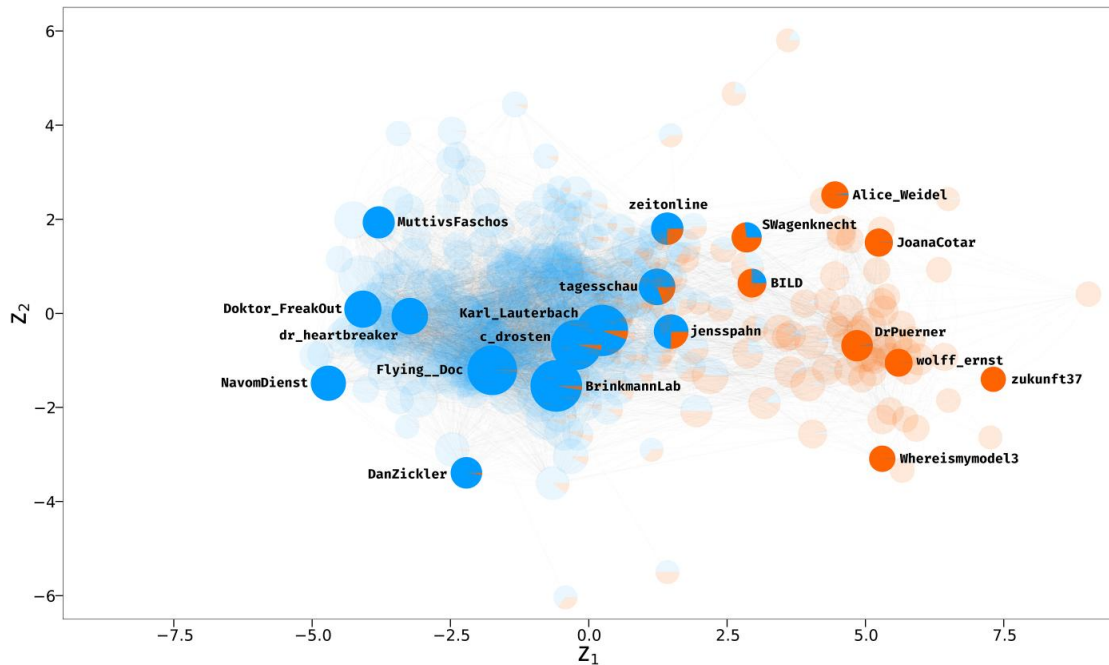


Fig. 3. Graphical representation of the latent positions of the actors in the network of COVID-19 Twitter elites estimated via the latent cluster random effects model, where the node size for each actor is scaled by its degree. A number of notable users are highlighted. The axes correspond to the two latent dimensions Z_1 and Z_2 , while the estimated posterior probabilities for each user to belong to the "pro vaccine mandates" (blue) or "anti compulsory vaccination" (orange) cluster are depicted through the node-specific pie charts. Major German media outlets are found between the two communities.

we can also find *wolff_ernst*, a self-described journalist and writer, who has openly associated himself with COVID-related and general conspiracy theories (45). We also labeled two more peripheral nodes in this cluster, namely users *zukunfft37* and *whereismymodel3*, anonymous accounts who openly voice their vaccine skepticism and opposition to government mandates. Two elected members of the aforementioned AfD, namely *Alice_Weidel*, who has been the leader of the party in the Bundestag (German Federal Parliament) since October 2017, and *JoanaCotar*, another member of the Bundestag who was part of AfD for the whole studied period and until late 2022, are also part of the orange community. Unsurprisingly, the two are close in the latent space, reflecting their similar policy stances. Perhaps more surprisingly, their estimated latent positions are not far from that of *Sahra Wagenknecht* (*swagenknecht*), member of the Bundestag for "The Left" ("Die Linke") since 2009, and former parliamentary leader of that same party. Despite being on the other end of the political spectrum, she also opposes general vaccination mandates (46). She is located more towards the middle of the plot and has substantial uncertainty in her community membership, with a posterior probability of approx. 75% to belong to the orange community. Another actor whose community membership is uncertain is Christian Democratic Union politician *Jens Spahn*, who served as health minister for most of the analyzed period, i.e. until December 8th, 2021 (*jensspahn*). He is not far in the space from his successor *Karl Lauterbach* but lies a bit more on the right: He is classified in the blue community but has a posterior probability of approximately 25% to belong to the orange one. This is in line with the fact that, while he is a proponent of widespread vaccination, he is opposed to the idea of compulsory

vaccination for all (47). To conclude our overview of the space, we highlight some other notable accounts located in between the two clusters, namely those belonging to prominent news outlets. Given that we expect them to have a diverse following due to their authority status, their central positioning makes intuitive sense. But even between media outlets, the model is able to draw a distinction: *zeitonline* and *tagesschau*, generally reputable news sources, are closer to the center of the space, and, although with substantial uncertainty, labeled as blue. On the other hand, *BILD*, the most prominent German boulevard newspaper, is located more towards the right, and has a higher probability of belonging to the orange group.

Discussion

In this article, we identified and modeled the network of users leading the conversation revolving around the COVID-19 pandemic on Twitter. More specifically, we made use of the latent cluster random effects model to map these elite users into a 2D Euclidean social space, in which users that are closer to each other have a higher likelihood to connect, i.e. to follow each other. The results suggest the emergence of a natural partition of the network into two dense macro-communities, which are only loosely connected with their opposing counterparts. By focusing on a number of notable users, such as politicians, activists, and news outlets, we show how those two communities can be interpreted as "generally pro" and "generally against" public interventions and vaccine mandates. This finding corroborates recent research demonstrating the polarized nature of pandemic-related online discourse, especially concerning vaccination (19–23). But a deeper inspection of

the latent space further reveals that users within communities are only partially homogeneous in their stances. To the contrary, the model is able to uncover a nuanced, continuous spectrum of pandemic-related beliefs and policy positions, ranging from demanding radical containment measures all the way to vaccine skepticism and COVID-denying conspiracy theories, covering everything in between those two extremes. In this context, neutral actors, such as mainstream news outlets, are positioned between the two clusters, which makes intuitive sense given their authority status. From the latent positions of users whose political inclination is known, we can also appreciate how attitudes toward governmental interventions tend to follow political inclination, with left- and right-wing respectively corresponding to more favorable or unfavorable positions towards restrictions and vaccine mandates. This finding echoes recent research showing how ideology can shape trust in scientists and attitudes towards vaccines (48, 49). The importance of vaccination as a subtheme within the pandemic-related discussion is corroborated by the fact that “vaccine” is one of the words appearing more often in the data (while not being used as a filtering mechanism), as shown in the [supplementary material](#) (Fig. S3).

A particular feature of the employed methodology is the ability to combine “classical” community detection, which alone would be insufficient to gain a proper understanding of the network at hand, with more refined, continuous latent space modeling. This allows to map the underlying latent social space with the necessary nuance while simultaneously returning a partition of the network into subgroups, which can be useful for understanding the network at a coarser resolution, or for classification purposes. The modeling results thus allow us to obtain a clearer picture of the network as a whole and can be used for garnering insight on single (politically unaffiliated) users.

We note that the studied network is fairly small as a result of the relatively restrictive popularity threshold we chose for defining a popular tweet: It would thus be possible to decrease the threshold to obtain a larger network. We also note, however, that using a lower value somehow “loosens” the definition of an elite, as users that are less popular on average would make it into the network. Experimenting with the threshold, we also observed that using different values almost only impacts the size of the network’s periphery and does not change the overall picture. Results of alternative analyses with different threshold values and inclusion criteria are provided in the [supplementary material](#) (Figs. S1 and S2) and corroborate the robustness of our findings. Moreover, a stricter definition of elites incidentally makes the network size more manageable, which is relevant given that model estimation, as it is currently implemented in the R package `latentnet`, only scales well up to a few thousand nodes. Nonetheless, while latent space models do pose serious computational challenges, different approaches to estimate them for larger networks have been proposed (50, 51). We also note that, as we here only model the behavior of elites, we cannot *a priori* assume our results to be valid for the overall discussion. While, given the well-documented strong influence of popular users in the conversation, it is reasonable to believe that many of the results could extend to the general Twitter population, further research would be needed to confirm this. Furthermore, there may be different patterns in how elite and nonelite actors follow other users. For example, whereas nonelites are likely to use their follows primarily instrumentally, i.e. to see tweets they are interested in on their timeline, elites could also use theirs for signaling, i.e. to publicly show support or endorsement towards other users, and may thus curate their follows more carefully. Similarly, highly active

elites could be more likely than nonelites to enter conflicts with each other and block opposing elites. On the one hand, these strategic follows (or nonfollows) are indeed relevant to our analysis, as they give information on the potential factions at play in the network, and aid us in identifying them. On the other hand, as a result of these mechanisms, polarization in the elite network may be higher than in the complete one. The latter consideration strengthens the notion that polarization, although undoubtedly present to some extent, is not the only determining factor in network formation, and that the different groups exist on a continuous spectrum rather than being completely isolated from one another.

We further emphasize that our approach is purely unsupervised and completely based on network structure, without including any element of natural language processing. In other words, this means that the two groups emerge only from using information on who follows whom. In this sense, we could have simply labeled the two clusters as “blue” and “orange”, or “left” and “right”. The description of the communities with respect to their attitudes towards vaccination, and, more in general, pandemic management, was done after the modeling, to shed some additional light on the data-driven cluster selection, and alternative characterizations would also be viable. While it would certainly be possible to make use of the tweets’ text content to obtain further insight into the users, we here explicitly chose to focus solely on the network component, thus demonstrating how tightly the users’ personal networks are intertwined with their beliefs. Indeed, given that the latent positions of the actors are estimated by the model solely using their follows and followers within the network, it is quite remarkable how consistently actors neighboring each other in the estimated latent space are also near in their stances on COVID-19 and its management, and how closely the space is able to track the belief spectrum. The echo chamber effect is well documented in the literature: Users tend to follow those who share similar ideas, and are thus rarely exposed to contrasting views. This, in turn, leads the users’ beliefs to become self-reinforcing (52, 53). However, our analysis demonstrates how this behavior is not only prevalent at the extremes of the socio-political spectrum but also towards the center of the belief space. On the one hand, the phenomenon implies that users with radical ideas will tend to follow people with similarly extreme beliefs, leading to further polarization; On the other hand, it also means that users following more moderate voices will also tend to gravitate towards more nuanced views. Central actors, which have the ability to act as a bridge between the two communities, are thus uniquely positioned to mitigate the polarization loop.

The fact that following behavior is so closely related to beliefs and attitudes paves the way for latent space models as powerful tools for drawing maps of social media landscapes, which can, in turn, be used to increase our understanding of the underlying social and behavioral structures. Indeed, while we here applied the methodology to map the discussion revolving around COVID-19, it is possible to perform similar types of analysis on other topics of public relevance. Given its explanatory and predictive power, we believe latent space modeling of elite social media networks to have the potential for improving our general understanding of the online landscape, ultimately aiding policymakers in making more informed decisions in their quests against polarization and misinformation worldwide.

Note

^aAt the time of publication, the Twitter social media platform is in the process of rebranding to “X”.

Acknowledgments

This manuscript was posted on arXiv as a preprint: [arXiv:2207.13352](https://arxiv.org/abs/2207.13352).

Supplementary material

[Supplementary material](#) is available at PNAS Nexus online.

Funding

This research was partially funded by the Elite Network of Bavaria (ESG Data Science).

Author contributions

G.D.N., V.H.T.M. and G.K. designed research; G.D.N. and V.H.T.M. performed research; G.D.N. and V.H.T.M. analyzed data; G.D.N., V.H.T.M. and G.K. wrote the paper.

Data availability

The network data used in this article and the code to reproduce the analysis are publicly available on our GitHub repository, at <https://github.com/gdenicola/latent-space-covid-twitter-elites>.

References

- Cinelli M, et al. 2020. The COVID-19 social media infodemic. *Sci Rep.* 10:16598.
- Zarocostas J. 2020. How to fight an infodemic. *Lancet.* 395(10225): 676.
- Gollust SE, Nagler RH, Fowler EF. 2020. The emergence of COVID-19 in the US: a public health and political communication crisis. *J Health Polit Policy Law.* 45(6):967–981.
- Gallagher RJ, Doroshenko L, Shugars S, Lazer D, Welles BF. 2021. Sustained online amplification of COVID-19 elites in the United States. *Soc Media Soc.* 7(2). doi:10.1177/205630512111024957
- Molyneux L, McGregor SC. 2021. Legitimizing a platform: evidence of journalists' role in transferring authority to Twitter. *Inf Commun Soc.* doi:10.1080/1369118X.2021.1874037
- Leader AE, Burke-Garcia A, Massey PM, Roark JB. 2021. Understanding the messages and motivation of vaccine hesitant or refusing social media influencers. *Vaccine.* 39(2):350–356.
- Finkel EJ, et al. 2020. Political sectarianism in America. *Science.* 370(6516):533–536.
- Fink K. 2019. The biggest challenge facing journalism: a lack of trust. *Journalism.* 20(1):40–43.
- Bridgman A, et al. 2020. The causes and consequences of COVID-19 misperceptions: understanding the role of news and social media. *Harvard Kennedy School Misinformation Review.* doi:10.37016/mr-2020-028
- Donovan J. 2020. Social-media companies must flatten the curve of misinformation. *Nature.* doi:10.1038/d41586-020-01107-z
- Johnson NF, et al. 2020. The online competition between pro- and anti-vaccination views. *Nature.* 582(7811):230–233.
- Hoff PD, Raftery AE, Handcock MS. 2002. Latent space approaches to social network analysis. *J Am Stat Assoc.* 97(460): 1090–1098.
- Krivitsky PN, Handcock MS, Raftery AE, Hoff PD. 2009. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Soc Netw.* 31(3):204–213.
- Caldarelli G, De Nicola R, Del Vigna F, Petrocchi M, Saracco F. 2020. The role of bot squads in the political propaganda on Twitter. *Commun Phys.* 3(1):81.
- Conover M, et al. 2011. Political polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 5(1):89–96.
- Garimella VRK, Weber I. 2017. A long-term analysis of polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media.* 11(1):528–531.
- Del Vicario M, et al. 2016. Echo chambers: emotional contagion and group polarization on Facebook. *Sci Rep.* 6(1):37825.
- Del Vicario M, Zollo F, Caldarelli G, Scala A, Quattrociocchi W. 2017. Mapping social dynamics on facebook: the brexit debate. *Soc Networks.* 50:6–16.
- Jiang X, et al. 2021. Polarization over vaccination: ideological differences in Twitter expression about COVID-19 vaccine favorability and specific hesitancy concerns. *Soc Media Soc.* 7(3). doi:10.1177/205630512111048413
- Reiter-Haas M, Klösch B, Hadler M, Lex E. 2022. Polarization of opinions on COVID-19 measures: integrating Twitter and survey data. *Soc Sci Comput Rev.* doi:10.1177/08944393221087662
- SteelFisher GK, Blendon RJ, Caporello H. 2021. An uncertain public—encouraging acceptance of Covid-19 vaccines. *N Engl J Med.* 384(16):1483–1487.
- Cowan SK, Mark N, Reich JA. 2021. COVID-19 vaccine hesitancy is the new terrain for political division among Americans. *Socius.* 7: 237802312110236.
- Mønsted B, Lehmann S. 2022. Characterizing polarization in online vaccine discourse—a large-scale study. *PLoS One.* 17(2): e0263746.
- Dubois E, Gaffney D. 2014. The multiple facets of influence: identifying political influentials and opinion leaders on Twitter. *Am Behav Sci.* 58(10):1260–1277.
- Banda JM, et al. 2021. A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration. *Epidemiologia.* 2(3):315–324.
- Twitter, Volume streams [accessed 2022 Oct 16]. <https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/introduction>
- Edgerly S, Vraga EK. 2019. The blue check of credibility: does account verification matter when evaluating news on Twitter? *Cyberpsychol Behav Soc Netw.* 22(4):283–287.
- Hu Y. 2005. Efficient, high-quality force-directed graph drawing. *Math J.* 10(1):37–71.
- Rivera MT, Soderstrom SB, Uzzi B. 2010. Dynamics of dyads in social networks: assortative, relational, and proximity mechanisms. *Annu Rev Sociol.* 36:91–115.
- Handcock MS, Raftery AE, Tantrum JM. 2007. Model-based clustering for social networks. *J R Stat Soc A (Stat. Soc.).* 170(2):301–354.
- Krivitsky PN, Handcock MS. 2008. Fitting position latent cluster models for social networks with latentnet. *J Stat Softw.* 24(5):1–23.
- Caldarelli G, Capocci A, De Los Rios P, Munoz MA. 2002. Scale-free networks from varying vertex intrinsic fitness. *Phys Rev Lett.* 89(25):258702.
- Bianconi G, Barabási A-L. 2001. Competition and multiscaling in evolving networks. *Europhys Lett.* 54(4):436–442.
- Lou T, Tang J, Hopcroft J, Fang Z, Ding X. 2013. Learning to predict reciprocity and triadic closure in social networks. *ACM Trans Knowl Discov Data.* 7(2):1–25.
- Colleoni E, Rozza A, Arvidsson A. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *J Commun.* 64(2):317–332.

- 36 Fortunato S, Hric D. 2016. Community detection in networks: a user guide. *Phys Rep.* 659:1–44.
- 37 Newman MEJ. 2006. Modularity and community structure in networks. *Proc Natl Acad Sci USA.* 103(23):8577–8582.
- 38 De Nicola G, Sischka B, Kauermann G. 2022. Mixture models and networks: the stochastic blockmodel. *Stat Modelling.* 22(1–2):67–94.
- 39 Guerra P, Meira Jr W, Cardie C, Kleinberg R. 2013. A measure of polarization on social media networks based on community boundaries. *Proceedings of the International AAAI Conference on Web and Social Media.* 7(1):215–224.
- 40 Sosa J, Betancourt B. 2022. A latent space model for multilayer network data. *Comput Stat Data Anal.* 169:107432.
- 41 Zimmermann K. 2022. Karl Lauterbach sieht gescheiterte Impfpflicht als herbe Niederlage. *ZeitOnline* [accessed 2022 Oct 16]. <https://zeit.de/politik/deutschland/2022-04/karl-lauterbach-impfpflicht-niederlage-impfkampagne>.
- 42 Henley J. 2020. Coronavirus: meet the scientists who are now household names. *The Guardian* [accessed 2022 Oct 16]. <https://theguardian.com/world/2020/mar/22/coronavirus-meet-the-scientists-who-are-now-household-names>.
- 43 Baumann M, et al. 2021. Eine neue proaktive Zielsetzung für Deutschland zur Bekämpfung von SARS-CoV-2. ifo Institute [accessed 2022 Oct 16]. <https://ifo.de/en/publikationen/2021/monograph-authorship/proaktive-zielsetzung-bekaempfung-sars-cov-2-handlungsoptionen>.
- 44 Stoepler T. 2021. Thesen vom Amtsarzt. *Sueddeutsche Zeitung* [accessed 2022 Oct 16]. <https://sueddeutsche.de/bayern/bayern-corona-amtsarzt-friedrich-puerner-buch-1.5466845>.
- 45 Ayyadi K. 2021. Wenn ein selbsterklärter “Ökonom” mit Antisemitismus Corona erklären will. *Belltower News* [accessed 2022 Oct 16]. <https://belltower.news/youtube-wenn-ein-selbsterklaeter-oekonom-mit-antisemitismus-corona-erklaren-will-97409/>.
- 46 Wagenknecht S. 2022. Deutsche Politik hat sich bei der Impfpflicht verrannt. *FOCUS Online* [accessed 2022 Oct 16]. <https://focus.de/politik/deutschland/weitergedacht/weitergedacht-die-wagenknecht-kolumne-deutsche-politik-hat-sich-bei-der-impfpflicht-verrannt-id/40754360.html>.
- 47 Kubitz M. 2021. Spahn will im Bundestag nicht für allgemeine Impfpflicht stimmen. *BR24* [accessed 2022 Oct 16]. <https://br.de/nachrichten/deutschland-welt/spahn-will-im-bundestag-gegen-allgemeine-impfpflicht-stimmen,SqWKYhF>.
- 48 Featherstone JD, Bell RA, Ruiz JB. 2019. Relationship of people’s sources of health information and political ideology with acceptance of conspiratorial beliefs about vaccines. *Vaccine.* 37(23):2993–2997.
- 49 Kossowska M, Szwed P, Czarnek G. 2021. Ideology shapes trust in scientists and attitudes towards vaccines during the COVID-19 pandemic. *Group Process Intergr Relat.* 24(5):720–737.
- 50 Raftery AE, Niu X, Hoff PD, Yeung KY. 2012. Fast inference for the latent space network model using a case-control approximate likelihood. *J Comput Graph Stat.* 21(4):901–919.
- 51 Yin J, Ho Q, Xing EP. 2013. A scalable approach to probabilistic latent space inference of large-scale networks. *Adv Neural Inf Process Syst.* 26:422–430.
- 52 Cinelli M, De Francisci Morales G, Galeazzi A, Quattrociocchi W, Starnini M. 2021. The echo chamber effect on social media. *Proc Natl Acad Sci USA.* 118(9):e2023301118.
- 53 Nguyen CT. 2020. Echo chambers and epistemic bubbles. *Episteme.* 17(2):141–161.

Part III.

Modeling and monitoring epidemics

9. Nowcasting fatal COVID-19 infections on a regional level in Germany

Contributing article

Schneble, M., De Nicola, G., Kauermann, G., and Berger, U. (2021). Nowcasting fatal COVID-19 infections on a regional level in Germany. *Biometrical Journal*, 63(3):471–489. <https://doi.org/10.1002/binj.202000143>.

Data and code

Available at <https://github.com/MarcSchneble/Nowcasting-Fatal-COVID-19-Infections>.

Copyright information

This article is distributed under a Creative Commons 4.0 International license (CC BY 4.0).

Author contributions

The idea of modeling fatal COVID-19 infections to effectively monitor the course and the severity of the pandemic can be attributed to Giacomo De Nicola and Göran Kauermann. Marc Schneble had the idea of nowcasting fatal infections, and Göran Kauermann substantiated the respective model. Giacomo De Nicola and Marc Schneble then drafted the manuscript together, with significant contributions from Göran Kauermann and Ursula Berger. More specifically, Giacomo De Nicola especially contributed by writing Sections 1, 2, 3, 4.1, 5 and 7. Giacomo De Nicola further supported Marc Schneble in the implementation of the model in the R language and in the data visualization. All authors contributed through fruitful comments and extensive proofreading of the manuscript.

Nowcasting fatal COVID-19 infections on a regional level in Germany

Marc Schneble¹ | Giacomo De Nicola¹ | Göran Kauermann¹ | Ursula Berger²

¹ Department of Statistics,
Ludwig-Maximilians-University Munich,
Munich, Germany

² Institute for Medical Information
Processing, Biometry, and Epidemiology,
Ludwig-Maximilians-University Munich,
Munich, Germany

Correspondence

Marc Schneble, Department of Statistics,
Ludwig-Maximilians-University Munich,
Ludwigstr. 33, 80539 Munich, Germany.
Email: marc.schneble@stat.
uni-muenchen.de



This article has earned an open data badge
“**Reproducible Research**” for mak-
ing publicly available the code necessary
to reproduce the reported results. The
results reported in this article could fully be
reproduced.

Abstract

We analyse the temporal and regional structure in mortality rates related to COVID-19 infections, making use of the openly available data on registered cases in Germany published by the Robert Koch Institute on a daily basis. Estimates for the number of present-day infections that will, at a later date, prove to be fatal are derived through a nowcasting model, which relates the day of death of each deceased patient to the corresponding day of registration of the infection. Our district-level modelling approach for fatal infections disentangles spatial variation into a global pattern for Germany, district-specific long-term effects and short-term dynamics, while also taking the age and gender structure of the regional population into account. This enables to highlight areas with unexpectedly high disease activity. The analysis of death counts contributes to a better understanding of the spread of the disease while being, to some extent, less dependent on testing strategy and capacity in comparison to infection counts. The proposed approach and the presented results thus provide reliable insight into the state and the dynamics of the pandemic during the early phases of the infection wave in spring 2020 in Germany, when little was known about the disease and limited data were available.

KEYWORDS

COVID-19, disease mapping, generalized regression model, nowcasting

1 | INTRODUCTION

In March 2020, COVID-19 became a global pandemic. From Wuhan, China, the virus spread across the whole world, and with its diffusion more and more data became available to scientists for analytical purposes. In daily reports, the WHO provides the number of registered infections as well as the daily death toll globally (<https://www.who.int/>). It is inevitable for the number of registered infections to depend on the testing strategy in each country (see, e.g., Cohen & Kupferschmidt, 2020). This has a direct influence on the number of undetected infections (see, e.g., Li et al., 2020), and first empirical analyses aim to quantify how detected and undetected infections are related (see, e.g., Niehus, De Salazar, Taylor, & Lipsitch, 2020). Though similar issues with respect to data quality hold for the reported number of fatalities

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

(see, e.g., Baud et al., 2020), the number of deaths can overall be considered a more reliable source of information than the number of registered infections. The results of the 'Heinsberg study' in Germany point in the same direction (Streeck et al., 2020). A thorough analysis of death counts can in turn generate insights on changes in infections as proposed in Flaxman et al. (2020) (see also Ferguson et al., 2020). In this paper, we pursue the idea of directly modelling registered death counts related to COVID-19 instead of registered infections. In other words, we restrict our analysis to fatal COVID-19 cases only, omitting recovered or symptom-free infections. We analyse data from Germany and break down the analyses to a regional level. Such regional view is apparently immensely important, considering the local nature of some of the outbreaks, for example in Italy (see, e.g. Grasselli, Pesenti, & Cecconi, 2020; Grasselli, Zangrillo, & Zanella, 2020), France (see, e.g., Massonnaud, Roux, & Crépey, 2020) or Spain and can assist local health authorities in monitoring the disease and planning infection control measures.

The analysis of fatalities has, however, an inevitable time delay and requires to take the course of the disease of COVID-19 patients into account. In particular, in this paper we consider the timespan between the registration of the infection through local health authorities and the report of its deadly outcome by the Robert Koch Institute (RKI). A first approach on modelling and analysing the time from illness and onset of symptoms to reporting and further to death is given in Jung et al. (2020) (see also Linton et al., 2020). Understanding the delay between onset and registration of an infection and, for severe cases, the time between registered infection and death, can be of vital importance. Knowledge on those timespans allows us to obtain estimates for the number of infections that are expected to be fatal based on the number of infections registered on the present day. The statistical technique to obtain such estimates is called nowcasting (see, e.g., Höhle & an der Heiden, 2014) and traces back to Zeger, See, and Diggle (1989) or Lawless (1994). Nowcasting in COVID-19 data analyses is not novel and is, for instance, used in Günther, Bender, Küchenhoff, Katz, and Höhle (2020) for nowcasting daily infection counts in Germany, that is to adjust daily reported new infections to include infections which occurred the same day but were not yet reported. Altmejd, Rocklöv, and Wallin (2020) apply nowcasting techniques to Swedish data and Bird and Nielsen (2020) provide nowcasting fatalities in English hospitals. We extend this approach to model the duration between the registration date of an infection and its fatal outcome, accounting for additional covariates. To do so, we combine a nowcasting model with a spatio-temporal regression model.

We analyse the number of fatal cases of COVID-19 infections in Germany using district-level data. The data are provided by the RKI (www.rki.de), the German federal government agency and scientific institute responsible for health reporting, disease control and prevention in humans. They report the cumulative number of deaths in different gender and age groups for each of the 412 administrative districts in Germany, together with the date of registration of the infection. The data are available in dynamic form through daily downloads of the updated cumulated numbers of deaths. Comparing two consecutive daily downloads allows to construct a new dataset which contains both the date at which a COVID-19 disease is registered and the date at which a fatality is reported to the RKI, with the latter usually being reported at a later time point. We employ flexible statistical models with smooth components (see, e.g., Wood, 2017), assuming the district-specific number of fatalities to be negatively binomial distributed, which permits to also account for possible overdispersion in the data. The spatial structure in the death rate is incorporated in two ways: First, we assume a spatial correlation of the number of deaths by including a long-range smooth spatial death intensity. This allows to map a general pattern of the spread of the disease over Germany, which shows that regions of Germany are affected to different extents. On top of this long-range effect, we include two types of unstructured region-specific effects. An overall region-specific effect reflects the situation of a district as a whole, while a short-term effect mirrors region-specific variations of fatalities over time and captures local outbreaks as happened in, for example Heinsberg (North-Rhine-Westphalia) or Tirschenreuth (Bavaria). This effect can be seen as an unstructured time-space interaction. In addition to the spatial components, we include an overall temporal effect to capture dynamic changes in the number of fatal infections for Germany. The latter effect mirrors the overall flattening of the infectious situation in the considered time period, that is spring 2020. Besides the spatio-temporal character, our modelling approach further adjusts for the district-specific age and gender structure.

Modelling infectious diseases is a well-developed field in statistics, and we refer to Held, Meyer, and Bracher (2017) for a general overview of the different models. We also refer to the powerful R package *surveillance* (Meyer, Held, & Höhle, 2017). Since our focus is on analysing district-specific dynamics, both structured and unstructured, as well as dynamic behaviour of fatal infections, we prefer to make use of generalized additive regressions implemented in the *mgcv* package in R, which also allows to decompose the spatial component in more depth.

The paper is organized as follows. In Section 2 we describe the data. Section 3 introduces our model, while Section 4 discusses the necessity of incorporating a nowcasting model. Section 5 shows the results of our analysis which are then refined to subgroups of the data in Section 6. Section 7 concludes the paper by also discussing the limitations of our modelling exercise.

TABLE 1 Illustration of the data structure, showing downloads of the data from April 25 and April 26, 2020 as an example. To facilitate reproducibility, the original column names used in the RKI datasets are given in brackets below our English notation

	District (Landkreis)	Age Group (Altersgruppe)	Gender (Geschlecht)	Infections (Anzahl Fall)	Fatal Infections (Anzahl Todesfall)	Registration Date (Meldedatum)	Reporting Date (Datenstand)
Data downloaded on April 25, 2020	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Munich City	60–79	F	3	1	April 22, 2020	April 25, 2020
	Munich City	60–79	M	5	1	April 22, 2020	April 25, 2020
	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Data downloaded on April 26, 2020	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Munich City	60–79	F	6	2	April 22, 2020	April 26, 2020
	Munich City	60–79	M	5	1	April 22, 2020	April 26, 2020
	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2 | DATA

We make use of the COVID-19 dataset (Esri Deutschland GmbH, 2020) provided by the RKI on a daily basis for the 412 districts in Germany (which also include the 12 districts of Berlin separately). The data are collected by the RKI, but originate from the district-based health authorities (*Gesundheitsämter*). Due to different population sizes in the districts, and certainly also because of different local situations, some health authorities transmit the daily numbers to the RKI with a delay. This happens in particular over the weekend, a fact that we need to take into account in our model. We have daily downloads of the data since March 27, 2020. We here choose to focus on a phase of the COVID-19 pandemic in which the death toll in Germany was high. The subsequent analysis was thus conducted with data up to May 14, 2020, and was performed considering only deadly infections with registration dates from March 26, 2020 until May 13, 2020 (the day before that of the analysis).

Table 1 illustrates an exemplary extract of the data that are available. For each of the 412 districts, the data contain the cumulated number of laboratory-confirmed COVID-19 infections as well as the cumulated number of deaths related to COVID-19 for each district of Germany, stratified by age group (15–34, 35–59, 60–79 or 80+), gender, and the date of registration of the infection by the local public health authorities. The time stamp for a fatal outcome always refers to the registration date of the infection and *not* to the individual’s date of death. Therefore, the numbers in the column ‘Fatal infections’ cannot exceed the numbers shown in the column ‘Infections’. Even though the time point of infection obviously precedes that of death, registration of an infection can also occur after death, for example when a post-mortem test is conducted, or when test results arrive after the patient has passed away. In the former case, the registration date are set to the day of death by the local health authority. Also note that it is not indicated in the dataset whether a fatal infection resulted from a post-mortem test, and that no information on whether the patient has died *with* or *because* of a COVID-19 infection is included.

The cumulative numbers are reported on a daily basis by the RKI, which is mirrored in the column ‘Reporting date’ in Table 1. The reporting date always corresponds to the query date and the download date of the data. In Table 1, we see that the number of reported infections with registration date April 22, 2020, which relate to females in the age group 60–79 living in the city of Munich, increases by three from April 25, 2020 to the following day. In the same period, the number of fatal infections increased by one. Thus, we can deduce that three registered infections in this sub-population were reported with a delay of 4 days. The single newly reported fatal infection belongs to an individual of this sub-population for which the time between registration by the local health authorities and reported death amounts to 4 days. In this paper, we are especially interested in the latter quantity, which we model as a duration time. It is of importance to note that we can derive such information only due to daily downloads of the dataset, which are not being provided retrospectively.

We refrain from providing general descriptive statistics on the spatio-temporal distribution of confirmed COVID-19 infections here, since these numbers are already visualized on the RKI dashboard (Robert Koch-Institut, 2020; see also StaBLab, LMU Munich, 2020). However, the number of fatal infections is less often taken into account. Thus, in Figure 1 we show the empirical duration between the day of registration as COVID-19 infected by the local health authorities and the day on which the death has been reported by the RKI (based on the data until May 14, 2020). Due to the aforementioned reporting delay, the minimum duration is 1 day. Note that these plots show stapled bar charts, highlighting the counts by gender. We see that considerably more fatal infections originate from the age group 80+. Regarding the age group 80– (aggregated age groups 15–39, 40–59 and 60–79), we see that males are much more affected than females, whereas in the age group 80+ the counts are more balanced. Finally, in both age groups there are a small number of deaths,

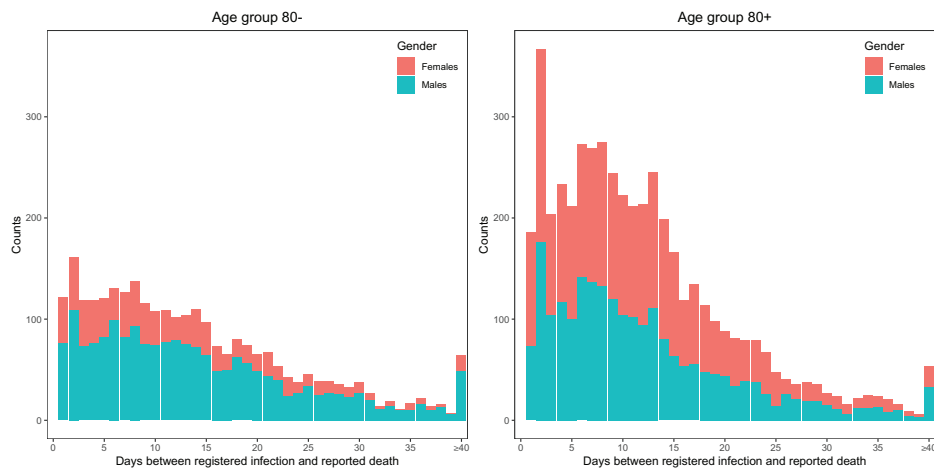


FIGURE 1 Stacked bar chart of the counts of fatal infections depending on days between registered infection and reported death. Only data reported until May 14, 2020 is considered here (left panel: age group 80– (less than 80 years), right panel: age group 80+ (80 years or older))

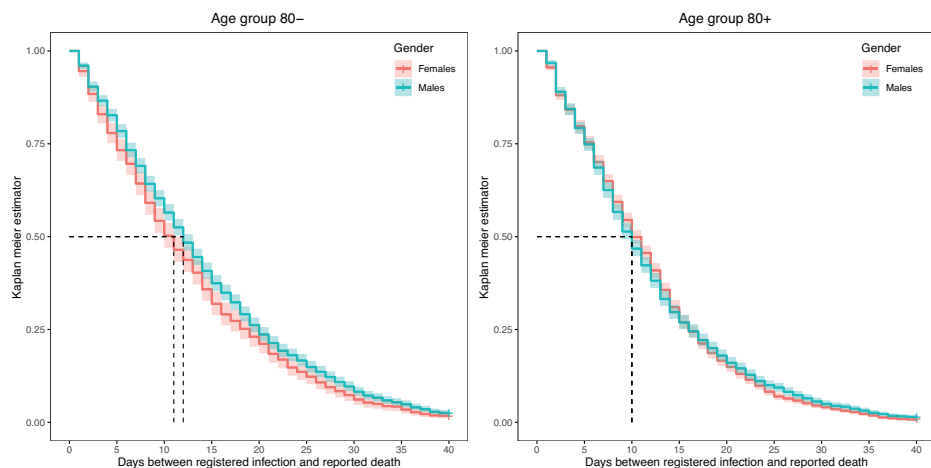


FIGURE 2 Kaplan–Meier estimators of the data shown in Figure 1 with 95% confidence intervals

which were reported 40 or more days after the registration of the COVID-19 infection. Kaplan–Meier estimators of the duration between registered infection and reported death are shown in Figure 2 for age groups 80– and 80+ by gender. Here we especially see that the median duration time of elderly patients is slightly shorter when compared to the younger age groups.

3 | MORTALITY MODEL

Let $Y_{t,r,c}$ denote the number of deaths due to COVID-19 with time point of registration $t = 0, \dots, T$ in district/region r and cohort c , where the cohort c is characterized by age group and gender of the deceased. Time index $t = T$ corresponds to the day of analysis, which is May 14, 2020, and $t = 0$ corresponds to March 26, 2020. Not all fatalities with registered infection at time point t have been observed at time T , as some deaths will occur later. We therefore need a model for nowcasting, which is discussed in the next section.

For now, we assume all $Y_{t,r,c}$ to be known. A family of discrete distributions which is supported on the set of nonnegative integers and also allows to account for possible overdispersion in the data is the negative binomial distribution. Therefore, we model those numbers as according to

$$Y_{t,r,c} \sim \text{NB}(\lambda_{t,r,c}, \phi), \tag{1}$$

where $\mathbb{E}(Y_{t,r,c}) = \lambda_{t,r,c}$ and the constant dispersion parameter ϕ relates to the variance by $\text{Var}(Y_{t,r,c}) = \lambda_{t,r,c} + \phi\lambda_{t,r,c}^2$. We model the mean $\lambda_{t,r,c}$ of the response $Y_{t,r,c}$ through a regression model and specify

$$\begin{aligned} \lambda_{t,r,c} = & \exp\{\beta_0 + \text{age}_c\beta_{\text{age}} + \text{gender}_c\beta_{\text{gender}} \\ & + \text{age}_c\text{gender}_c\beta_{\text{age, gender}} + \text{weekday}_r\beta_{\text{weekday}} \\ & + m_1(t) + m_2(s_r) + u_{r0} + \mathbb{1}_{\{t \geq T-14\}}u_{r1} + \log(\text{pop}_{r,c})\}, \end{aligned} \tag{2}$$

where the linear predictor is composed as follows:

- β_0 is the intercept.
- β_{age} and β_{gender} are the age- and gender-related regression coefficients, and $\beta_{\text{age, gender}}$ is the coefficient that models the interaction between age and gender.
- β_{weekday} are regression coefficients, which relate to the weekday of the registration date as COVID-19 infected.
- $m_1(t)$ is an overall smooth time trend, with no prior structure imposed on it.
- $m_2(s_r)$ is a smooth spatial effect, where s_r is the geographical centroid of district/region r .
- u_{r0} and u_{r1} are district-/region-specific random effects, which are independently and identically distributed (i.i.d.) and follow a normal prior probability model. While u_{r0} specifies an overall level of the death rate for district r over the entire observation time, u_{r1} is a spatio-temporal effect that reveals region-specific dynamics by allowing the regional effects to differ for the last 14 days.
- $\text{pop}_{r,c}$ is the gender and age group-specific population size in district/region r and serves as an offset in our model.

We here emphasize that we fit spatial effects of different types: We model a smooth spatial effect, that is $m_2(s_r)$, which takes the correlation between the fatal infections of neighbouring districts/regions into account and gives a global overview of the spatial distribution of fatal infections. In addition to that we also have unstructured district-/region-specific effects $\mathbf{u}_r = (u_{r0}, u_{r1})^\top$, which capture local behaviour related to single districts only. While u_{r0} captures the corresponding long-term effect, u_{r1} captures the short-term effect of the last 14 days; see (2). This means that we also model a dichotomous and unstructured interaction of space with time. The district-specific effects \mathbf{u}_r are considered as random, with prior structure

$$\mathbf{u}_r \sim \mathcal{N}(\mathbf{0}, \Sigma_{ii}) \text{ i.i.d} \tag{3}$$

for $r = 1, \dots, 412$. The prior variance matrix Σ_{ii} is estimated from the data. The predicted values $\hat{\mathbf{u}}_r$ (i.e. the posterior mode) exhibit districts that show unexpectedly high or low death tolls when adjusted for the global spatial structure and for age- and gender-specific population sizes.

While model (2) is complex and highly structured, note that no autoregressive components are included in the linear predictor in (2). We will demonstrate in Section 6.4 below that auto-correlation is of negligible size, and that time dependence is fully captured by $m_1(t)$ as well as the unstructured effects u_{r1} .

The mortality model defined through (1) and (2) belongs to the model class of generalized additive mixed model (see, e.g., Wood, 2017). The smooth functions are estimated by penalized splines without restrictions on the number of degrees of freedom, with a quadratic penalty that can be comprehended as a normal prior (see, e.g., Wand, 2003). The same type of prior structure holds for the region-specific random effects \mathbf{u}_r . In other words, smooth estimation and random effect estimation can be accommodated in one fitting routine, which is implemented in the R package `mgcv`. This package has been used to fit the model, so that no extra software implementation was necessary. This demonstrates the practicability of the proposed method. Our analysis is completely reproducible, with code and data openly available and downloadable from our GitHub repository.¹

¹<https://github.com/MarcSchneble/Nowcasting-Fatal-COVID-19-Infections>

4 | NOWCASTING MODEL

4.1 | Model description

The above model cannot be fitted directly to the available data, since we need to take the course of the disease on the individual level into account. This means that the final number of fatal outcomes for infections registered on date $t < T$ is not known at the time point of analysis $t = T$, since not all patients with a fatal outcome of the disease have died yet. This requires the implementation of nowcasting. Due to the sparsity of the data, we perform the nowcast on a national level, that is we cumulate the numbers over district/region r . For reasons of notation, we temporarily drop the gender and age-related subscript g , and we simply notate the cumulated number of deaths with registered infections at day t with Y_t .

Let $N_{t,d}$ denote the number of deaths reported on day $t + d$ for infections registered on day t . Assuming that the true date of death is at $t + d$, or at least close to it, we ignore any time delays between time of death and its notification to the health authorities. We call d the duration in days between the registration date as a COVID-19 patient and the reported day of death, where $d = 1, \dots, d_{\max}$. Here, d_{\max} is a fixed reasonable maximum duration, which we set to 40 days (see, e.g., Wilson, Kvalsvig, Barnard, & Baker, 2020). This is also motivated by the means of Figure 1. The minimum duration is one day, since the RKI daily reports the new numbers, which they have received from the public health departments the day before. In nowcasting, we are interested in the cumulated number of deaths for infections registered on day t , which we define as

$$Y_t = \sum_{d=1}^{d_{\max}} N_{t,d}.$$

Therefore, the total number of deaths with a registered infection at t becomes available only after d_{\max} days. In other words, only after d_{\max} days we know exactly how many deaths occurred due to an infection which was registered on day t . We define the partial cumulated sum of deaths as

$$C_{t,d} = \sum_{l=1}^d N_{t,l}$$

so that by definition $C_{t,d_{\max}} = Y_t$.

On day $t = T$, when the nowcasting is performed, we are faced with the following data constellation, where NA stands for not (yet) available:

t	d				Reported deaths
	1	2	...	d_{\max}	
0	$N_{0,1}$	$N_{0,2}$...	$N_{0,d_{\max}}$	Y_0
1	$N_{1,1}$	$N_{1,2}$...	$N_{1,d_{\max}}$	Y_1
⋮	⋮	⋮	⋮	⋮	⋮
$T - d_{\max}$	$N_{T-d_{\max},1}$	$N_{T-d_{\max},2}$...	$N_{T-d_{\max},d_{\max}}$	$Y_{T-d_{\max}}$
$T - d_{\max} + 1$	$N_{T-d_{\max}+1,1}$	$N_{T-d_{\max}+1,2}$...	NA	$C_{T-d_{\max}-1,d_{\max}-1}$
⋮	⋮	⋮	⋮	⋮	⋮
$T - 2$	$N_{T-2,1}$	$N_{T-2,2}$	NA	NA	$C_{T-2,2}$
$T - 1$	$N_{T-1,1}$	NA	NA	NA	$C_{T-1,1}$

We may consider the timespan between registered infection and (reported) death as a discrete duration time taking values $d = 1, \dots, d_{\max}$. Let D be the random duration time, which by construction is a multinomial random variable. In principle, for each death we can consider the pairs (D_i, t_i) as i.i.d. and we aim to find a suitable regression model for D_i given t_i , including potential additional covariates $x_{t,d}$. We make use of the sequential multinomial model (see Agresti, 2010) and define

$$\pi(d; t, x_{t,d}) = P(D = d | D \leq d; t, x_{t,d}).$$

Let $F_t(d)$ denote the corresponding cumulated distribution function of D which relates to probabilities $\pi(\cdot)$ through

$$\begin{aligned}
 F_t(d) &= P_t(D \leq d) = P(D \leq d | D \leq d + 1) \cdot P(D \leq d + 1) \\
 &= (1 - \pi(d + 1; \cdot)) \cdot (1 - \pi(d + 2; \cdot)) \cdot \dots \cdot (1 - \pi(d_{\max}; \cdot)) \\
 &= \prod_{k=d+1}^{d_{\max}} (1 - \pi(k; \cdot))
 \end{aligned}
 \tag{4}$$

for $d = 1, \dots, d_{\max} - 1$ and $F_t(d_{\max}) = 1$.

We generalize notation again by including the subscript g , which in the nowcasting model only distinguishes between the two age groups 80– and 80+. The available data on cumulated death counts now allow us to estimate the conditional probabilities $\pi(d; \cdot)$ for $d = 2, \dots, d_{\max}$. In fact, the sequential multinomial model allows to look at binary data such that

$$N_{t,d,c} \sim (\text{quasi-})\text{Binomial}(C_{t,d,c}, \pi(d; t, c, x_{t,d}))
 \tag{5}$$

with

$$\text{logit}(\pi(d; t, c, x_{t,d})) = s_1(t) + s_2(d) + s_3(d) \cdot \mathbb{1}_{\text{age}\{80+\}} + x_{t,d}\gamma,
 \tag{6}$$

where

- $s_1(t)$ is an overall smooth time trend over calendar days.
- $s_2(d)$ is a smooth duration effect, capturing the course of the disease.
- $s_3(d)$ is a varying smooth duration effect, capturing interaction between the dynamics of the disease and age, particularly for the age group 80+. Note that with effect $s_3(d)$ we take into account that for infections with a fatal outcome, the individual course of the disease for elderly patients might differ compared to younger patients.
- $x_{t,d}$ are covariates which may be time and duration specific.

By utilizing a quasi-likelihood model (Fahrmeir, Kneib, Lang, & Marx, 2007) as in (5), we account for possible overdispersion in the data, which results in adjusted standard errors of the parameter estimates, while, however, the estimates themselves are the same when compared to the fit of a binomial model.

Assuming that D , the duration between a registered fatal infection and its reported death, is independent of the number of fatal COVID-19 infections, we obtain the relationship

$$\mathbb{E}(C_{t,d,c}) = F_{t,c}(d) \cdot \mathbb{E}(Y_{t,c}).
 \tag{7}$$

Note further that if we model $Y_{t,c}$ with a negative binomial model as presented in the previous section, we have no final observation $Y_{t,c}$ for time points $t > T - d_{\max}$. Instead, we have observed $C_{t,T-t,c}$, which relates to the mean of $Y_{t,c}$ through (7) by $C_{t,T-t,c} = F_{t,c}(T - t) \cdot \mathbb{E}(Y_{t,c})$. Including therefore $\log F_{t,c}(T - t)$ as additional offset in model (2) allows to fit the model as before, but with the nowcasted number of fatal infections included. That means, instead of $\lambda_{t,r,c}$ as in (2), the expected number of fatal infections are now parameterized by $\lambda_{t,r,c}^* = \lambda_{t,r,c} \exp(\log F_{t,c}(T - t))$, where the latter multiplicative term is included as additional offset in the model.

4.2 | Results for nowcasting

We fit the nowcasting model (5) with parameterization (6). We include a weekday effect for the registration date of the infection with reference category ‘Monday’. The estimates of the fixed linear effects are shown in Table 2. The fitted smooth effects are shown in Figure 3. The top panel shows the effect over calendar time, which is very weak and confirms that the individual course of the disease hardly varies over time. This is supported by the fact that the German healthcare

TABLE 2 Estimated fixed linear effects (standard errors in brackets) in the nowcasting model (6). Parameters and their standard errors are given on the log scale. The relative risk is given together with 95% confidence intervals. The reference for the weekdays is Monday

	Effect (SE)	exp(Effect) Relative risk	95% Confidence interval of relative risk
Intercept	-3.12 (0.045)	0.04	[0.04, 0.05]
Tuesday	0.06 (0.060)	1.06	[0.94, 1.19]
Wednesday	0.11 (0.059)	1.12	[0.99, 1.25]
Thursday	0.20 (0.058)	1.23	[1.09, 1.38]
Friday	0.26 (0.059)	1.30	[1.16, 1.45]
Saturday	0.27 (0.063)	1.31	[1.16, 1.48]
Sunday	0.20 (0.068)	1.22	[1.07, 1.40]

system remained stable over the considered period, and hence survival did not depend on the date on which the infection was notified.

The bottom panel of Figure 3 shows the course of the disease as a smooth effect over the time between registration of the infection and death. We see that the probabilities $\pi(d; \cdot)$ decrease in d , where this effect is the strongest in the first days after registration. Thus, most of the COVID-19 patients with fatal infections are expected to die not long after their registration date. We also see no overall significant difference in the duration effect between the age groups 80– and 80+, since the fitted curves $s_2(d)$ and $s_2(d) + s_3(d)$ hardly differ. To some extent, this was already visible from Figure 1. This shows that, given that a registered case ends with a fatal outcome, the individual's course of the disease does not depend on the age group. The effect of d becomes easier to interpret by visualizing the resulting distribution function $F_{t,c}(d)$, where here g refers to the age group 80+. This is shown in Figure 4 for two different values of t , that is April 13 and May 13. The plot also shows how the course of the disease hardly varies over calendar time: In fact, the small differences between the two distribution functions is dominated by the weekday effect, since the red curve is related to a Monday while the blue one is from a Wednesday.

4.3 | Nowcasted number of fatal infections

On the day of analysis, we do not observe the total counts of deaths for recently registered infections. This means that there are an unknown number of currently infected people which will die at a future point in time. We therefore nowcast those numbers, that is we predict the prospective deaths which can be attributed to all registration dates up to today. This is done on a national level, and the resulting nowcast of fatal infections for Germany is shown in Figure 5. For example, on May 14, 2020 there are 25 deaths reported where the infection was registered on May 5 (red bullets on May 5). We expect this number to increase to about 50 when all deaths due to COVID-19 for this registration date will have been reported (green triangles on May 5). Naturally, the closer a date is to the present, the larger the uncertainty in the nowcast will be. This is shown by the shaded bands. Details on how the statistical uncertainty has been quantified are provided below. In Section 5, we incorporate the nowcasting results into the mortality model as discussed before, but the nowcast results are interesting in their own right. The curve confirms that the number of fatal infections is decreasing since the beginning of April. Note that the curve also mirrors the 'weekend effect' in registration, as less infections are reported on Sundays.

Since we are now more than $d_{\max} = 40$ days after the day of analysis (May 14, 2020), we can assess the predictive accuracy of our nowcast. Therefore, we also show in Figure 5 the counts of fatal infections, which we observe 40 days after the respective registration date. We see that our nowcast performs in general very well. However, there are a handful of registration dates for which the nowcasted values were clearly outside of the prediction intervals. Most remarkably, the cumulative number of fatal infections for registered infections on April 8, 2020 has dropped after May 14, 2020. This happens in the rare case in which the database has been modified retrospectively by the local health authorities.

4.4 | Uncertainty quantification in nowcasting

In Figure 5, we have shown the nowcasting results along with uncertainty intervals shaded in grey. These were constructed using a bootstrap approach as follows. Given the fitted model, we simulate $n = 10,000$ times from the asymptotic joint

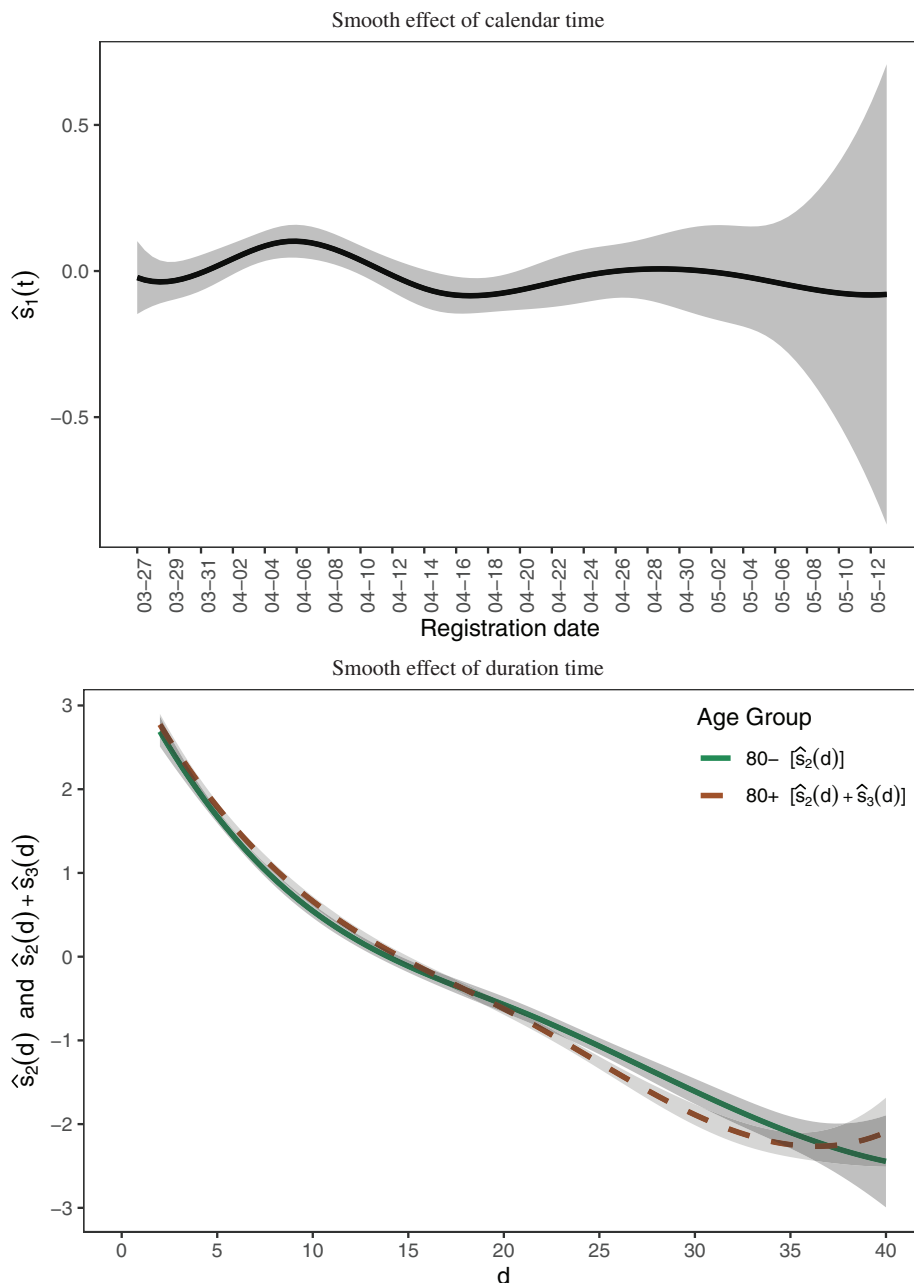


FIGURE 3 Estimates of smooth effects in the nowcasting model

normal distribution of the estimated model parameters which results through (4). This leads to a set of bootstrapped distribution functions $\mathcal{F} = \{\hat{F}_t^{(i)}(T - t), i = 1, \dots, n; t = T - d_{\max} + 1, \dots, T - 1\}$. This set is used to compute the simulated nowcasts $\hat{Y}_t^{(i)} = C_{t, T-t} / \hat{F}_t^{(i)}(T - t)$ applying (7), where $C_{t, T-t}$ is the observed partial cumulated sum of deaths at time point $T - t$ with registration date t . The point-wise lower and upper bounds of the 95% prediction intervals for the nowcast for Y_t are then given by the 2.5 and the 97.5 quantiles of the set $\{\hat{Y}_t^{(i)}, i = 1, \dots, n\}$, respectively.

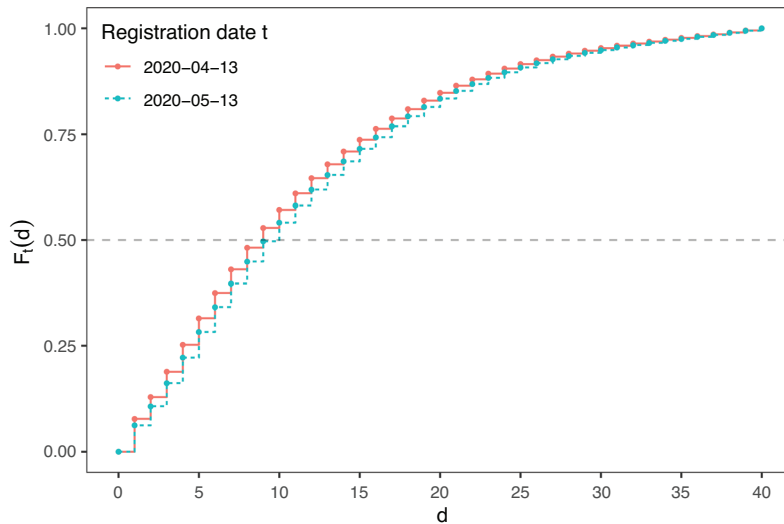


FIGURE 4 Fitted distribution function $F_t(d)$ for the age groups 80+ and 80–, where t corresponds to Wednesday, May 13, 2020

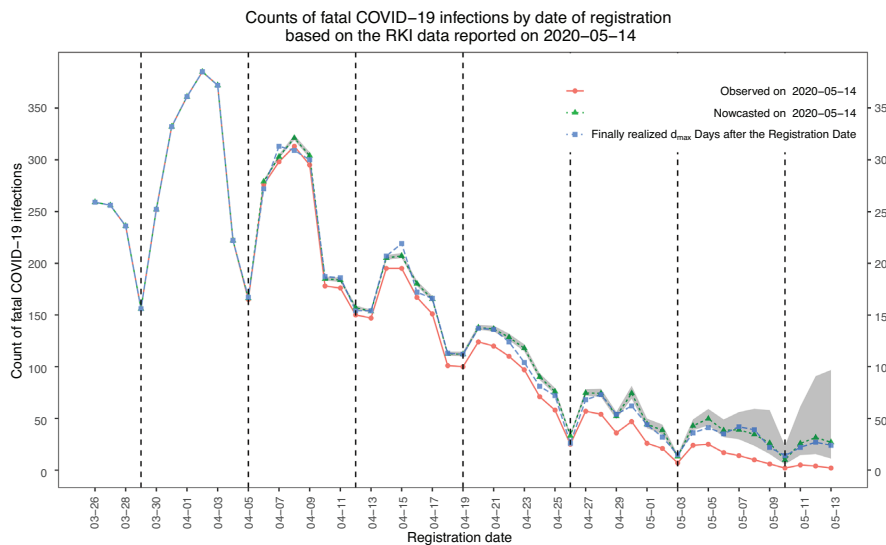


FIGURE 5 Observed (red line) and nowcasted (blue line) of daily death counts due to a COVID-19 infection on May 14, 2020 including 95% prediction intervals (shaded areas). Sundays are marked by a dashed vertical line. Finally realized death counts (d_{max} after the respective registration date) are shown as blue squares

5 | RESULTS OF THE MORTALITY MODEL

We first discuss the estimates of the fixed linear effects included in model (2), which are shown in Table 3. We see that both age and gender play a major role when estimating the numbers of fatal infections. Elderly people exhibit a much higher death rate from COVID-19, which is, for males (females) in the age group 80+, around 80 times ($148 \approx \exp(4.39 + 0.61)$ times) higher than in the reference age group 35–59. This already hints at a remarkable difference between genders, where the expected death rate for females in the reference age group is around 60% ($\approx 1 - \exp(-0.94)$) lower than the

TABLE 3 Estimated fixed linear effects (standard errors in brackets) in the mortality model (2). Parameters and their standard errors are given on the log scale. The relative risk is given together with 95% confidence intervals. The reference category for age is the age group 35–59. The reference for the weekdays is Monday

		Effect (S.E.)	exp(Effect) Relative risk	95% Confidence interval of relative risk
	Intercept	−15.90 (0.095)	$1.27 \cdot 10^{-7}$	$[1.05 \cdot 10^{-7}, 1.53 \cdot 10^{-7}]$
Patient related	Female	−0.94 (0.142)	0.39	[0.29, 0.51]
	Age 15–34	−2.53 (0.325)	0.08	[0.04, 0.15]
	Age 15–34 Female	−0.18 (0.674)	0.84	[0.22, 3.18]
	Age 60–79	2.61 (0.081)	13.58	[11.60, 15.90]
	Age 60–79 Female	0.07 (0.151)	1.07	[0.80, 1.45]
	Age 80+	4.41 (0.080)	81.9	[70.20, 95.90]
	Age 80+ Female	0.61 (0.147)	1.83	[1.38, 2.45]
Reporting related	Tuesday	0.20 (0.051)	1.22	[1.10, 1.35]
	Wednesday	0.23 (0.052)	1.26	[1.14, 1.39]
	Thursday	0.24 (0.050)	1.28	[1.16, 1.41]
	Friday	0.10 (0.051)	1.10	[1.00, 1.22]
	Saturday	−0.12 (0.054)	0.88	[0.79, 0.98]
	Sunday	−0.41 (0.058)	0.66	[0.59, 0.74]

corresponding death rate for males. When considering the total gender-related numbers of fatal infections in the age group 80+ (see Figure 1), the difference between the genders is seemingly very small. However, by respecting the district-, gender- and age-specific population sizes in our model we see that the death rate of females in the age group 80+ is still around 28% ($\approx 1 - \exp(-0.94 + 0.61)$) lower when compared to the male population in this age group. Furthermore, we see that significantly less deaths are attributed to infections registered on Sundays compared to weekdays, due to the existing reporting delay during weekends.

Our model includes a global smooth time trend representing changes in the death rate since March 26. This is visualized in Figure 6. The plotted death rate is scaled to give the expected number of deaths per 100,000 people in an average district for the reference group, that is males in the age group 35–59. Overall, we see a peak in the death rate on April 3 and a downwards slope until the end of April. However, our nowcast reveals that the rate remains constant since beginning of May. Note that such developments cannot be seen by simply displaying the raw death counts of these days. The nowcasting step inevitably carries statistical uncertainty, which is taken into account in Figure 6 by including best and worst case scenarios. The latter are based on bootstrapped confidence intervals, where details are provided in Section 6.3 later in the paper.

Our aim is to investigate spatial variation and regional dynamics. To do so, we combine a global geographic trend for Germany with unstructured region-specific effects, where the latter uncovers local behaviour. In Figure 7, we combine these different components and map the fitted nowcasted death counts related to COVID-19 for the different districts of Germany, cumulated over the last 14 days before the day of analysis, that is May 14, 2020. While in most districts of Germany, the death rate is relatively low, some hotspots can be identified. Among those, Traunstein and Rosenheim (in the south-east part of Bavaria) are the most evident, but Greiz and Sonneberg (east and south part of Thuringia) stand out as well, to mention a few. A deeper investigation of the spatial structure is provided in Section 6, where we show the global geographic trend and provide maps that allow to detect new hotspot areas, after correcting for the overall spatial distribution of the infection.

6 | MORE RESULTS AND MODEL EVALUATION

6.1 | Spatial effects

It is of general interest to disentangle the two spatial components that we introduced in Section 3. We visualize the fitted global geographic trend $m_2(\cdot)$ for Germany in Figure 8. The plot confirms that, up to mid May 2020, the northern parts

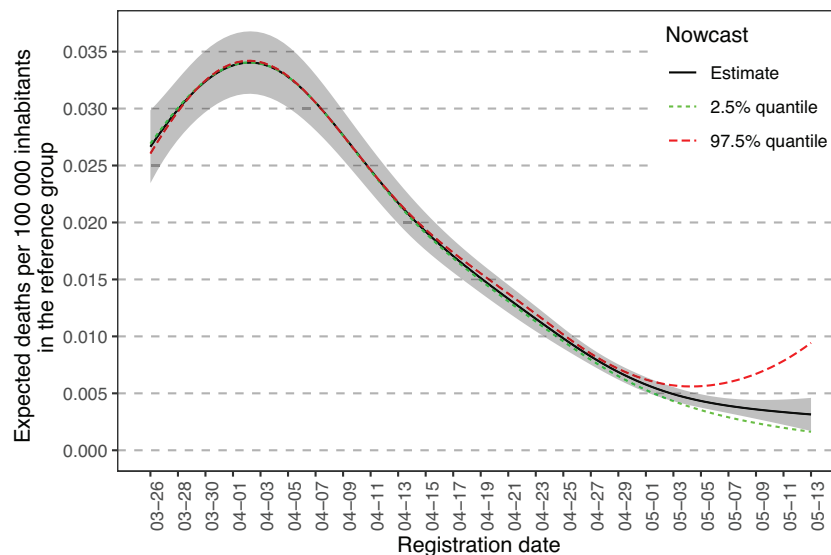


FIGURE 6 Fitted smoothed expected fatal COVID-19 infections per 100,000 inhabitants in the reference group (males aged between 35 and 59 in an average district) by registration date including 95% confidence bands as shaded area. Uncertainty resulting from the nowcast model is shown as dashed coloured lines

of the country are less affected by the disease in comparison to the southern states. The two plots in Figure 9 map the region-specific effects, that is the predicted long-term level of a district u_{r0} (left-hand side) and the predicted short-term dynamics u_{r1} (right-hand side). Both plots uncover quite some region-specific variability. In particular, the short-term dynamics u_{r1} (right plot) pinpoint districts with unexpectedly high nowcasted death rates in the last two weeks, after correcting for the global geographic trend and the long-term effect of the district. Some of the noticeable districts have already been highlighted in Section 3 above, but we can here detect further districts which are less evident in Figure 6: For instance, Steinfurt (in the north-west of North Rhine-Westphalia), Olpe (southern North Rhine-Westphalia) and Gotha (center of Thuringia) all show a relatively high rate of fatal infections.

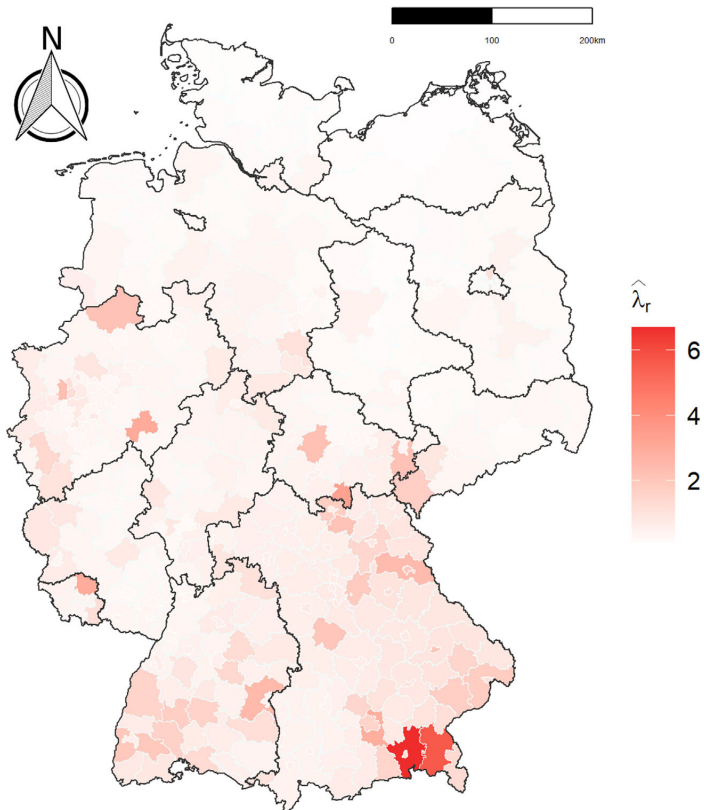
6.2 | Age group-specific analyses

A large portion of the registered fatal infections related to COVID-19 stems from people in the age group 80+. Locally, high numbers are often caused by an outbreak in a retirement home. Such outbreaks apparently have a different effect on the spread of the disease, and the risk of an epidemic infection caused by outbreaks in this age group is limited. Thus, the death rate among elderly people could vary differently across districts when compared to regional peaks in the death rate of the rest of the population. In order to respect this, we decompose the district-specific effects \mathbf{u}_r in (2) into $\mathbf{u}_r^{80-} = (u_{r0}^{80-}, u_{r1}^{80-})^\top$ for the age group 80– and $\mathbf{u}_r^{80+} = (u_{r0}^{80+}, u_{r1}^{80+})^\top$ for the age group 80+, where the age group 80– consists of the aggregated age groups 15–34, 35–59 and 60–79. We put the same prior assumption on the random effects as we did in (3), but now the variance matrix that needs to be estimated from the data has dimension 4×4 .

The fitted age group-specific random effects are shown in Figure 10, where the \mathbf{u}_r^{80-} are shown in the top panel and the \mathbf{u}_r^{80+} in the bottom panel. Most evidently, the variation of the random effects is much higher in the age group 80+ when compared to the younger age groups, as more districts occur which are coloured dark blue or dark red, respectively. When comparing the district-specific short-term dynamics of the last 14 days (u_{r1}) in Figure 10 to those in Figure 9, we recognize that in most of the districts which recently experienced very high death intensities (with respect to the whole period of analysis), these stem from the age group 80+. As mentioned before, this can often be explained by outbreaks in retirement homes.

FIGURE 7 Nowcasted fatal COVID-19 infections per 100,000 inhabitants in each district in the timespan from Thursday, April 30 until Wednesday, May 13, 2020

Nowcasted fatal infections per 100 000 inhabitants with registration dates from 2020-04-30 until 2020-05-13



Based on data reported up to 2020-05-14.
Model includes registration dates from
2020-03-26 until 2020-05-13.

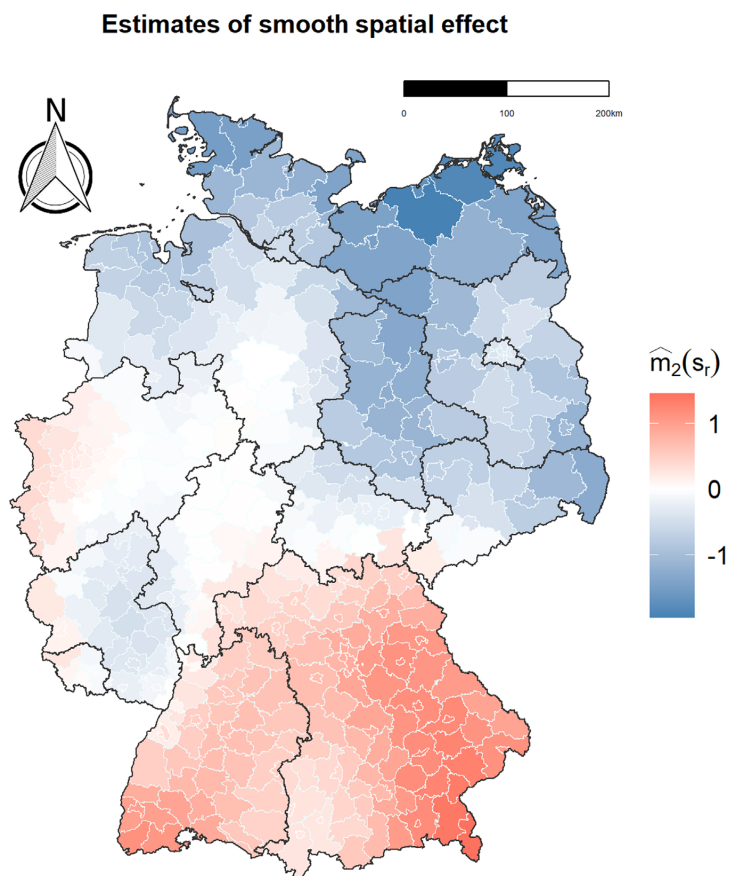
6.3 | Additional uncertainty in the mortality through the nowcast

When fitting the mortality model (1), we included the fitted nowcast model as offset parameter. This apparently neglects the estimation variability in the nowcasting model, which we explored via bootstrap as explained in Section 4.4 and visualized in Figure 5. In order to also incorporate this uncertainty in the fit of the mortality model, we refitted the model using (a) the upper end and (b) the lower end of the prediction intervals shown in Figure 5. It appears that there is little (and hardly any visible) effect on the spatial components, which is therefore not shown here. But the time trend shown in Figure 6 does change, which is visualized by including the two fitted functions corresponding to the 2.5% and 97.5% quantile of the offset function. We can see that the estimated uncertainty of the nowcast model mostly affects the last 10 days, with a strong potential increase in the death rate mirroring a possible worst case scenario.

6.4 | Auto-correlation of residuals in the mortality model

In the mortality model (2), we did not include an epidemic component accounting for possible temporal auto-correlation, as it is often done in endemic-epidemic models (see, e.g., Meyer et al., 2017). To check for possibly omitted auto-correlation

FIGURE 8 Smooth spatial effect of the death rate in Germany



Based on data reported up to 2020-05-14.
Model includes registration dates from
2020-03-26 until 2020-05-13.

in our model, we explore the temporal correlation of the Pearson residuals in the mortality model (2). To do so, we compute the auto-correlation function (ACF) for all lags $k = 0, \dots, T - 1$. The corresponding ACF plot is shown in Figure 11. Apparently, the results do not show any pattern of auto-correlation and support the suitability of our model. We emphasize, however, that infection dynamics are included in the model through the time trend $m_1(t)$. Moreover, even if we ignore possibly existing auto-correlation, this time trend $m_1(t)$ is still estimated unbiased with penalized spline smoothing, which is robust against misspecification of the auto-correlation structure (Krivobokova & Kauermann, 2007).

We also think that the epidemic component is generally less impactful when modelling fatal infections in comparison to modelling the number of registered infections. The time between person-to-person transmission of the virus and a fatal outcome of a COVID-19 infection is much larger than the time until the registration of the infection, as shown in Figure 1, and hence any auto-correlation is rather indistinct for fatal cases.

7 | DISCUSSION

The paper presents a general approach for monitoring the dynamic behaviour of COVID-19 infections on a small-area level purely based on the analysis of the number of observed death counts. This in turn means that the results are less dependent on testing strategies, which may vary by region and over time.

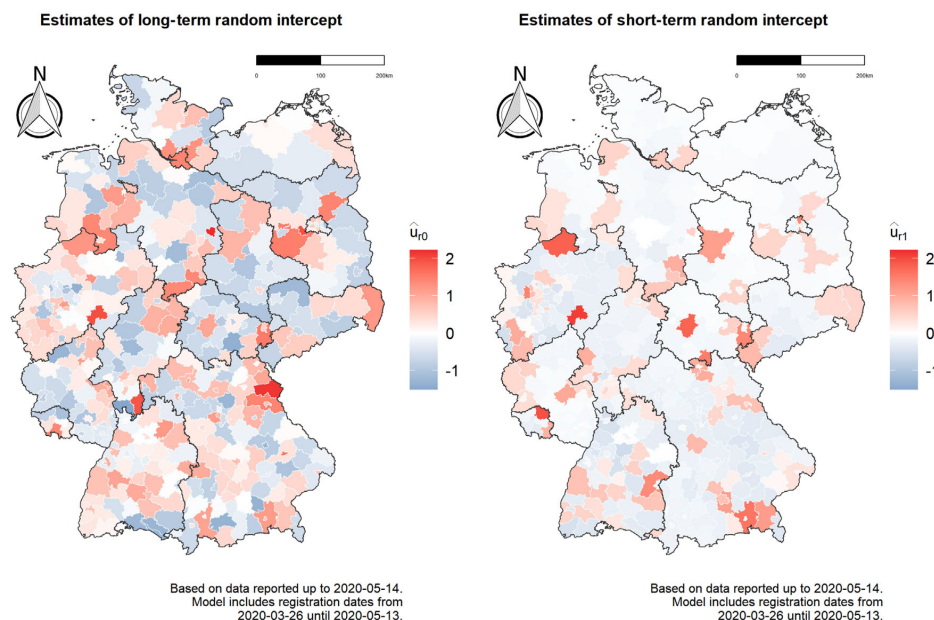


FIGURE 9 Region-specific long-term level (left-hand side) and short-term dynamics of the 14 days prior to May 14, 2020 (right-hand side) of fatal COVID-19 infections

In addition, patients with fatal infections typically require intensive medical care and are therefore relevant in the planning of clinical capacities of the local health system. An analysis of fatal infections is especially interesting in situations in which reliable information on hospitalization is not available, as in the considered timespan of the COVID-19 pandemic in Germany.

The described nowcasting approach enables us to estimate the number of deaths following a registered infection even if the fatal outcome has not occurred yet, providing an up-to-date picture of the situation. The results of the nowcasting model confirm that the individual course of the disease for fatal infections did not change over calendar time nor did it differ by gender. More in particular, it uncovers that in Germany, during the considered timespan, elderly patients had, in the case of fatal infections, about the same course of the disease as younger patients.

Our analysis of the nowcasted number of fatal infections on a regional level allows to draw conclusions on the current dynamics of the disease on the spatial dimension. By separately estimating, for each district, a long-range effect which mirrors the overall situation as well as a short-term dynamic effect, we can timely identify districts with unexpectedly high nowcasted death rates. An additional interaction for elderly people allows us to distinguish between outbreaks which might be attributed to activity in retirement homes and those due to unexpected activities in the general population. Mapping the general pattern of the spread of the disease in Germany confirms that different regions are affected to different extents, with southern and western regions being generally more affected than northern states. In addition, a global smooth time trend captures the changes in death rate, showing the peak at the beginning of April and a constant decrease since then. Thanks to the implemented nowcasting, the time trend can be estimated up to the date of analysis. This spatially differentiated picture would not be achievable through a simple monitoring of district-specific observed deaths.

A natural next step would now be to consider the nowcasted fatal infections in relation to the number of newly registered infections, which is, in contrast, highly dependent on both testing strategy and capacity. We consider this as possible future research, but the proposed model allows to explore data in this direction. This might ultimately help us in shedding light on the relationship between registered and undetected infections as well as on the effectiveness of different testing strategies.

There are several limitations to this study, which we want to address as well. First and foremost, even though death counts are, with respect to cases counts, less dependent on testing strategies, they are not completely independent from them. This applies in particular to the handling of post-mortem tests. We therefore do not claim that our analysis of death counts is completely unaffected by testing strategies. Second, a fundamental assumption in the model is the independence between

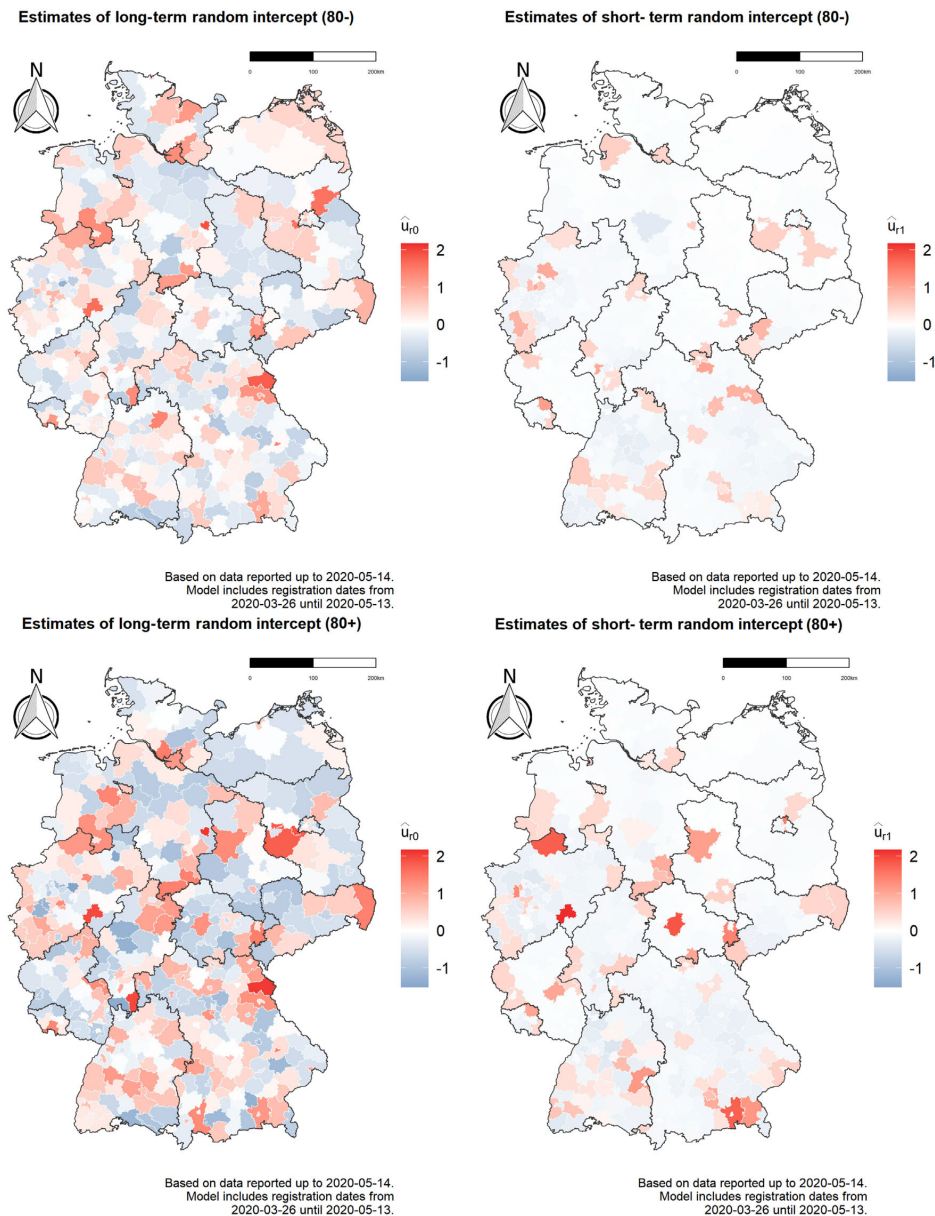


FIGURE 10 Region-specific long-term level (left-hand column) and short-term dynamics of the 14 days prior to May 14, 2020 (right-hand column) of fatal COVID-19 infections for the age groups under 80 (80–, upper row) and above 80 (80+, bottom row)

the course of the disease (on the population level) and the number of infections. Overall, if the local health systems have sufficient capacity and triage can be avoided, this assumption seems plausible, but it is difficult or even impossible to prove the assumption formally. However, the results of the nowcasting model empirically show a rather stable course of the disease, supporting our assumption. Furthermore, the registration of a COVID-19 case is related to the district of residence, while the infection does not necessarily occur in the district where the infected person resides. However, due to a lack of data we cannot explore this point further. Also, in the considered timespan, the mobility in the population has

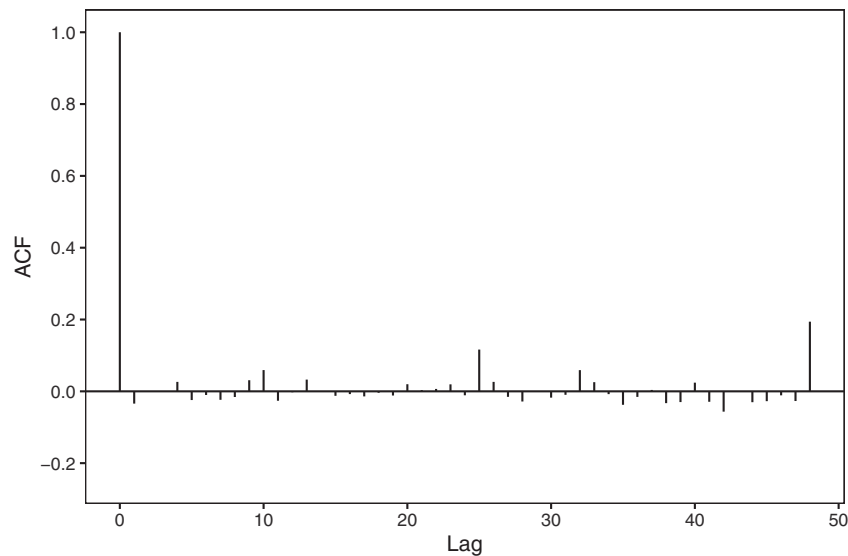


FIGURE 11 ACF plot of the Pearson residuals in the mortality model

been rather low due to governmental restrictions. Even though the model focuses on regional aspects of the pandemic, the nowcasting itself is carried out on a national level, due to sparse data. Given that our results show that the course of the disease from registration to death in Germany did not notably depend on age or gender, we do not expect it to depend on place of residence either.

A general limitation results through the availability of information. Our analyses are based on available data of all registered COVID-19 infections in Germany together with the information on fatalities, which is published daily by the RKI. While these data allow for an analysis of the occurrence of the disease in Germany, it lacks further detailed patient-specific information, for example on clinical aspects or on the differentiation between death with or because of COVID-19. This issue is shared with many other public disease registers. Note also that the methods we are proposing in this paper are not necessarily restricted to the use case of German COVID-19 data. For the purpose of applying our methodology to other countries, the data need to be in the same format as illustrated in Table 1, that is death counts need to be available in an aggregated form stratified by age (group), gender and district. For an appropriate interpretation of the results, it is critical that the reference date of every infection with a fatal outcome (here: registration date) corresponds to a time point at the early stages of the course of the disease. This could also be the date of infection with COVID-19, if known. The second date, which is needed for our nowcasting approach, is the reporting day of each fatal infection. While in Germany, this information can be deduced by considering the COVID-19 database daily over a longer period, the health authorities in other countries might supply historical reporting dates in a consecutively updated database.

Finally, the proposed approach demonstrates that valuable insight into the state and the dynamic of the disease can be obtained by disentangling spatial variation into a global pattern, district-specific long-term effects and current short-term dynamics in a spatio-temporal model. A particular virtue of the presented modelling approach over other proposals is that it also adjusts for the age and gender structure of the local population. This can provide relevant support for the monitoring of this new disease and can assist local health authorities in the planning of infection control measures as well as healthcare system capacities, in a further step towards the understanding and control of the COVID-19 pandemic.

ACKNOWLEDGEMENTS


We want to thank Maximilian Weigert and Andreas Bender for introducing us to the art of producing geographic maps with **R**. Moreover, we would like to thank all members of the Corona Data Analysis Group (CoDAG) at LMU Munich for fruitful discussions.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Marc Schneble  <https://orcid.org/0000-0001-9523-4173>

REFERENCES

- Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). Hoboken, NJ: Wiley.
- Altmejd, A., Rocklöv, J., & Wallin, J. (2020). Nowcasting COVID-19 statistics reported with delay: A case-study of Sweden. Preprint arXiv:2006.06840v1.
- Baud, D., Qi, X., Nielsen-Saines, K., Musso, D., Pomar, L., & Favre, G. (2020). Real estimates of mortality following COVID-19 infection. *The Lancet. Infectious Diseases*, 20(7), 773. [https://doi.org/10.1016/S1473-3099\(20\)30195-X](https://doi.org/10.1016/S1473-3099(20)30195-X).
- Bird, S., & Nielsen, B. (2020). Now-casting of COVID-19 deaths in English hospitals. Retrieved from <http://users.ox.ac.uk/nuff0078/Covid/index.htm>.
- Cohen, J., & Kupferschmidt, K. (2020). Countries test tactics in “war” against COVID-19. *Science*, 367(6484), 1287–1288.
- Esri Deutschland GmbH (2020). Daily COVID-19 case numbers provided by the Robert-Koch-Institute. Retrieved from <https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74>. Accessed: 30/09/2020.
- Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2007). *Regression*. Berlin, Germany: Springer.
- Ferguson, N., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., ... Ghani, A. C. (2020). Report 9 - Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. London: Imperial College London. <https://doi.org/10.25561/77482>.
- Flaxman, S., Mishra, S., Gandy, A., Unwin, H., Coupland, H., Mellan, T., ... Bhatt, S. (2020). Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on COVID-19 in 11 European countries. Retrieved from <https://spiral.imperial.ac.uk/8443/handle/10044/1/77731>.
- Grasselli, G., Pesenti, A., & Cecconi, M. (2020). Critical care utilization for the COVID-19 outbreak in Lombardy, Italy: Early experience and forecast during an emergency response. *JAMA*, 323(16), 1545–1546.
- Grasselli, G., Zangrillo, A., & Zanella, A. (2020). Baseline characteristics and outcomes of 1591 patients infected with SARS-CoV-2. *JAMA*, 323(16), 1574–1581.
- Günther, F., Bender, A., Katz, K., Küchenhoff, H., & Höhle, M. (2020). Nowcasting the COVID-19 pandemic in Bavaria. *Biometrical Journal*, 1–13. <https://doi.org/10.1002/bimj.202000112>.
- Held, L., Meyer, S., & Bracher, J. (2017). Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. *Statistics in Medicine*, 36(22), 3443–3460.
- Höhle, M., & an der Heiden, M. (2014). Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics*, 70, 993–1002.
- Jung, S.-M., Akhmetzhanov, A., Hayashi, K., Linton, N., Yang, Y., Yuan, B., ... Nishiura, H. (2020). Real-time estimation of the risk of death from novel coronavirus (COVID-19) infection: Inference using exported cases. *Journal of Clinical Medicine*, 9, 523.
- Krivobokova, T., & Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, 102(480), 1328–1337.
- Lawless, J. (1994). Adjustment for reporting delays and the prediction of occurred but not reported events. *Canadian Journal of Statistics*, 22(1), 15–31.
- Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 368(6490), 489–493.
- Linton, N. M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A. R., Jung, S.-m., ... Nishiura, H. (2020). Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine*, 9(2), 538.
- Massonnaud, C., Roux, J., & Crépey, P. (2020). COVID-19: Forecasting short term hospital needs in France. *medRxiv* 2020.03.16.20036939. <https://doi.org/10.1101/2020.03.16.20036939>.
- Meyer, S., Held, L., & Höhle, M. (2017). Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *Journal of Statistical Software, Articles*, 77(11), 1–55.
- Niehus, R., De Salazar, P. M., Taylor, A., & Lipsitch, M. (2020). Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan, China that depend on reported cases in international travelers. *medRxiv* 2020.02.13.20022707. <https://doi.org/10.1101/2020.02.13.20022707>.

- Robert Koch-Institut (2020). COVID-19-dashboard. Retrieved from <https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bfd4>.
- StaBLab, LMU Munich (2020). CoronaMaps. Retrieved from <https://corona.stat.uni-muenchen.de/maps/>.
- Streeck, H., Schulte, B., Kümmerer, B. M., Richter, E., Höller, T., Fuhrmann, C., ... Hartmann, G. (2020). Infection fatality rate of SARS-CoV-2 infection in a German community with a super-spreading event. *medRxiv* 2020.05.04.20090076. <https://doi.org/10.1101/2020.05.04.20090076>.
- Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, 18(2), 223–249.
- Wilson, N., Kvalsvig, A., Barnard, L. T., & Baker, M. G. (2020). Case-fatality risk estimates for COVID-19 calculated by using a lag time for fatality. *Emerging Infectious Diseases*, 20(6), 1339–1441.
- Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Boca Raton, FL: CRC Press.
- Zeger, S. L., See, L. C., & Diggle, P. J. (1989). Statistical methods for monitoring the AIDS epidemic. *Statistics in Medicine*, 8, 3–21.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Schneble M, De Nicola G, Kauermann G, Berger U. Nowcasting fatal COVID-19 infections on a regional level in Germany. *Biometrical Journal*. 2021;63:471–489. <https://doi.org/10.1002/bimj.202000143>

10. Regional now- and forecasting for data reported with delay: toward surveillance of COVID-19 infections

Contributing article

De Nicola, G., Schneble, M., Kauermann, G., and Berger, U. (2022). Regional now- and forecasting for data reported with delay: Towards surveillance of COVID-19 infections. *ASTA Advances in Statistical Analysis*, 106:407–426. <https://doi.org/10.1007/s10182-021-00433-5>.

Data and code

Available at <https://github.com/gdenicola/Now-and-Forecasting-COVID-19-Infections>.

Copyright information

This article is distributed under a Creative Commons 4.0 International license (CC BY 4.0).

Supplementary material

Supplementary material is available at Springer online.

Author contributions

The idea of now- and forecasting COVID-19 infections can be attributed to Göran Kauermann, who also substantiated the model. Giacomo De Nicola then structured and wrote the paper, and was further responsible for data analysis and visualization. All authors contributed through fruitful comments and extensive proofreading of the manuscript.

AStA Advances in Statistical Analysis (2022) 106:407–426
https://doi.org/10.1007/s10182-021-00433-5

ORIGINAL PAPER



Regional now- and forecasting for data reported with delay: toward surveillance of COVID-19 infections

Giacomo De Nicola¹ · Marc Schneble¹ · Göran Kauermann¹ · Ursula Berger²

Received: 18 February 2021 / Accepted: 17 December 2021 / Published online: 18 January 2022
© The Author(s) 2022

Abstract

Governments around the world continue to act to contain and mitigate the spread of COVID-19. The rapidly evolving situation compels officials and executives to continuously adapt policies and social distancing measures depending on the current state of the spread of the disease. In this context, it is crucial for policymakers to have a firm grasp on what the current state of the pandemic is, and to envision how the number of infections is going to evolve over the next days. However, as in many other situations involving compulsory registration of sensitive data, cases are reported with delay to a central register, with this delay deferring an up-to-date view of the state of things. We provide a stable tool for monitoring current infection levels as well as predicting infection numbers in the immediate future at the regional level. We accomplish this through nowcasting of cases that have not yet been reported as well as through predictions of future infections. We apply our model to German data, for which our focus lies in predicting and explain infectious behavior by district.

Keywords Nowcasting · Forecasting · COVID-19 · Generalized regression models · Delayed reporting · Disease mapping

1 Introduction

The infectious disease known as COVID-19 hit the planet in tsunami-like fashion. The first cases were identified in December 2019 in the city of Wuhan, China, and by March 2020 infections had already spread over the entire world. Nearly all of the affected countries progressively implemented measures to slow down the spread of the virus, ranging from recommended social distancing to almost complete

✉ Giacomo De Nicola
giacomo.denicola@stat.uni-muenchen.de

¹ Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany

² Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-Universität München, Munich, Germany

lockdowns of social and economic activity. These measures eventually proved to be effective, as the number of infections could be slowed down (see e.g., Flaxman et al. 2020 and Roux et al. 2020). This allowed numerous states to relax restrictions, in an attempt to gradually return to normality. At the same time, with the threat posed by the virus still looming, decision makers are forced to strike a balance between epidemiological risk and allowance of socioeconomic activity. In this context, surveillance of the number of new infections became increasingly important, and particularly so on a regional level. Given the local nature of the phenomenon (see e.g., Gatto et al. 2020 and Li et al. 2020), such regional view appears to be of crucial importance. One of the difficulties lies in the fact that exact numbers of infections detected on a particular day are only available with a reporting delay of, in some cases, several days, which occurs along the reporting line from local health authorities to the central registers. The following paper provides a stable tool for monitoring current infection levels, correcting for incompleteness of the data due to reporting delays. This approach is also extended toward predicting new infections for the immediate future at the regional level.

More specifically, the scope of our model is threefold: Firstly, we aim to understand the current epidemiological situation as well as to comprehend the association between detected infections, demographic characteristics and geographical location. Secondly, our goal is to nowcast infections that have already been observed but have not yet been included in the official numbers. New infections are detected through tests and registered by the local health authorities, which in turn will report the numbers to national authorities with an inevitable delay. Since we observe reports of infections for each day, we are able to model this delay, which indeed allows to nowcast infection numbers correcting for infections which have not yet been reported. Note that we are not modeling the incubation period (Qin et al. 2020; McAloon et al. 2020), nor the time passing from the onset of symptoms to detection and registration by the local health authority (Lima et al. 2020), as those are beyond the scope of this paper. We instead focus solely on the delay which occurs along the reporting chain from local to national authorities. Lastly, our aim is also to forecast the epidemiological situation for the immediate future. We here want to stress that our model is not aiming to exactly predict future infection numbers, as that would not be realistic. The goal is rather to give a general idea of what is going to happen in the next days in the different districts, and, perhaps most importantly, help in identifying which districts are going to be the most problematic. This could also help policymakers in making decisions regarding the implementation of safety measures at the regional level. We apply our modeling approach to explain and predict numbers of registered COVID-19 infections for Germany by district, age group and gender. While the regional component is of evident and paramount importance, the age group and gender distinctions are also very relevant, given the powerful interaction of demography and current age-specific mortality for COVID-19 (Dowd et al. 2020).

Our nowcasting approach can also be used to obtain up-to-date measures of the 7-days incidence, both at the local as well as at the national level. This quantity is often used by authorities to assess how hard a specific area is currently hit by the pandemic, and sometimes, as is the case for Germany, it is also employed as a criterion to decide which containment measures are appropriate (Bundesministerium

der Justiz 2021). It is especially important to have up-to-date infection numbers when computing such a measure, as it is inherently evolving on a daily basis. At the time of writing, the index is calculated by German officials with reference to the date of report of each infection by the local health authorities. Given that, as already stated, there are significant delays in the reporting of cases from local authorities to national ones, the resulting figures are consistently underestimating the actual incidence, with the error being potentially quite large and problematic. Our nowcasts offer a simple and stable solution to this issue, providing infection numbers that are already corrected for expected delays.

The statistical modeling of infectious diseases is a well developed scientific field. We refer to Held et al. (2017) for a general overview of the different models. Modeling and forecasting COVID-19 infections has been tackled by numerous research groups using different models. Panovska-Griffiths (2020) discusses whether one or multiple models may be useful for COVID-19 data analytics. Stübinger and Schneider (2020) make use of time warping to forecast COVID-19 infections for different countries (see also Cintra et al. 2020), while Dehesh et al. (2020) utilize ARIMA time series models. Ray et al. (2020) combine forecasts from several different models to obtain robust short-term forecasts for deaths related to COVID-19. Fritz et al. (2021) present a multimodal learning approach combining statistical regression and machine learning models for predicting COVID-19 cases in Germany at the local level. Early references dating back to the first stages of the pandemic are Anastassopoulou et al. (2020) and Petropoulos and Makridakis (2020). In this paper we make use of negative binomial regression models implemented in the `mgcv` package in R (Wood 2017). This allows us to decompose the spatial component in depth, and obtain district-level nowcasts and forecasts for Germany. Our results confirm the dynamic and highly local nature of outbreaks, highlighting the need for continuous regional surveillance on a small area level.

The rest of the paper is structured as follows: Sect. 2 describes the data, while Sect. 3 frames the problem, presents our model and compares the performance of different model specifications over time, motivating our modeling choices. Section 4 exemplifies surveillance and describes how predictions are performed in practice, showing the results for exemplary dates. Finally, Sect. 5 concludes the paper, highlighting the limitations of this study and adding some concluding remarks.

2 Data

As previously anticipated, we focus our analyses on German data. To do so, we make use of the COVID-19 dataset published by the Robert-Koch-Institute (RKI) on a daily basis. The RKI is a German federal government agency and scientific institute responsible for health reporting and for disease control and prevention. It maintains the national register for COVID-19, where all identified cases of the disease are reported from the local health authorities to the RKI. In our analysis we make use of daily downloads of the data, which we have at our disposal starting from April 12, 2020 until December 29, 2020.

Table 1 shows an excerpt of the data we are confronted with. Every morning, the database containing all registered COVID-19 infections is updated and released to the public, downloadable from the Robert-Koch-Institute's repository¹. The dataset contains, for each of the 412 districts, the cumulated number of confirmed cases of COVID-19 infections stratified by age group (00-04, 05-14, 15-34, 35-59, 60-79 or 80+) and gender, updated to that day. The dataset is also stratified by the date of registration of each case by the local public health authorities (*Gesundheitsämter*). Through the merging of daily downloads of this RKI report, we can construct the full dataset as sketched in Table 1, where the release date is defined in the column "Reporting Date". This full data format is necessary to trace the reporting delay for each observation. It can sometimes indeed take several days for the data to get from the local health authorities to the nation-wide central one, and we thus define the reporting delay as the number of days between registration date and reporting date. In Fig. 1 we show the empirical cumulative distribution function of the reporting delay observed during the three weeks prior to two exemplary dates close to the extremes of our examined time period. From the plot we can appreciate how the delays were slightly lower in December than in May, possibly due to improvements along the reporting chain. Nonetheless, the delay remains significant across all of our sample. Note that since the RKI reports data every morning, all reported cases will have a delay of at least one day. The delay is especially high during weekends, a fact for which we account in our model. Due to the delayed nature of reporting, the number of registered COVID-19 cases which refer to a specific registration date might change with the reporting date, as exemplified in Table 1. On September 25, 2020, the RKI has reported three registered infections of females in the age group from 60-79 living in the city of Munich, which were registered on September 22, 2020. Due to delayed reporting, this number increased to six in the report of September 26, 2020. The three newly reported cases have therefore been reported with a delay of four days. Note once again that the RKI dataset available for download only contains the information up to the current date, thus making daily downloads of the datasets necessary to determine reporting delay.

For the sake of brevity, we here do not provide general descriptive statistics of the data, since these numbers can be easily obtained from many other sources. Among others, we refer to the RKI webpage², which also includes a dashboard to visualize the data (see also CoronaMaps³).

3 Surveillance model

3.1 Framing

We start motivating the model by first reformulating the data structure in a way that is suitable for the analysis. Let $N_{t,d}$ denote the newly registered infections at day t

¹ <https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74>

² <https://www.rki.de/covid-19-en>

³ <https://corona.stat.uni-muenchen.de/maps>

10. Regional now- and forecasting for data reported with delay: toward surveillance of COVID-19 infections

Table 1 Illustration of the raw data structure, showing downloads of the data from September 25 and September 26, 2020 as an example. To facilitate reproducibility, the original column names used in the RKI datasets are given in brackets below our English notation

Data downloaded on	District (Landkreis)	Age Group (Altersgruppe)	Gender (Geschlecht)	Infections (Anzahl Fall)	Registration Date (Meldedatum)	Reporting Date (Datenstand)
	September 25, 2020
	Munich City	60-79	F	3	September 22, 2020	September 25, 2020
	Munich City	60-79	M	5	September 22, 2020	September 25, 2020

Data downloaded on	District	Age Group	Gender	Infections	Registration Date	Reporting Date
	September 26, 2020
	Munich City	60-79	F	6	September 22, 2020	September 26, 2020
	Munich City	60-79	M	5	September 22, 2020	September 26, 2020

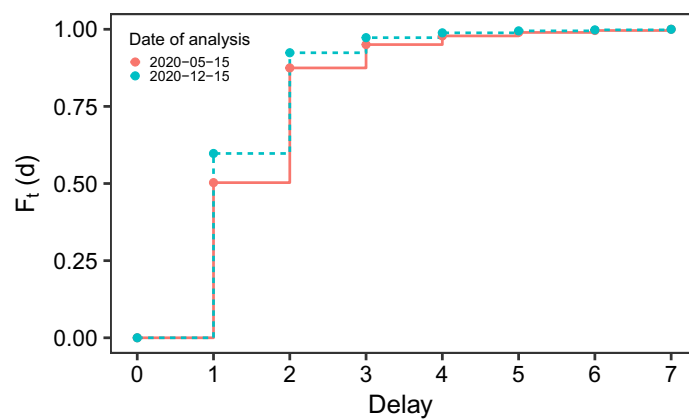


Fig. 1 Empirical cumulative distribution function $F_t(d)$ of reporting delays observed during the three weeks preceding May 15 and December 15, 2020

which are reported with delay d and hence included in the database from day $t + d$. The minimum possible delay is one day, and we assume the maximum delay to be equal to d_{max} days. In our analysis we set $d_{max} = 7$, which corresponds to a week. In other words, we assume delayed reporting to happen within a week. If we define T as the time point of the analysis, the data available at that moment will take the form shown in Table 2.

The bottom right triangle of the data is missing, so that the structure of the available data is akin to that of a guillotine blade. This comparison can be helpful to understand prediction of future values, since predicting by reporting date corresponds to making the blade fall down by one or more days. In other words, one of our goals will be to predict the diagonal edge of the blade, which corresponds to the prediction for cases to be reported on day $T + 1$. To better explain our prediction strategy, we give a sketch of this idea in Fig. 2. In the sketch, the green dots represent data that are already observed at time T (the day of analysis), while the crosses represent entries that are not yet observed and that we aim to predict with our model. This is done in three steps, which are described below. To be specific, we pursue *nowcasting*, *forecasting* and the combination of both,

Table 2 Reformulated data structure for a single district, age group and gender, explicitly including delay. Available data are akin to a guillotine blade

t	d				
	1	2	...	d_{max}	
1	$N_{1,1}$	$N_{1,2}$...	$N_{1,d_{max}}$	
2	$N_{2,1}$	$N_{2,2}$...	$N_{2,d_{max}}$	
\vdots	\vdots	\vdots	\vdots	\vdots	
$T - d_{max}$	$N_{T-d_{max},1}$	$N_{T-d_{max},2}$...	$N_{T-d_{max},d_{max}}$	
$T - d_{max} + 1$	$N_{T-d_{max}+1,1}$	$N_{T-d_{max}+1,2}$...	NA	
\vdots	\vdots	\vdots	\vdots	\vdots	
$T - 1$	$N_{T-1,1}$	NA	NA	NA	
T	NA	NA	NA	NA	

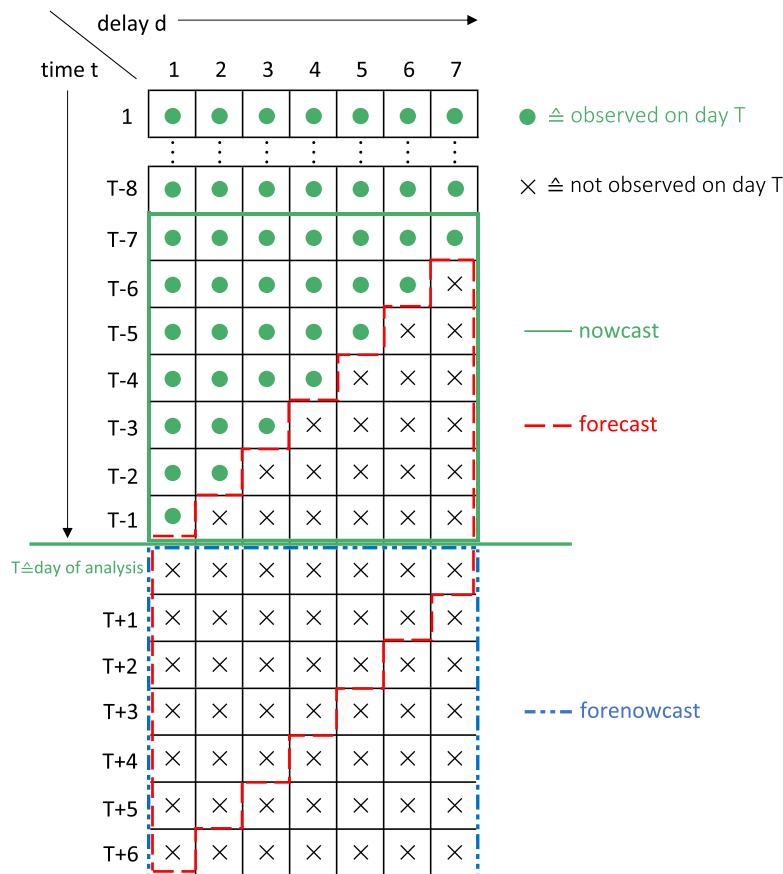


Fig. 2 Sketch of the reformulated data structure showing how nowcasting, forecasting and forenowcasting are performed

which we name *forenowcasting*. Note that forecasting and forenowcasting can be defined, in short, respectively, as "forecasting of reported cases" and "forecasting of registered cases".

Nowcasting: Each row of the matrix contains cases registered on a single date and reported with different delays. To obtain the amount of cases registered on that day regardless of the delay with which they were reported we therefore need to take the row sum. If the goal is to obtain predictions by registration date for several days, we then just sum the cases over the corresponding rows. In Fig. 2 we highlight this type of prediction with a green square, which represents a weekly nowcast, that is the number of cases with registration dates over the past week. This comprises numbers that have already been observed as well as the predictions for cases from past days that have not yet been reported.

Forecasting If we shift the focus from predicting by registration date to reporting date, that is, if the aim is to predict reported numbers regardless of when the reported infections were actually first discovered, we cannot sum the entries of the matrix row-wise, but we need to do so diagonally. This is because the reported number on day T is comprised of the sum of cases registered on day $T - 1$ reported with delay 1, cases registered on day $T - 2$ reported with delay 2, and so on and so forth, up until cases registered on day $T - d_{max}$ reported with delay d_{max} . The red parallelogram in Fig. 2 thus represents the cumulated weekly forecast, that is, the predicted number of infections to be reported over the next seven days. Here all entries are unobserved and will need to be predicted through our model, which will be uncovered in the following section.

Forenowcasting We can also combine the two aspects and predict the number of infections that will be registered in the next week, regardless of their reporting date. We call this process “forenowcasting”. While the previously described forecasting (i.e., predicting by reporting date) is useful to get a picture of the numbers that will be reported each day, what really gives a picture of the ongoing situation are infection numbers based on registration date. This weekly prediction corresponds to the blue square in Fig. 2 and in fact is a combination of forecasting and nowcasting. We will demonstrate that the three types of predictions can be carried out with a single model.

3.2 Statistical model

As already stated in Sect. 2, the cumulative numbers of registered COVID-19 infections are, other than by registration date, also stratified by district, age group and gender. To accommodate for this additional information, we extend the notation from above and define with $N_{t,d,r,g}$ the number of newly registered infections on day t in region/district r and gender and age group g , reported by the RKI on day $t + d$ (thus with delay d). Row-wise cumulated numbers are defined through

$$C_{t,d,r,g} = \sum_{j=1}^d N_{t+j,r,g} \tag{1}$$

which represents the group- and district-specific cumulated number of cases with registration date t and delay up to d . We define with z_r the geo-coordinates of district/region r and generally denote covariates with x , where varying subscripts

indicate dependence on either gender- and age group g , region r , time point t or delay d .

We assume the counts $N_{t,d,r,g}$ to follow a negative binomial distribution with mean $\mu_{t,d,r,g}$ and variance $\mu_{t,d,r,g} + \theta\mu_{t,d,r,g}^2$, where $\theta > 0$ and the limit $\theta \rightarrow 0$ leads to a Poisson distribution. More specifically, we set

$$\begin{aligned} \mu_{t,d,r,g} = & \exp\{s_1(t) + s_2(z_r) + \gamma_d + \mathbf{x}_{t,d}\boldsymbol{\alpha} + \mathbf{x}_g\boldsymbol{\beta} + \mathbf{x}_t\mathbf{u}_r \\ & + \phi \log(1 + C_{t-1,d,r,g}) + \delta \log(1 + C_{t,d-1,r,g}) + \text{offset}_{r,g}\}. \end{aligned} \quad (2)$$

Here $s_1(t)$ is a global smooth time trend, and $s_2(z_r)$ is a smooth spatial effect over the districts of Germany. The parameters $\boldsymbol{\gamma}_d = (\gamma_1, \dots, \gamma_{d_{\max}})$ capture the delay effect for each delay d , while the parameters contained in $\boldsymbol{\alpha}$ capture effects related to time and delay, which in our case will be weekday effects. Gender and age effects are included in $\boldsymbol{\beta}$, and \mathbf{u}_r are unstructured regional effects which will be subsequently specified in more detail. Coefficient ϕ captures the time-related autoregressive (AR) component of the process, indicating the effect of cases from the same district and gender- and age group which were registered on the previous day. Coefficient δ expresses the effect of infections registered on the same day which were reported with delay up to $d - 1$, or in other words a delay-related autoregressive component. Finally, the offset is set to the logarithm of the regional population size in the different gender and age groups, enabling us to model the infection rate. Using a population offset is quite standard in disease mapping and in count time series analyses of rare infectious diseases (see e.g., Bauer and Wakefield 2018). The offset defined this way also allows to incorporate the size of the susceptible population in each region, showing that this type of modeling is practicable at different stages of the pandemic. In this case, the population size would need to be replaced by the number of susceptible in region r , incorporating the SIR (susceptible-infected-removed) model or other similar ones (see e.g., Allen 1994). This is not particularly relevant at the time point chosen for the analysis, as the number of susceptible corresponds more or less to the population size due to the small (and unknown) size of the immune populations in each district (note that vaccines were not yet available during the analyzed time period).

The previously mentioned spatial effect is comprised of two components: An overall smooth effect $s_2(z_r)$ mirroring the fact that different parts of Germany are differently affected, and a region-specific component accounting for infection rates that are particularly high or low in single districts with respect to the neighbouring situation. To be more specific, $s_2(\cdot)$ is a smooth spatial function of the geo-coordinates z_r for region r , while the \mathbf{u}_r are unstructured region-specific effects, interacting with the time dependent covariates \mathbf{x}_t . We put a normal prior on \mathbf{u}_r , i.e., we model $\mathbf{u}_r = (u_{r0}, u_{r1})^\top$ as random effects, where u_{r0} is a general random intercept capturing the long-term level (from $t = 1, \dots, T$) of the epidemiological situation in the different districts, while u_{r1} is a second random intercept estimated exclusively over the last k days, expressing the short-term dynamics (within k days prior to $t = T$) of infections. In our analysis we set $k = 7$. For \mathbf{u}_r we assume the structure

$$\mathbf{u}_r \stackrel{iid}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_u) \tag{3}$$

for $r = 1, \dots, 412$, with the posterior variance matrix $\boldsymbol{\Sigma}_u$ being estimated from the data. The predicted values $\hat{\mathbf{u}}_r$ (i.e., the posterior mode) measure how much and in which direction the infection rate of each district deviates from the global spatial structure, controlling for covariates and age- and gender-specific population sizes.

3.3 Model selection and performance

Model (3.2) includes several components. In this section we aim at assessing whether the inclusion of some of those components is beneficial in terms of predictive performance, and to generally evaluate the overall performance of the final model. Note that we fit the model including only infections with registration dates within 21 days of the day of analysis in the training set. This is because, while on the one hand we would like to use as much data as possible for the fitting, the data-generating process (i.e the spread of the disease) is subject to exogenous changes over time. In other words, we must strike a balance between having a large enough training set and keeping the model as loyal to the current data-generating process as possible. We therefore fit our model using data from a rolling window of 21 days. This choice is motivated more precisely in the supplementary material, where plots comparing the predictive accuracy of the model using different fitting windows are included. The choice of a shorter fitting window also allows to keep other components of the model, such as the smooth spatial effect, constant over time: Such effects are not, in general, time constant, and if we used the whole dataset for the model, we would need to have them interact with the temporal dimension. The rolling window thus also enables the use of a simpler model.

In this section, we are specifically interested in seeing how the unstructured random effects $\mathbf{x}_t \mathbf{u}_r$ and the autoregressive components $\phi \log(1 + C_{t-1,d,r,g})$ and $\delta \log(1 + C_{t,d-1,r,g})$ impact predictive accuracy. To do so, we consider the realized absolute prediction error with regards to nowcasts, forecasts and forenowcasts, cumulated for each district over a period of seven days using different model specifications, to compare performance over time through a weekly rolling window approach. The specifics of how predictions are performed will be described in detail in Sect. 4.

Starting with nowcasting, let therefore $Y_{T,r}^{(n)}$ denote the cumulated number of registered infections in district r over $k = 7$ days prior to the day of analysis at time T , that is

$$Y_{T,r}^{(n)} = \sum_{t=1}^k \sum_g C_{T-t,d_{max},r,g}.$$

This corresponds to the sum of all numbers in the green square in Fig. 2. Accordingly, we define with $\hat{Y}_{T,r}^{(n)}$ the corresponding prediction based on the fitted model as described above. For forecasting, we modify the definition and look at the cumulated number of cases

$$Y_{T,r}^{(f)} = \sum_{t=1}^k \sum_{d=1}^{d_{max}} \sum_g N_{T+t-d,d,r,g}$$

which corresponds to the red parallelogram in Fig. 2. Again, the corresponding predicted value is notated as $\widehat{Y}_{T,r}^{(f)}$. Finally, for forenowcasting we concentrate on the cumulated numbers in the blue square, and set

$$Y_{T,t}^{(fn)} = \sum_{t=1}^k \sum_g C_{T+t-1,d_{max},r,g}$$

with matching prediction $\widehat{Y}_{T,t}^{(fn)}$ based on the fitted model. With the notation just given, we can define the relative district-specific prediction error (standardized per 100,000 inhabitants) simply as

$$\text{RPE}_{T,r}^{(\cdot)} = 100\,000 \frac{Y_{T,r}^{(\cdot)} - \widehat{Y}_{T,r}^{(\cdot)}}{\text{pop}_r}$$

where pop_r is the population size in district r , and the dot refers to nowcasting, forecasting or forenowcasting, respectively. It should be clear that, setting $k = d_{max} = 7$, the numbers defined above are only observable on day $T + 7$ for nowcasting and forecasting, and on day $T + 14$ for forenowcasting.

To obtain a measure of the overall predictive performance of the model for a certain fitting date T , we take the mean of $\text{RPE}_{T,r}^{(\cdot)}$ in absolute value over all districts, which we call Mean Absolute Relative Prediction Error (MARPE):

$$\text{MARPE}_T^{(\cdot)} = \frac{1}{412} \sum_{r=1}^{412} |\text{RPE}_{T,r}^{(\cdot)}|$$

To get a sense of the average bias of predictions over time, we also plot the Mean Relative Prediction Error (MRPE), which takes the mean of relative errors without considering them in absolute value:

$$\text{MRPE}_T^{(\cdot)} = \frac{1}{412} \sum_{r=1}^{412} \text{RPE}_{T,r}^{(\cdot)}$$

This last measure will be positive if the model tends to underpredict on average over the districts, and negative otherwise.

To evaluate the predictive accuracy of different model specifications, we compute $\text{MARPE}_T^{(\cdot)}$ and $\text{MRPE}_T^{(\cdot)}$ over time by fitting the model weekly for each of the considered specifications, in a rolling window approach. In particular, we consider the following model variations:

- Full model as in (3.2);
- Model without the time-related autoregressive component, $C_{t-1,d,r,g}$;
- Model without the delay-related autoregressive component, $C_{t,d-1,r,g}$;

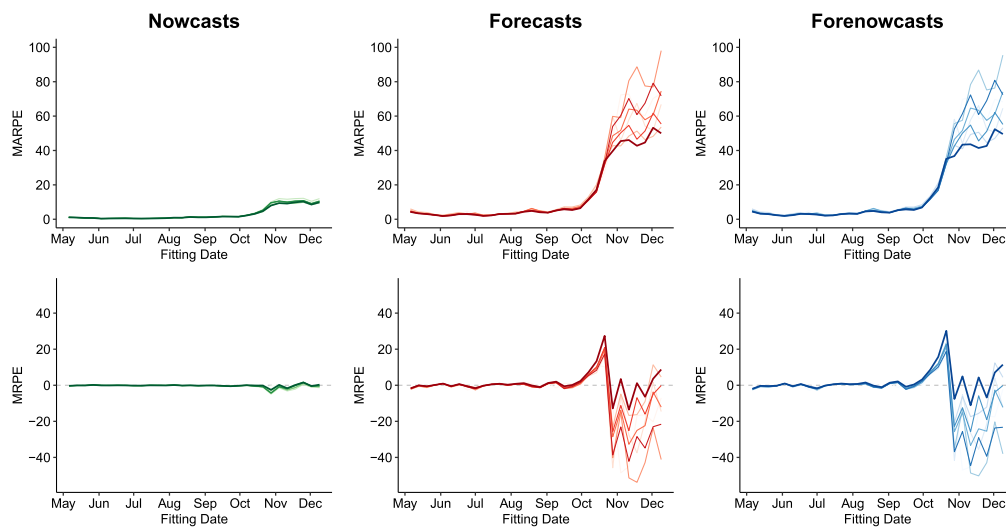


Fig. 3 Mean absolute relative prediction error ($MARPE_{T,r}^{(i)}$, top panel) and Mean Relative Prediction Error ($MRPE_{T,r}^{(i)}$, bottom panel) for all districts in Germany, calculated over time for different model specifications, respectively, for nowcasts (green), forecasts (red) and forenowcasts (blue). Different color shadings refer to model alternatives. The thicker line indicates the selected model, which corresponds to the full model with the exclusion of the time-related AR component, $C_{t-1,d,r,g}$

- Model without the autoregressive components, $C_{t-1,d,r,g}$ and $C_{t,d-1,r,g}$;
- Model without the short-term district-specific random intercept, u_{r1} ;
- Model without the unstructured district-specific random effects u_r ;
- Model without the short-term district-specific random intercept, u_{r1} and the autoregressive components, $C_{t-1,d,r,g}$ and $C_{t,d-1,r,g}$;
- Model without the unstructured district-specific random effects u_r and the autoregressive components, $C_{t-1,d,r,g}$ and $C_{t,d-1,r,g}$;

Figure 3 plots the MARPE and the MRPE by model fitting date for nowcasts, forecasts and forenowcasts, respectively. The plots already reveal several aspects of the goodness of fit of our model. Looking at the MARPE (top panel), it immediately stands out how the errors for nowcasts are, as expected, much smaller than for forecasts and forenowcasts. Secondly, we can see how prediction errors are remarkably small for the first five months of model fitting. Those months coincide with the late spring and summer months, during which infection numbers were relatively under control in Germany. Our model was thus able to capture most of the variability in the process, resulting in precise predictions not only for nowcasts, but also for forecasts and forenowcasts. Finally, we notice how there is a large increase in MARPE for all fitted models starting from October, which coincides with the beginning of the second wave of COVID-19 in Germany. This is due to the fact that in that period the infection dynamics changed and the numbers got much larger, thus also leading to an increase in prediction errors. The model variant that performed the best during this later period is the full model with the exclusion of the time-related autoregressive component $C_{t-1,d,r,g}$, highlighted with a thicker line. The plots for the MRPE (bottom panel) confirm this fact and help explaining the reasons behind it. For both

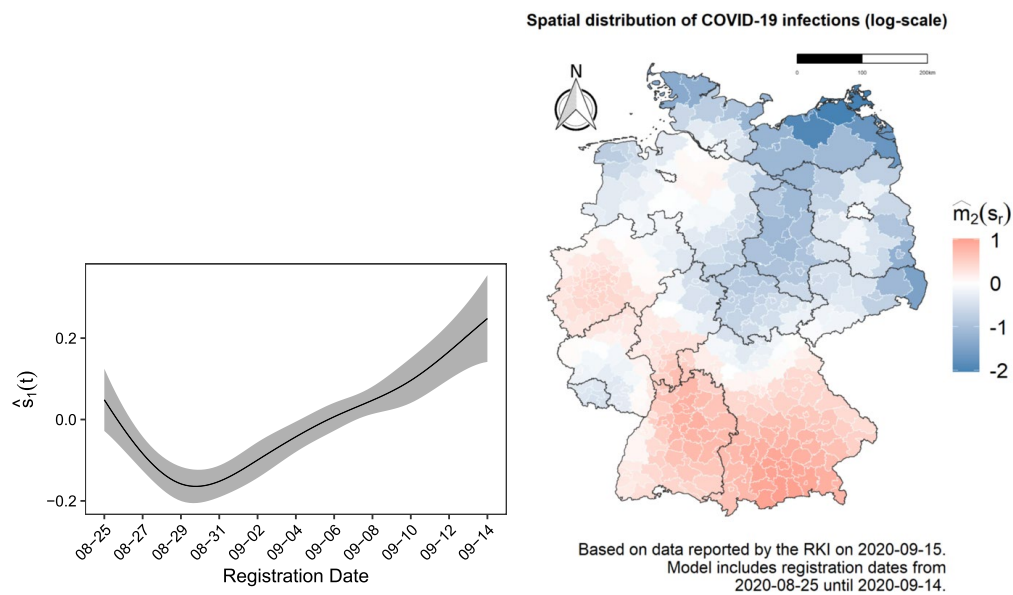


Fig. 4 Estimated smooth effects $s_1(t)$ and $s_2(z_r)$, respectively the fitted smooth effect of time and the fitted smooth spatial effect for the prevalence of COVID-19 infections in Germany (measured on the log scale). Both effects are estimated over the 21 days prior to September 15, 2020

forecasts and forenowcasts, it is apparent how at the beginning of the second wave all models tend to underpredict, while they overpredict from November onward. The chosen model without the AR component is actually the one which tends to underpredict the most (even though it is not performing worse than the others in terms of MARPE), while it then becomes by far the least overpredicting one in later months. This is because infection numbers grew very fast in October, and models including the autoregressive component were better able to capture the quick increase. In contrast though, after new infections somewhat stabilized, the models including the autoregressive component were still projecting the increase of past months on new ones, causing large overestimation. The chosen model is instead more conservative in its predictions, resulting in better overall predictive performance.

4 Applied surveillance

Given that what we propose is a monitoring tool, the results change over time. We here give an exemplary snapshot of the estimates and how predictions can be obtained using Tuesday, September 15, 2020, as date of the analysis. This date was chosen as it lies just before the beginning of second wave of COVID-19 infections in Germany. As an additional remark, note that our analysis is completely reproducible for different dates as well, with code and data openly available and downloadable from our GitHub repository⁴.

⁴ <https://github.com/gdenicola/Now-and-Forecasting-COVID-19-Infections>

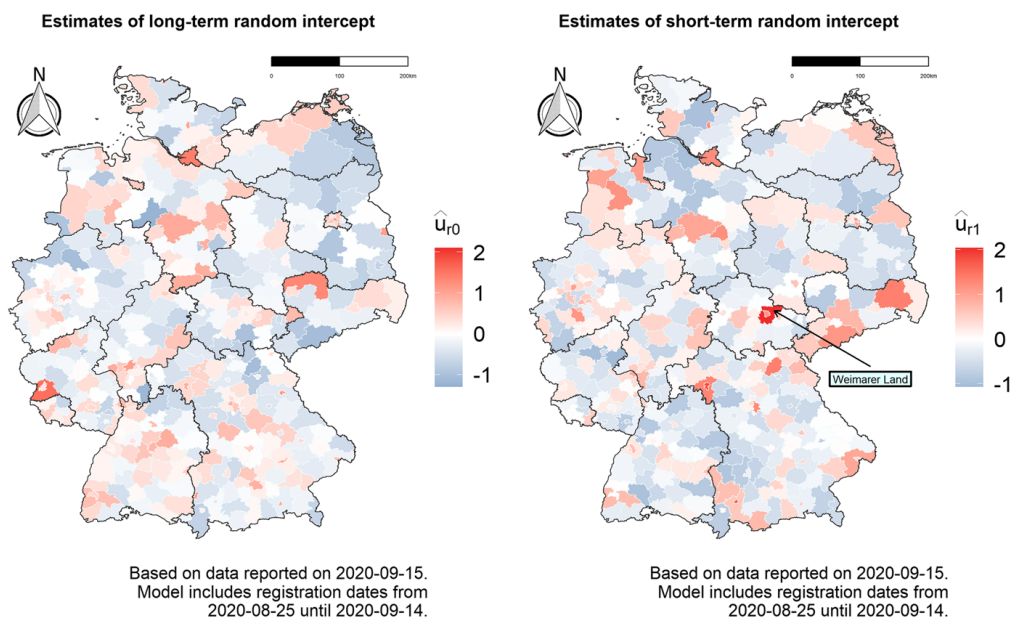


Fig. 5 Region specific level (left) and dynamics (right) of COVID-19 infections, controlling for the smooth spatial effect on the right hand side of Fig. 4

4.1 Model-based monitoring

In addition to giving proper predictions (nowcasts, forecasts and forenowcasts), which will be shown in the next section, our model also estimates linear coefficients, which are given in table form in the supplementary material, and fits smooth components over time and space, which are visualized in Fig. 4. The left hand side shows the estimated infection rate over time for the three weeks prior to the day of analysis. We notice how the rate of registered infections has been dropping until the end of August, while in the following weeks numbers started rising again, leading to a reversal and a steady increase in the smooth spline. The map on the right hand side depicts the smooth spatial effect estimated as a function of longitude and latitude, on the log scale. From the plot we can see how the regions of Bavaria and Baden-Württemberg in the south of Germany were generally the most affected during the observed period. We also observe that the west was also, on average, more affected than the east.

The two maps in Fig. 5 show further spatial components of the model, namely the district-specific random intercepts. Those reflect the situation in single districts controlling for the previously shown smooth spatial effect, that is, in comparison to the average of the neighboring areas. More specifically, the map on the left displays the overall district-specific long-term random intercept, depicting the relative infection situation in the 21 days prior to the day of analysis, while the map on the right hand side shows the additional short-term random intercept which enters the linear predictor only over the last 7 days, giving an idea of the more recent infection dynamics. We can thus see that, for example, the district of Weimarer Land in the region of Thuringia has had the most rapidly evolving number of cases in the 7 days

prior to the day of analysis controlling for the situation in its surroundings, reflecting the outbreak that happened in the region during the analyzed period. This second map can already be regarded as a first way of monitoring infection dynamics at a local level, even before looking at the predicted numbers: If a district has a very high short term random effect, it probably means that the affected area deserves further consideration.

4.2 Predictions

As previously explained, our model can be used to directly nowcast (correct reports from previous days for delay), forecast (predict the number of cases reported in the next days) and forenowcast (predict the number of infections that will be registered for the next days). The obtained predictions can be used to get a picture of how the pandemic is going to unfold in the short term. In the following, we explain how we obtain those predictions from our model.

Nowcasting In our case, nowcasting is equivalent to filling all NA (missing) entries of the matrix in Table 2, turning the trapezoid shape of the data into a full rectangle. This is also equivalent to completing the green square in Fig. 2. Given that we model delay d as a stand-alone variable in our generalized additive model, we are able to simply predict the missing cells directly by setting the delay d to the necessary value in the data vector used for predictions alongside all other covariates. We can thus nowcast infections for each delay, day, district, gender and age group. If the autoregressive terms $C_{t-1,d,r,g}$ and $C_{t,d-1,r,g}$ are included in the model, the predictions are dependent on them. Those terms are in general not yet known at the day of analysis (except when predicting the first diagonal of the red parallelogram in Fig. 2). We therefore perform the prediction of the black crosses in Fig. 2 iteratively, by utilizing the predictions of the previous diagonal as the autoregressive components. Based on the model, we can also take uncertainty into account by simulating data from a negative binomial distribution with the corresponding mean and variance structure. More precisely, we apply the same strategy as above, but instead of using the mean value we now plug counts simulated from the model into the autoregressive components, and repeat this procedure $n = 1000$ times. This parametric bootstrap approach easily allows us to compute lower and upper bounds of the prediction intervals.

Forecasting The model also allows to directly predict cases for future dates. With T denoting the time point of data analysis, we can obtain predictions for the number of reported cases on days $T, T + 1, \dots, T + k - 1$. Let us start with the predictions for cases with reporting date T . Referring once again to the guillotine blade structure in Table 2, we proceed as follows: For $d = 1$, i.e., at the leftmost point of the blade, we take the fitted mean values as prediction, while keeping the smooth function of time constant, that is, setting $s(t + 1) \equiv s(t)$ for the sake of stability. For the remaining $d_{max} - 1$ elements of the blade edge we take the mean value by setting $d = d + 1$. To get predictions for the numbers of infections reported on days $T + 1, \dots, T + k - 1$ we can then proceed in an analogous way, using the values just predicted to update the autoregressive components ($C_{t-1,d,r,g}$ and $C_{t,d-1,r,g}$). Figure 2

visualizes the strategy, with cumulated predictions for the number of cases reported on days $T, T + 1, \dots, T + 6$ being represented by the red parallelogram. Similarly as we did for the nowcasting, we can take uncertainty into account through simulations, sampling from a negative-binomial model with the estimated group-specific mean and variance structure.

Forenowcasting Predicting by registration date, i.e., forenowcasting, is equivalent to filling the blue square on the bottom of Fig. 2. This is done by computing forecasts as described in the previous subsection, and then performing nowcasting on the forecasted numbers. We also obtain uncertainty estimates in an analogous way as for forecasts and nowcasts.

4.3 Retrospective surveillance

It is also possible to utilize the proposed model as a surveillance tool retrospectively. After a certain period of time has passed from the day of analysis, we are able to compare predictions with infections observed in the corresponding time span. If the predictions are aggregated on a weekly basis and we keep the maximum delay set as $d_{max} = 7$, the waiting time to observe realized infection numbers will be equal to seven days for nowcasts and forecasts and fourteen days for forenowcasts. Figure 6 shows predictions of all three kinds and corresponding infections observed *a posteriori* for two exemplary days of analysis, namely September 15 (left hand panel) and November 11, 2020 (right hand panel).

From the plots we can observe how nowcasts tend to be, in general, quite precise, as already seen from Fig. 3. We can also immediately notice how performance is very different for the two dates, especially for forecasts and forenowcasts: We see that the predictions for September 15 are relatively precise and unbiased, while for November 11 we observe quite a strong tendency toward overprediction.

Focusing first on the forecast and forenowcasts from September (during a “stable” phase of the pandemic), we see that the biggest prediction errors appear for the districts of Kaufbeuren, Bavaria (overprediction) and Cloppenburg, Lower Saxony (underprediction). In the first case, there was an outbreak in a nursing home in the week preceding the forecasted one. This outbreak initially leads to an increase in the infection numbers, but was subsequently contained very quickly, therefore leading the model to overpredict. The underprediction in Cloppenburg, in contrast, was the product of a sudden increase in cases during the forecasted week. More specifically, the higher numbers resulted from cases in schools and the contagion of an almost complete football team in the small city of Lönigen. All in all, we can see that in general the prediction errors are not massive, and in line with what we would expect simply due to the inherent randomness of the process.

The situation is different when looking at the predictions for November 11. This is because, while the September date belongs to a period in which the pandemic was relatively stable in Germany, the second one lies at the heart of the second wave. Moreover, the latter date was immediately successive to the sudden increase in new infections in October and to the consequent implementation of social distancing

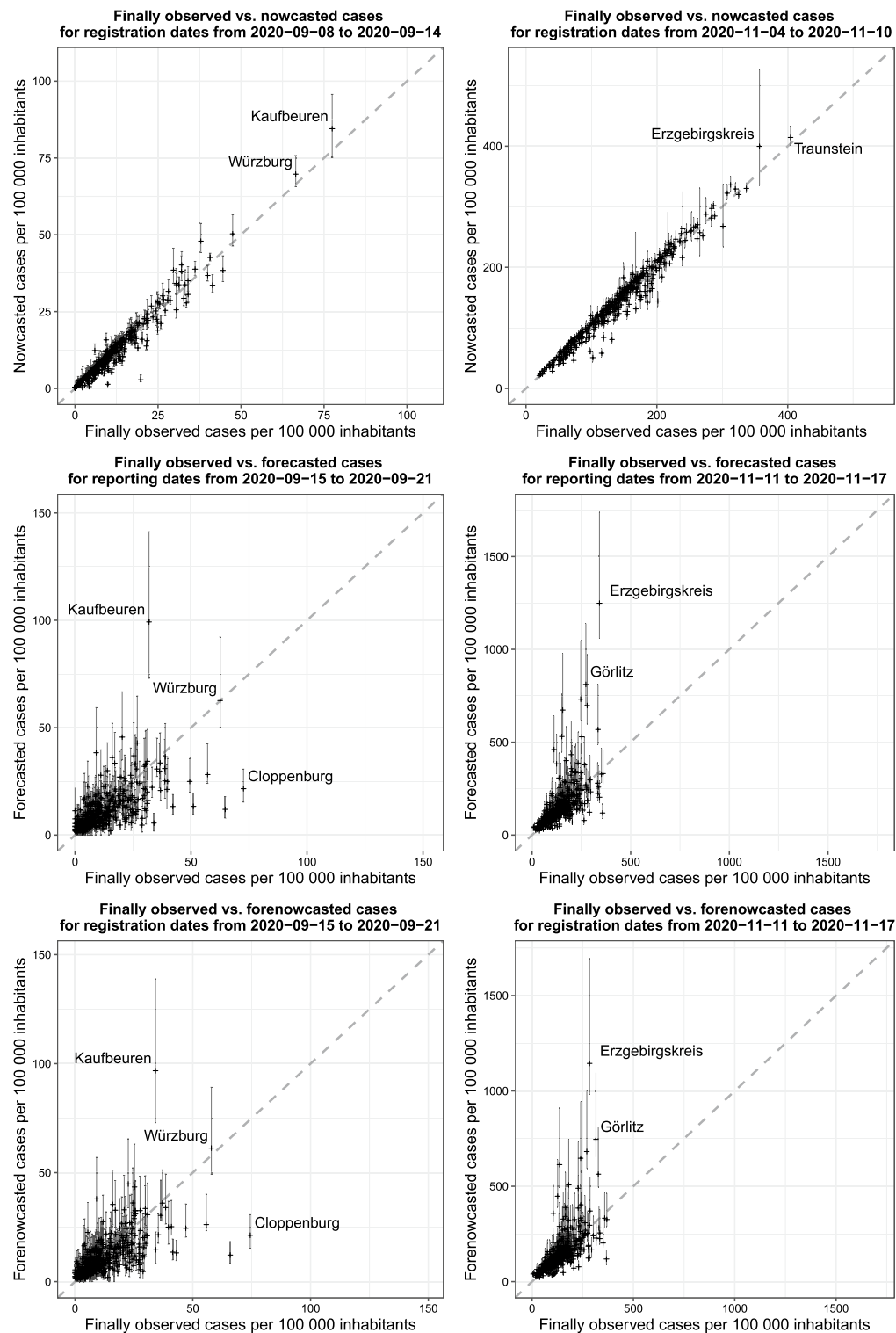


Fig. 6 Nowcasts (top), forecasts (middle) and forenowcasts (bottom) of cumulated infections over a week, cumulated by district, plotted against values observed *a posteriori*. The model is fitted with data available on the dates September 15, 2020 (left) and November 11, 2020 (right). Vertical lines represent prediction intervals computed at the 90% level

measures (the so-called “lockdown light”) from the beginning of November. However, our model does not include anything regarding exogenous governmental interventions and general changes in population behavior. This means that the predictions are to be interpreted assuming that everything else stays the same as in the three weeks used to fit the model, leading to overprediction for areas in which measures are indeed imposed, and possibly underpredictions after those measures are softened or lifted. As a result, forecasted numbers for November 11 suffer from severe overprediction in many districts. The most extreme example is the district of Erzgebirgskreis, Saxony, which saw a rapid increase in cases between the end of October and the beginning of November, which lead to the district being the one with the highest incidence in the whole country for a short period of time. The infection numbers then stabilized in the following week, leading the model to overpredict. As a side remark, note that the prediction errors for both dates are not majorly spatially correlated. This is in line with our expectations, as both a smooth spatial effect as well as two district specific random effects are included in the model. Maps of the prediction errors by district for both dates analyzed are included in the supplementary material.

While the inability to capture governmental intervention and sudden changes in the population behavior is certainly a limitation of our approach, it can on the other hand also be seen as a feature of the model, which in a sense provides potential future “counterfactual” scenarios in which no action was taken by decision makers. This can thus be used to try to quantify the effect of social distancing policies and interventions, in specific districts as well as at a broader level. This also applies in the case of sudden outbreaks: If a rapid spike in cases in a specific district is observed, and that outbreak was not yet known to health authorities at the time of the analysis, the model will naturally underpredict infection numbers in that district. Severe underpredictions observed *a posteriori* can also be used as an indicator for “true” outbreaks, revealing if they were explainable by past data or not. This “counterfactual” use of our model can thus be seen as an additional feature, which becomes available in retrospect, to measure the effect of NPIs (Non-Pharmaceutical Interventions) and to assess the nature of outbreaks.

5 Discussion

We proposed a modeling tool to nowcast and forecast COVID-19 cases reported with delay. This allows to perform surveillance by gender and age group at the regional level, providing an up-to-date and detailed picture of the pandemic, as well as giving insight into the dynamics of the near future. Our model can be used for computing inherently dynamic index measures, such as the 7-days incidence, both at the regional and national level, and it can also aid governments in the implementation of more targeted area- and population-specific containment strategies. However, as previously mentioned, this approach does not come without limitations, which we also want to address.

The number of detected cases greatly depends on local testing strategies and capacities. This implies that comparisons between different states or regions are not

straightforward. As our model makes use of reported infections, direct comparisons between outputs should be limited to areas for which it is reasonable to assume that testing has been carried out in a similar manner.

Another important thing to note is that our model only addresses the delay in reporting from local to national health authorities, and not the time that occurs between each test and its (positive) result. This would be useful for our application as it would give an even more up-to date picture of the current situation, but it is not pursued due to a lack of data.

An eminent limitation of our approach is the inability to capture new outbreaks related to specific phenomena that are not yet known to the health authorities. An example of this would be the outbreaks in slaughterhouses which happened during the summer of 2020 in Coesfeld and Gütersloh, North-Rhine-Westphalia. On the other hand, as previously discussed, severe underpredictions observed *a posteriori* can also be used in retrospect as an indicator for outbreaks that are localized and not explainable by past data, while overpredictions can signal and quantify the effectiveness of social distancing measures.

Taking into account the previously mentioned limitations, the model is able to capture a good chunk of the variability that is present. The methodology that we employed is quite general, and, if suitable data are available, can easily be adapted to other countries as well. Moreover, we only employed standard tools for software implementation, and this makes adapting and enriching the model, e.g., with more covariates, relatively straightforward. Our analysis focuses more on now- and forecasting rather than on increasing our understanding of the spread of the disease, and in this context the random effects enable us to capture unobserved heterogeneity fairly well, so the addition of more (time-constant) covariates is not paramount to our goals. Nonetheless it could be fruitful to include more covariates available for specific cases in the model. For the analyzed case of Germany we pursue this in the supplementary material, by adding to the model the German Indexes of Multiple Deprivation, which measure material and social differences at the regional level in Germany (Maier 2017). The results do not differ greatly from what was obtained without this inclusion.

We complete our discussion by emphasizing that the proposed methodology is flexible and applicable to any data constellation in which reporting delay plays a role. In other words, one can easily adopt the proposed model to any guillotine blade-like data structures, i.e., data where t_i denotes the time point of an event and d_i the delay with which the event is reported. Moreover, our approach can not only be applied to correct for the delay between registration of an event and its reporting, but also, for example, to bridge the delay between disease onset and registration of its positive test result. Data in guillotine blade-like form also occur in areas beyond epidemiology, e.g., when cases of unemployment are reported from regional offices to a central state register. The generality of the data structure supports the proposed modeling approach, where corrections for the missing data structure are directly incorporated in the model. In particular, however, the modeling exercise exhibits promising performance for COVID-19 infections, and may therefore be incorporated into a general surveillance tool to assist health authorities and policymakers in their efforts to contain the spread.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10182-021-00433-5>.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access 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/>.

References

- Allen, L.J.: Some discrete-time SI, SIR, and SIS epidemic models. *Math. Biosci.* **124**(1), 83–105 (1994)
- Anastassopoulou, C., Russo, L., Tsakris, A., Siettos, C.: Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLoS ONE* **15**(3), e0230405 (2020)
- Bauer, C., Wakefield, J.: Stratified space-time infectious disease modelling, with an application to hand, foot and mouth disease in China. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **67**(5), 1379–1398 (2018)
- Bundesministerium der Justiz: Gesetz zur Verhütung und Bekämpfung von Infektionskrankheiten beim Menschen (Infektionsschutzgesetz - IfSG) \§ 28a – Besondere Schutzmaßnahmen zur Verhinderung der Verbreitung der Coronavirus-Krankheit-2019 (COVID-19) (2021)
- Cintra, P., Citeli, M., Fontinele, F.: Mathematical models for describing and predicting the COVID-19 pandemic crisis. [arXiv:200602507](https://arxiv.org/abs/2006.02507) (2020)
- Dehesh, T., Mardani-Fard, H., Dehesh, P.: Forecasting of covid-19 confirmed cases in different countries with arima models. *medRxiv*. <https://doi.org/10.1101/2020.03.13.20035345> (2020)
- Dowd, J.B., Andriano, L., Brazel, D.M., Rotondi, V., Block, P., Ding, X., Liu, Y., Mills, M.C.: Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proc. Natl. Acad. Sci.* **117**(18), 9696–9698 (2020)
- Flaxman, S., Mishra, S., Gandy, A., Unwin, H.J.T., Mellan, T.A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J.W., et al.: Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**(7820), 257–261 (2020)
- Fritz, C., Dorigatti, E., Rügamer, D.: Combining graph neural networks and spatio-temporal disease models to predict COVID-19 cases in Germany. [arXiv:210100661](https://arxiv.org/abs/210100661) (2021)
- Gatto, M., Bertuzzo, E., Mari, L., Miccoli, S., Carraro, L., Casagrandi, R., Rinaldo, A.: Spread and dynamics of the COVID-19 epidemic in Italy: effects of emergency containment measures. *Proc. Natl. Acad. Sci.* **117**(19), 10484–10491 (2020)
- Held, L., Meyer, S., Bracher, J.: Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. *Stat. Med.* **36**(22), 3443–3460 (2017)
- Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., Shaman, J.: Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science* **368**(6490), 489–493 (2020)
- Lima, F.E.T., de Albuquerque, N.L.S., Florencio, S.D.S.G., Fontenele, M.G.M., Queiroz, A.P.O., Lima, G.A., de Figueiredo, L.M., Amorim, S.M.C., Barbosa, L.P.: Time interval between onset of symptoms and COVID-19 testing in Brazilian state capitals, August 2020. *Epidemiologia e Serviços de Saúde* **30**(1) (2020)

- Maier, W.: Indices of multiple deprivation for the analysis of regional health disparities in Germany: experiences from epidemiology and healthcare research. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz* **60**(12), 1403–1412 (2017)
- McAloon, C., Collins, Á., Hunt, K., Barber, A., Byrne, A.W., Butler, F., Casey, M., Griffin, J., Lane, E., McEvoy, D. et al.: Incubation period of COVID-19: a rapid systematic review and meta-analysis of observational research. *BMJ Open* **10**(8), e039652 (2020)
- Panovska-Griffiths, J.: Can mathematical modelling solve the current Covid-19 crisis? *BMC Public Health* **20**, 551 (2020)
- Petropoulos, F., Makridakis, S.: Forecasting the novel coronavirus COVID-19. *PLoS ONE* **15**(3), e0231236 (2020)
- Qin, J., You, C., Lin, Q., Hu, T., Yu, S., Zhou, X.H.: Estimation of incubation period distribution of COVID-19 using disease onset forward time: a novel cross-sectional and forward follow-up study. *Sci. Adv* **6**(33), eabc1202 (2020)
- Ray, E.L., Wattanachit, N., Niemi, J., Kanji, A.H., House, K., Cramer, E.Y., Bracher, J., Zheng, A., Yamana, T.K., Xiong, X., et al.: Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the U.S. medRxiv. <https://doi.org/10.1101/2020.08.19.20177493> (2020)
- Roux, J., Massonnaud, C., Crépey, P.: COVID-19: one-month impact of the French lockdown on the epidemic burden. medRxiv. <https://doi.org/10.1101/2020.04.22.20075705> (2020)
- Stübinger, J., Schneider, L.: Epidemiology of coronavirus COVID-19: forecasting the future incidence in different countries. *Healthcare* **8**(2), 99 (2020)
- Wood, S.N.: *Generalized additive models: an introduction with R*. CRC Press, London (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

11. A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020

Contributing article

Schneble, M., De Nicola, G., Kauermann, G., and Berger, U. (2021). A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020. *Biometrical Journal*, 63(8):1623–1632. <https://doi.org/10.1002/bimj.202100125>.

Data and code

Available under “Supporting Information” at <https://onlinelibrary.wiley.com/doi/10.1002/bimj.202100125>.

Copyright information

This article is distributed under a Creative Commons 4.0 International license (CC BY 4.0).

Supplementary material

Supplementary material is available at Wiley online.

Author contributions

The idea of investigating the COVID-19 case-detection ratio can be attributed to Giacomo De Nicola, while the idea of relating registered COVID-19 infections and deaths related to COVID-19 can be attributed to Göran Kauermann. The latter also formulated the model to estimate the change of the case detection ratio over time. Marc Schneble was responsible for implementing the model in R and for the visualization of the results, in addition to writing major parts of the manuscript together with Göran Kauermann. Giacomo De Nicola further contributed by writing significant parts of the manuscript, especially with regards to Sections 1 and 5. All authors contributed through fruitful comments and extensive proofreading of the manuscript.



Received: 22 April 2021 | Revised: 21 June 2021 | Accepted: 16 July 2021

DOI: 10.1002/bimj.202100125

Biometrical Journal

RESEARCH ARTICLE

A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020

Marc Schneble¹ | Giacomo De Nicola¹ | Göran Kauermann¹ | Ursula Berger²

¹ Department of Statistics, LMU Munich, Munich, Germany

² Institute for Medical Information Processing, Biometry and Epidemiology, LMU Munich, Munich, Germany

Correspondence

Marc Schneble, Department of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich, Germany.

Email:

marc.schneble@stat.uni-muenchen.de



This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

The case detection ratio of coronavirus disease 2019 (COVID-19) infections varies over time due to changing testing capacities, different testing strategies, and the evolving underlying number of infections itself. This note shows a way of quantifying these dynamics by jointly modeling the reported number of detected COVID-19 infections with nonfatal and fatal outcomes. The proposed methodology also allows to explore the temporal development of the actual number of infections, both detected and undetected, thereby shedding light on the infection dynamics. We exemplify our approach by analyzing German data from 2020, making only use of data available since the beginning of the pandemic. Our modeling approach can be used to quantify the effect of different testing strategies, visualize the dynamics in the case detection ratio over time, and obtain information about the underlying true infection numbers, thus enabling us to get a clearer picture of the course of the COVID-19 pandemic in 2020.

KEYWORDS

case detection ratio, COVID-19, dark figure of infections, generalized additive models, penalized splines

1 | INTRODUCTION

Originating from Wuhan, China, coronavirus disease 2019 (COVID-19) developed to become a worldwide pandemic in the spring of 2020 (Velavan & Meyer, 2020). Starting from the very beginning of this unprecedented health crisis, the issue of case detection, while always being at the center of scientific and public discourse, has been all but transparent. Knowing how many infections are really present in the population would be of paramount importance, and researchers have tried to tackle the problem in several different ways. Early in the epidemic wave, the ratio of undetected COVID-19 cases was likely to be high, that is, 5–20 times higher than the number of confirmed cases (e.g., Li et al., 2020 or Wu et al., 2020). The problem of discovering the case detection ratio (CDR) is tightly intertwined with the issue of uncovering the true fatality ratio of the disease, as knowledge on one of those two unknown quantities would provide information about the other. A natural experiment that allowed to obtain initial estimates of both the fatality ratio and the CDR occurred with the outbreak on the cruise ship “Diamond Princess” (Mizumoto et al., 2020). During the early stages of the pandemic, the actual percentage of the population infected for 11 European countries was deduced from early estimates of the mortality

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2020 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

rates (Flaxman et al., 2020). Moreover, Aspelund et al. (2020) used Bayes arguments applied to testing data from Ireland to estimate the CDR in the order of 7–11% at the beginning of the pandemic, and in the order of 10–20% after that. The argument is based on relating the number of tests and the share of positive tests. A similar approach has been pursued making use of Canadian data (Benatia et al., 2020). The problem of estimating the true numbers of COVID-19 infections has also been discussed from a purely statistical point of view, where the CDR was related to the fatality ratio (Manski & Molinari, 2020). A capture–recapture approach to estimate the total number of COVID-19 cases was proposed by Böhning et al. (2020) and Rocchetti et al. (2020), where the latter derive an upper bound for the cumulative number in mid-April for 10 European countries. The ratio of the upper bound and the observed number of cases ranges from around 4 (Greece) to around 8 (France). The capture–recapture method makes only use of publicly available data on COVID-19 cases and deaths, which also holds for the method that we present in this note. Here, we assume that the number of infected can be split into detected and undetected infections. In SIDARTHE models (Giordano et al., 2020), there is additional distinction into either asymptomatic or symptomatic cases, which we ignore here since the database that we use does not reliably contain these numbers. However, it should be noted that pre- and asymptomatic individuals have a significant impact on the spread of a pandemic disease, especially in the younger population (Stella et al., 2020). Thereby, presymptomatic individuals play a more significant role than asymptomatic ones (Buitrago-Garcia et al., 2020). Nonetheless, the number of asymptomatic cases can reduce the reproduction value of a disease because a background immunity is established, as shown for influenza transmission (Mathews et al., 2007).

Overall, underreporting appears to be an overarching problem, which plays a central role when estimating the CDR for COVID-19 (Russell et al., 2020). The importance of assessing the detection ratio and its effect on predictions of future infections has been demonstrated in mathematical simulation studies (Fuhrmann & Barbarossa, 2020). In this context, different national underreporting ratios have been compared (e.g., Rahmandad et al., 2020 or Jagodnik et al., 2020) and a general discussion and survey on assessing the infection fatality ratio (IFR) was conducted (Levin et al., 2020). In general, it is clear that the CDR changes greatly over time depending on testing strategy and capacities, which vary over time and across different regions. In Germany, the number of tests has increased considerably since the pandemic outbreak in March 2020. The testing strategy has also been adjusted several times: In the beginning, mainly individuals with symptoms were being tested, whereas in later phases, a very high number of tests have been performed on travelers returning from foreign countries and contact persons of COVID-19-positive individuals.

In this note, we explore the dynamics in the CDR using publicly available registry data on COVID-19 infections in Germany from March to December 2020 provided by the Robert-Koch-Institute (RKI). It is important to mention that in Germany's first months of the pandemic, no mass or systematic testing of the population had taken place. Our model therefore only makes use of a limited amount of information. We propose to jointly model fatal and nonfatal infections using a dynamic generalized linear mixed model with smooth random effects (see, e.g., Durbán et al., 2005; Durban & Aguilera-Morillo, 2017; Wood, 2017). The major advantage of our approach is that it only relies on the assumption that age-specific COVID-19 fatality ratios, while unknown, have not substantially changed over time. Whether this assumption is valid is currently discussed (Harris, 2020; Kip et al., 2020) and the possibility of differing fatality ratios in the second wave has been considered as well (Aspelund et al., 2020; Kenyon, 2020). To assess the impact of this assumption on our results, we provide sensitivity analyses and a simulation study in the Supporting Information, which demonstrate that our approach is sufficiently robust if there is no abrupt change in the infection fatality ratio.

Overall, our approach allows investigating the following. First, we explore how the case detection rate has changed over time, how it varies among different age groups, and if and how it changes in different regions of Germany, depending on infection dynamics and different testing strategies. Second, the model also provides an estimate of the dynamics in the true number of infections, regardless of whether they have been detected or not. All in all, this provides insight into the course of the COVID-19 pandemic, built exclusively on registry data.

The remainder of the paper is structured as follows. We describe the data constellation in depth in Section 2, and we propose our model in Section 3. In Section 4, we show the results of our analyses and provide extensive interpretations, whereas Section 5 concludes the paper with some implications and limitations of our study.

2 | DATA

We make use of COVID-19 data openly provided by the RKI, the German federal government agency and scientific institute responsible for health reporting, disease control, and prevention in humans (Esri Deutschland GmbH, 2020). The data, exemplified in Table 1, contain cumulated counts of newly registered, laboratory-confirmed COVID-19 cases in Germany

11. A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020

TABLE 1 Illustration of the data structure. To facilitate reproducibility, the original column names used in the RKI dataset are given in brackets below our English notation

District (Landkreis)	Age group (Altersgruppe)	Gender (Geschlecht)	Cases (Anzahl Fall)	Deaths (Anzahl Todesfall)	Registration date (Meldedatum)
⋮	⋮	⋮	⋮	⋮	⋮
Munich City	60–79	F	26	0	September 8, 2020
Munich City	60–79	M	21	1	September 8, 2020
⋮	⋮	⋮	⋮	⋮	⋮

for each calendar day stratified by age group (0–4, 5–14, 15–34, 35–59, 60–79, or 80+ years), gender (male/female), and district (412 in total). Furthermore, for all registration dates and strata, the number of deaths associated with COVID-19 transmitted to the RKI by the local health authorities of the respective district is recorded. Note that the date of death is not provided, but for each death, we have the date when the infection was detected and confirmed by a (PCR) test. The database of the RKI is updated every morning with the new numbers transmitted to it from the local health authorities.

In this study, we only consider data entries with registration dates ranging from calendar week (CW) 10 (mid-March) to CW 53 (end of December) of the year 2020. For earlier weeks, the number of tests being positive was not large enough to draw conclusive results. On the other hand, the German vaccination campaign started at the very end of 2020. As this increasingly reduces the IFR, we only include infections that were registered in 2020. Consequently, the final outcome of almost all of these infections is known today. Moreover, although the data are given on a daily resolution, we here aggregate it into weekly data, which renders reporting delays occurring over the weekends and weekly reporting cycles irrelevant to our analysis, leading to more stable results. Since for children aged 14 years and younger, barely, any fatalities have been recorded, we excluded these age groups from our analysis.

To give a first insight into the data at hand, we plot in Figure 1 the raw numbers of cases reported by the official health authorities over time together with the raw number of fatalities stratified by age group. This is shown in the top four plots on a log-scale. Both the number of registered cases and that of fatal cases (indexed by registration date of the infection, and not by day of death) peak in CW 13 for the two younger age groups and in CW 14 for the two oldest age groups, respectively. Over the following weeks, these numbers decrease. The small peak in CW 25 was caused by an outbreak in the district of Gütersloh, which is explored in more depth later on in the paper. From CW 28 onward, we resume seeing an exponential increase of registered cases, whereas the numbers of registered fatal cases only start to rise 7 weeks later, also exponentially. By the end of the year 2020, we see a slight decrease in registered infections.

The raw case fatality ratio, calculated as the ratio of fatal cases over total registered cases, stratified by age group, is shown at the bottom of Figure 1. The raw case fatality ratio for the age group 80+ generally dropped from CW 10 onward and fluctuated mostly between 10% and 15% from week 25 onward. However, since CW 40 the case fatality ratio in this age group steadily climbed up to more than 20%. For the age group 60–79, the case fatality ratio has peaked in CW 16 and gradually decreased to 2.5%. Here, we also observe a steady increase toward the end of 2020, which results in more than a doubling of the case fatality ratio within 10 weeks. All other age groups exhibit relatively low raw case fatality ratios throughout.

Note that the raw data do not contain undetected cases, and therefore cannot provide a complete picture of the actual infection numbers, nor do these plots provide any information about the CDR. In the following, we develop a statistical model that enables us to estimate the relative changes in the CDR and the true infection numbers over time.

3 | METHODS

When describing the dynamics of the COVID-19 pandemic, the number of interest is the true count of newly infected persons in a cohort, which shall be denoted by I_t for week $t = 1, \dots, T$. Note that I_t remains unobservable. However, the number can be decomposed into the number of detected and reported cases D_t and the unknown number of newly infected persons, who have not been tested and remain undetected, which we can call the “dark number,” U_t . Hence, we have $I_t = D_t + U_t$, and D_t/I_t defines the CDR, which, however, remains unknown due to U_t being unknown.

Note that the index t indicates the time point on which the infection took place, which is usually unknown. The infection is eventually detected through a positive test at a later time point $\tilde{t} = t + d$. As d is often unknown, in particular, if the spread of the disease is diffuse, we will conceptually omit d in the following, which means that we set t equal to the

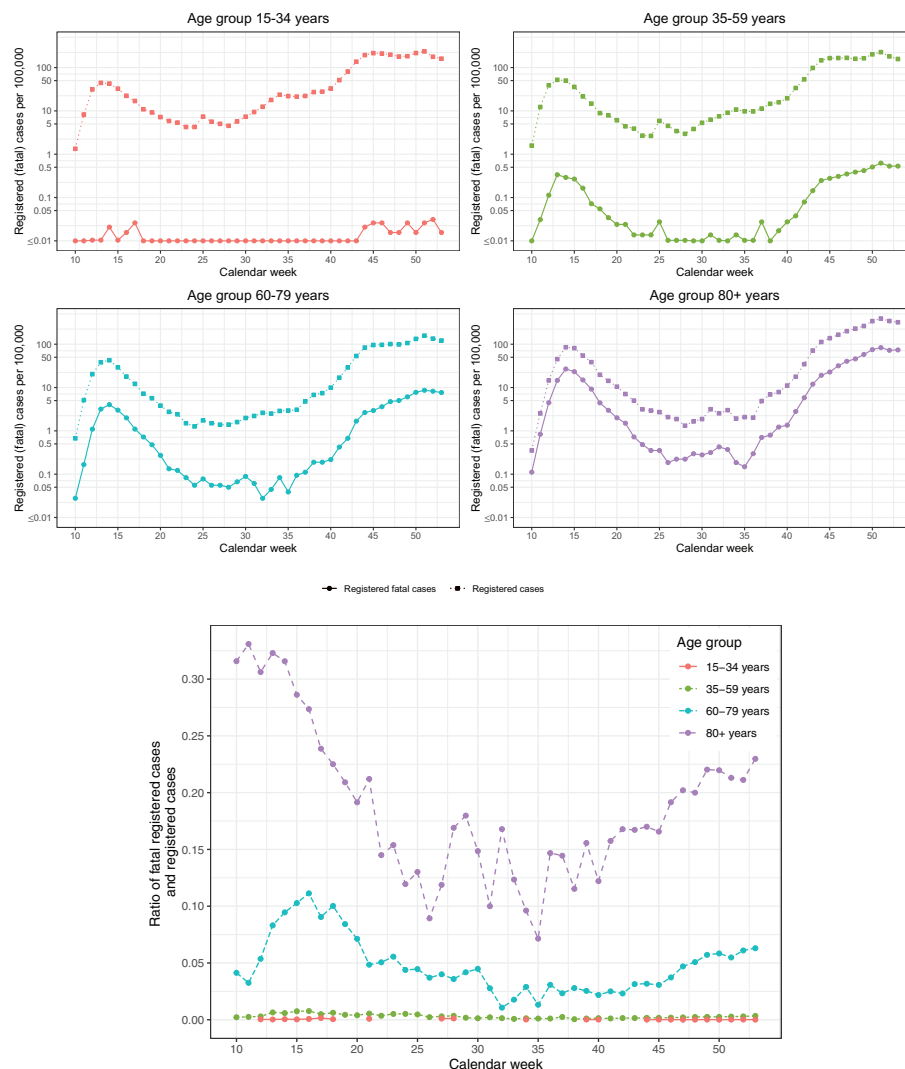


FIGURE 1 Raw data: registered cases of COVID-19 infections and registered fatal cases on a weekly basis for Germany. Top figure: Absolute numbers on a log-scale stratified by age group. Bottom figure: Case fatality ratios (= fatal cases / registered cases) stratified by age group

registration date when an infection is confirmed through a test. This time point is the registration date described in the previous section. Generally, this approach is justifiable for COVID-19 infections because the range of delay d is small compared to the time range T of our data analysis (Mallet et al., 2020).

From today's perspective, we have uncensored knowledge on the outcomes of all reported cases D_t . That is, we know if they ended fatally or if they recovered. Consequently, the reported cases are composed of recovered (nonfatal) outcomes R_t and fatal outcomes F_t , that is, $D_t = R_t + F_t$. Given this, the total number of infected persons splits into $I_t = R_t + F_t + U_t$.

The expected number of reported fatal cases F_t as well as the expected number of recovered cases R_t are fractions of the total number of infections I_t . This leads to

$$\mathbb{E}(F_t | I_t) = I_t a \text{ and } \mathbb{E}(R_t | I_t) = I_t c_t, \quad (1)$$

where $0 < (a + c_t) < 1$. Here, quantity a defines the infection fatality ratio (IFR), whereas c_t is the CDR of nonfatal (recovered) infections. Note that these nonfatal infections also include mild and symptom-free cases. Thus, if testing capacities are increased or the testing strategy is changed, c_t will change as well, which is incorporated in the notation by time index t . In contrast, the IFR a will be assumed to remain constant over time. This can be justified by the fact that fatal cases, due to their severeness, are likely to be detected independently of any testing policy. This also includes, to some extent, postmortem tests.

With this notation, we obtain the time-dependent case detection ratio $CDR_t = a + c_t$. Note that for the dark number, that is, the latent number of undetected infections U_t , it holds that $\mathbb{E}(U_t | I_t) = (1 - CDR_t)I_t$. It would, of course, be favorable to estimate the number of undetected infections U_t via estimation of a and c_t . However, when only the reported fatal and nonfatal cases F_t and R_t are known, these two ratios cannot be estimated due to nonidentifiability issues, which we will demonstrate below. Nonetheless, with the data at hand, we are able to estimate the ratio c_t/a . To see this, we rewrite the above model in an equivalent form by defining a binary covariate $x \in \{0, 1\}$ and by specifying the response variable Y_t through

$$Y_t | x = \begin{cases} F_t & \text{for } x = 0 \\ R_t & \text{for } x = 1. \end{cases}$$

This notational trick allows us to rewrite the above relations (1) as a regression model

$$\mathbb{E}(Y_t | I_t, x = 0) = \mathbb{E}(F_t | I_t) = \exp\{\log(I_t a)\} = \exp\{V_t + \alpha\}, \tag{2}$$

$$\mathbb{E}(Y_t | I_t, x = 1) = \mathbb{E}(R_t | I_t) = \exp\{V_t + \gamma_t\}, \tag{3}$$

where $V_t = \log(I_t)$, $\alpha = \log(a)$, and $\gamma_t = \log(c_t)$. Equations (2) and (3) can, in turn, be summarized into a single regression model formula

$$\mathbb{E}(Y_t | V_t, x) = \exp\{V_t + \alpha + x(\gamma_t - \alpha)\}. \tag{4}$$

Note that I_t and hence $V_t = \log(I_t)$ remain unobserved. We employ a Bayesian view and model V_t as normally distributed random effects $V_t \sim N(\mu_t, \sigma^2)$. Still, the parameters in model (4) are not identifiable, because any shift in μ_t and a matching negative shift in α and γ_t , respectively, results in the same model. This demonstrates the identifiability problem, which we have mentioned above. Hence, we are neither able to estimate the fatality ratio $a = \exp(\alpha)$ nor the time-dependent ratio $c_t = \exp(\gamma_t)$ with the data at hand. However, we can shift μ_t such that the integral of $\tilde{\mu}_t = \mu_t - k$ is equal to zero and define the global intercept $\beta_0 = \alpha + k$, which allows to rewrite (4) in an identifiable form (see Wood, 2017) to obtain the final regression model

$$\mathbb{E}(Y_t | V_t, x) = \exp(V_t + \beta_0 + x\beta_t) \text{ and } V_t \sim N(\tilde{\mu}_t, \sigma^2) \text{ for } t = 1, \dots, T, \tag{5}$$

where $\beta_t = \gamma_t - \alpha$ and $\exp(\beta_t) = c_t/a$. With this model, we can now explore the dynamics in the CDR. For two different time points t_1 and t_2 , we have using the small $o()$ notation

$$\frac{CDR_{t_2}}{CDR_{t_1}} = \frac{c_{t_2} + a}{c_{t_1} + a} = \frac{c_{t_2}}{c_{t_1}} \{1 + o(a)\} = \frac{\exp(\beta_{t_2})}{\exp(\beta_{t_1})} \{1 + o(a)\} \approx \frac{\exp(\beta_{t_2})}{\exp(\beta_{t_1})}. \tag{6}$$

The latter approximation in (6) holds as long as the fatality rate a is small, which holds for COVID-19. Consequently, $\beta_{t_2} - \beta_{t_1}$ can serve as a proxy for $\log(CDR_{t_2}) - \log(CDR_{t_1})$, and $\exp(\beta_{t_2} - \beta_{t_1})$ is a proxy for the relative change in the case detection ratio CDR_{t_2}/CDR_{t_1} .

Based on these considerations, we see that it is necessary to model the dynamics in time t more appropriately to derive stable estimates for the CDR. It is natural to assume that changes in the CDR over time do not occur suddenly but gradually. For instance, test capacities are slowly increased and test strategies are gradually changed. To accommodate this in our model (5), we fit β_t by a smooth function in time leading to a time-varying coefficient model (Hastie & Tibshirani, 1993). We also induce smooth dynamics on the random component, leading to a time-varying random effect (Durban &

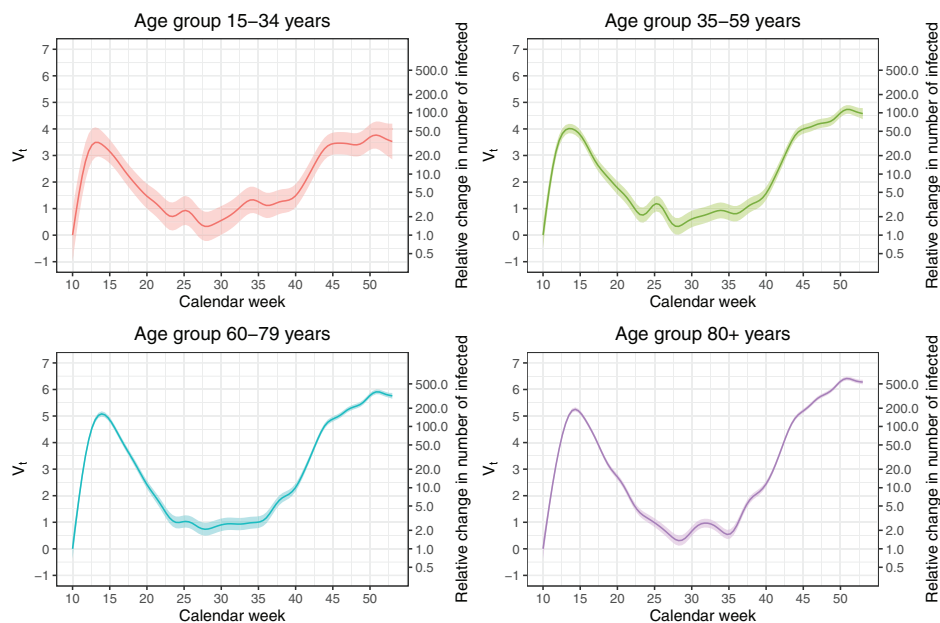


FIGURE 2 Dynamics of the true infection numbers on the log-scale for different age groups: The smooth random effects V_t . The shaded areas represent 95% confidence bands

Aguilera-Morillo, 2017). These modifications lead to an identifiable and dynamic mixed regression model, for which we use a negative-binomial distribution for Y_t with a constant dispersion factor. The entire model can be fitted with standard software: All of our analyses were performed in **R** (R Core Team, 2013) and the dynamic mixed regression model is fitted using the **R**-package **mgcv** (Wood, 2017).

We apply this modeling approach using the reported data from CW 10 (beginning of March) up to CW 53 (final week of 2020), stratified by different age groups, to visualize the dynamics in the real infection numbers and the CDR from the beginning of the pandemic up to the beginning of the second wave. To assess the robustness of the approach concerning the assumption of time-constant and age-specific fatality ratios, we also refit the model when subdividing the data into different time frames. The results of this analysis are shown in the Supporting Information.

4 | RESULTS

4.1 | Model estimates

As the IFR a depends on age, we fit separate models for each of the relevant age groups defined by the RKI, that is, 15–34, 35–59, 60–79, and 80+ years. The dynamics in the true infection numbers on the log-scale, represented by the fitted smooth dynamic random effects V_t , are displayed in Figure 2. These curves mirror the relative change in the actual number of infected (detected and undetected) over time. Note that the absolute numbers cannot be interpreted on their own due to the mentioned identifiability issues. We therefore shift the curves such that $V_{CW10} = 0$. We can see that the relative course of the pandemic was very similar across all age groups, where a peak is reached around CW 14. However, the peak for the younger age groups is estimated to be around 1 week earlier than for the older age groups, that is, in CW 13. An explanation for this finding is that the younger age groups have been more affected by the lockdown, which started in Germany in CW 12. Looking at the difference between the maximum $\max_t V_t$ and the minimum of V_t during the summer months, that is, $\min_{20 \leq t \leq 40} V_t$, we see that this difference increases with age, that is, the relative decline in true infections numbers after the first wave and the relative increase toward the second wave, respectively, was less pronounced in the younger age groups. Also eye-catching is the increase in infections around CW 25 for people below 60 years of age. This is

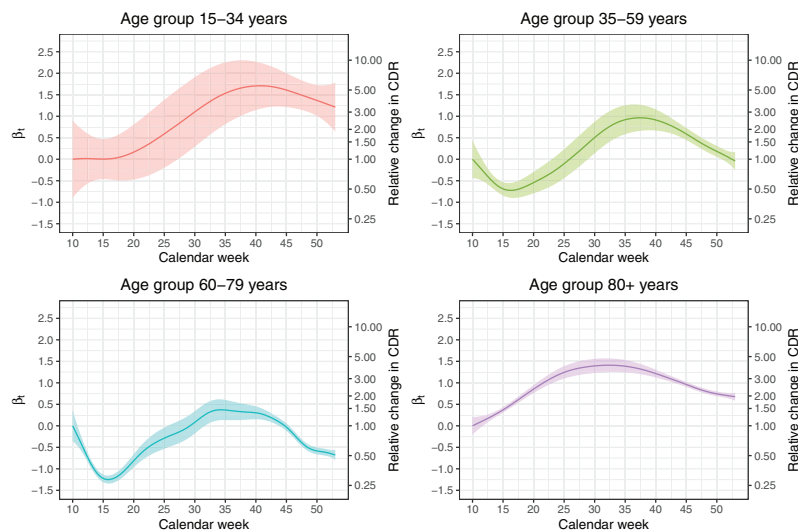


FIGURE 3 Dynamics in the case-detection ratio for different age groups: The normalized time-varying coefficients β_t . The function values on the exp-scale (right y-axes) are the relative change in the case-detection ratio (CDR) with respect to calendar week 10

the aforementioned outbreak in the district of Gütersloh, which occurred in an industrial slaughterhouse and has mainly affected people of the working age. From CW 35 (end of August), all curves start rising steadily, where the steepest rise is seen for the oldest age group, whereas the rise is flatter for the younger age group. This shows that the second wave of the pandemic had already begun around CW 35. Moreover, Figure 2 shows that in all age groups but the youngest one, the peak of the second wave has surpassed the peak of the first wave.

Next, we look at the dynamics in the CDR. Figure 3 shows the fitted time-varying coefficients β_t , together with corresponding 95% confidence bands. Again, the absolute level is not identifiable, so these curves are normalized such that $\beta_{CW10} = 0$. Hence, the function values on the exp-scale (right y-axes) give the relative change in the CDR with respect to CW 10. The CDR in the age group 80+ has risen monotonically since the beginning of the pandemic up to CW 33, where our model estimates the CDR to be more than four times higher as in mid-March. Note that in later weeks, the CDR among the elderly decreased again to the level of April/May. In contrast, for people aged 60–79, the CDR first dropped by about 70%, reaching its bottom as the pandemic passed its peak in Germany in CW 16. We subsequently see a monotonic increase, with the CDR becoming 1.5 times higher compared to the beginning of the pandemic. However, in this age group, the CDR has been more than halved from CW 40 up to the end of 2020 again. The dynamics in the CDR in the population aged 35–59 years are similar to those of the 60–79 years old: After a drop during March and April (CW 10–CW 16), the CDR increases, in mid-September, to nearly three times what it was in CW 10. For the youngest age group (aged 15–34), we also see a rise in the CDR over time, which seems substantial. However, the confidence bands in this age group are relatively wide because this age group is not as prone to fatal outcomes as older age groups.

4.2 | Interpretations

For the population aged 80 years and older, the CDR had increased until late summer, when it started to stagnate before slightly decreasing again. As the CDR can be at most 100%, and given that the relative change in this age group was about as high as a factor of 4 in CW 33 compared to March, we can conclude that at the beginning of the pandemic, the CDR among the population of 80 years and older could not have been more than 25%. Moreover, considering the relative change in the CDR, we can adjust the numbers from the peak in the first wave to be comparable, for example, to the numbers in week 40. To exemplify this, note that in week 40, the CDR for the age group 80+ was 2.3 times higher as in CW 15, at the peak of the first wave. This ratio results from the plot in Figure 3 (bottom right) by taking $\beta_{CW15} = 0.4$ and $\beta_{CW40} = 1.25$ and calculating the ratio $\exp(1.25 - 0.4) = 2.3$. In week 40, we had about 11 new infections per week per 100,000 reported

in this age group. In CW 15, this number had become 80. However, in week 15, the CDR was much lower as in CW 40, and thus, we would have seen $2.3 \cdot 80 = 184$ cases per 100,000 in this age group 80+ if we had the same CDR in CW 15 as in CW 40.

For the population aged 60–79 years, the CDR between the minimum in CW 16 and its maximum in calendar week 34 changed by a factor of around 5. From this, we can deduce that around the peak of the first wave in Germany, at most 20% of the infections were detected, whereas at least 80% remained unseen. To be able to compare numbers from the first wave to those in autumn, we apply a similar calculation as above. This results in an estimated number of at least $5 \cdot 17 = 85$ cases per 100,000, where only 16 cases per 100,000 have been observed in CW 16.

In the age group 35–59, the relative change of CDR during the minimum in CW 16 and the maximum in CW 36 was as high as a factor of 5 as well. Again, the same calculation shows that the 22 detected infections per 100,000 in week 16 would increase to $5 \cdot 22 = 110$ cases per 100,000 if we would have had the CDR in week 16 as it was in week 36.

A general question in the pandemic is whether extensive testing leads to a high CDR. Applying our model to regional data allows us to investigate this question. The Supporting Information compares separate model fits for the two most populous German states, North-Rhine-Westphalia and Bavaria. The two states implemented different testing strategies over the summer months. Although in Bavaria, public test stations were opened in summer, particularly at the borders on the motorways, such fine screening of holiday returnees was not pursued in North-Rhine-Westphalia. Our model allows assessing and, in particular, quantifying how such different testing strategies lead to different CDRs in these two regions. The results quantify by how much the dark figure was reduced in relationship with the Bavarian testing strategy.

5 | DISCUSSION

Raw reported case numbers and measures derived from them, such as the case fatality ratio, are prone to changes in testing strategies and test capacities, which also influence the CDR. Comparisons between raw case numbers over time therefore need to be interpreted with care. The case-fatality ratio, calculated from the raw number of reported deaths related to COVID-19 divided by reported cases, is also impaired because deaths occur with a time delay after registration, meaning that deaths registered today correspond to infections that have been reported up to several weeks ago. Our method allows us to uncover relative changes in the CDR over different pandemic phases. Moreover, by shedding light on the number of undetected cases, we can describe the dynamics in the true number of COVID-19 infections for Germany from March 2020 until December 2020. The approach is based on publicly available data on registered cases and does not rely on simulations or additional survey data. We make use of the fact that, for each fatal outcome, the registration date of the infection is included in the data. This allows us to jointly model the number of registered nonfatal cases and that of fatal infections in a dynamic mixed model, leading to an assessment of the dynamics taking place in real infection numbers. Based on the available information on the relative change in the CDR over time, we are able to compare numbers from the first wave of the pandemic in spring with numbers from the second wave in autumn, adjusting for the difference in the proportion of undetected cases.

A general limitation of our approach is that it suffers from an identifiability issue and hence does not derive absolute values of the CDR. One may, however, combine our results with findings from seroepidemiological studies, which aim to assess the prevalence of COVID-19 in the general population by screening a representative sample. A list of current seroepidemiological studies in Germany is provided by the RKI (Robert-Koch-Institute, 2020). Although these studies provide crucial information on the current situation of the spread of the disease, they can only give a snapshot of the instantaneous situation when the study was conducted. With the knowledge of the dynamics in new infections given by our approach, the findings of such studies can be used to estimate the situation at other time points. For example, we look at the Prospective Covid-19 Cohort Study Munich (KoCo19, Radon et al., 2020). They report a CDR of about 25%, where the survey was run between May and June 2020 in the city of Munich. We can deduce that the CDR for October to be about three times higher for the 35–59 age group. More precise calculations would require age-specific numbers in the study as well as a regional refit of our model. A nationwide seroprevalence study was conducted between the beginning of July and mid-August of 2020, which yielded a CDR of around 55% in the adult population (ifo Institut & fors, 2020). Nonetheless, the authors admit that the fading of COVID-19 antibodies could influence their findings sometime after the infection. A seroprevalence study, which is also nationwide but on a larger scale, is currently being carried out, but the results are not yet available.¹ In principle, however, this demonstrates that the combination of seroepidemiological studies and our approach allows obtaining estimates for absolute numbers of the CDR instead of relative comparisons only.

¹https://www.rki.de/DE/Content/Gesundheitsmonitoring/Studien/lid/lid_node.html?jsessionid=02C6FAB6F407B92315BDA5C1650F4D3A.internet072

A critical assumption of our model is that we assume the IFR a to be constant over time for a given age group and negligibly small compared to the detection ratio of nonfatal cases. The latter is certainly valid for the numbers we looked at. Staerk et al. (2021) show that most of the dynamics in the effective IFR of the German population can be explained by the varying age distribution of COVID-19 cases. As the age distribution within the RKI age categories varies as well, the IFR a within each age group might slightly change over time that, however, occurs not abruptly but smoothly over time. The sensitivity analysis, which can be found in the Supporting Information, provides evidence that our assumption of a being constant is, for the most part, fulfilled. With increasing vaccination levels in the population starting from January 2021, the assumption of a constant case fatality ratio becomes invalid. This eventually prevents the application of our model to later stages of the pandemic.


CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at <https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74>.

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the [Supporting Information](#) section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Marc Schneble  <https://orcid.org/0000-0001-9523-4173>

REFERENCES

- Aspelund, K., Droste, M., Stock, J. H. & Walker, C. D. (2020). Identification and estimation of undetected COVID-19 cases using testing data from Iceland. NBER Working Paper w27528.
- Benatia, D., Godefroy, R., & Lewis, J. (2020). Estimates of COVID-19 cases across four Canadian provinces. *Canadian Public Policy*, 46(S3), S203–S216.
- Böhning, D., Rocchetti, I., Maruotti, A., & Holling, H. (2020). Estimating the undetected infections in the COVID-19 outbreak by harnessing capture–recapture methods. *International Journal of Infectious Diseases*, 97, 197–201.
- Buitrago-Garcia, D., Egli-Gany, D., Counotte, M. J., Hossmann, S., Imeri, H., Ipekci, A. M., Salanti, G., & Low, N. (2020). Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2 infections: A living systematic review and meta-analysis. *PLoS Medicine*, 17(9), e1003346.
- De Boor, C. (1972). On calculating with B-splines. *Journal of Approximation Theory*, 6(1), 50–62.
- Drikvandi, R., Verbeke, G., & Molenberghs, G. (2017). Diagnosing misspecification of the random-effects distribution in mixed models. *Biometrics*, 73(1), 63–71.
- Durban, M., & Aguilera-Morillo, M. C. (2017). On the estimation of functional random effects. *Statistical Modelling*, 17(1–2), 50–58.
- Durbán, M., Harezlak, J., Wand, M., & Carroll, R. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24(8), 1153–1167.
- Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2), 89–121.
- Esri Deutschland GmbH. (2020). Daily COVID-19 case numbers provided by the Robert-Koch-Institute. <https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74>
- Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., Monod, M., Ghani, C. A., Donnelly, A. C., Riley, S., Vollmer, M. A. C., Ferguson, N. M., Okell, L. C., & Bhatt, S. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820), 257–261.
- Fuhrmann, J., & Barbarossa, M. V. (2020). The significance of case detection ratios for predictions on the outcome of an epidemic - A message from mathematical modelers. *Archives of Public Health*, 78(63).
- Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, 26(6), 855–860.
- Harris, J. E. (2020). COVID-19 case mortality rates continue to decline in Florida. *medRxiv*.
- Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4), 757–779.

- ifo Institut, & forsa. (2020). Die Deutschen und Corona - Schlussbericht der BMG-“Corona-BUND-Studie”. <https://www.ifo.de/publikationen/2020/monographie-autorenschaft/die-deutschen-und-corona>
- Jagodnik, K. M., Ray, F., Giorgi, F. M., & Lachmann, A. (2020). Correcting under-reported COVID-19 case numbers: Estimating the true scale of the pandemic. *medRxiv*.
- Kenyon, C. (2020). Flattening-the-curve associated with reduced COVID-19 case fatality rates-an ecological analysis of 65 countries. *Journal of Infection*, *81*(1), e98–e99.
- Kip, K. E., Snyder, G., Yealy, D. M., Mellors, Minnier, T., Donahoe, M. P., McKibben, J., Collins, K., & Marroquin, O. C. (2020). Temporal changes in clinical practice with COVID-19 hospitalized patients: Potential explanations for better in-hospital outcomes. *medRxiv*.
- Levin, A., Hanage, W., Owusu-Boaitey, N., Cochran, B., Walsh, S. P., & Meyerowitz-Katz, G. (2020). Assessing the age specificity of infection fatality rates for COVID-19: Systematic review, meta-analysis, and public policy implications. *European Journal of Epidemiology*, *35*, 1123–1138.
- Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, *368*(6490), 489–493.
- Mallett, S., Allen, A. J., Graziadio, S., Taylor, S. A., Sakai, N. S., Green, K., Suklan, J., Hyde, C., Shinkins, B., Zhelev, Z., Peters, J., Turner, P. J., Roberts, N. W., di Ruffano, L. F., Wolff, R., Whiting, P., Winter, A., Bhatnagar, G., Nicholson, B. D., & Halligan, S. (2020). At what times during infection is SARS-CoV-2 detectable and no longer detectable using rt-pcr-based tests? A systematic review of individual participant data. *BMC Medicine*, *18*.
- Manski, C. F., & Molinari, F. (2020). Estimating the COVID-19 infection rate: Anatomy of an inference problem. *Journal of Econometrics*, *220*, 181–192.
- Mathews, J. D., McCaw, C. T., McVernon, J., McBryde, E. S., & McCaw, J. M. (2007). A biological model for influenza transmission: pandemic planning implications of asymptomatic infection and immunity. *PLoS One*, *2*(11), e1220.
- Mizumoto, K., Kagaya, K., Zarebski, A., & Chowell, G. (2020). Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. *Eurosurveillance*, *25*(10), 2000180.
- R Core Team. (2013). R: A language and environment for statistical computing.
- Radon, K., Saathoff, E., Pritsch, M., Guggenbühl, N., Jessica, M., Kroidl, I., Olbrich, L., Thiel, V., Diefenbach, M., Riess, F., Forster, F., Theis, F., Wieser, A., Hoelscher, M., Bakuli, A., Eckstein, J., Froeschl, G., Geisenberger, O., Geldmacher, C. ... Schwetmann, L. (2020). Protocol of a population-based prospective COVID-19 cohort study Munich, Germany (KoCo19). *medRxiv*.
- Rahmandad, H., Lim, T. Y., & Sterman, J. (2020). Estimating COVID-19 under-reporting across 86 nations: Implications for projections and control. Available at SSRN 3635047.
- Robert-Koch-Institute. (2020). Seroepidemiological studies in the general population. https://www.rki.de/EN/Content/infections/epidemiology/outbreaks/COVID-19/AK-Studien-english/Sero_General.html
- Rocchetti, I., Böhning, D., Holling, H., & Maruotti, A. (2020). Estimating the size of undetected cases of the COVID-19 outbreak in Europe: An upper bound estimator. *Epidemiologic Methods*, *9*(s1).
- Russell, T. W., Hellewell, J., Abbott, S., Jarvis, C., van Zandvoort, K., Ratnayake, R., CMMID nCov working group, Flasche, S., Eggo, R., Edmunds, W. J., & Kucharski, A. J. (2020). *Using a delay-adjusted case fatality ratio to estimate under-reporting*. Centre for Mathematical Modelling of Infectious Diseases Repository.
- Staerk, C., Wistuba, T., & Mayr, A. (2021). Estimating effective infection fatality rates during the course of the COVID-19 pandemic in Germany. *BMC Public Health*, *21*(1073).
- Stella, L., Martínez, A. P., Bauso, D., & Colaneri, P. (2020). *The role of asymptomatic individuals in the covid-19 pandemic via complex networks*. arXiv preprint arXiv:2009.03649.
- Velavan, T. P., & Meyer, C. G. (2020). The COVID-19 epidemic. *Tropical Medicine & International Health*, *25*(3), 278–280.
- Wood, S. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman & Hall/CRC Texts in Statistical Science. CRC Press.
- Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563.
- Wu, S. L., Mertens, A. N., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., Seth, A., Hsiang, M. S., Colford, J. M., Reingold, A., Arnold, B. F., Hubbard, A., & Benjamin-Chung, J. (2020). Substantial underestimation of SARS-CoV-2 infection in the United States. *Nature Communications*, *11*(1), 1–10.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Schneble, M., De Nicola, G., Kauermann, G., & Berger, U. A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020. *Biometrical Journal*. 2021;63:1623–1632. <https://doi.org/10.1002/bimj.202100125>

12. Statistical modelling of COVID-19 data: Putting generalized additive models to work

Contributing article

Fritz, C., De Nicola, G., Rave, M., Weigert, M., Khazaei, Y., Berger, U., Küchenhoff, H. and Kauermann, G. (2022). Statistical modelling of COVID-19 data: Putting generalized additive models to work. *Statistical Modelling (OnlineFirst)*. <https://doi.org/10.1177/1471082X221124628>.

Copyright information

This article is distributed under a Creative Commons NonCommercial 4.0 International license (CC BY-NC 4.0).

Supplementary material

[Supplementary material](#) is available online.

Author contributions

The concept of a research article highlighting the use of generalized additive models in the analysis of COVID-19 data was developed by Göran Kauermann. The manuscript, which consists of three independent applications of generalized additive models in the context of COVID-19 data analysis, was jointly drafted by Cornelius Fritz, Giacomo De Nicola, Martje Rave and Maximilian Weigert. Giacomo De Nicola was responsible for designing and drafting Sections 1 and 2. Furthermore, Giacomo De Nicola substantially contributed to writing Section 4 together with Maximilian Weigert, who carried out the modeling and data analysis for this section. The analysis on associations between infections from different age groups was conducted and drafted by Cornelius Fritz, the analysis on ICU occupancy by Martje Rave. Ursula Berger, Helmut Küchenhoff and Göran Kauermann monitored and supervised the entire research process. All authors contributed through fruitful comments and extensive proofreading of the manuscript.



Statistical Modelling xxxx; xx(x): 1–24

Statistical modelling of COVID-19 data: Putting generalized additive models to work

Cornelius Fritz,¹ Giacomo De Nicola,¹ Martje Rave,¹ Maximilian Weigert,¹ Yeganeh Khazaei,¹ Ursula Berger,² Helmut Küchenhoff¹ and Göran Kauermann¹

¹Department of Statistics, Ludwig-Maximilians-University Munich, Munich, Germany

²Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-University Munich, Munich, Germany

Abstract: Over the course of the COVID-19 pandemic, Generalized Additive Models (GAMs) have been successfully employed on numerous occasions to obtain vital data-driven insights. In this article we further substantiate the success story of GAMs, demonstrating their flexibility by focusing on three relevant pandemic-related issues. First, we examine the interdependency among infections in different age groups, concentrating on school children. In this context, we derive the setting under which parameter estimates are independent of the (unknown) case-detection ratio, which plays an important role in COVID-19 surveillance data. Second, we model the incidence of hospitalizations, for which data is only available with a temporal delay. We illustrate how correcting for this reporting delay through a nowcasting procedure can be naturally incorporated into the GAM framework as an offset term. Third, we propose a multinomial model for the weekly occupancy of intensive care units (ICU), where we distinguish between the number of COVID-19 patients, other patients and vacant beds. With these three examples, we aim to showcase the practical and ‘off-the-shelf’ applicability of GAMs to gain new insights from real-world data.

Key words: Case-detection ratio, COVID-19, generalized additive models, modelling icu occupancy, nowcasting

Received December 2021; revised June 2022; accepted June 2022

1 Introduction

From the early stages of the COVID-19 crisis, it became clear that looking at the raw data would only provide an incomplete picture of the situation, and that the application of principled statistical knowledge would be necessary to understand the manifold facets of the disease and its implications (Panovska-Griffiths, 2020; Pearce et al., 2020). Statistical modelling has played an important role in providing decision-makers with robust, data-driven insights in this context. In this article, we specifically highlight the versatility and practicality of Generalized Additive Models (GAMs). GAMs constitute a well-known model class, dating back to Hastie and Tibshirani (1987), who extended classical Generalized Linear Models (Nelder and Wedderburn, 1972) to include non-parametric

Address for correspondence: Göran Kauermann, Department of Statistics, Ludwig-Maximilians-University Munich, Ludwigstr. 33, 80539 München, Germany.
E-mail: goeran.kauermann@stat.uni-muenchen.de



© 2022 The Author(s)

10.1177/1471082X221124628

smooth components. This framework allows the practitioner to model arbitrary target variables that follow a distribution from the exponential family to depend on covariates in a flexible manner. Due to the duality between spline smoothing and normal random effects, mixed models with Gaussian random effects are also encompassed in this model class (Kimeldorf and Wahba, 1970). One can justifiably claim that the model class is one of the main work-horses in statistical modelling (see Wood, 2017 and Wood, 2020 for a comprehensive overview of the most recent advances) and numerous authors have already used this model class for COVID-19-related data analyses. As research on topics related to COVID-19 is still developing rapidly, a complete survey of applications is impossible; hence, we here only highlight selected applications, sorted according to the topic they investigate. Many applications analyse the possibly non-linear and delayed effect of meteorological factors (including, e.g., temperature, humidity, and rainfall) on COVID-19 cases and deaths (see Goswami et al., 2020; Prata et al., 2020; Ward et al., 2020; Xie and Zhu, 2020). While the results for cold temperatures are consistent across publications in that the risk of dying of or being infected with COVID-19 increases, the findings for high temperatures diverge between studies from no effects (Xie and Zhu, 2020) to U-shaped effects (Ma et al., 2020). Logistic regression with a smooth temporal effect, on the other hand, was used to identify adequate risk factors for severe COVID-19 cases in a matched case-control study in Scotland (McKeigue et al., 2020). In the field of demographic research, Basellini and Camarda (2021) investigate regional differences in mortality during the first infection wave in Italy through a Poisson GAM with Gaussian random effects that account for regional heterogeneities. With fine-grained district-level data, Fritz and Kauermann (2022) present an analysis confirming that mobility and social connectivity affect the spread of COVID-19 in Germany. Wood (2021) shows that UK data strongly suggest that the decline in infections began before the first full lockdown, implying that the measures preceding the lockdown may have been sufficient to bring the epidemic under control. This list of applications illustrates how GAMs have been successfully employed to obtain data-driven insights into the societal and healthcare-related implications of the crisis.

We contribute to this success story by focusing on three applications to demonstrate the ‘off-the-shelf’ usability of GAMs. First, we investigate how infections of children influence the infection dynamics in other age groups. In this context, we detail in which setting the unknown case-detection ratio does not affect the (multiplicative) parameter estimates of interest. Second, we show how correcting for a reporting delay through a nowcasting procedure akin to that proposed by Lawless (1994) can be naturally incorporated in a GAM as an offset term. Here, the application case focuses on the reporting delay of hospitalizations. Third, we propose a prediction model for the occupancy of Intensive Care Units (ICU) in hospitals with COVID-19 and non-COVID-19 patients. We thereby provide authorities with interpretable, reliable and robust tools to better manage healthcare resources.

The remainder of the article is organized as follows: Section 2 shortly describes the available data on infections, hospitalizations and ICU capacities that we use in the subsequent analyses, which are presented in Sections 3, 4 and 5, respectively. We conclude the article in Section 6.

2 Data

For our analyses, we use data from official sources, which we describe below. Note that our applications are limited to Germany although all of our analyses could be extended to other countries given

data availability. We pursue all subsequent analyses on the spatial level of German federal districts, which we henceforth refer to as ‘districts’. This spatial unit corresponds to NUTS 3, the third and most fine-grained category of the NUTS European standard (Nomenclature of Territorial Units for Statistics). We refer to Annex A for a graphical depiction of the spatial resolution of the data.

Infections and hospitalizations For investigating infection dynamics across different age groups, we use data provided by the Bavarian Health and Food Safety Authority (Landesamt für Gesundheit und Lebensmittelsicherheit, LGL). This statewide register includes, the registration date for all COVID-19 infections reported in Bavaria, as well as information on the patient’s age and gender. Infection data for Germany is also published daily by the RKI (Robert Koch Institute, 2021), the German federal government agency and scientific institute responsible for health reporting and disease control. Due to privacy protection, the RKI groups patients in broad age categories, which inhibits the analysis of the group of school children. As this is necessary for our first application in Section 3.3, we restrict the analysis to Bavarian data and use LGL data where not stated otherwise.

In addition, the LGL dataset includes information on the hospitalization status of each patient, which is not included in the RKI data, that is, whether or not a case has been hospitalized and the date of hospitalization, if this had occurred. We determine the date on which a hospitalized case is reported to the health authorities by matching the cases across the downloads available on different dates. This is necessary in order to derive the reporting delay for each hospitalization, which is of interest in Section 4.

Intensive care unit occupancy Data on the daily occupancy of ICU beds in Germany, on the other hand, is made publicly available by the German Interdisciplinary Association for ICU Medicine and Emergency Medicine (Deutsche interdisziplinäre Vereinigung für Intensiv und Notfallmedizin, DIVI, 2021). Using this dataset we obtain information on the number of high and low care ICU-beds occupied by patients infected with COVID-19 and patients not infected with COVID-19. As a third category, there are also the vacant beds. In contrast to the infection data, no information is available on the age or gender composition of the occupied beds.

Population data In conjunction with the data sources described above, we use demographic data on the German population at the administrative district level, provided by the German Federal Statistical Office (DESTATIS). Since the raw numbers on infections and hospitalizations are strongly influenced by the number of people living in a particular district, we use this population data to transform the absolute infection and hospitalizations to incidence rates. In general, we use the term incidence rates to refer to infection incidence rates, and hospitalization incidence rates when writing about hospitalizations. While we effectively model the incidence rate in Section 3 and the hospitalization incidence rate in Section 4, we incorporate the incidence rate per 100.000 inhabitants as a regressor in Section 5.

3 Analysing associations between infections from different age groups

A central focus during the COVID-19 pandemic is to identify the main transmission patterns of the infection dynamics and their driving factors. In this context, the role of children in schools for the

general incidence poses an important question with many socio-economic and psychological implications to it (see Andrew et al., 2020; Luijten et al., 2021). Since findings from previous influenza epidemics have tended to identify the younger population, children aged between 5 and 17, as the key ‘drivers’ of the disease (Worby et al., 2015), the German government ordered school closures throughout the course of the pandemic between spring 2020 and 2021 to contain the pandemic. However, whether these measures were necessary or effective in the case of COVID-19 is still subject to current research (e.g., Perra, 2021). In particular, several studies investigated the global effect of infections among school children, but a general conclusion could not be drawn (see Flasche and Edmunds, 2021; Hippich et al., 2021; Hoch et al., 2021; Im Kampe et al., 2020). In general, we would like to remark that in many studies the main goal was to arrive at conclusions about the susceptibility, severity, and transmissibility of COVID-19 for children (Gaythorpe et al., 2021). On the other hand, we are here primarily interested in quantifying how the incidences of children are associated with the incidences in other age groups. Therefore, we want to assess whether children are key ‘drivers’ of the pandemic. Our analysis is based on aggregated data on the macro level, as opposed to the data on the individual level, which is needed to answer hypotheses, for example, about the susceptibility of a particular child.

3.1 Autoregressive model for incidences

To tackle this problem from a statistical point of view, we propose to analyse the infection data using a time-series approach (Fokianos and Kedem, 2004). Let therefore $Y_{w,r,a}$ denote the number of infections in week w in district r and age group a . For simplicity, we assume independent developments among the districts and let $Y_{w,r,a}$ depend on the incidences in all age groups from the previous week $w - 1$. Put differently, we include $Y_{w-1,r} = (Y_{w-1,r,1}, \dots, Y_{w-1,r,A})$ as covariates, where $1, \dots, A$ indexes all A considered age groups. Among the components of $Y_{w,r}$ we then postulate independence conditional on $Y_{w-1,r}$. For illustration, Figure 1 depicts the assumed dependence structure. As for the distributional assumption, we make use of a negative binomial distribution with mean structure

$$\mathbb{E}(Y_{w,r,a} | Y_{w-1,r}) = \exp\{\eta_{w,r,a} + o_{r,a}\} \quad (3.1)$$

where $o_{r,a}$ serves as offset and η gives the linear predictor. To be specific, we set $o_{r,a} = \log(x_{\text{pop},r,a})$, where $x_{\text{pop},r,a}$ is the time-constant population size in district r and age group a . Note that we implicitly model the incidences by incorporating this offset term, since the incidences $I_{w,r,a}$ relate to the counts through $Y_{w,r,a} = I_{w,r,a} x_{\text{pop},r,a}$. The linear predictor is now defined as

$$\eta_{w,r,a} = \theta_w + \sum_{k=1}^A \log(Y_{w-1,r,k} + \delta) \theta_{a,k}, \quad (3.2)$$

where θ_w serves as week-specific intercept, $\theta_{a,k}$ is the coefficient weighting the influence of lagged infections of age group k on the infections in age group a and δ is a small constant, which is included

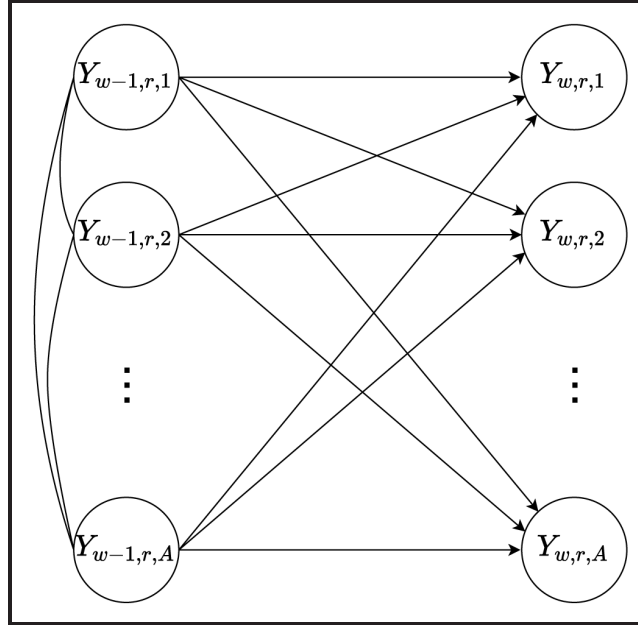


Figure 1 Assumed temporal dependence structure visualized as a directed acyclic graph (DAG)

for numerical stability to cope with zero infections, . We set δ to 1 in the calculation but omit the term subsequently for a less cluttered notation.

3.2 Robustness under time-varying case-detection ratio

Model (3.1) has the important methodological advantage of being able to cope with an unknown case-detection ratio, which is inevitable if there are under-reported cases. This is a key problem in COVID-19 surveillance as not all infections are reported (Li et al., 2020); hence the case-detection ratio (CDR) is typically less than one. Various approaches have been pursued to quantify the number of unreported cases, for example, by estimating the proportion of current infections which are not detected by PCR tests (Schneble et al., 2021a). For demonstration, assume that $\tilde{Y}_{w,r,a}$ are the detected infections in week w in district r for age group a , while $Y_{w,r,a}$ are the true infections. Apparently $\tilde{Y}_{w,r,a} \leq Y_{w,r,a}$ holds if we assume under-reporting. We assume multiplicative under-reporting and denote with $0 < R_{w,r,a} \leq 1$ the multiplicative CDR in district r in age group a and set with $R_{w,r} = (R_{w,r,1}, \dots, R_{w,r,A})$ the joint CDRs for all A available age groups. In this setting, we observe

$$\tilde{Y}_{w,r,a} = R_{w,r,a} Y_{w,r,a} \tag{3.3}$$

infections in the corresponding week w , district r , and age group a from the $Y_{w,r,a}$ true infections. Apparently, integrity for $Y_{w,r,a}$ is not guaranteed with (3.3), which we could, however, impose by rounding. We further assume that $R_{w,r,a}$ and $Y_{w,r,a}$ are independent of each other, conditional on the

previous week's data. We further assume that $R_{w,r,a}$ are independent random draws for the different districts, thus the case-detection ratio may vary between the districts. Assuming further an i.i.d. setting such that $\mathbb{E}(R_{w,r,a}) = \pi_{w,a}$ yields for model (3.1) under (3.3):

$$\begin{aligned} \mathbb{E}(\tilde{Y}_{w,r,a} | \tilde{Y}_{w-1,r}) &= \mathbb{E}_{R_w, R_{w-1}} \left(\mathbb{E}_{Y_w} (R_{w,r,a} Y_{w,r,a} | \tilde{Y}_{w-1,r}, R_{w,r,a}, R_{w-1,r}) \right) \\ &= \mathbb{E}_{R_w, R_{w-1}} \left(R_{w,r,a} \mathbb{E}_{Y_w} (Y_{w,r,a} | Y_{w-1,r}) \right) \\ &= \pi_{w,a} \mathbb{E}_{R_{w-1}} \left(\exp\{\eta_{w,r,a}\} \right) \exp\{o_{r,a}\} \end{aligned} \quad (3.4)$$

where for clarity we include the random variable as an index in the notation of the expectation. Note that

$$\begin{aligned} \mathbb{E}_{R_{w-1}} \left(\exp\{\eta_{w-1,r,a}\} \right) &= \mathbb{E}_{R_{w-1}} \left(\exp \left\{ \sum_{k=1}^A \log(R_{w-1,r,k}^{-1} \tilde{Y}_{w-1,r,k}) \theta_{a,k} + \theta_w \right\} \right) \\ &= \exp \{ \tilde{\eta}_{w,r,a} \} \mathbb{E}_{R_{w-1}} \left(\exp \left\{ \sum_{k=1}^A \log(R_{w-1,r,k}^{-1}) \theta_{a,k} + \theta_w \right\} \right) \\ &= \exp \{ \tilde{\eta}_{w,r,a} + \tilde{\theta}_w \}, \end{aligned} \quad (3.5)$$

where

$$\tilde{\eta}_{w,r,a} = \sum_{k=1}^A \log(\tilde{Y}_{w-1,r,k}) \theta_{a,k}$$

and

$$\tilde{\theta}_w = \theta_w + \log \left(\mathbb{E}_{R_{w-1}} \left(\exp \left\{ \sum_{k=1}^A \log(R_{w-1,r,k}^{-1}) \theta_{a,k} \right\} \right) \right).$$

Hence, combining (3.4) and (3.5) shows that if we fit the model (3.2) to the observed data, which are affected by unreported cases, we obtain the same autoregressive coefficients $\theta_{a,k}$ for $k = 1, \dots, A$ as for the model trained with the true (unknown) infection numbers. All effects related to undetected cases accumulate in the intercept, which is of no particular interest in this context. In summary, if we assume that the CDR does not depend on the number of infections but might be different between age groups and different weeks, we obtain valid estimates for the autoregressive coefficients even if (multiplicative) under-reporting is present. While the independence assumptions made are generally

questionable, it is reasonable to assume these for a short time interval. Note that a similar argument holds for an additive CDR under epidemiological models proposed by Meyer and Held (2017) and Held et al. (2005).

3.3 Infection dynamics for school children

We can now investigate the infection dynamics between different age groups to answer the question brought up at the beginning of Section 3.1. Since the age groups provided by the RKI are too coarse for this purpose, we rely on the data provided by the LGL for Bavaria. For this dataset, we have the age for each recorded case, which, in turn, enables us to define customized age groups. To be specific, we define the age groups of the younger population in line with the proposal of the WHO and UNICEF (2020): 0–4, 5–11, 12–20, 21–39, 40–65, +65. For this analysis, we estimate model (3.1) with data on infections which were registered between 1 and 27 March 2021. The data was downloaded in May 2021; hence reporting delays should have no relevant impact on the analysis. We employ model (3.1) separately for all five analysed age groups to assess how all age groups affect each other. The fitted autoregressive coefficients $\theta_{a,k}$ are visualized in Figure 2 including their 95% confidence intervals. The partition of the x-axis refers to index a , while index k , the influence of the other age groups, is indicated by the different colours and drawn from left (5–11) to right (65+). For instance, the label ‘Model 5–11’ shows all interpretable effects where the target variable is the incidence of people aged between 5 and 11. Note that the only interpretative results of our model concern the effects between the age groups. Thus we omit the weekly intercept estimates from (3.2) in Figure 2, which lose all interpretative power in the context of under-reporting as argued in Section 3.2.

In general, we observe that the autoregressive effects for the own age group, that is, $a = k$ (drawn as triangles in Figure 2) are among the essential predictors in all age-group-specific models. Regarding the effects between age groups, the association of 5–11-year-olds (yellow, most left coefficient) with all other age groups is relatively small and, in most cases, not significant. In contrast, the age groups of working people aged between 21–39 (blue, middle) and 40–65 years (green, second right) have the highest relative effect on the incidences for all age groups (except for the autoregressive coefficients). For instance, we see that the effects of the children and adolescents (5–11 and 12–20 years) on the incidences of 21–39 and 40–65-year-olds, albeit sometimes being significantly different from 0, affect the prediction far less than the incidences of the working population. In this respect, the results confirm previous analyses concluding that increasing incidences in children and adolescents are weakly associated with the incidences of other age groups. Vice versa, we find empirical evidence that people between 21 and 65 are the main drivers of infection dynamics.

The results do not come without limitations. First of all, note that the data is observational, not experimental. Hence, we can only draw associative and not causal conclusions from the data without additional assumptions. Moreover, we rely on the given assumptions on the under-reporting. Still, rerunning the analyses for other weeks, shown in the Supplementary Material, yielded similar results, supporting the robustness of our approach and findings. Further, by the beginning of March 2021 around 2.2 million people predominantly from the 65+ age group were already fully vaccinated against COVID-19, which may have an effect on the estimates.

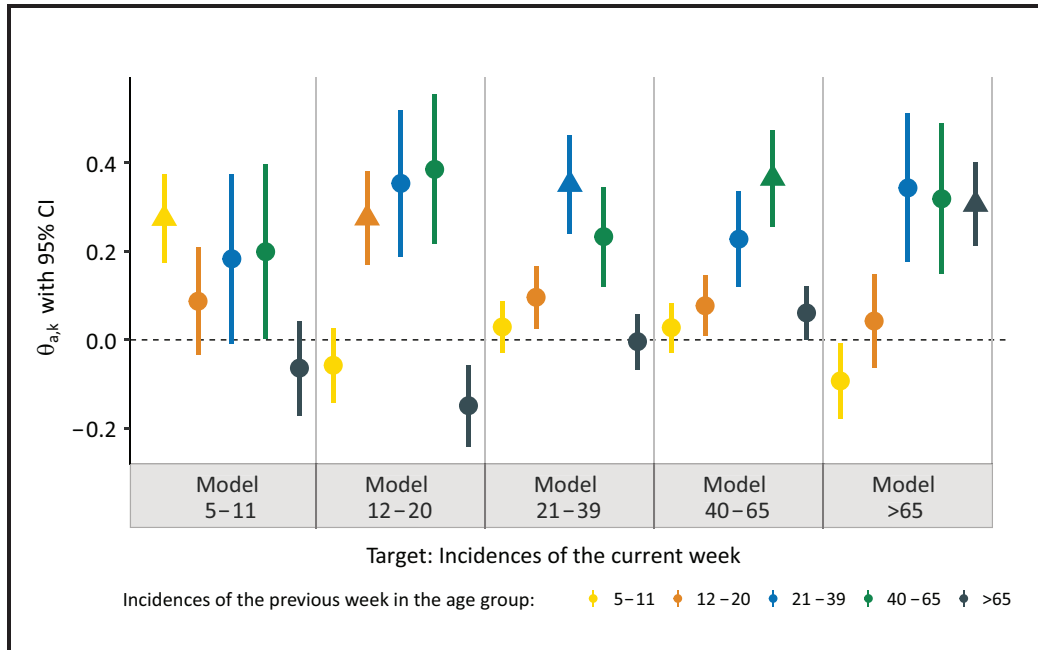


Figure 2 Association of previous week's incidences in different age groups (colour-coded) with the current-week incidences for calendar weeks 9–12 in 2021 stratified by age group (5 age groups correspond to 5 distinct Models)

4 Modelling hospitalizations accounting for reporting delay

A relevant number of COVID-19 infections lead to hospitalizations, and the incidence of patients hospitalized in relation to COVID-19 is of paramount importance to policymakers for several reasons. First, hospitalized cases are most likely to result in very severe illnesses and deaths, the minimization of which is generally the primary aim of healthcare management efforts. In addition, knowing the number of hospitalized patients is crucial to adequately assess the current state of the healthcare system. Finally, while the number of detected infections depends considerably on testing strategy and capacity, the number of hospitalizations provides a more precise picture of the current situation. For these reasons, hospitalization incidence has been deemed increasingly more relevant by scientists and decisionmakers over the course of the pandemic, and finally became the central indicator for pandemic management in Germany from September 2021, complementing the incidence of reported infections.

The central problem in calculating the hospitalization incidence with current data is that hospitalizations are often reported with a delay. Such late registrations occur along reporting chains (from local authorities to central registers), but also due to data validity checking at different levels. Visual proof of the degree of this phenomenon is given in Figure 3, which depicts the empirical distribution function of the time (in days) between the date on which a patient is admitted into a Bavarian hospital and the date on which the hospitalization is included in the central Bavarian register.

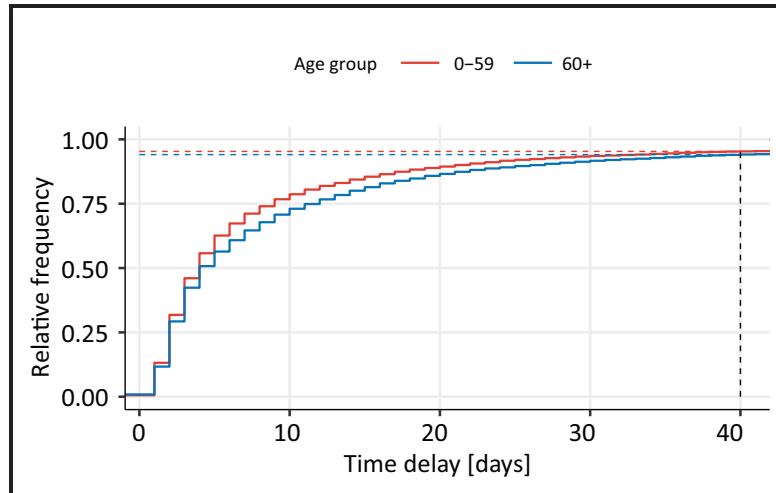


Figure 3 Cumulative distribution function of the time delay (in days) between hospitalization and its reporting, calculated with data from 1 January to 18 November of 2021, shown separately for the age groups 0–59 and 60+. The curves for both age groups are truncated at a delay of 40 days, when approximately 94.6% of all hospitalizations have been reported

In 2021, only 12.3% of hospitalized cases in Bavaria are known the day after admission, and about two thirds of them (67.2%) are reported within seven days. Moreover, the duration tends to be slightly shorter for patients younger than 60 than older patients.

Modelling and interpreting current data with only partially observed hospitalization incidences can lead to biased estimates and misleading conclusions, especially if one is interested in the temporal dynamics. To correct for such reporting delays, we utilize ‘nowcasting’ techniques, loosely defined as ‘[t]he problem of predicting the present, the very near future, and the very recent past’ (p. 193, Bańbura et al., 2012). Related methods have been extensively treated in the statistical literature (see, e.g., Höhle and An Der Heiden, 2014; Lawless, 1994) and successfully applied to infections and fatalities data during the current health crisis (De Nicola et al., 2022; Günther et al., 2020; Schneble et al., 2021b). In contrast to these approaches, we here focus on modelling the hospitalization incidences, correcting for delayed reporting through a nowcasting procedure based on the work of Schneble et al. (2021b).

We denote by $R_{t,r,g}$ the hospitalization incidence on day t for district r and age/gender group g , while the absolute count of hospitalizations in the same cohort is defined by $H_{t,r,g}$. Naturally, those two quantities related to one another through

$$R_{t,r,g} = \frac{H_{t,r,g}}{x_{\text{pop},r,g}}. \quad (4.1)$$

To account for the delayed registration of hospitalizations in $H_{t,r,g}$ when modelling $R_{t,r,g}$, we pursue a two-step approach, consisting of a nowcasting and a modelling step. In the former step, we nowcast the hospitalizations that are expected but not yet reported, while in the latter step we model $R_{t,r,g}$

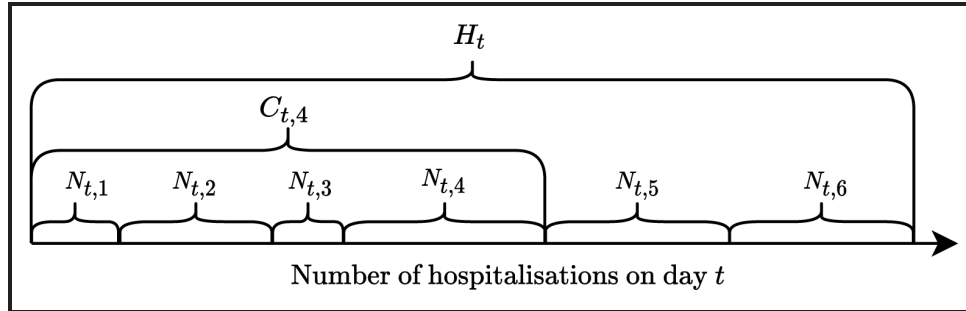


Figure 4 Illustration of the data setting for $d_{\max} = 6$. $N_{t,d}$ indicates hospitalizations reported with a specific delay d , while $C_{t,d}$ denotes all those reported with delay up to d . H_t denotes the final number of hospitalized cases regardless of the delay with which they were reported, that is with a delay up to the maximum possible, d_{\max}

as a function of several covariates, which will allow us to gain insights into the geographic and sociodemographic drivers of the pandemic. We describe the two steps below.

4.1 Nowcasting model

In this first step, we estimate the final number of hospitalized patients on day t , denoted by H_t , factoring in the expected reporting delay. Note that, while we do have data available at the district level, at this stage we aggregate hospitalizations across Bavaria due to the sparsity of the data. If we are performing the analysis on day T , we can compute the cumulative hospitalization counts $C_{t,d} = \sum_{l=1}^d N_{t,l}$, where $N_{t,d}$ is the number of hospitalizations on day t reported with delay d , for every $t \in \{1, \dots, T\}$ and $d \in \{1, \dots, T-t\}$. Assuming a maximal reporting delay of d_{\max} days, we denote the complete distribution of delayed registrations of cases with hospitalization on day t by $N_t = (N_{t,1}, \dots, N_{t,d_{\max}}) \in \mathbb{N}^{d_{\max}}$ with $\sum_{d=1}^{d_{\max}} N_{t,d} = H_t$. We graphically demonstrate how $N_{t,d}$, $C_{t,d}$, and H_t relate to one another in Figure 4. By design, N_t follows a multinomial distribution:

$$N_t \sim \text{Multinomial}(H_t, \pi_t), \quad (4.2)$$

where $\pi_t = (\mathbb{P}(D_t = 1; t), \dots, \mathbb{P}(D_t = d_{\max}; t))$ are the proportions of hospitalizations on day t with a specific delay, and D_t is a random variable describing the reporting delay of a single hospitalization which occurred at time t . For this application, we do not directly model those probabilities but instead opt for a variant of the sequential multinomial model proposed by Tutz (1991). In particular, we define the conditional probabilities through

$$p_t(d|x_t) := \mathbb{P}(D_t = d | D_t \leq d; x_t), \quad (4.3)$$

conditional on covariates x_t . It follows that the cumulative distribution function of D can be written as:

$$\begin{aligned}
 F_t(d|x_t) &= \mathbb{P}(D_t \leq d; x_{t,a}) \\
 &= \mathbb{P}(D_t \leq d | D_t \leq d+1; x_t) \mathbb{P}(D_t \leq d+1; x_t) \\
 &= \prod_{k=d}^{d_{\max}-1} \mathbb{P}(D_t \leq k | D_t \leq k+1; x_t) \\
 &= \prod_{k=d}^{d_{\max}-1} (1 - \mathbb{P}(D_t = k+1 | D_t \leq k+1; x_t)) \\
 &= \prod_{k=d+1}^{d_{\max}} (1 - \mathbb{P}(D_t = k | D_t \leq k; x_t)) \\
 &= \prod_{k=d+1}^{d_{\max}} (1 - p_t(k|x_t)). \tag{4.4}
 \end{aligned}$$

Combining (4.2) and (4.3) allows us to model the delay distribution with incomplete data. We do this separately for two age groups, which we denote by an additional index a . This leads to the model

$$N_{t,a,d} \sim \text{Binomial}(C_{t,d}, p_{t,a}(d|x_{t,a,d})) \tag{4.5}$$

with the structural assumption

$$\log \left(\frac{p_{t,a}(d|x_{t,a,d})}{1 - p_{t,a}(d|x_{t,a,d})} \right) = \theta_0 + s_1(t) + s_2(d) + s_3(d) \cdot \mathbb{I}(60+) + x_{t,d}^\top \theta,$$

where θ_0 is the intercept, $s_1(t) = \theta_1 t + \sum_{l=1}^L \alpha_l \cdot (t - 28l)_+$ is the piece-wise linear time effect, $s_2(d)$ the smooth duration effect, $s_3(d)$ a varying smooth duration effect for the age group 60+, and $x_{t,d}$ are additional covariates depending on t and the delay d , that is, a weekday effect for t and $t+d$.

From Figure 4, one can also derive that the proportion of $H_{t,a}$ included in $C_{t,a,d}$ can be comprehended as the probability that a hospitalization on day t in age group a has a reporting delay smaller than or equal to d , that is, $F_{t,a}(d|x_{t,a})$. Assuming independence of $H_{t,a}$ from $D_{t,a}$ then yields:

$$\mathbb{E}(H_{t,a}) F_{t,a}(d|x_{t,a}) = \mathbb{E}(C_{t,a,d}), \tag{4.6}$$

meaning that the expected number of patients from age group a hospitalized on day t can finally be obtained as

$$\mathbb{E}(H_{t,a}) = \frac{\mathbb{E}(C_{t,a,d})}{F_{t,a}(d|x_{t,a})}. \tag{4.7}$$

This equation holds for any delay $d \leq T - t$ which is already observed at the date of analysis. Thus, it is possible to express the expected numbers of hospitalized patients through the ratio between the number of already reported patients up to delay d and the cumulative distribution function F .

In summary, we can fit the logistic regression model given by (4.5) with the available data on hospitalizations. Based on this model, we exploit (4.7) to obtain an estimate for the expected number of hospitalizations from age group a on day t . Uncertainty intervals for the estimated nowcasts can then be obtained, for example, through a parametric bootstrapping approach relying on the asymptotic multivariate normal distribution of the estimated model coefficients.

4.2 Hospitalization model

In the second step, we propose a model for the expected value of $R_{t,r,g}$, the hospitalization incidence on day t in district r and age/gender group g , conditional on covariates $x_{t,r,g}$. To be specific we set

$$\begin{aligned} \mathbb{E}(R_{t,r,g}|x_{t,r,g}) &= \exp\{\theta_0 + \theta_{\text{age}}x_{\text{age},g} + \theta_{\text{gender}}x_{\text{gender},g} + \theta_{\text{gender:age}}x_{\text{age},g}x_{\text{gender},g} + \\ &\quad \theta_{\text{weekday}}x_{\text{weekday},t} + s_1(t) + s_2(x_{\text{Lon},r}, x_{\text{Lat},r}) + u_r\} \\ &= \exp\{\eta_{t,r,g}\}, \end{aligned} \quad (4.8)$$

where the linear predictor $\eta_{t,r,g}$ includes, in addition to the intercept θ_0 , effects for the age/gender groups through the main and interaction effects θ_{age} , θ_{gender} and $\theta_{\text{gender:age}}$. Additionally, we include dummy effects θ_{weekday} for each day of the week to account for potentially different hospitalization rates over the course of the week. Furthermore, the hospitalization incidences are allowed to vary over time through the smooth term $s_1(t)$. Finally to account for spatial heterogeneity, we add a smooth spatial effect of each district's average longitude and latitude $s_2(r)$ and a Gaussian random effect to capture random deviations from this smooth effect, that is, $u_r \sim N(0, \tau^2)$ with $\tau^2 \in \mathbb{R}^+$.

Note that, on any given day $t > T - d_{\text{max}}$, we do not yet observe the final hospitalization counts $H_{t,r,g}$, but only the ones already reported at this time, that is $C_{t,r,g,T-t}$, indicating the cumulative observations on day t in district r reported with a delay of up to $d = T - t$ days for age/gender group g . The age/gender group indexed by g extends the coarse (binary) age categorization a used in Section 4.1, which only differentiates between cases younger and older than 60 years. Exploiting (4.7) and the definition (4.1) of the incidence leads to the final model

$$\mathbb{E}(R_{t,r,g}|x_{t,r,g}) = \frac{\mathbb{E}(C_{t,r,g,T-t}|x_{t,r,g})}{x_{\text{pop},r,g}F_{t,g}(T-t|x_{t,g})}, \quad (4.9)$$

where we set $C_{t,r,g,T-t} = H_{t,r,g}$ if $T - t \geq d_{\text{max}}$. Rearranging (4.9) shows that modelling the count variable $C_{t,r,g,T-t}$ with the offset term $\log(x_{\text{pop},r,g}F_{t,g}(T-t|x_{t,g}))$ is equivalent to modelling $R_{t,r,g}$ as in (4.8), since

$$\mathbb{E}(C_{t,r,g,T-t}|x_{t,r,g}) = \exp\{\eta_{t,r,g} + \log(x_{\text{pop},r,g}F_{t,g}(T-t|x_{t,g}))\} = \mu_{t,r,g} \quad (4.10)$$

holds. In practice we thereby replace the unknown quantities in the offset with their estimates derived in the previous section. In other words, the delayed reporting is accommodated through an offset in

the model using only the reported data $C_{t,r,g,T-t}$. We can then complete the model by making use of a negative binomial model to account for possible overdispersion:

$$C_{t,r,g,T-t} | x_{t,r,g} \sim \text{NB}(\mu_{t,r,g}, \sigma^2),$$

with $\mu_{t,r,g}$ parametrized as in (4.10) and (4.8), and the dispersion parameter σ^2 is estimated from the data.

As an additional note, we point out that accounting for late registrations works analogously for any model within the endemic–epidemic framework originating in Held et al. (2005). The only difference to the approach presented here is that the exact functional form of the expected value must be adequately accounted for. For instance, if $\mu_{t,r,g}$ consists of the sum of non-negative endemic and epidemic terms, one should incorporate the offset in both terms.

4.3 Application to the fourth COVID-19 wave in Bavaria

For the application, we focus on the first two months of the fourth wave of the pandemic in Bavaria, which began towards the end of September 2021. In particular, we consider hospitalizations between 24 September and 18 November, using data reported as of 18 November 2021. We set $d_{\max} = 40$ days to be the maximum possible duration between hospitalization and its reporting in the central Bavarian register. We derive this choice from the empirical delay distribution in Figure 3, proving that since the beginning of 2021, around 94% of the hospitalizations have been reported within 40 days of their occurrence. We have no information on the date of hospital admission for about 9.6% of all hospitalizations related to COVID infections that were reported between 24 September and 19 November. For those cases, we replace the date of hospitalization with the respective COVID-19 infection date as reported by the local health authorities. For brevity, we only present a comparison of the nowcasted and raw hospitalization counts for the nowcasting model and the age/gender group-specific and spatial effects of the hospitalization model. We refer to the Supplementary Material for additional results.

Figure 5 maps the raw and corrected rolling weekly sums of hospitalization counts accompanied by the 95% confidence intervals for the whole population as well as separately for the two age groups under consideration. While reported numbers indicate a relatively stable or even slightly decreasing development over the last two weeks of observed data, the nowcast reveals a continuous upward trend since the beginning of October. Comparing both age-stratified populations, the increase for those over 60 years (the more vulnerable) is steeper. The figure also plots the realized hospitalization counts observed after 40 days have passed since 19 November 2021. The comparison of our nowcast with those realized figures observed *a posteriori* shows that our model tends to slightly overestimate the reported cases for the younger population. This might be due to the beginning of the Delta curve with rapidly increasing hospitalizations since October 2021 after a phase with rather low hospitalization numbers. Nevertheless, our nowcast estimates show a clear improvement in terms of reflecting the true dynamics of hospitalized cases compared to the curve of the reported values. These results emphasize the need to adjust reported hospitalization counts, as they tend to systematically underestimate the number of recently occurred hospitalizations, which can lead to inaccurate conclusions about the current state of the pandemic.

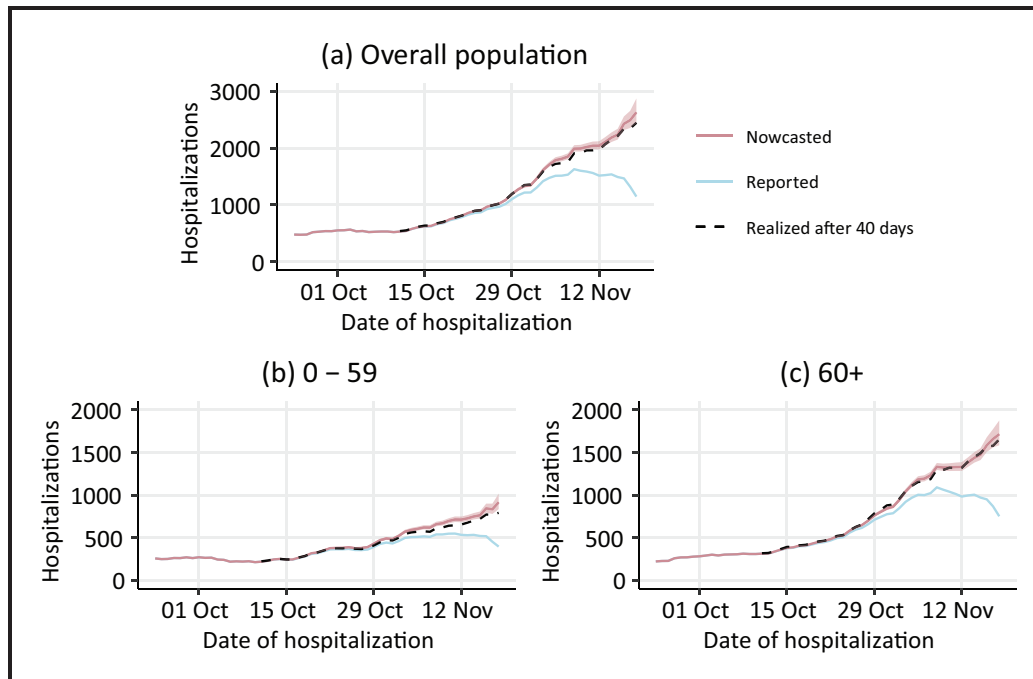


Figure 5 Comparison of nowcasted (red) and reported (blue) rolling weekly sums of hospitalization counts between 24 September and 18 November 2021, based on data reported as of 19 November 2021. Note: 95% confidence intervals of the nowcast estimates are indicated by the shaded areas. The dashed black lines show the realized weekly sums of hospitalization after 40 days, that is, the maximum delay assumed in our nowcasting model. Results are displayed for the overall population (a) as well as separately for age groups 0–59 (b) and 60+ (c)

Turning to the results of the hospitalization model proposed in Section 4.2, the estimated coefficients for all age and gender combinations can be seen in Figure 6. Those estimates reveal considerably lower hospitalization rates for people younger than 35 than all other age groups. We generally observe a positive correlation between age and risk of hospitalization for both genders, that is, older people are more likely to be hospitalized. The only exception to this intuitive finding is seen for men over 80 years, whose expected hospitalization rates are slightly lower than men aged 60 to 79. Statistically significant differences between men and women are visible across all age groups. While women in the youngest and oldest age group tend to have a (slightly) higher hospitalization rate than men, the opposite holds for the other groups.

Figure 7 depicts the random and smooth spatial effects (on the log-scale). The smooth effect in Figure 7 (a) paints a clear spatial pattern, with generally higher hospitalization rates in the eastern parts of Bavaria and lower rates in the north-western districts. This structure reflects the pandemic situation in Bavaria during autumn 2021, where we observed the most severe dynamics in those eastern districts. Districts with unexpectedly high or low hospitalization rates (when compared to their neighbouring areas) can be located on the map of the district-specific random intercepts in

12. Statistical modelling of COVID-19 data: Putting generalized additive models to work

Statistical modelling of COVID-19 data 15

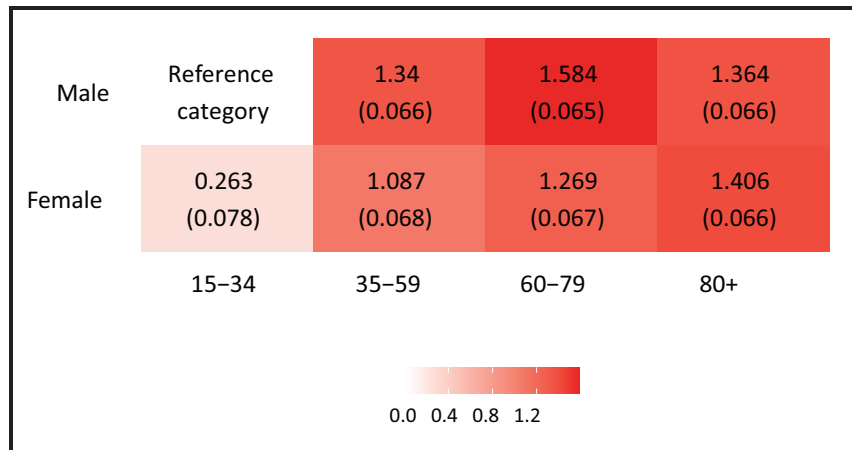


Figure 6 Estimated linear effects for different age and gender groups in the hospitalization model, where males aged 15–34 are the reference category. Note: Estimated standard deviations are written in brackets

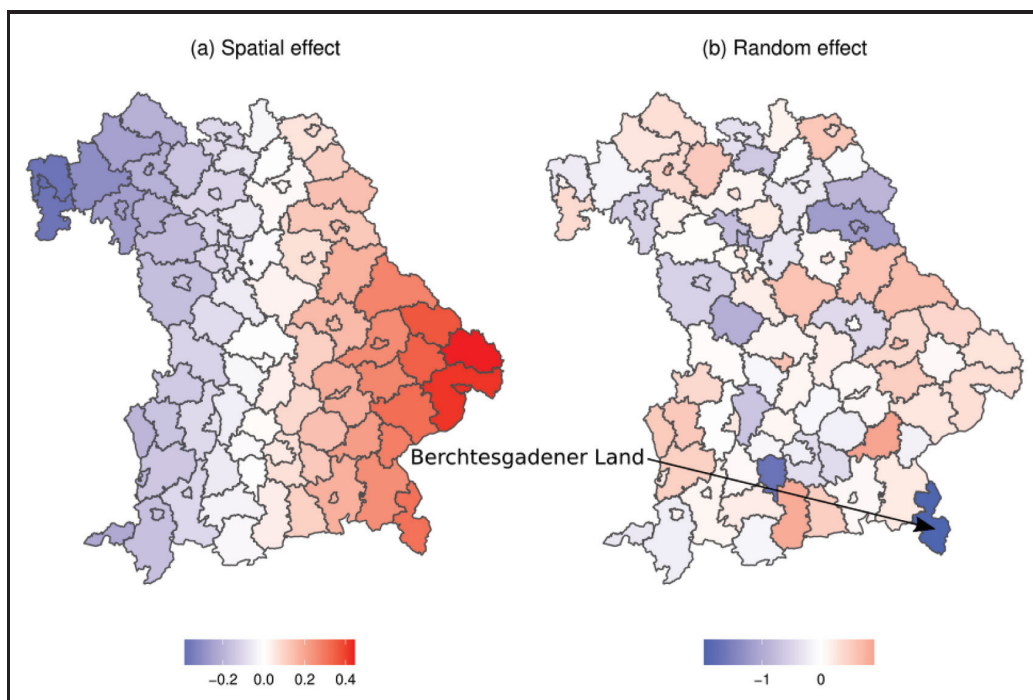


Figure 7 Estimated smooth spatial effect (a) and district-specific random effect (b) in the hospitalization model

Statistical Modelling xxxx; xx(x): 1–24

Figure 7 (b). Contrary to its role as a hotspot during the second wave in autumn 2020, the district with the lowest random effect is Berchtesgadener Land. We estimate an overall variance of $\tau^2 = 0.274$ for the district-specific random effects.

5 Modelling ICU occupancy

The primary aims of healthcare management efforts during a pandemic include minimizing very severe and fatal cases, as well as preventing the overload and collapse of the healthcare system. Information on these very severe cases, among other quantities of interest, can be captured by the ICU occupancy, which is the focus of our third application case.

5.1 Multinomial model

We consider the occupancy of ICUs where, as described in Section 2, beds are categorized into the number of vacant beds ($Z_{w,r,1}$), number of beds occupied by patients not infected with COVID-19 ($Z_{w,r,2}$), and number of beds occupied by patients infected with COVID-19 ($Z_{w,r,3}$). Further, we denote by $Z_{w,r} = (Z_{w,r,1}, Z_{w,r,2}, Z_{w,r,3})$ the vector of length three expressing the average number of ICU-bed occupancy in week w and district r . The canonical GAM for this type of data is a multinomial model; hence the distributional assumption is:

$$Z_{w,r} \sim \text{Multinomial}(N_{w,r}, \pi_{w,r}), \quad (5.1)$$

where $N_{w,r} = \sum_{j=1}^3 Z_{w,r,j}$ is the known number of available beds in district r and week w and $\pi_{w,r} = (\pi_{w,r,1}, \pi_{w,r,2}, \pi_{w,r,3})$ defines the proportion of occupied beds in the respective categories.

One advantage of this multinomial approach is that we implicitly account for displacement effects commonly observed for ICU occupancy data. Over time, as the number of beds occupied by patients infected with COVID-19 rise, both free beds and beds occupied by patients not infected with COVID-19 decrease almost simultaneously. In particular, the ‘displacement’ may be caused by practices such as rescheduling non-urgent operations or other treatments which would have required an ICU stay, which were already common during the first wave of COVID-19 (Stöß et al., 2020). These effects lead to negative correlations between the entries in $Z_{w,r}$, which is naturally accounted for in model (5.1) as the covariance between arbitrary counts $Z_{w,r,k}$ and $Z_{w,r,l}$ is $-N_{w,r} \pi_{w,r,k} \pi_{w,r,l} \forall k, l \in \{1, 2, 3\}, k \neq l$.

Taking the number of beds occupied by patients infected with COVID-19 as the reference category, we effectively parametrize pairwise comparisons via

$$\log \left(\frac{\pi_{w,r,j}}{\pi_{w,r,3}} \right) = \eta_{w,r,j} \forall j = 1, 2, \quad (5.2)$$

where the linear predictors $\eta_{w,r,j}$ are functions of covariates labeled as $x_{w,r}$ and defined by:

$$\eta_{w,r,j} = \theta_{0,j} + \theta_{AR(1),j}^\top (\tilde{Z}_{w-1,r,1}, \tilde{Z}_{w-1,r,2})^\top + \theta_{I,j}^\top \log(Y_{w-1,r} + \delta) + s_j(x_{Lon,r}, x_{Lat,r}) + u_{r,j} \quad \forall j = 1, 2, \quad (5.3)$$

where $\theta_{0,j}$ is the intercept term. Further, we incorporate an autoregressive component in (5.3) by including the relative ICU occupancy observed in the previous week as a regressor. We denote the distribution of the different occupancies of the previous week as $\tilde{Z}_{w-1,d} = (Z_{w-1,r,1}, Z_{w-1,r,2}) / (\sum_{j=1}^3 Z_{w-1,r,j})$, and the respective effect is denoted by $\theta_{AR(1),j}$ for the j th linear predictor. We also let (5.3) depend on the previous week's district and age-specific infections per 100.000 inhabitants (incidences) denoted by $Y_{w-1,r,a}$, that are weighted by the coefficient $\theta_{I,j} \quad \forall j = 1, 2$. To control for district-specific heterogeneity, we include Gaussian random effects, that is, $u_{r,j} \sim N(0, \tau^2) \quad \forall r \in \{1, \dots, R\} \quad \forall j = 1, 2$. For smooth spatial deviations from these random effects, we add a bivariate function $s_j(\cdot, \cdot) \quad \forall j = 1, 2$ parametrized by thin-plate splines that take the longitude and latitude of each district as arguments (see Wood, 2003, for more details). For notational brevity, let θ denote the joint parameter vector of (5.3) $\forall j = 1, 2$.

5.2 Quantification of uncertainty

As stated, the multinomial model has the beneficial property of automatically accounting for displacement effects. Note, however, that patients' expected length of stay in intensive care may exceed our time unit of one week, as the average stay of COVID-19 patients is about 13 days (see Vekaria et al., 2021). This means that not all beds are completely redistributed at every time point of observation. However, apart from including the previous week's occupancy in the covariates, our proposed model does not adequately account for this stochastic variability.

We therefore pursue a Bayesian view and let $N_{w,r}$ be the number of ICU beds in district r in week w . This number is known, and we assume that each week only a fixed but unknown proportion α of beds in the three categories become disposable, where $0 < \alpha < 1$. That is to say that $\alpha N_{w,r}$ beds are redistributed among the three categories, where integrity is assumed but not explicitly included in the notation for simplicity. We assume that this new allocation is independent of the previous status of the beds and denote the newly allocated beds with the three-dimensional vector $A_{w,r} = (A_{w,r,1}, A_{w,r,2}, A_{w,r,3})$. This setting translates to:

$$Z_{w,r} = (1 - \alpha)Z_{w-1,r} + A_{w,r}.$$

For the newly allocated beds we still assume a multinomial model:

$$A_{w,r} \sim \text{Multinomial}(\alpha N_{w,r}, \pi_{w,r}), \quad (5.4)$$

with $\pi_{w,r}$ specified in (5.3). Note, however, that we do not know α and that no information is provided in the data concerning the length of stay or the number of beds changing their status. To account for that data deficiency, we impose a Dirichlet distribution on the vector $\pi_{w,r}$, where the

prior information is determined by the available beds, that is,

$$f_{\pi}(\pi_{w,r}) \propto \prod_{j=1}^3 \pi_{w,r,j}^{(1-\alpha)Z_{w-1,r,j}}. \quad (5.5)$$

Combining the prior (5.5) with the likelihood from (5.4), leads to the posterior

$$f_{\pi}(\pi_{w,r} | A_{w,d}) \propto \prod_{j=1}^3 \pi_{w,r,j}^{A_{w,r,j} + (1-\alpha)Z_{w-1,r,j}} = \prod_{j=1}^3 \pi_{w,r,j}^{Z_{w,r,j}} \quad (5.6)$$

This, in turn, equals the likelihood resulting from the multinomial model and justifies the use of model (5.2) even though not all beds are allocated weekly. Nevertheless, the central assumption of independent observations in standard uncertainty quantification in GAMs (Wood, 2006) is violated. To correct for this bias, we substitute the canonical covariance of the estimators with the robust sandwich estimator based on M-estimators defined by:

$$\mathbf{V}(\theta) = \mathbf{A}(\theta)^{-1} \mathbf{B}(\theta) \mathbf{A}(\theta)^{-1}, \quad (5.7)$$

where we set $\mathbf{A}(\theta) = \mathbb{E} \left(-\frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta} \ell(\theta) \right)$, $\mathbf{B}(\theta) = \text{Var} \left(\frac{\partial}{\partial \theta} \ell(\theta) \right)$, and $\ell(\theta)$ is the logarithmic likelihood resulting from (5.1) or equivalently the logarithm of the posterior of (5.3). See also Stefanski and Boos (2002) and Zeileis (2006).

5.3 Application to the third wave

We now employ the multinomial logistic regression (5.1) to ICU data recorded during the third wave between March and June 2021. For the incidence data used in the covariates, we employ the RKI data; hence we set $A = 4$ and the age groups are: 15–34, 35–59, 60–79 and 80+. Further, we normalize all non-binary covariates:

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\sqrt{\frac{1}{n} \sum_j^n (x_j - \bar{x})^2}} \quad \text{with} \quad \bar{x} = \frac{\sum_j^n x_j}{n}. \quad (5.8)$$

This way, we facilitate the interpretation of associations and guarantee a meaningful comparison between the covariates. Due to space restrictions, we here only present the linear effects from (5.3) and refer to the Supplementary Material for the random and smooth estimates.

In Figure 8, we visualize the estimated coefficients, including their confidence intervals. The reference category in both pairwise comparisons is COVID-beds; thus, we refer to the two models as free vs COVID beds and non-COVID vs COVID beds. In particular, the coefficients relate to the association between the covariates and the logarithmic odds of a bed not being occupied compared to being occupied by a patient with COVID-19, shown with blue dots in Figure 8. Analogously, the orange triangles in Figure 8 illustrate the estimated association between the covariates and the logarithmic odds of a bed being occupied by a patient not infected with COVID-19 in comparison to a

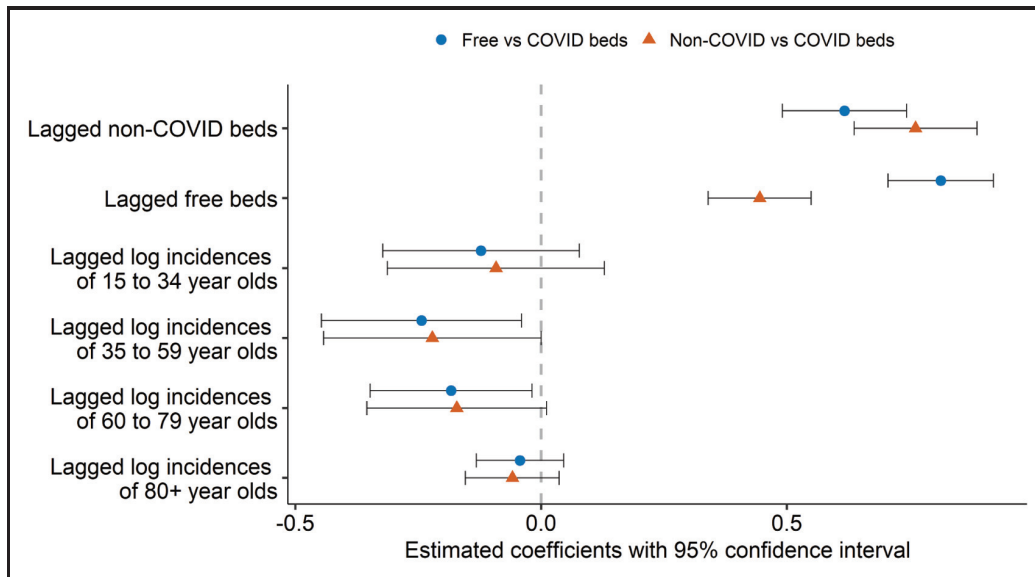


Figure 8 Estimated coefficients with confidence interval of the associations between normalized linear covariates included in the multinomial model and the logarithmic odds of a bed being free vs occupied by a patient infected with COVID-19 (blue dots) and the logarithmic odds of a bed being occupied by a patient not infected with COVID-19 vs a patient infected with COVID-19 (orange triangles)

bed being occupied by a patient infected with COVID-19. To demonstrate the uncertainty of each estimate, a 95% confidence interval is added. Keeping the other variables constant, the normalized lagged log-incidences of all age groups generally have a negative effect on the logarithmic odds of both pairwise comparisons. This translates to the finding that an increase in the incidences leads to a decrease in the proportion of non-COVID and free-beds in when compared to COVID beds. The lagged normalized proportion of free and non-COVID beds is estimated to have a stronger, positive association with the logarithmic odds of both pairwise comparisons. We, therefore, expect a higher number of non-COVID beds in the previous week to be followed by a higher number of non-COVID beds in the next week.

The model can be extended to a forecasting model, as shown in the supplementary material. In particular, we demonstrate how forecasting performance changes over the different waves of the pandemic. In principle, we could also incorporate further covariates like district-specific proportions of vaccinated people. Unfortunately, these numbers are not very reliable and require sophisticated cleaning, so we prefer not to present results in this direction here.

6 Discussion

The COVID-19 pandemic poses numerous complex challenges to scientists from different disciplines. Statisticians and epidemiologists, in particular, face the problem of extracting meaningful in-

formation from imperfect, incomplete and rapidly changing data. Generalized additive models are a powerful tool that, if used correctly, can help solving some of these challenges. In this work, we have addressed three such challenges where the utilization of GAMs provided meaningful insight.

1. We investigated whether children are the main drivers of the pandemic under a time-varying case-detection ratio.
2. We modelled hospitalization incidences controlling for delayed registrations, thereby providing both up to dates estimates of current hospitalization numbers as well as insight on the demographic and spatio-temporal drivers of COVID-19.
3. We developed an interpretable predictive tool for ICU bed occupancy that is actively used by the Bavarian government.

We achieved all of those results by using GAMs with different methodological extensions. Nevertheless, the use of our proposed models to extract novel information from the data provided is still subject to both data-related and methodological limitations. In general, our data sources are subject to exogenous shocks (e.g., policy changes) that lead to sudden changes in population behaviour and pose a danger to the validity of our results. Regarding the study of infection dynamics of school kids, revised testing policies hinder the long-range comparability of our findings. In the hospitalization data, the exact date of hospitalization is missing for about 10% of the hospitalized cases, which we impute by the given registration date of the infection. Furthermore, the records on the ICU-bed occupancy do not include intrinsic constraints, as the capacity of beds available to COVID-19 patients does not equate to the capacity of beds available to patients not infected with COVID-19. There are also methodological limitations. First of all, note that the data is observational, not experimental. Additionally, the set of covariates in our model can easily be extended to control for other factors, such as meteorological and socioeconomic ones.

We close this work by emphasizing that the nowcasting model can also be used as a stand-alone model. In the German COVID-19 Nowcast Hub (KIT), the described model is used among other nowcasting methods, including the work of Günther et al. (2020) and van de Kasstele et al. (2019), to estimate hospitalization counts on the national and federal state level in Germany. Apart from a systematic evaluation of the different approaches, one of the main goals of this project is to combine individual nowcasts to an ensemble nowcast, which may lead to more accurate estimates.

Supplementary materials

Supplementary materials for this article are available online, including additional information on the three application cases. The replication code is available in the following repository: <https://github.com/corneliusfritz/Statistical-modelling-of-COVID-19-data>.

Acknowledgements

We would like to thank Manfred Wildner and Katharina Katz on behalf of the staff of the IfSG Reporting Office of the Bavarian Health and Food Safety Authority (LGL) for cooperatively providing the data used for Sections 3 and 4 and for fruitful discussions on the analysis of the COVID-19 pandemic. We would also like to thank all COVID-19 Data Analysis Group (CODAG)

members at LMU Munich for countless beneficial conversations and Constanze Schmaling for proofreading. Moreover, we would like to thank the two anonymous reviewers whose valuable and constructive comments were highly appreciated and led to an improvement of the manuscript.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The work has been partially supported by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. We also acknowledge support of the Deutsche Forschungsgemeinschaft (KA 1188/13-1) and the Bavarian Health and Food Safety Authority (LGL).

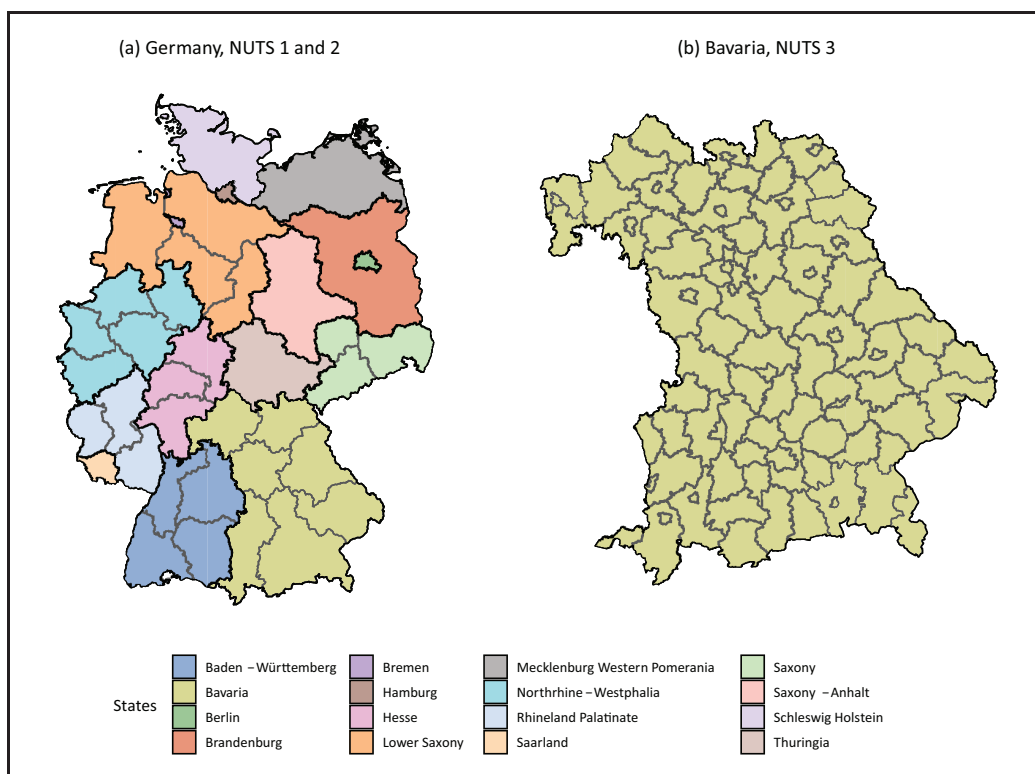


Figure A.1 (a): Map of Germany, where the NUTS 1 regions are indicated by the black borders and the different colours. The NUTS 2 regions, on the other hand, are drawn in grey. Note that all NUTS 1 region borders are also NUTS 2 region borders. (b): Map of Bavaria where also the NUTS 3 regions are marked. In the legend, we state the names of each NUTS 1 region

Appendix: A Spatial unit

We carried out most modelling endeavours presented in this article on the NUTS 3 level, which is shown on the right side of Figure A.1. The only exception is the Nowcasting model from Section 4.1, where we aggregate all data onto the NUTS 1 level in Bavaria. Moreover, NUTS 1 regions, depicted on the left side of Figure A.1, are the federal states in Germany and Bavaria is one of them. In Section 3 and 4, we are only analysing data from Bavaria, while we employ data from complete Germany in Section 5.

References

- Andrew A, Cattan S, Costa Dias M, Farquharson C, Kraftman L, Krutikova S, Phimister A and Sevilla A (2020) Inequalities in children's experiences of home learning during the COVID-19 lockdown in England. *Fiscal Studies*, **41**, 653–83.
- Bañbura M, Giannone D and Reichlin L (2012) Nowcasting. In *The Oxford Handbook of Economic Forecasting*, edited by MP Clements and DF Hendry, pages 193–224. Oxford University Press.
- Basellini U and Camarda GC (2021) Explaining regional differences in mortality during the first wave of COVID-19 in Italy. *Population Studies*, **76**, 99–118.
- De Nicola G, Schneble M, Kauermann G and Berger U (2022) Regional now-and forecasting for data reported with delay: toward surveillance of COVID-19 infections. *AStA Advances in Statistical Analysis*, **106**, 407–26.
- DIVI (2021) Daily ICU occupancy data for COVID-19 and non-COVID-19 patients. <https://www.divi.de/register/tagesreport>. (Accessed on June 17, 2022).
- Flasche S and Edmunds WJ (2021) The role of schools and school-aged children in SARS-CoV-2 transmission. *The Lancet Infectious Diseases*, **21**, 298–9.
- Fokianos K and Kedem B (2004) Partial likelihood inference for time series following generalized linear models. *Journal of Time Series Analysis*, **25**, 173–97.
- Fritz C and Kauermann G (2022) On the interplay of regional mobility, social connectivity, and the spread of COVID-19 in Germany. *Journal of the Royal Statistical Society, Series A*, **185**, 400–24.
- Gaythorpe KA, Bhatia S, Mangal T, Unwin HJT, Imai N, Cuomo-Dannenburg G, Walters CE, Jauneikaite E, Bayley H, Kont MD, Mousa A, Whittles LK, Riley S and Ferguson NM (2021) Children's role in the COVID-19 pandemic: A systematic review of early surveillance data on susceptibility, severity, and transmissibility. *Scientific Reports*, **11**.
- Goswami K, Bharali S and Hazarika J (2020) Projections for COVID-19 pandemic in india and effect of temperature and humidity. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, **14**, 801–5.
- Günther F, Bender A, Katz K, Küchenhoff H and Höhle M (2020) Nowcasting the COVID-19 pandemic in Bavaria. *Biometrical Journal*, **63**, 490–502.
- Hastie T and Tibshirani R (1987) Generalized additive models: Some applications. *Journal of the American Statistical Association*, **82**, 371–386.
- Held L, Höhle M and Hofmann M (2005) A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, **5**, 187–99.
- Hippich M, Sift P, Zapardiel-Gonzalo J, Böhmer MM, Lampasona V, Bonifacio E and Ziegler AG (2021) A public health antibody screening indicates a marked increase of SARS-CoV-2 exposure rate in children during the second wave. *Med*, **2**, 571–2.
- Hoch M, Vogel S, Kolberg L, Dick E, Fingerle V, Eberle U, Ackermann N, Sing A, Huebner J,

- Rack-Hoch A, Schober T and von Both U (2021) Weekly SARS-CoV-2 sentinel surveillance in primary schools, kindergartens, and nurseries, Germany, June–November 2020. *Emerging Infectious Diseases*, **27**, 2192–6.
- Höhle M and An Der Heiden M (2014) Bayesian nowcasting during the STEC O104: H4 outbreak in Germany, 2011. *Biometrics*, **70**, 993–1002.
- Im Kampe EO, Lehfeld AS, Buda S, Buchholz U and Haas W (2020) Surveillance of COVID-19 school outbreaks, Germany, March to August 2020. *Eurosurveillance*, **25**.
- Kimeldorf GS and Wahba G (1970) A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, **41**, 495–502.
- KIT. Nowcasts of the hospitalization incidence in Germany (COVID-19). <https://covid19nowcasthub.de/index.html>. (Accessed: June 17, 2022).
- Lawless J (1994) Adjustments for reporting delays and the prediction of occurred but not reported events. *Canadian Journal of Statistics*, **22**, 15–31.
- Li R, Pei S, Chen B, Song Y, Zhang T, Yang W and Shaman J (2020) Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science*, **368**, 489–93.
- Luijten MA, van Muilekom MM, Teela L, Polderman TJ, Terwee CB, Zijlmans J, Klaufus L, Popma A, Oostrom KJ, van Oers HA and Haverman L (2021) The impact of lockdown during the COVID-19 pandemic on mental and social health of children and adolescents. *Quality of Life Research*, **30**, 2795–804.
- Ma Y, Zhao Y, Liu J, He X, Wang B, Fu S, Yan J, Niu J, Zhou J and Luo B (2020) Effects of temperature variation and humidity on the death of COVID-19 in wuhan, china. *Science of The Total Environment*, **724**.
- McKeigue PM, Weir A, Bishop J, McGurnaghan SJ, Kennedy S, McAllister D, Robertson C, Wood R, Lone N, Murray J, Caparrotta TM, Smith-Palmer A, Goldberg D, McMenamin J, Ramsay C, Hutchinson S and Colhoun HM (2020) Rapid epidemiological analysis of comorbidities and treatments as risk factors for COVID-19 in Scotland (REACT-SCOT): A population-based case-control study. *PLOS Medicine*, **17**, 1–17.
- Meyer S and Held L (2017) Incorporating social contact data in spatio-temporal models for infectious disease spread. *Biostatistics*, **18**, 338–51.
- Nelder JA and Wedderburn RWM (1972) Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, 370.
- Panovska-Griffiths J (2020) Can mathematical modelling solve the current COVID-19 crisis? *BMC Public Health*, **20**, 551.
- Pearce N, Vandenbroucke JP, VanderWeele TJ and Greenland S (2020) Accurate statistics on covid-19 are essential for policy guidance and decisions. *American Journal of Public Health*, **110**, 949–51.
- Perra N (2021) Non-pharmaceutical interventions during the COVID-19 pandemic: A review. *Physics Reports*, **913**, 1–52.
- Prata DN, Rodrigues W and Bermejo PH (2020) Temperature significantly changes COVID-19 transmission in (sub)tropical cities of brazil. *Science of The Total Environment*, **729**.
- Robert Koch Institute (2021). Daily COVID-19 cases data. <https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74>. (Accessed: June 17, 2022).
- Schneble M, De Nicola G, Kauermann G and Berger U (2021a) A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020. *Biometrical Journal*, **63**, 1623–32.
- Schneble M, De Nicola G, Kauermann G and Berger U (2021b) Nowcasting fatal COVID-19 infections on a regional level in Germany. *Biometrical Journal*, **63**, 471–89.
- Stefanski LA and Boos DD (2002) The calculus of M-estimation. *American Statistician*, **56**, 29–38.
- Stöß C, Steffani M, Kohlhaw K, Rudroff C, Staib L, Hartmann D, Friess H and Müller MW (2020) The COVID-19 pandemic: Impact on surgical departments of non-university hospitals. *BMC Surgery*, **20**, 1–9.

- Tutz G (1991) Sequential models in categorical regression. *Computational Statistics and Data Analysis*, **11**, 275–95.
- van de Kasstele J, Eilers PH and Wallinga J (2019) Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained p-spline smoothing. *Epidemiology*, **30**, 737–45.
- Vekaria B, Overton C, Wiśniowski A, Ahmad S, Aparicio-Castro A, Curran-Sebastian J, Edleston J, Hanley NA, House T, Kim J, Olsen W, Pampaka M, Pellis L, Ruiz DP, Schofield J, Shryane N and Elliot MJ (2021) Hospital length of stay for COVID-19 patients: Data-driven methods for forward planning. *BMC Infectious Diseases*, **21**.
- Ward MP, Xiao S and Zhang Z (2020) The role of climate during the COVID-19 epidemic in new south wales, australia. *Transboundary and Emerging Diseases*, **67**, 2313–17.
- WHO and UNICEF (2020). Advice on the use of masks for children in the community in the context of COVID-19: Annex to the advice on the use of masks in the context of COVID-19, 21 August 2020. Technical report. URL <https://apps.who.int/iris/handle/10665/333919>. (Accessed: June 17, 2022).
- Wood SN (2003) Thin plate regression splines. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **65**, 95–114.
- Wood SN (2006) On confidence intervals for generalized additive models based on penalized regression splines. *Australian and New Zealand Journal of Statistics*, **48**, 445–64.
- Wood SN (2017) *Generalized additive models: An introduction with R*. Boca Raton: CRC press.
- Wood SN (2020) Inference and computation with generalized additive models and their extensions. *Test*, **29**, 307–39.
- Wood SN (2021) Inferring UK COVID-19 fatal infection trajectories from daily mortality data: Were infections already in decline before the uk lockdowns? *Biometrics*.
- Worby CJ, Chaves SS, Wallinga J, Lipsitch M, Finelli L and Goldstein E (2015) On the relative role of different age groups in influenza epidemics. *Epidemics*, **13**, 10–6.
- Xie J and Zhu Y (2020) Association between ambient temperature and COVID-19 infection in 122 cities from china. *Science of The Total Environment*, **724**, 138201.
- Zeileis A (2006) Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, **16**, 1–16.

Part IV.

Estimating excess mortality during crises

13. On assessing excess mortality in Germany during the COVID-19 pandemic

Contributing article

De Nicola, G., Kauermann, G., and Höhle, M. (2022). On assessing excess mortality in Germany during the COVID-19 pandemic. *ASTA Wirtschafts-und Sozialstatistisches Archiv*, 16:5–20. <https://doi.org/10.1007/s11943-021-00297-w>.

Copyright information

This article is distributed under a Creative Commons 4.0 International license ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).

Author contributions

The idea of writing a paper addressing the issue of age-adjustment for calculating excess mortality can be attributed to Göran Kauermann and Giacomo De Nicola. The two joined forces with Michael Höhle, who was independently working on the same problem from a different angle. Göran Kauermann designed and implemented the model for calculating excess mortality at the yearly level. Giacomo De Nicola was responsible for the related data management, analysis and visualization. Further, Giacomo De Nicola contributed by writing major parts of Sections 1, 2 and 4. Michael Höhle was mainly responsible for writing Section 3, and for carrying out the related data analysis and visualization. All authors contributed through fruitful comments and extensive proofreading of the manuscript.



On assessing excess mortality in Germany during the COVID-19 pandemic

Giacomo De Nicola · Göran Kauermann · Michael Höhle

Received: 25 June 2021 / Accepted: 22 November 2021 / Published online: 10 January 2022
© The Author(s) 2022

Abstract Coronavirus disease 2019 (COVID-19) is associated with a very high number of casualties in the general population. Assessing the exact magnitude of this number is a non-trivial problem, as relying only on officially reported COVID-19 associated fatalities runs the risk of incurring in several kinds of biases. One of the ways to approach the issue is to compare overall mortality during the pandemic with expected mortality computed using the observed mortality figures of previous years. In this paper, we build on existing methodology and propose two ways to compute expected as well as excess mortality, namely at the weekly and at the yearly level. Particular focus is put on the role of age, which plays a central part in both COVID-19-associated and overall mortality. We illustrate our methods by making use of age-stratified mortality data from the years 2016 to 2020 in Germany to compute age group-specific excess mortality during the COVID-19 pandemic in 2020.

Keywords COVID-19 · Excess mortality · Expected mortality · Standardized mortality rate

Giacomo De Nicola (✉) · Göran Kauermann
Department of Statistics, Ludwig-Maximilians-Universität München, Munich, Germany
E-Mail: giacomo.denicola@stat.uni-muenchen.de

Michael Höhle
Department of Mathematics, University of Stockholm, Stockholm, Sweden

Zur Berechnung der Übersterblichkeit in Deutschland während der COVID-19-Pandemie

Zusammenfassung Die Corona-Pandemie (COVID-19) ist mit einer erhöhten Zahl an Todesfällen in der Bevölkerung verbunden. Die Quantifizierung der Übersterblichkeit ist ein nicht triviales Problem, denn wenn man sich nur auf die öffentlich gemeldeten COVID-19-assoziierten Todesfälle stützt, besteht die Gefahr von Verzerrungen. Eine Möglichkeit, das Problem zu umgehen, ist der Vergleich der Gesamtsterblichkeit während der Pandemie mit der erwarteten Sterblichkeit, welche aus den beobachteten Sterblichkeitszahlen der Vorjahre berechnet werden kann. In unserem Artikel bauen wir auf dieser Methodik auf und schlagen zwei Methoden zur Berechnung der erwarteten Sterblichkeit und damit der Übersterblichkeit vor, nämlich auf wöchentlicher und auf Jahresebene. Besonderes Augenmerk liegt auf dem Einfluss des Alters auf die Sterblichkeit, welches eine zentrale Rolle bei COVID-19-assoziierten Todesfällen spielt. Wir veranschaulichen unsere Methoden anhand von Sterbedaten aus den Jahren 2016 bis 2020 in Deutschland und zeigen wie altersgruppenspezifische Übersterblichkeit während der COVID-19-Pandemie im Jahr 2020 berechnet werden kann.

Schlüsselwörter COVID-19 · Übersterblichkeit · Erwartete Sterblichkeit · Standardisierte Mortalitätsrate

1 Introduction

First identified in Wuhan, China, in December 2019, the Coronavirus disease 2019 (COVID-19) caused by the SARS-CoV-2 virus developed into a worldwide pandemic during the spring of 2020 (Velavan and Meyer 2020). One of the challenges for scientists has been to evaluate its impact in terms of life loss across different countries and regions of the world. A possible way to do this is through directly looking at the number of people who died while they were confirmed to be infected. This measure, often defined as COVID-19-associated mortality, is certainly more robust than other pandemic-related quantities such as, for example, the number of reported COVID-19 cases, for which it has become clear that there is a non-negligible discrepancy between cases detected through tests and the number of individuals who were infected (Lau et al. 2021; Schneble et al. 2021). Nonetheless, the raw number of COVID-related fatalities can also be subject to interpretative issues and biases due to underreporting and misclassification. In particular, this number might be biased downwards, as COVID-19 cases can still remain unreported until and after the point of death. Moreover, it is not always straightforward to identify if COVID-19 was the primary cause of death: Some patients might have a SARS-CoV-2 infection, but the actual contribution of the virus to the death might be minimal (Vincent and Taccone 2020). To deal with these issues, comparing all-cause mortality is generally considered a more robust alternative for assessing the damage done by the pandemic, and to compare its impact between regions or countries. A first look at this matter for Germany was provided by Stang et al. (2020), who looked at data

from the first wave ranging from calendar weeks 10 to 26 in 2020. The authors came to the conclusion that a moderate excess mortality was observable for this period in Germany, in particular for the elderly. Morfeld et al. (2021) consider regional variation in mortality in Germany during the first wave (see also Morfeld et al. 2020). A calculation of the years of life lost over the course of the pandemic in Germany in 2020 was pursued by Rommel et al. (2021). International analyses on excess mortality due to COVID-19 include e.g. Krieger et al. (2020) looking at data from Massachusetts, Vandoros (2020) who focuses on England and Wales, and Michelozzi et al. (2020) investigating mortality in Italian cities. Global analyses in this direction were pursued by Karlinsky and Kobak (2021) and Aburto et al. (2021).

Monitoring excess mortality has a long tradition as part of analysing the impact of pandemics (Johnson and Mueller 2002; Simonsen et al. 2013). With the EuroMOMO project, Europe also runs an early-warning system specifically dedicated to mortality monitoring (Mazick et al. 2007). However, no unified methodological definition exists for deciding if the currently observed death counts are higher than what would be expected. A very simple approach is to compare the currently observed deaths for a selected time-period with the average of death counts for a similar period in previous years¹. Alternatively, the expected value can be computed by an underlying time-series model based on past values, e.g. including seasonality and excluding past phases of excess, as done in the EuroMOMO project (see e.g. Vestergaard et al. 2020; Nørgaard et al. 2021). These approaches, however, do not come without problems, as the age structure within a population can change significantly over time. Given that both general and COVID-related mortality are heavily dependent on age (Dowd et al. 2020; Levin et al. 2020), comparisons between different years based only on raw data will often lead to biased estimates. More specifically, using such techniques will lead to overestimating excess mortality for aging populations (such as those, e.g., in western Europe), and underestimating it for populations that get progressively younger. More sophisticated approaches thus need to adjust for different or changing age structures in the population. The latter point is of particular relevance when looking at aging populations (Kanasi et al. 2016) and the infectious risks for the elderly (Kline and Bowdish 2016). Such age-adjustments have a long tradition in demography when comparing mortality across different regions with different age-structure (Keiding and Clayton 2014; Kitagawa 1964). A general discussion on aging populations and mortality can be found in Crimmins and Zhang (2019).

In this paper, we build on existing methodology to propose two ways of calculating expected mortality taking age into account, respectively at the weekly and at the yearly level. These methods are compared to the existing benchmarks on data from Germany over the years 2016–2019, for which age-stratified information is available. We furthermore apply those methods to assess age group-specific excess mortality in Germany during the COVID-19 pandemic in 2020. The remainder of the manuscript is structured as follows. In Sect. 2 we look at yearly expected mor-

¹ <https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Sterbefaelle-Lebenserwartung/sterbefallzahlen.html>.

tality, while the weekly view is pursued in Sect. 3. Sect. 4 ends the paper with some interpretative caveats and concluding remarks.

2 Yearly Excess Mortality

We first look at yearly data, and tackle the question of whether there was excess mortality in Germany in 2020. In order to obtain an age adjustment for mortality data, we calculate expected deaths based on official life tables. Life tables give the probability q_x of a person who has completed x years of age to die before completing their next life-year, i.e. before their $x + 1^{\text{th}}$ birthday. In our analysis we consider the life table provided for the year 2017/2019 by the Federal Statistical Office of Germany (Destatis 2020). The calculation of a life table, as simple as it sounds, is not straightforward, and is an age-old actuarial problem. First references date far back, to Price (1771) and Dale (1772). A historical digest of the topic is provided by Keiding (1987). Over the last decades, the calculation of the German life-tables made use of different methods proposed in Becker (1874), Raths (1909) and Farr (1859). We will come back to this point and demonstrate that further adjustments are recommendable to relate the expected number of deaths to recently observed ones. In particular, with increasing life expectancy, the average age of the German population has been steadily increasing (see e.g. Buttler 2003), and this has an effect on the validity of life tables, as discussed in Dinkel (2002). Generally, an aging population leads to increasingly high yearly death tolls (see e.g. Klenk et al. 2007). To quantify excess mortality one therefore needs to account for age effects, leading to the computation of standardized quantities such as the standardized mortality ratio (SMR, see e.g. Rothman et al. 2008). The SMR is defined as the ratio of observed death counts over expected deaths, and thus allows for an age adjusted view, meaning that instead of pure death counts one takes the (dynamic) age structure into account.

Calculating excess mortality on a yearly basis requires to calculate expected fatalities using life tables provided by the relevant statistical bureau. We make use of data provided by the Federal Statistical Office of Germany (Destatis 2020). A straightforward way of obtaining the expected number of deaths for age group A in year y is to calculate

$$e_{A,y} = \sum_{x \in A} q_x P_{x,y} \quad (1)$$

where $P_{x,y}$ is the population size of individuals aged x years at the beginning of year y and q_x are the age-specific death probabilities, e.g. those found in the most recent German life table from the years 2017/19, calculated following Raths (1909). More specifically, let D_x be the cumulated number of individuals that died at x years old, i.e. before their $x + 1$ -th birthday in the considered years 2017 to 2019. Let $P_{x,y}$ denote the population size of x year old individuals on December 31st in year

$y \in \{2016, 2017, 2018, 2019\}$. q_x provided in the German life-tables is then defined as

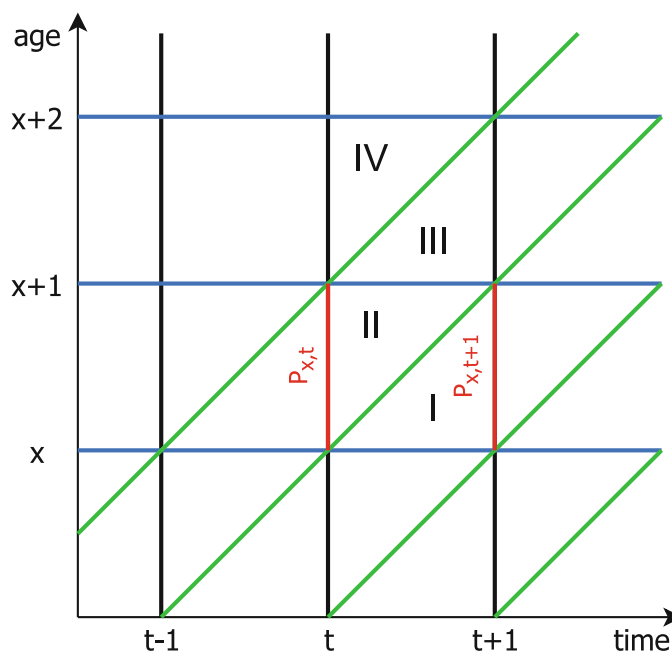
$$q_x = \frac{D_x}{\sum_{y=2016}^{2018} \frac{P_{x,y} + P_{x,y+1}}{2} + \frac{D_x}{2}} \tag{2}$$

We label (1) in combination with (2) as Method 1 below. We now show that this quantity is biased for estimating the expected number of deaths of x year old people in year y . To motivate this we look at the Lexis diagram in Fig. 1, and for simplicity we replace the calculation in (2) by looking at a single year only, i.e from $y = t$ to $y = t + 1$. This leads to $D_x = I + II$, where I and II refer to the observed deaths in the two triangles in Fig. 1. Note that following the calculation principle (2) of the Statistisches Bundesamt we would obtain q_x as

$$q_x = \frac{D_x}{\frac{P_{x,t} + P_{x,t+1}}{2} + \frac{D_x}{2}} \tag{3}$$

where $P_{x,t}$ and $P_{x,t+1}$ are the population sizes of x year olds indicated in Fig. 1. That is q_x is the probability of dying in triangles I and II . Let us define with \tilde{q}_x the probability of an individual aged x years at the beginning of year t (i.e. on December 31st in year $t - 1$) to die before year $t + 1$ starts. In other words \tilde{q}_x is the probability of dying in triangles II and III . In fact, this is the probability we are interested in. It is easy to see that $\tilde{q}_x \neq q_x$. Assuming that the probability of dying in triangle I is roughly equal to the probability of dying in triangle II , and

Fig. 1 Lexis Diagram indicating the different quantities to be estimated



assuming the same relationship for triangles *III* and *IV* holds, we can conclude the approximate equivalence

$$\tilde{q}_x = \frac{1}{2}q_x + \frac{1}{2}q_{x+1} \quad (4)$$

which leads to the expected number of deaths

$$\tilde{e}_{A,y} = \sum_{x \in A} \tilde{q}_x P_{x,y}. \quad (5)$$

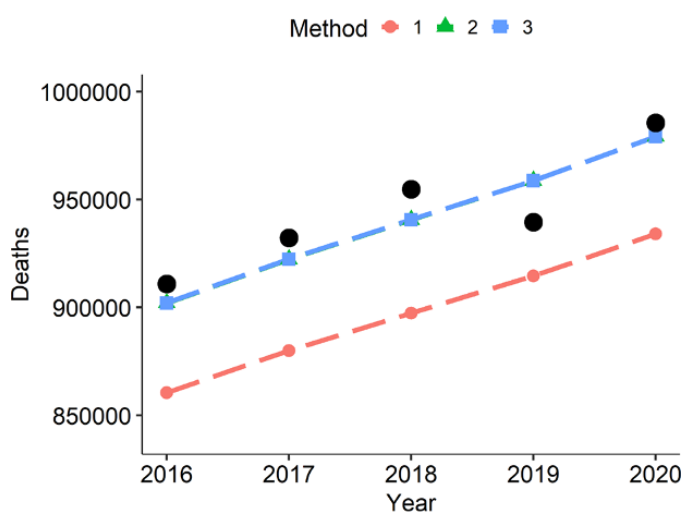
We label (5) as Method 2 below. The adjustment is still not complete, and in fact it can be shown that (5) is still biased for delimited age groups (see Hartz et al. 1983). This is because individuals dying in triangle *III* count as $x + 1$ years old, so that part of the deaths contributes to an age group that is different from the target. We may now assume for simplicity that the probability of dying in triangles *II* and *III* is roughly the same, which leads to the following calculation. Let $A = [a_l, a_r]$,

$$\hat{e}_{A,y} = 0.5 \cdot \tilde{q}_{a_l-1} P_{a_l-1,y} + \sum_{x=a_l}^{a_r-1} \tilde{q}_x p_{x,y} + 0.5 \cdot \tilde{q}_{a_r} P_{a_r,y} \quad (6)$$

where A is the age group, a_l and a_r refer to the left and right age boundaries of the group, $\tilde{q}_{-1} = \tilde{q}_0$, and $P_{-1,y} = P_{0,y}$ gives the approximation for the youngest age group. Accordingly, for $a_r = \max(x)$ we take the full fraction of the last year, that is we add an additional $0.5 \cdot \tilde{q}_{a_r} p_{a_r,y}$ to the formula above. We label (6) as Method 3 below.

Based on this method we can now compare expected and observed fatalities over the last years using the same 2017/2019 life table as basis. Note that, when looking at different years, one may more accurately also consider different life tables to account for changing life expectancy. We omit this point for simplicity since we only look at five years, and changes in life expectancy over this short period were moderate (Wenau et al. 2019). This is equivalent to implicitly assuming constant age-specific hazards over the last five years (while we still, of course, account for the changing age structure). Fig. 2 gives a first overview of the results for all age groups combined. In the figures, alongside Method 3, we also show the results obtained with Methods 1 and 2. This is to demonstrate how impactful their previously underlined biases, which may seem small on paper, can be in practice. We plot the observed death counts as black dots, and we represent the expected death counts based on the different methods as dashed lines. We can see that Method 1, which uses (1), clearly underestimates the expected death counts. Method 2 and Method 3 perform equally well, which is not surprising, since we here do not take an age-specific view. The latter is carried out in Fig. 3 for all different age groups available from the data. This age-specific view shows how Methods 2 and 3 differ, and that overall Method 3 shows the better fit. We can quantify the empirical discrepancy between the three methods by calculating the mean absolute percentage error for the different age groups, where we explicitly exclude year 2020 due to the COVID-19 pandemic. The results of this can be found in Table 1.

Fig. 2 Expected deaths computed by calendar year with the three different methods described, for all age groups combined. Realized fatalities are shown as black dots. Note that Methods 2 and 3 are visually indistinguishable, as here all age groups are pooled together



Having seen that Method 3 empirically outperforms the other two over recent years, we can use the expected number of fatalities computed with this method for 2020 to quantify excess mortality during the first calendar year of the COVID-19 pandemic in Germany. Table 2 contains expected and observed mortality figures for all age groups in 2020, as well as the absolute and percentage differences between the two. From the table we can see that, for the entire population, the age-adjusted excess mortality was in the order of 1% in 2020. We stress that these results in terms of COVID-19 impact need to be interpreted with utmost care: We here focus on the methodological aspects, and defer the subject-matter discussion of the results to Sect. 4. Also note that, while this section focuses the attention on the difference between observed and expected mortality, one could also easily obtain the yearly SMRs by simply taking the ratio of those two quantities. We believe the (percentage) differences to be more interesting when looking at the data at the yearly level. Nonetheless, the real insight lies in estimating the expected number of deaths in a given period; once that is calculated, one can use any preferred method to quantify the excess.

In this Section we approached the problem of excess mortality from a yearly standpoint. A natural follow up would be to zoom into a monthly or weekly view. A way to move in this direction would be to divide the expected yearly mortality by the total number of weeks in a year, and computing a weekly “SMR” using weekly observed deaths. The main issue with this type of approach is that it does not allow to take within-year seasonality into account for the expected deaths. In the following section we therefore follow a different approach based on standardization, which can account for seasonality and is more model-free.

3 Weekly Excess Mortality

To tackle the question of weekly excess mortality, classical standardization approaches such as direct and indirect standardization can be used to adjust the ob-

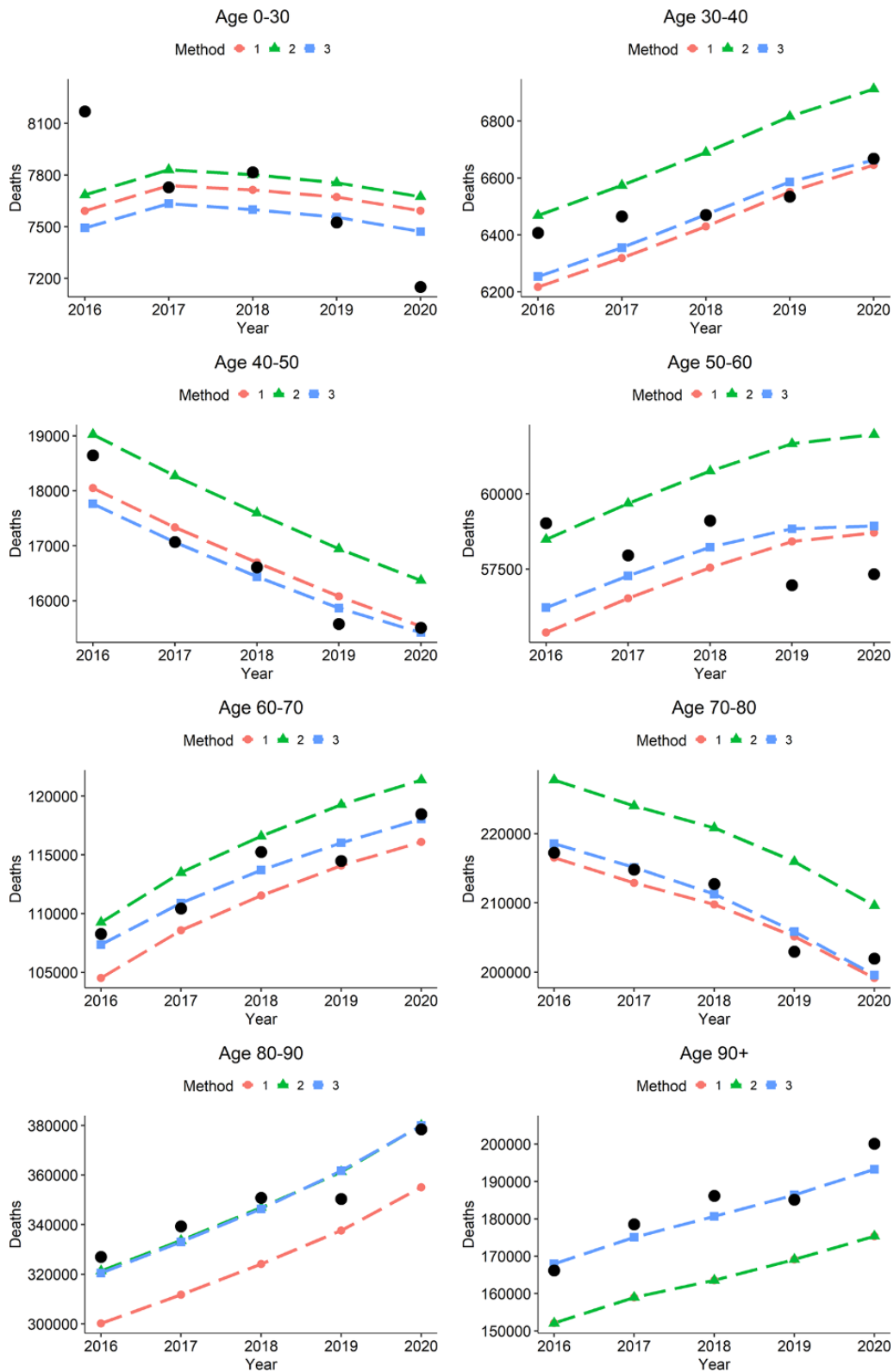


Fig. 3 Expected deaths by calendar year and age group computed with the three different methods described. Realized fatalities are shown as black dots

13. On assessing excess mortality in Germany during the COVID-19 pandemic

On assessing excess mortality in Germany during the COVID-19 pandemic

13

Table 1 Age-specific mean absolute percentage error for expected yearly fatalities calculated with different methods over the years 2016 to 2019. Year 2020 is excluded due to the COVID-19 pandemic. The smallest value for each age group is highlighted in bold

	0–30	30–40	40–50	50–60	60–70	70–80	80–90	90+	Overall
Method 1	2.74%	1.56%	2.12%	3.57%	2.24%	0.93%	7.44%	11.22%	5.23%
Method 2	2.69%	2.50%	5.56%	3.54%	2.20%	4.60%	1.90%	11.22%	1.39%
Method 3	3.37%	1.25%	1.96%	2.72%	0.99%	0.71%	2.08%	1.68%	1.39%

Table 2 Expected and observed yearly mortality in 2020 for each of the six age groups, computed with Method 3

Age group	Expected 2020	Observed 2020	Absolute diff.	Relative diff.
[00,30)	7471	7150	−321	−4%
[30,40)	6663	6668	5	+0%
[40,50)	15420	15507	87	+1%
[50,60)	58929	57331	−1598	−3%
[60,70)	118047	118460	413	+0%
[70,80)	199569	201957	2388	+1%
[80,90)	379917	378406	−1511	−0%
[90, ∞)	193238	200093	6855	+4%
Total	979255	985572	6317	+1%

served values for age effects, see e.g. Kitagawa (1964). We will focus on indirect standardization, but given an appropriate choice of reference population, direct standardization approaches are straightforward adaptations.

Let $q_{t,x}$ be the mortality probability specific to age x and time period t . In what follows, the considered time period will be one International Organization for Standardization (ISO) week, but other intervals (e.g. months) are also imaginable. We estimate $q_{t,x}$ by dividing the number of observed deaths at age x during time period t , defined as $D_{t,x}$, by the corresponding population at the beginning of the time period, i.e. $P_{t,x}$. To be specific, we define

$$\hat{q}_{t,x} = \frac{D_{t,x}}{P_{t,x}}. \quad (7)$$

Since the age-stratified population is only available as a point estimate for December 31st of each year, we use linear interpolation to estimate $P_{t,x}$. The corresponding estimates of weekly mortality probabilities (7) are shown in Fig. 4. We see that in the age groups ≥ 50 years old a substantial weekly excess mortality is observable from week 45 on, with more pronounced excess mortality for the elderly. Also note that the official 2020 population data are available since June 2021 and, hence, were used for the present retrospective analysis. However, when performing analyses in real time, recent population data might not (yet) be available, and projections would therefore be needed (see e.g. Ragnitz 2021; Höhle 2021).

A weekly SMR-based excess mortality measure for the entire year 2020 can now be computed as follows. Let t denote a specific ISO week in 2020, i.e. this will

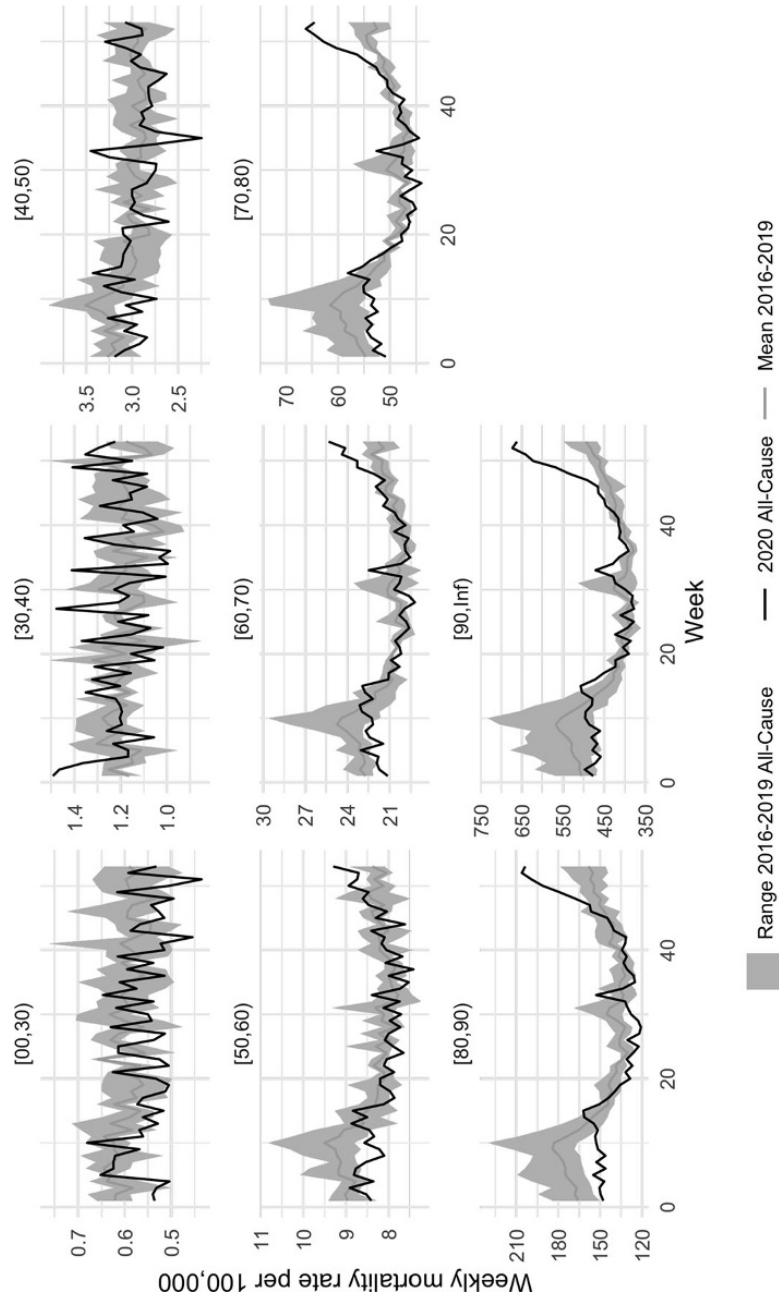


Fig. 4 Weekly mortality rate per 100,000 for the different age groups as well as the range (min-max) of the corresponding mortality rates of the past four years and their mean

serve as notational shorthand for ISO week 2020- W_t , where $t = 1, \dots, 53$. We form the expected age-time mortality probability for this week by computing the average of the mortality of the same week over the last 4 years, i.e.

$$\bar{q}_{t,x} = \frac{1}{4} \sum_{y=2016}^{2019} \hat{q}_{y-W_t,x}, \quad t = 1, \dots, 53.$$

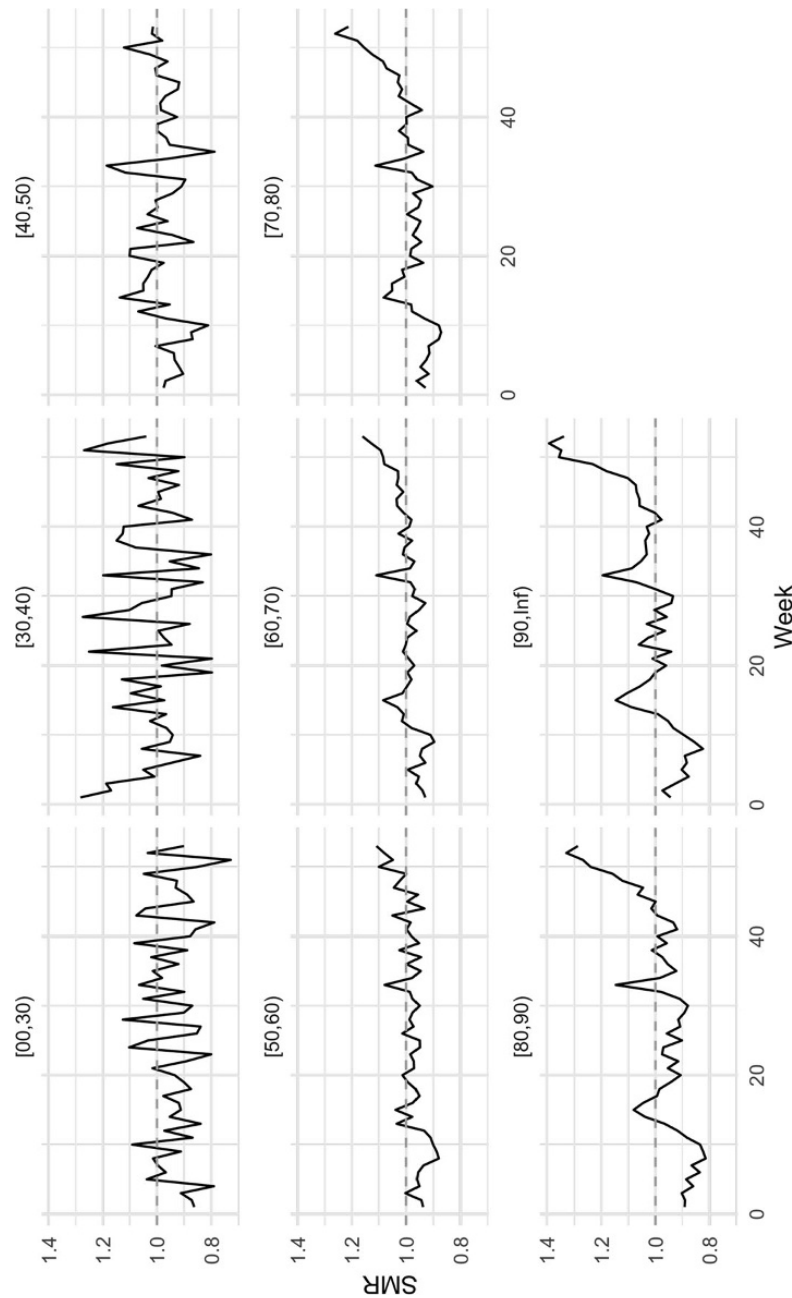


Fig. 5 Weekly SMR estimates for the different age groups

Because the years 2016–2019 do not have an ISO week 53, we define y -W53 for $y = 2016, \dots, 2019$ as $\frac{1}{2}(q_{y\text{-W52}} + q_{(y+1)\text{-W01}})$. The indirect standardization now computes the expected number of deaths for week t as

$$\bar{e}_{t,x} = \bar{q}_{t,x} \cdot P_{t,x}$$

This corresponds to the expected number of deaths in week t at age x , if the current population would have been subject to the average death probability calculated over the past four years. Since fatalities are not given with exact ages but rather by age

group, we indicate this by using $q_{t,A}$, $P_{t,A}$ and $e_{t,A}$, where A denotes the age classes. Fig. 4 shows $\widehat{q}_{t,A}$ as well as $\bar{q}_{t,x}$ for Germany for the available age classes. Also note that this computation is equivalent to computing, for each reference year y , the expected number of deaths for the relevant week in 2020, and then taking the average of the expected deaths. In other words: by applying the mortality probabilities for the same week of the reference year y to our study population (i.e. 2020- Wt) and then averaging the four expected fatalities, we get:

$$\bar{e}_{t,x} = \frac{1}{4} \sum_{y=2016}^{2019} q_{y-Wt,x} \cdot P_{t,x}.$$

One can now define the absolute excess mortality in week t and age-group A as $D_{t,A} - e_{t,A}$. Instead of focusing on absolute differences, it is better in terms of interpretation to look at relative estimates of excess mortality given by the standardized mortality ratio (SMR)

$$\text{SMR}_{t,A} = \frac{D_{t,A}}{\bar{e}_{t,A}}. \quad (8)$$

We plot the corresponding weekly estimate resulting from (8) for all age groups in Fig. 5. As already seen in the incidence plots, we note that in the older age groups the first approx. 10 weeks of the year had a rather low SMR, followed by a small increase consistent with the first COVID-19 wave. Furthermore, substantial increases are observed in the ≥ 50 years old age groups starting from week 45, coinciding with the 2nd wave, and reaching up to 40% more deaths than expected in certain weeks. Note that it would also be possible to aggregate the weekly numbers to generate yearly excess-mortality statements similar to those in Table 2. All in all, the results of the two methods at the yearly resolution are similar. We don't include the results of this aggregation here, but refer to Höhle (2021) for comparison.

4 Discussion

The COVID-19 pandemic posed numerous challenges to scientists. One of those challenges lies in estimating the number of fatalities brought upon by the pandemic. To tackle this issue, we pursued an approach based on comparing observed all-cause mortality in 2020 with the number of fatalities that would have been expected in the same year without the advent of COVID-19. Building on existing methodology, we proposed two simple ways of computing expected mortality, respectively at the yearly and at the weekly level. We then put those methods to work to obtain estimates for excess mortality in 2020 in Germany. The two approaches yield similar results at the aggregate level, and highlight how 2020 was characterized by an overall excess mortality of approximately 1%. The light excess mortality was apparently driven by a spike in fatalities related to COVID-19 at the end of the year in the older age groups.

Interpreting COVID-19 mortality has become a politically sensitive issue, where the same underlying data are used to either enhance or downplay the consequences of COVID-19 infections. We therefore stress that our interests are methodological, and that the presented results are restricted to the calendar year 2020 for Germany as a whole. Altogether, the mild mortality in the older age groups during the first weeks (e.g. due to a mild influenza season) balanced the excess in the higher age groups which came later in the year. Clearly noticeable is the second wave during Nov-Dec 2020, which also continued in the early months of 2021. To better account for such seasonality, excess mortality computations for influenza are often pursued by season instead of calendar year, i.e. in the northern hemisphere for the period from July in Year X to June in Year $X + 1$ (Nielsen et al. 2011). Similarly, the impact of COVID-19 cases and fatalities was not only temporally, but also spatially heterogeneous, with strong peaks in Dec 2020 in the federal states of Saxony, Brandenburg and Thuringia (Höhle 2021). Hence, using mortality aggregates over periods and regions only provides a partial picture of the impact of COVID-19. Furthermore, the mortality figures observed in 2020 naturally incorporate the effects of all types of pandemic management consequences, which include changes in the behavior of the population (voluntary or due to governmental interventions). Disentangling the complex effects of all-cause mortality and the COVID-19 pandemic is a delicate matter, which takes experts in several disciplines (demographers, statisticians, epidemiologists) to solve. Timely analysis of all-cause mortality data is just one building block of this process; Nevertheless, the pandemic has shown the need to do this in near real-time based on sound data while adjusting for age structure.

Our analysis was motivated by the fact that many of the methods that have been applied to tackle this issue so far fail to take the changing age structure of the population into account. This can lead to biased results, and especially so for the rapidly aging developed countries. In the case of Germany, for example, the absolute number of people aged 80 or more increased by approximately 20% from 2016 to 2020. Such a remarkable increase will naturally have an effect on overall mortality, and as such direct comparisons in the number of casualties across different years will lead to significant overestimation of the excess mortality. Our approaches are instead robust to such changes in population structure, and can be used regardless of the demographic context. Note that, for both of our approaches, it would also be possible to obtain confidence intervals through imposing distributional assumptions. This would, however, not be straightforward, for several reasons. First of all, the residual variability is well beyond what would be explainable through a Poisson distributional assumption. To solve this, one could, in principle, replace the Poisson distribution with a Negative Binomial one, or adopt an approach based on quasi-likelihood (see McCullagh 1983) and incorporate an additional overdispersion parameter. But in addition to this, stating confidence intervals would also require an understanding of which (super-)population parameters the confidence intervals make statements about. Since the all-cause excess mortality estimates are for the entire population of interest (the German population), some kind of repeated sampling setting would have to be assumed. For those reasons we refrain from pursuing this, and leave it for future research on the subject. The same methodologies could also be used to pursue a similar analysis for any country in which mortality data and a mortality

table are available, for any given year. A natural use for the proposed methodology would also be to assess the overall damages caused by the pandemic when it will be finally considered a thing of the past. All in all, we hope the proposed methods will help shedding light on the issue of computing the expected number of fatalities, and consequently in the assessment of (potential) general excess mortality.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access 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/>.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Aburto JM, Schöley J, Zhang L, Kashnitsky I, Rahal C, Missov TI, Mills MC, Dowd JB, Kashyap R (2021) Recent gains in life expectancy reversed by the COVID-19 pandemic. medRxiv
- Becker K (1874) Zur Berechnung von Sterbetafeln an die Bevölkerungsstatistik zu stellende Anforderungen: Gutachten über die Frage: welche Unterlagen hat die Statistik zu beschaffen, um richtige Mortalitätstafeln zu gewinnen? Verlag des Königlichen Statistischen Bureaus
- Buttler G (2003) Steigende Lebenserwartung – was verspricht die Demographie? *Z Gerontol Geriat* 36:90–94
- Crimmins EM, Zhang YS (2019) Aging populations, mortality, and life expectancy. *Annu Rev Sociol* 45(1):69–89
- Dale W (1772) Calculations deduced from first principles, in the most familiar manner, by plain arithmetic, for the use of the societies instituted for the benefit of old age: intended as an introduction to the study of the doctrine of annuities. J. Ridley, London (By a member of one of the societies)
- Destatis (2020) Sterbetafel 2017/2019. Tech. rep., Statistisches Bundesamt. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Sterbefaelle-Lebenserwartung/Tabellen/_tabellen-innen-lebenserwartung-sterbetafel.html. Accessed: 21 October 2021
- Dinkel RH (2002) Die langfristige Entwicklung der Sterblichkeit in Deutschland. *Z Gerontol Geriat* 35:400–405
- Dowd JB, Andriano L, Brazel DM, Rotondi V, Block P, Ding X, Liu Y, Mills MC (2020) Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proc Natl Acad Sci* 117(18):9696–9698
- Farr W (1859) On the construction of life-tables, illustrated by a new life-table of the healthy districts of England. *Philos Trans R Soc Lond* 149:837–878
- Hartz AJ, Giefer EE, Hoffmann RG (1983) A comparison of two methods for calculating expected mortality. *Statist Med* 2(3):381–386
- Höhle M (2021) Age-structure adjusted all-cause mortality. <https://staff.math.su.se/hoehle/blog/2021/03/01/mortadj.html>. Accessed: 21 October 2021
- Johnson NP, Mueller J (2002) Updating the accounts: global mortality of the 1918–1920 “Spanish” influenza pandemic. *Bull Hist Med* 76(1):105–115
- Kanasi E, Ayilavarapu S, Jones J (2016) The aging population: demographics and the biology of aging. *Periodontol* 2000 72(1):13–18
- Karlinsky A, Kobak D (2021) The World Mortality Dataset: tracking excess mortality across countries during the COVID-19 pandemic. medRxiv

- Keiding N (1987) The method of expected number of deaths, 1786–1886–1986. *Int Stat Rev* 55(1):1–20
- Keiding N, Clayton D (2014) Standardization and control for confounding in observational studies: a historical perspective. *Stat Sci* 29(4):529–558
- Kitagawa EM (1964) Standardized comparisons in population research. *Demography* 1(1):296–315
- Klenk J, Rapp K, Büchele G, Keil U, Weiland SK (2007) Increasing life expectancy in Germany: quantitative contributions from changes in age- and disease-specific mortality. *Eur J Public Health* 17(6):587–592
- Kline KA, Bowdish DM (2016) Infection in an aging population. *Curr Opin Microbiol* 29:63–67
- Krieger N, Chen JT, Waterman PD (2020) Excess mortality in men and women in Massachusetts during the COVID-19 pandemic. *Lancet* 395(10240):1829
- Lau H, Khosrawipour T, Kocbach P, Ichii H, Bania J, Khosrawipour V (2021) Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters. *Pulmonology* 27(2):110–115
- Levin AT, Hanage WP, Owusu-Boaitey N, Cochran KB, Walsh SP, Meyerowitz-Katz G (2020) Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications. *Eur J Epidemiol* 35(12):1123–1138
- Mazick A et al (2007) Monitoring excess mortality for public health action: potential for a future European network. *Wkly Releases (1997–2007)* 12(1):3107
- McCullagh P (1983) Quasi-likelihood functions. *Ann Stat* 11(1):59–67
- Michelozzi P, de’Donato F, Scortichini M, Pezzotti P, Stafoggia M, Sario MD, Costa G, Noccioli F, Riccardo F, Bella A, Demaria M, Rossi P, Brusaferro S, Rezza G, Davoli M (2020) Temporal dynamics in total excess mortality and COVID-19 deaths in Italian cities. *BMC Public Health* 20:1238
- Morfeld P, Timmermann B, Groß J, DeMattheis S, Lewis P, Cocco P, Erren T (2020) COVID-19: spatial resolution of excess mortality in Germany and Italy. *J Infect* 82(3):414–451
- Morfeld P, Timmermann B, Groß VJ, Lewis P, Erren TC (2021) COVID-19: Wie änderte sich die Sterblichkeit?–Mortalität von Frauen und Männern in Deutschland und seinen Bundesländern bis Oktober 2020. *Dtsch Med Wochenschr* 146(02):129–131
- Nielsen J, Mazick A, Glismann S, Mølbak K (2011) Excess mortality related to seasonal influenza and extreme temperatures in Denmark, 1994–2010. *BMC Infect Dis* 11(1):350
- Nørgaard SK, Vestergaard LS, Nielsen J, Richter L, Schmid D, Bustos N, Braye T, Athanasiadou M, Lytras T, Denissov G et al (2021) Real-time monitoring shows substantial excess all-cause mortality during second wave of COVID-19 in Europe, October to December 2020. *Euro Surveill* 26(2):2002023
- Price R (1771) Observations on reversionary payments; on schemes for providing annuities for widows, and for persons in old age; on the method of calculating the values of assurances on lives, and on the national debt. Cadell, London
- Ragnitz J (2021) Hat die Corona-Pandemie zu einer Übersterblichkeit in Deutschland geführt? - Aktualisierung 15.1.2021. ifo Institute, München (Tech. rep.)
- Raths J (1909) Die sterblichkeitsmessung in der allgemeinen bevölkerung. In: Denkschriften und Verhandlungen des 6. Internationalen Kongressess für Versicherungswissenschaften, Wien, pp 115–129
- Rommel A, von der Lippe E, Plaß D, Ziese T, Diercke M, Haller S, Wengler A et al (2021) COVID-19-Krankheitslast in Deutschland im Jahr 2020. Robert Koch-Institut, Berlin (Tech. rep.)
- Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology. Lippincott Williams & Wilkins, Philadelphia
- Schneble M, De Nicola G, Kauermann G, Berger U (2021) A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020. *Biom J* 63(8):1623–1632
- Simonsen L, Spreeuwenberg P, Lustig R, Taylor RJ, Fleming DM, Kroneman M, Van Kerkhove MD, Mounst AW, Paget WJ et al (2013) Global mortality estimates for the 2009 influenza pandemic from the GLaMOR project: a modeling study. *PLoS Med* 10(11):e1001558
- Stang A, Standl F, Kowall B, Brune B, Böttcher J, Brinkmann M, Dittmer U, Jöckel KH (2020) Excess mortality due to COVID-19 in Germany. *J Infect* 81(5):797–801
- Vandoros S (2020) Excess mortality during the COVID-19 pandemic: early evidence from England and Wales. *Soc Sci Med* 258:113101
- Velavan TP, Meyer CG (2020) The COVID-19 epidemic. *Trop Med Int Health* 25(3):278
- Vestergaard LS, Nielsen J, Richter L, Schmid D, Bustos N, Braeye T, Denissov G, Veideman T, Luomala O, Möttönen T et al (2020) Excess all-cause mortality during the COVID-19 pandemic in Europe—preliminary pooled estimates from the EuroMOMO network, March to April 2020. *Euro Surveill* 25(26):2001214
- Vincent JL, Taccone FS (2020) Understanding pathways to death in patients with COVID-19. *Lancet Respir Med* 8(5):430–432

Wenau G, Grigoriev P, Shkolnikov V (2019) Socioeconomic disparities in life expectancy gains among retired German men, 1997–2016. *J Epidemiol Community Health* 73(7):605–611

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

14. An update on excess mortality in the second year of the COVID-19 pandemic in Germany

Contributing article

De Nicola, G., and Kauermann, G. (2022). An update on excess mortality in the second year of the COVID-19 pandemic in Germany. *AStA Wirtschafts-und Sozialstatistisches Archiv*, 16:21–24. <https://doi.org/10.1007/s11943-022-00303-9>.

Copyright information

This article is distributed under a Creative Commons 4.0 International license ([CC BY 4.0](#)).

Author contributions

Göran Kauermann had the idea of writing a follow-up article extending the estimation of excess mortality in Germany to the year 2021. The paper was then written by Giacomo De Nicola, who was also responsible for data analysis and visualization. Göran Kauermann contributed through fruitful comments and extensive proofreading of the manuscript.



An update on excess mortality in the second year of the COVID-19 pandemic in Germany

Giacomo De Nicola · Göran Kauermann

Received: 2 February 2022 / Accepted: 9 February 2022 / Published online: 15 March 2022
© The Author(s) 2022

Abstract In this short note, we apply the method of De Nicola et al. (2022) to the most recent available data, thereby providing up-to-date estimates of all-cause excess mortality in Germany for 2021. The analysis reveals a preliminary excess mortality of approximately 2.3% for the calendar year considered. The excess is mainly driven by significantly higher excess mortality in the 60–79 age group.

Keywords COVID-19 · Excess mortality · Expected mortality · Standardized mortality rate

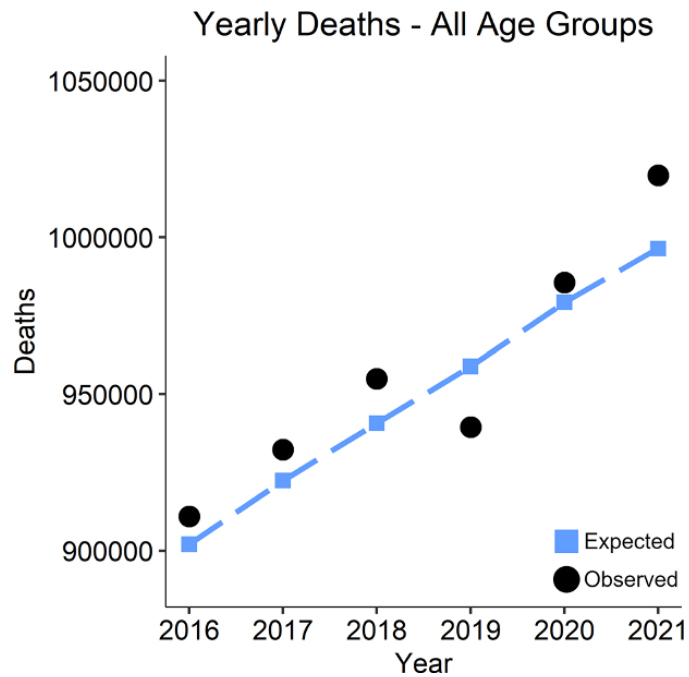
Ein Update zur Übersterblichkeit im zweiten Jahr der COVID-19 Pandemie in Deutschland

Zusammenfassung In diesem kurzen Beitrag wenden wir die Methode von De Nicola et al. (2022) auf die neuesten verfügbaren Daten an und zeigen aktuelle Schätzungen der Gesamt-Übersterblichkeit in Deutschland für das Jahr 2021. Die Analyse zeigt eine vorläufige Übersterblichkeit von etwa 2,3% für das betrachtete Kalenderjahr. Dieser Wert ist hauptsächlich auf eine deutlich höhere Übersterblichkeit in der Altersgruppe der 60–79-Jährigen zurückzuführen.

Schlüsselwörter COVID-19 · Übersterblichkeit · Erwartete Sterblichkeit · Standardisierte Mortalitätsrate

Giacomo De Nicola (✉) · Göran Kauermann
Ludwig-Maximilians-Universität München, Munich, Germany
E-Mail: giacomo.denicola@stat.uni-muenchen.de

Fig. 1 Expected deaths by year, represented by blue squares, plotted against observed fatalities, depicted by black dots. Overall excess mortality in 2021 was more pronounced than in 2020



In our article (De Nicola et al. 2022) in this issue, we presented a simple and novel method to compute excess mortality in a given calendar year while effectively taking the age structure of the population into account. We then applied our method to age-stratified mortality data to obtain estimates for general and age group-specific excess mortality for Germany in 2020, the first year of the COVID-19 pandemic. As we enter 2022, mortality figures from 2021 are starting to become available. With this short note, we thereby aim to provide the reader with up-to-date estimates of excess mortality for the second consecutive year of the pandemic. Mortality data are provided by the German Federal Statistical Office (Destatis 2022). Figures for 2021 are, at time point of submission of this note, not final, and numbers will presumably increase due to data corrections. We leave this problem aside here, and work with data as of February 1, 2022.

Table 1 Expected and observed yearly mortality in 2021 for each age group

Age group	Expected 2021	Observed 2021	Absolute diff.	Relative diff.
[00,30)	7383	7386	3	+0%
[30,40)	6696	6910	214	+3%
[40,50)	15 107	16 190	1083	+7%
[50,60)	58 593	59 221	628	+1%
[60,70)	120 356	126 183	5827	+5%
[70,80)	193 669	203 732	10 063	+5%
[80,90)	397 875	396 578	-1297	-0%
[90, ∞)	196 878	203 609	6731	+3%
Total	996 410	1 019 809	23 399	+2%

14. An update on excess mortality in the second year of the COVID-19 pandemic in Germany

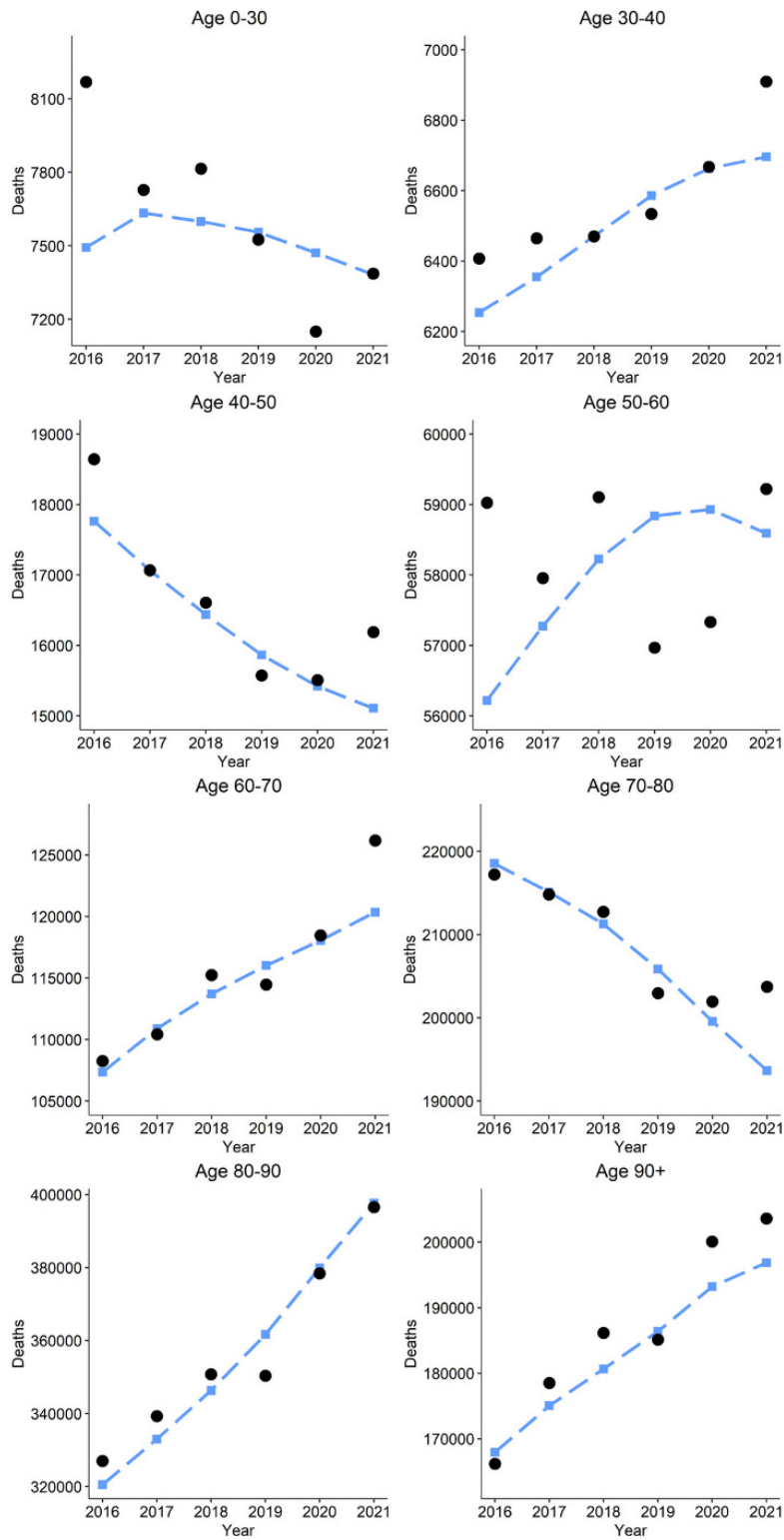


Fig. 2 Expected deaths per year, represented by blue squares, plotted against observed fatalities, depicted by black dots, shown separately for each age group. Relative excess mortality in 2021 was most pronounced in the 40–50, 60–70 and 70–80 age categories

Fig. 1 gives an overview of the results for all age groups combined. We plot the expected death counts for each year as blue squares (see De Nicola et al. 2022 for details), and the observed death counts as black dots. We can see that overall excess mortality in 2021 was more pronounced than in 2020. More specifically, as of February 1, 2022, a total of 1 019 809 deaths were registered in Germany for the year 2021, i.e. 23 399 deaths more than expected. This corresponds to an estimated overall excess mortality of approximately 2.3%.

Table 1 and Fig. 2 give a more complete picture of the mortality observed in 2021 for the different age groups. We observe that the most pronounced relative excess mortality was observed in the age groups 40–50, 60–70 and 70–80. We can also see how, in general, excess mortality was more driven by deaths in the 60–79 age category rather than in the 80+ group.

As a concluding note, we emphasise that all results presented here are based on provisional data, as the final death tolls for 2021 in Germany are not yet available at the time of writing. We can therefore expect some more deaths to be registered in the coming months. Based on past experience, those late registration should produce an increase of a few thousand units in the final toll (last year 982 489 deaths were registered for 2020 as of January 29, 2021, while the final, official toll amounted to 985 572). All in all, we can conclude that excess mortality for 2021 in Germany can, with data up to February 1, 2022, be estimated at a minimum of 2.3%, and that the final estimate will most likely be higher by a few decimal points.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access 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/>.

References

- De Nicola G, Kauermann G, Höhle M (2022) On assessing excess mortality in Germany during the COVID-19 pandemic. *AStA Wirtsch Sozialstat Arch*. <https://doi.org/10.1007/s11943-021-00297-w>
- Destatis (2022) Sterbefälle – Fallzahlen nach Tagen, Wochen, Monaten, Altersgruppen, Geschlecht und Bundesländern für Deutschland. www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Sterbefaelle-Lebenserwartung/Tabellen/sonderauswertung-sterbefaelle.html. Accessed 1 Feb 2022

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

15. Estimating excess mortality in high-income countries during the COVID-19 pandemic

Contributing article

De Nicola, G., and Kauermann, G. (2024). Estimating excess mortality in high-income countries during the COVID-19 pandemic. *Journal of the Royal Statistical Society Series A: Statistics in Society*, qnae031. <https://doi.org/10.1093/jrssa/qnae031>.

Data and code

Available at <https://github.com/gdenicola/excess-mortality-world>.

Copyright information

© (RSS) Royal Statistical Society 2024. The full article is included, as a license to reuse it in this dissertation for non-commercial purposes has been obtained by the author.

Supplementary material

Supplementary material is available at JRSSA online.

Author contributions

Giacomo De Nicola had the idea of estimating excess mortality for all countries with openly available data. Moreover, Giacomo De Nicola extended the expected mortality model by substantiating a method of uncertainty quantification for it. Giacomo De Nicola further designed and wrote the entire paper, and was responsible for data management, analysis and visualization. Göran Kauermann contributed through fruitful comments and extensive proofreading of the manuscript.

Journal of the Royal Statistical Society Series A: Statistics in Society, 2024, 00, 1–18
<https://doi.org/10.1093/jrssa/qnae031>



Original Article

Estimating excess mortality in high-income countries during the COVID-19 pandemic

Giacomo De Nicola  and Göran Kauermann 

Department of Statistics, LMU Munich, Ludwigstr. 33, Munich 80539, Bavaria, Germany

Address for correspondence: Giacomo De Nicola, Department of Statistics, LMU Munich, Ludwigstr. 33, Munich 80539, Bavaria, Germany. Email: giacomo.denicola@stat.uni-muenchen.de

Abstract

Quantifying the number of deaths caused by the COVID-19 crisis has been an ongoing challenge for scientists, and no golden standard to do so has yet been established. We propose a principled approach to calculate age-adjusted yearly excess mortality and apply it to obtain estimates and uncertainty bounds for 30 countries with publicly available data. The results uncover considerable variation in pandemic outcomes across different countries. We further compare our findings with existing estimates published in other major scientific outlets, highlighting the importance of proper age adjustment to obtain unbiased figures.

Keywords: age adjustment, COVID-19, excess mortality, expected mortality, standardization, uncertainty

1 Introduction

The COVID-19 pandemic has caused a tragically large number of casualties in the general population. Accurately quantifying the magnitude of this number has been a challenge for scientists since the start of the crisis. Knowing how many fatalities were caused by the pandemic is crucial for understanding the factors that govern its spread and severity, and to be able to evaluate the effectiveness of government responses to it. However, death tolls officially related to the virus can only paint an incomplete picture of the situation, as many fatal cases of COVID-19 went undetected in official reports from 2020 to 2021, because of limited testing capacity and misclassification of causes of death (Acosta, 2023). Moreover, the reliability of reported deaths varies massively between locations and over time, rendering comparisons largely ineffective. For these reasons, all-cause excess mortality is generally considered to be a more reliable way of assessing the death toll extracted by the pandemic (Beaney et al., 2020; Leon et al., 2020).

Excess mortality can generally be defined as the number of deaths from all causes during a crisis beyond what we would have expected to see under ‘normal’ conditions (Checchi & Roberts, 2005). Specifically, our interest here lies in comparing all-cause mortality observed during the COVID-19 pandemic with the overall number of deaths that would have been expected in its absence. If correctly estimated, excess mortality allows to go beyond confirmed COVID-19 deaths, also capturing fatalities that were not correctly diagnosed and reported as well as deaths from other causes that are attributable to the overall crisis conditions. The concept of excess mortality is well established and has long been utilized for analysing the impact of wars, natural disasters, and pandemics (Johnson & Mueller, 2002; Simonsen et al., 2013), with its application dating as far back as the Great Plague of London in 1665 (see Boka & Wainer, 2020). Today, the concept is routinely employed by governments around the world, with e.g. Europe running an early-warning system specifically dedicated to mortality monitoring (the EuroMOMO project, see Mazick, 2007). Despite this long tradition, however, estimating excess mortality remains a challenge, and no single, unified method for doing so has yet been established (Acosta, 2023; Nepomuceno et al., 2022). The difficulty lies in

Received: May 30, 2023. Revised: December 19, 2023. Accepted: March 4, 2024

© The Royal Statistical Society 2024. All rights reserved. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.



Figure 1. Population pyramids in 2015 and 2020 for Germany and the U.S. Black and coloured contours indicate population levels in 2015 and 2020, respectively. While the German population is rapidly ageing and its structure is irregular, the US pyramid is fairly smooth and stable over time. (a) Germany and (b) USA.

estimating the ‘counterfactual’ expected mortality, i.e. the number of deaths that would have been expected had the crisis not occurred. This is a hard feat as mortality rates, trends, and data availability vary greatly across different regions and periods of time. Estimating expected mortality thus requires (i) choosing a reference period and (ii) using some model to project mortality rates from the reference period to the period of interest. This second point also holds for methods that may seem completely ‘model-free’, such as the basic approach of simply using the mean number of deaths during the reference period as the expected mortality for the crisis period, widely used in media reports at earlier stages of the pandemic. This method implicitly assumes the total (expected) number of deaths to be constant over both the reference and the crisis period, disregarding other factors that may influence mortality, such as varying life expectancy, due to e.g. changes in living conditions, and shifts in the age structure of the population over time. Ignoring the role of age can be particularly damning, as the age structure within a population can change considerably over short periods of time. Moreover, countries can show large variation in how their populations evolve over time, even when their income levels are comparable. To showcase this, [Figure 1](#) depicts the population pyramids of Germany (left panel) and the U.S. (right panel) during the years 2015 (black contour) and 2020 (coloured), respectively. From the German pyramid, it is apparent how the population aged considerably over the 5 years period prior to the start of the pandemic. In particular, the number of people aged 80 and older increased by approximately 25% in just 5 years. Since the probability of death increases sharply at higher ages, the number of deaths in Germany is expected to rise by about 2% per year as a result of ageing alone, as documented by the German Federal Statistical Office in its annual report ([Destatis, 2021](#)). This means that, for these countries, comparing mortality during the pandemic directly with the reference period will lead to severely underestimating expected mortality, and thus overestimating the excess. If we instead shift our focus to the US pyramid in the right panel of [Figure 1](#), we can appreciate how the American population is ageing at a much slower rate than the German one, and that the pyramid is fairly smooth. This means that age adjustment will have a smaller impact on expected and excess mortality estimates. Unfortunately, the case of the U.S. is not the norm in modern day high-income countries. To the contrary, what we here showed for Germany is true for many other nations,

rapidly ageing as a result of declining fertility rates (Bloom et al., 2015). Due to this, and given that both overall and COVID-related mortality are heavily dependent on age (Dowd et al., 2020; Levin et al., 2020), it is crucial to take age into account to avoid systematic bias in the estimates.

Several high profile studies tackling the estimation of excess mortality in multiple countries try to capture trends in mortality by fitting various regression models to the data in the reference period, and then extrapolating the trend to the period of the crisis to obtain expected mortality figures for that period. Karlinsky and Kobak (2021) use linear fits over the years 2015–2019 to obtain excess mortality for 103 countries. The authors use these estimates to compile the World Mortality Dataset, an important source which is also used by the well know outlet Our World in Data (Giattino et al., 2023). The Economist also runs a page dedicated to tracking excess mortality in the different countries of the world by using a mix of boosted gradient, random forest and bootstrapping (The Economist, 2023). Wang et al. (2022) provide estimates of excess deaths due to the pandemic in the 2020–2021 period for 191 countries and territories by using an ensemble of six different regression models, including spline regression. These estimates are also the ones officially reported by the IHME (Institute for Health Metrics & Evaluation, 2022). While incorporating a trend can account for some of the variation in expected mortality rates over the years, the approach is still not free of problems, as not explicitly accounting for age in the model implicitly assumes the age pyramid to be smooth. This is often not the case for many modern-day European countries, where demographic traces of World War II are still visible. As an example of this, the German age pyramid depicted in the left panel of Figure 1 clearly shows a bulge in the cohorts entering their ninth decade of life, which causes large non-linear variation in the age-structure of the population over time. For this reason, a trend alone is often not capable of capturing the effect of age. Furthermore, incorporating country-specific trends in the estimation has the effect of projecting any evolution in the overall death rate observed during the reference period on the period of interest, including those due to factors other than age. While, on the one hand, capturing true long term trends in mortality would be desirable, mortality rates have been shown to exhibit large variance across short periods even in the same region (Bergeron-Boucher & Kjærgaard, 2022). This large variation can lead to predicting large decreases or increases in expected mortality based on variance alone, especially if the trend estimate is based on a period of only 3–5 years, ultimately resulting in overly sensitive and less stable estimates (see Ioannidis et al., 2023; Levitt et al., 2023 for more detail).

Another prominent study is that by Msemburi et al. (2023), which produced the figures officially presented by the WHO (World Health Organization, 2022). The authors use Poisson-based spline regression to fit trends to the reference period, with the exact model specification varying depending on data availability. Excess mortality figures are then obtained by extrapolating a smooth spline based on mortality trends from the 5 years prior to the pandemic (see Knutson et al., 2023, for methodological details). While trend extrapolation, as mentioned, can already be problematic when using linear models, the author's extrapolation is based on smoothing splines, which tend to react strongly to short-term fluctuations. Consequently, in countries with weak influenza seasons in 2019, such as many European ones, a steep decline in (expected) mortality is predicted based on a single data point. Extrapolations from smooth splines are generally very sensitive to the last observation (see Carballo et al., 2021). Problems with this method have been acknowledged by the authors (Van Noorden, 2022), who are also working to incorporate age into their analysis and correct their figures (Acosta, 2023).

Given the major role that age plays in mortality, many have argued explicit age-adjustment to be a sensible way forward (Gianicolo et al., 2021; Levitt et al., 2022; Nepomuceno et al., 2022; Stang et al., 2020). Several prominent multi-nation studies also do take age into account in their analysis. Islam et al. (2021) provide a comparative study on excess mortality in 29 high-income countries in 2020. Their estimates are based on Poisson regression models including the age group (0–14, 15–64, 65–74, 75–85, and >85 years old) as a covariate and show that different states had excess mortality in different weeks in 2020. Konstantinoudis et al. (2022) also fit Poisson regression models including different age groups, and provide overall figures for five European countries at a fine regional level. Levitt et al. (2022) compare different global excess mortality evaluations, and show the importance of age adjustment by comparing raw calculations with their proposed method for obtaining age-adjusted estimates, showing that the results differ strongly depending on the method used. The author's age adjustment procedure consists in simply dividing the population into the five

age strata (the same ones used by [Islam et al., 2021](#)) and calculating raw excess mortality for each strata, before summing them up to obtain the overall figure.

Distinguishing between age classes as done by the studies mentioned above reduces the magnitude of the age-induced bias and is thus certainly better than not accounting for age at all; on the other hand, simply partitioning the population into age groups is equivalent to assuming age structure to be homogeneous within those age groups, which is unrealistic for, e.g. the age group 15–64 in ageing European countries, where the older part of the group (i.e. close to 64 years of age) is decisively more numerous than the younger part (i.e. close to 15 years). This is also visible in the German age pyramid depicted in [Figure 1](#). To eradicate bias from the estimates, it is thus necessary to perform age-standardization by partitioning the age-pyramid into finer age classes, when data to do so is available. In this article, we present a simple approach to tackle this issue. Building on [De Nicola et al. \(2022a\)](#), we propose a method to estimate yearly excess mortality figures with fine-grained age adjustment, accounting for any changes occurring in the age structure of the target population over time. Unlike most previously described approaches, our method is not regression-based, and instead directly uses a corrected version of the life tables from the reference period to directly compute expected mortality based on hazards and population by age. This allows to avoid the problems with leverage points mentioned above, and makes the estimates more stable. In addition to point estimates, our approach also allows to compute plausible excess mortality ranges based on the age-specific death rates observed in the years prior to the crisis.

After showcasing our method, we go on to apply it to obtain excess mortality estimates and ranges for 30 countries for which the necessary data, i.e. life tables, population pyramids, and yearly deaths tolls, is available. This includes the majority of the world's high income countries. Our results show that, out of these 30 countries, 10 suffered from considerable excess mortality over the years 2020–2021, and 8 displayed a sizeable mortality deficit. In the remaining 12 countries, age-adjusted mortality did not substantially deviate from levels observed during the 5-year period preceding the pandemic. After presenting our estimates, we compare them with those obtained by five other prominent multi-country studies previously discussed, namely those from the IHME, WMD, Economist, WHO, and [Levitt et al. \(2022\)](#). The comparison showcases the impact of the employed methodology on the results, as the estimates differ greatly based on the method used, and especially so in countries where population is ageing at higher pace. In addition to presenting our results, we further make data and code to reproduce all of our analyses available in our public GitHub repository ([De Nicola, 2023](#)). This is done to enhance the reproducibility of our results, as well as to facilitate researchers in employing and adapting our methods for further application.

The remainder of the article is organized as follows. [Section 2](#) describes the employed methods, and [Section 3](#) introduces the various data sources used. We present our empirical results in [Section 4](#), while [Section 5](#) is dedicated to the comparison with other prominent studies. [Section 6](#) concludes the article with an extended discussion of our contributions.

2 Methods

The main challenge in estimating excess mortality during any crisis lies estimating the ‘counterfactual’ expected mortality, i.e. the number of deaths that would have been expected had the crisis not occurred. One way to do this is to consider mortality rates observed shortly before the crisis and project them onto the period of interest. A natural approach in this regard is to consider age-specific mortality data contained in official life tables, which give the probability q_x of a person who has completed x years of age to die before completing their next life-year, i.e. before their $x + 1$ th birthday. The calculation of a life table, as simple as it sounds, is not straightforward, and is an age-old actuarial problem. First references date far back, to [Price \(1771\)](#) and [Dale \(1772\)](#). A historical digest of the topic is provided by [Keiding \(1987\)](#). To calculate expected mortality in 2020 and 2021, we here make use of the 2015–2019 5-year period life tables provided by the Human Mortality Database (HMD), one of the most comprehensive and up to date databases on mortality freely available to the public ([HMD, 2023](#)). These life tables are calculated as described in [Section 7.1](#) of the HMD Methods Protocol ([Wilmoth et al., 2021](#)). The calculation method is akin to that of traditional life tables (see e.g. [Raths, 1909](#)), with the sensible addition of smoothing the death rates via a logistic function for old ages (80+). As demonstrated in [De Nicola et al. \(2022a\)](#), further adjustments to the tables are recommendable to relate the

expected number of deaths to recently observed ones, especially for countries in which the age pyramid is not smooth, i.e. where the assumption of a stationary population (see Wilmoth et al., 2021, p.36) does not hold. In particular, population data and life tables need to be appropriately matched, since life tables count the number of deaths of x -year-old people over the course of a year, while population data typically gives the number of x -year-old people at a fixed time point (typically the beginning of the year). This requires correction (1), which accounts for the fact that a person that dies at x years of age in a given year t was either x years old or $x - 1$ years old at the beginning of the year, i.e. the time point used for the population data. We therefore apply this additional correction, which consists in calculating the adjusted age-specific death probabilities \tilde{q}_x at age x as

$$\tilde{q}_x = \frac{1}{2}q_x + \frac{1}{2}q_{x+1}, \tag{1}$$

where q_x are the death probabilities for age x contained in the life tables before the adjustment. Assuming the maximum possible age to be 110 years, as done in the HMD life tables, we can then compute the overall expected number of deaths in year t as

$$E_t = \sum_{x=1}^{110} \tilde{q}_x P_{x,t}, \tag{2}$$

where $P_{x,t}$ is the population aged x at the beginning of year t . We can then obtain excess mortality estimate Δ_t for a given year t by simply subtracting the expected mortality estimate E_t from the observed death toll O_t in the same year:

$$\Delta_t = O_t - E_t.$$

Computing Δ_t with $t = 2020, 2021$ using 5-year 2015–2019 life tables yields our excess mortality estimates for the first 2 years of the COVID-19 pandemic in a given country.

Note that the choice of a 5-year reference period is somewhat arbitrary, but driven by the following principles. In general, it is desirable to have a reference period that is (a) long enough to provide robust data evidence and not fall prey of variance and (b) short enough to be as similar to the period of interest as possible. Given that the HMD provides life tables calculated on either 1-year, 5-year, or 10-year periods, we picked the 5-year one as it strikes a balance between duration and similarity to the pandemic period. Indeed, using only a single year as the reference period is generally problematic, as yearly death rates exhibit considerable variation, well beyond what can be explained by underlying changes in life expectancy over time. This is particularly evident in our application, as the year 2019, immediately preceding the pandemic, was characterized by relatively low mortality levels in Europe, due to e.g. a mild influenza season. On the other hand, using a reference period as long as 10 years is problematic due to the fact that baseline mortality levels can change over such a wide time window. In fact, using data from 2010 to calculate expected mortality 10 years later would downweigh real gains in life expectancy due to e.g. changes in living condition and advances in medical technology over time. As such, 5-year life tables seem to be a reasonable choice to strike a balance between bias and variance. We further note that, as shown by Levitt et al. (2023), while the choice of the reference period does influence the absolute value of the estimates, it tends not to strongly impact how countries rank relative to one another in terms of excess mortality. In the interest of completeness, we also provide alternative estimates calculated with a 3-year reference period (i.e. 2017–2019) and compare them with the 5-year ones in the supplementary material (Section S.3). The comparison highlights how the shorter reference period leads to slightly higher excess mortality estimates, while leaving country rankings largely unaffected.

An open issue in assessing excess mortality is the quantification of uncertainty. Probability models do not seem very useful here, as variation in mortality is in large part driven by external factors, such as, e.g. the strength of an influenza wave and other exogenous shocks. Because of that, residual variability is well beyond what would be explainable via standard distributional

assumptions. For this reason, we explicitly refrain from pursuing model-based approaches, and instead propose a data-driven empirical assessment of variability. Specifically, we make use of age-specific single-year mortality rates to provide what we can call a ‘plausible range’ for expected mortality. To do so, we consider the single-year life tables for each year of the reference period, and use them to calculate expected mortality for the years of interest in the same way as above, i.e. using (1) and (2). Assuming the reference period to contain a total of K years (in our case $K = 5$), we will obtain K different excess mortality estimates. We can then take the lowest and highest resulting estimates as the upper and lower bound of our plausible expected mortality range. In other words, we use mortality rates from the ‘worst’ and ‘best’ years of the reference period to obtain a plausible range for expected mortality in the years of interest. To be more precise, the upper mortality bound for year t can be written as:

$$E_t^{\text{upper}} = \max(\tilde{E}_{t,1}, \tilde{E}_{t,2}, \dots, \tilde{E}_{t,K}), \quad (3)$$

where $\tilde{E}_{t,k}$ represents expected mortality for year t calculated using the (corrected) single-year life tables from year k . Analogously, the lower bound can be defined as

$$E_t^{\text{lower}} = \min(\tilde{E}_{t,1}, \tilde{E}_{t,2}, \dots, \tilde{E}_{t,K}). \quad (4)$$

These expected mortality bounds can then be used to obtain excess mortality intervals in a straightforward manner. Note that these bounds do not give a probabilistic measure of uncertainty, as no distributional model is used. Instead, they provide us with plausible high-mortality and low-mortality scenarios for expected mortality in the years of interest based on levels observed during reference years. In a sense, this is akin to the multiverse approach proposed by [Levitt et al. \(2023\)](#), whereas instead of presenting all possible universes we only present the average one, the best one and the worst one.

3 Data

To compute expected mortality for a given year in a single country/region, our method needs (i) life tables for the reference period and (ii) population data by age for the year of interest. Once expected mortality is calculated, to compute the excess we also need (iii) the yearly death toll for the year of interest. In our case, the period of interest is given by the first 2 years of the pandemic, i.e. 2020 and 2021, while we set the reference period to be 2015–2019. As such, we included in our analysis all countries for which these three pieces of information were readily available at the time of the analysis. We briefly summarize where each of the data pieces was sourced from below, with additional details given in the [supplementary material](#).

Life tables: A great source for population data by year and life tables is given by the Human Mortality Database (HMD), one of the most comprehensive and up to date databases on mortality freely available to the public ([HMD, 2023](#)). All life tables used here were sourced from the HMD, and in fact, HMD availability was used as our first inclusion criterion: only countries for which life tables up to 2019 are present in the HMD at the time of the analysis were included in our study. More specifically, 5-year life tables from 2015 to 2019 were used to calculate average expected mortality, while single-year life tables from 2015 to 2019 were used to calculate plausible intervals, as detailed in Section 2. Note that all life tables were calculated as described in the HMD’s method protocol ([Wilmoth et al., 2021](#)), and subsequently adjusted as described in the Methods section.

Population: Population data by single year of age were also sourced from the Human Mortality Database. The presence of this data for the years 2020 and 2021 was the second inclusion criterion for our analysis, as it is needed to calculate expected mortality. Exceptions were made for Italy and Austria, as both countries were central in the COVID-19 debate, especially in the early stages of the pandemic. For both countries, population pyramids were downloaded from the websites of the respective national statistical offices. More details on those sources are given in the [supplementary material](#).

Deaths: Overall death tolls by country for the years 2020 and 2021 are needed to calculate excess mortality in those years. For EU countries, official death tolls were sourced from the Eurostat website (Eurostat, 2023). An exception was made for France, as the Eurostat tolls also include overseas territories; as the HMD life tables refer to mainland France, we sourced mainland death data from the website of the French national statistical office. We also obtained deaths data from the respective official sources for all non-EU countries included in our analysis: details on these sources can be found in the [supplementary material](#).

4 Results

Using the method detailed in Section 2 and following the inclusion criteria detailed in Section 3, we estimated excess mortality for the years 2020 and 2021 in a total of 30 countries. Table 1 shows all country-specific figures pooled for the years 2020 and 2021. In particular, the table shows, for each country, expected, observed, and excess mortality. The table additionally provides the percentage excess mortality, calculated as $\Delta_t^{\%} = \Delta_t/E_t$, as well as the percentage plausible range for the excess in the 2 years, calculated as detailed in Section 2. Similar tables with separate figures for 2020 and 2021 are given in the [supplementary material](#). From Table 1, we can see that the analysed countries exhibit considerable variation in pandemic outcomes, with relative figures ranging from an excess mortality of 22.8% in Bulgaria, all the way to a mortality deficit of 9.1% in South Korea. Within our sample, 17 countries had positive excess mortality, while 13 countries saw a mortality deficit during the analysed period. To check whether excess (or deficit) mortality was beyond the expected variation in a given country, we can make use of the estimated plausible range: Indeed, if the range is completely above zero, it means that mortality in the 2020–2021 period was higher than in any of the single years of the reference period, providing solid evidence towards the presence of sizeable excess mortality during the pandemic period. In contrast, a fully negative range implies that mortality during the pandemic period was lower than in any of the years of the 2015–2019 range, indicating a considerable mortality deficit during 2020 and 2021. Using this criterion, we can say that 10 of the 30 analysed countries had substantial excess mortality during the first 2 years of the pandemic. Those countries, ordered by total absolute excess, are: U.S., Italy, UK, Bulgaria, Czechia, Hungary, the Netherlands, Portugal, Croatia, and Lithuania. On the other hand, 8 countries enjoyed a considerable mortality deficit. These are, in order: Japan, South Korea, Taiwan, Australia, Hong Kong, New Zealand, Norway, and Iceland. In all other analysed countries, the two extremes of the range display opposite signs, meaning that pooled mortality from 2020 and 2021 was higher than in the lowest mortality year between 2015 and 2019, but lower than in the highest mortality year in the same range. For these countries, the data thus do not bring conclusive evidence of an excess or a deficit in mortality.

As mortality data are inherently spatial in nature, it is often useful to visualize it on a map, to allow for a better overall view as well as to recognize any spatial patterns that may emerge. Given that many of the analysed countries are in Europe, and given that our data cover most of the western part of the continent, we provide a heat-map of excess mortality in Europe during the 2020–2021 period in Figure 2. From the map, we can clearly see that Bulgaria stands out as the state with the highest excess mortality, with a value of 22.8% (as seen from Table 1). An important thing to note in this regard is that Bulgaria is the only country within our sample which is not high-income by World Bank standards but is rather found within the upper-middle income bracket (World Bank, 2022). This is relevant, as it provides an indication of how much harder lower income countries were hit by the pandemic in terms of life loss. While our study only focuses on countries for which the necessary data is fully available, i.e. primarily high-income countries, studies working with incomplete data such as those of Msemburi et al. (2023) and Karlinsky and Kobak (2021) corroborate this. From the map, a spatial pattern is also visible: Within Europe, southern and eastern countries suffered from excess mortality, while northern countries mostly display mortality deficits. Lower mortality in the Nordic countries may be due to a combination of campaigns delivering vaccines faster to more people than the European Union (EU) average, effective non-pharmaceutical public health interventions (NPIs) and high baseline capacities of the health care systems (Schöley et al., 2022). Lower population densities may also have played a role in stifling the spread of the disease (Rocklöv & Sjödin, 2020).

Table 1. Expected and observed mortality in the 2020–2021 period for each of the 30 countries included in the analysis

Country	Pop.	Expected	Observed	Excess	%Excess (%)	Plausible range
Australia	25.6M	358,397	332,769	−25,628	−7.2	(−10.4%, −3.8%)
Austria	8.9M	176,736	183,561	6,825	+3.9	(−0.5%, +7.1%)
Belgium	11.5M	232,757	239,227	6,470	+2.8	(−2.1%, +7.4%)
Bulgaria	6.9M	222,890	273,730	50,830	+22.8	(+19.5%, +25.5%)
Canada	37.9M	611,142	620,052	8,910	+1.5	(−0.3%, +4.1%)
Croatia	4.0M	111,092	119,735	8,642	+7.8	(+1.9%, +12.9%)
Czechia	10.5M	237,489	269,180	31,691	+13.3	(+8.7%, +16.7%)
Denmark	5.8M	115,349	111,797	−3,552	−3.1	(−5.2%, +0.1%)
Finland	5.5M	116,484	113,147	−3,337	−2.9	(−5.4%, +1.0%)
France	65.4M	1,268,041	1,298,800	30,759	+2.4	(−0.5%, +5.3%)
Germany	83.2M	2,005,161	2,009,259	4,098	+0.2	(−3.2%, +3.5%)
Hong Kong	7.5M	108,628	102,189	−6,439	−5.9	(−12.1%, −1.4%)
Hungary	9.8M	270,804	297,457	26,653	+9.8	(+6.0%, +13.0%)
Iceland	0.4M	4,926	4,637	−289	−5.8	(−11.2%, −2.0%)
Ireland	5.0M	68,404	65,445	−2,959	−4.3	(−8.6%, +0.4%)
Italy	59.4M	1,352,461	1,449,352	96,891	+7.2	(+2.0%, +10.8%)
Japan	123.5M	2,944,310	2,825,044	−119,266	−4.1	(−6.4%, −1.9%)
Lithuania	2.8M	83,293	91,293	8,000	+9.6	(+3.5%, +17.2%)
Luxembourg	0.6M	9,211	9,098	−113	−1.2	(−3.0%, +2.1%)
Netherlands	17.4M	325,475	339,650	14,175	+4.4	(+1.9%, +7.9%)
New Zealand	5.1M	72,746	67,545	−5,201	−7.2	(−9.3%, −5.4%)
Norway	5.4M	87,031	82,613	−4,418	−5.1	(−8.2%, −1.6%)
Portugal	10.3M	237,744	248,198	10,454	+4.4	(+1.9%, +7.8%)
South Korea	51.3M	684,662	622,628	−62,034	−9.1	(−14.4%, −2.6%)
Spain	47.4M	905,407	941,717	36,310	+4.0	(−1.2%, +9.4%)
Sweden	10.4M	192,763	190,082	−2,681	−1.4	(−4.2%, +4.1%)
Switzerland	8.6M	145,131	147,387	2,256	+1.6	(−4.3%, +5.2%)
Taiwan	23.6M	383,471	357,239	−26,232	−6.8	(−10.4%, −3.2%)
UK	66.9M	1,285,300	1,357,108	71,808	+5.6	(+2.4%, +9.5%)
U.S.	330.7M	5,921,695	6,842,426	920,731	+15.6	(+14.1%, +17.4%)

Zooming in on the four largest EU countries, [Figure 3](#) plots expected and observed mortality figures by calendar year in Germany (top-left), France (top-right), Italy (bottom-left), and Spain (bottom-right). Note that expected deaths from years 2015–2019 were calculated the same way as the 2020–2021 ones, i.e. using corrected 2015–2019 life tables. We plot the observed death counts for each year as black dots, while expected death counts are represented by blue squares. Furthermore, the light-blue bands mark the plausible expected mortality range for each year, calculated using (3) and (4). From the figure, we can immediately appreciate the importance of age adjustment for these four countries, as both expected and observed mortality tend to naturally increase year over year due to ageing populations. Calculating excess mortality using raw data would thus return inflated figures. Focusing on the single countries we can see how Germany did not suffer from sizeable excess mortality during the pandemic, with age-adjusted mortality in 2020 and 2021 being in line with previous years. On the other hand, the three Mediterranean countries depicted all suffered from varying degrees of increased mortality in 2020, with mortality converging back to more normal levels in the second pandemic year.

15. Estimating excess mortality in high-income countries during the COVID-19 pandemic

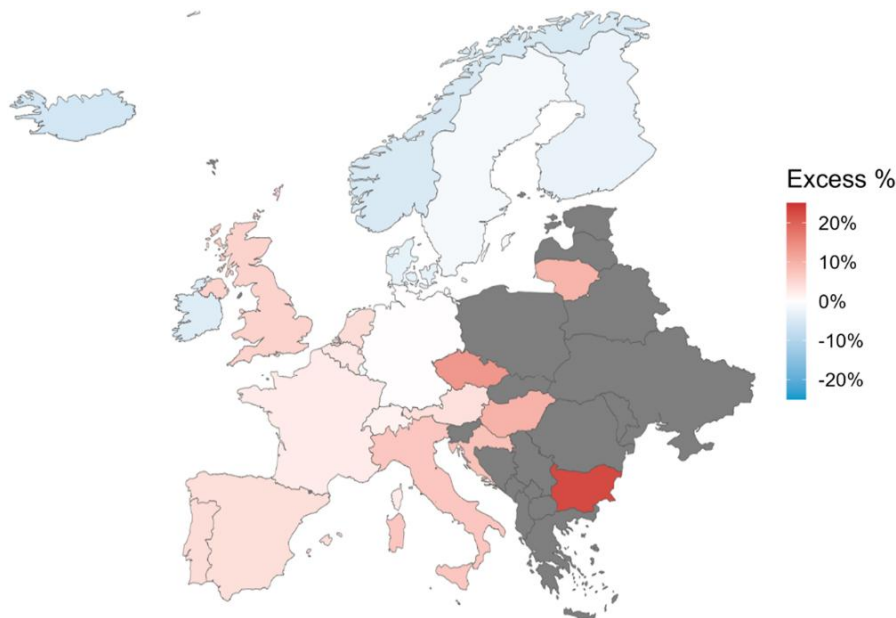


Figure 2. Heat-map of excess mortality in Europe in the 2020–2021 period. Dark grey indicates that the necessary data is not available at the time of the analysis.

Figure 4 shows the same plot for four of the largest non-EU countries included in our analysis, namely the U.S. (top-left), the UK (top-right), Japan (bottom-left) and Australia (bottom-right). From the U.S. plot, we can immediately notice the anomalous mortality levels that characterized the years 2020 and 2021—quantifiable in 14.5% and 16.6% excess mortality by year, respectively. Overall, total excess deaths in the country for those 2 years amount to almost one million—by far the largest figure within our sample. The deaths are mainly attributable to COVID-19, with outcomes worsened by lower vaccination uptake (Suthar et al., 2022) as well as conditions that may have resulted from delayed medical care and overwhelmed health systems (Woolf et al., 2021). From the US plot we can also see how mortality from COVID-19 is especially high among the elderly: after regularly increasing from 2015 to 2020, expected mortality remains almost constant from 2020 to 2021, as the victims of COVID-19 in 2020 were in large part pertaining to the elderly population, which disproportionately contributes to expected mortality. This change in growth rate in expected mortality is also visible in other countries, such as Italy and Spain in Figure 3, as well as the UK in the top-right panel of Figure 4. From the latter plot, we can see how the UK also suffered from increased mortality levels in both pandemic years, even though to a lesser extent than the U.S. On the other hand, the countries depicted in the bottom panels of the same figure, i.e. Japan and Australia, paint a completely different picture. Both nations withstood the first two pandemic years without incurring in excess mortality, and instead registered considerable mortality deficits in both years. Note that a mortality deficit during the pandemic years does not imply over-reporting of COVID-19 deaths, but rather that deaths avoided or postponed by NPIs and behavioural changes in the population outweighed deaths caused by the virus. These changes may have led to, e.g. reduced mortality from other respiratory infections as well as accidents (Barnes et al., 2022; Olsen et al., 2020). Geographical isolation of the two island countries may also have contributed to reduce the spread of COVID-19, as in the case of New Zealand (Kung et al., 2021). From the data in Table 1, we can see how similar levels of reduced mortality during the pandemic were observed, among others, in other East Asian regions (South Korea, Taiwan, Hong Kong) as well as other islands (Iceland and New Zealand), hinting at the presence of geographical patterns. Plots similar to those in Figures 3 and 4 for all other countries included in our analysis are provided in the supplementary material (Section S.5).

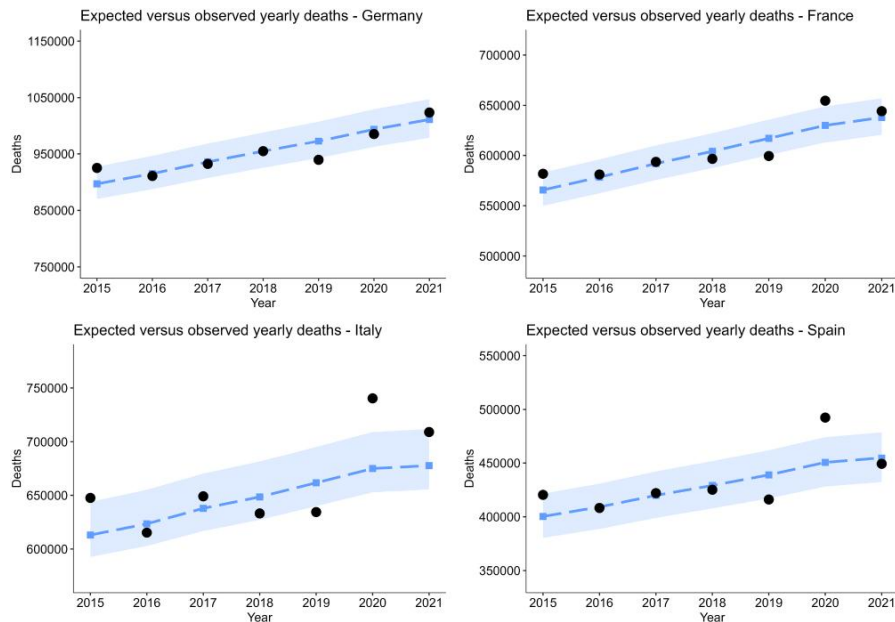


Figure 3. Expected and observed mortality figures by calendar year for the four largest EU countries: Germany, France, Italy, and Spain. Black dots indicate observed mortality in a given calendar year, while blue squares indicate estimated yearly expected mortality. Shaded bands represent the estimated expected mortality range.

5 Comparison with other estimates

This section is dedicated to comparing excess mortality estimates obtained through our method with figures produced by five other prominent multi-country studies. In particular, we contrast our results with those obtained in the already mentioned studies by Msemburi et al. (2023), Karlinsky and Kobak (2021), Wang et al. (2022), The Economist (2023), and Levitt et al. (2022). The reasoning behind the choice of these benchmarks is the following: All five studies are very high profile, and the first four were used by prominent institutions and media outlets as their official estimates. Respectively, the first one is used by the WHO (World Health Organization, 2022), the second one for the World Mortality Database (WMD) and Our World In Data (Giattino et al., 2023), the third one by the IHME (Institute for Health Metrics & Evaluation, 2022), and the fourth one by The Economist. As for the fifth method by Levitt et al., while it is not currently used by official sources, we include it as we want at least one method performing some type of explicit age adjustment, and because, not unlike the other methods, it is highly published and well cited. Note that, while our analysis encompasses 30 countries, countries for which estimates are not available for all six methods were excluded from the comparison. The results for the remaining 25 countries are shown in Figure 5, which depicts percentage excess mortality estimates for the period 2020–2021 calculated with the six different methods. Countries are shown in ascending order with respect to excess mortality computed with our method. Note that, as before, percentage excess mortality was calculated using expected mortality as the base, i.e. $\Delta_t^{\%} = \Delta_t/E_t$, and that our own measure of expected mortality was used as base for all six methods, for reasons of comparability and to ensure consistent rankings. Also note that, while results are here only presented graphically, tables containing figures for all six different methods are provided in the supplementary material (Section S.2).

From the figure, we can appreciate how different estimation methods result in sizeable differences in the estimates. In particular, several patterns emerge. Firstly, we can see how the IHME and Economist methods, which do not account for age at all, consistently produce the highest excess mortality estimates among the six methods. This is to be expected, due to the fact that all

15. Estimating excess mortality in high-income countries during the COVID-19 pandemic

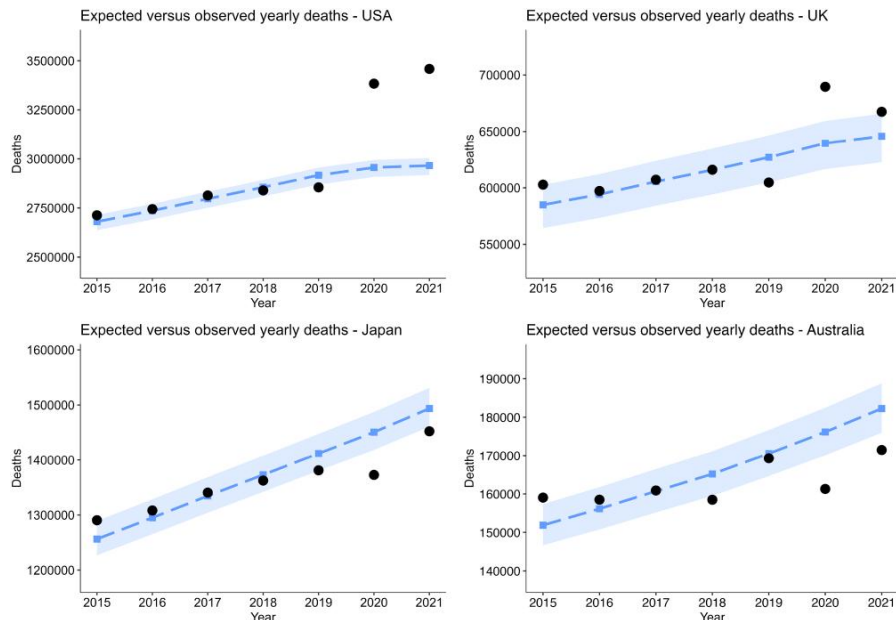


Figure 4. Expected and observed mortality figures by calendar year for four large non-EU high-income countries: U.S., UK, Japan, and Australia. Black dots indicate observed mortality in a given calendar year, while blue squares indicate estimated yearly expected mortality. Shaded bands represent the estimated expected mortality range.

populations considered are ageing to at least some extent. Not considering age thus causes upward bias in the estimates. This difference is also directly related to the extent to which the population is ageing, as it is smaller and almost negligible for countries with relatively stable age pyramids, such as, e.g. the U.S., while it gets larger for rapidly ageing countries, like e.g. Germany and France. From the plot, we also see how the WHO and WMD methods tend to produce estimates that are lower than those of IHME and Economist, but still considerably larger than those of the two explicitly age-adjusted methods considered. As already mentioned in Section 1, the WMD calculates expected mortality through use of a linear trend, and thus is only able to partially capture the effect of age, as age pyramids are generally not smooth. As for the WHO method, it uses Poisson-based spline regression to fit a smooth trend to the reference period and then extrapolates it to calculate expected mortality in the period of interest. While trend extrapolation can already be problematic when using linear models, splines have the additional issue of reacting strongly to short-term fluctuations. As 2019 was a generally a year of low mortality in many of the analysed countries, as visible, e.g. in Figure 3, it is likely for that to have led to overestimation in excess mortality. Shifting our focus to the remaining two methods, i.e. that of Levitt et al. and ours, we can see that the corresponding estimates not only tend to be lower than the other methods but are also closer to each other than to the other ones. Furthermore, the two methods also rank countries similarly. However, we can also observe considerable differences between the two, with the Levitt method tending to produce slightly higher excess mortality estimates. These differences are, in some cases, in the order of 5%. We believe this to be mainly due to two factors. Firstly, the Levitt method uses 2017–2019 as reference period, while we use 2015–2019. Given that mortality was overall slightly higher in 2015 and 2016 than in the 2017–2019 window, estimates of expected mortality based only on the latter years will naturally be slightly lower. In our opinion, both choices of reference period are equally valid, with each having pros and cons. Crucially, our choice of a 5-year reference period allows us to more effectively apply our uncertainty quantification approach by using yearly age-specific mortality rates. Nonetheless, we also recalculated our mean estimates using the same 2017–2019 window as the reference period and included the results in the [supplementary material \(Section S.3\)](#). The alternative reference period results in

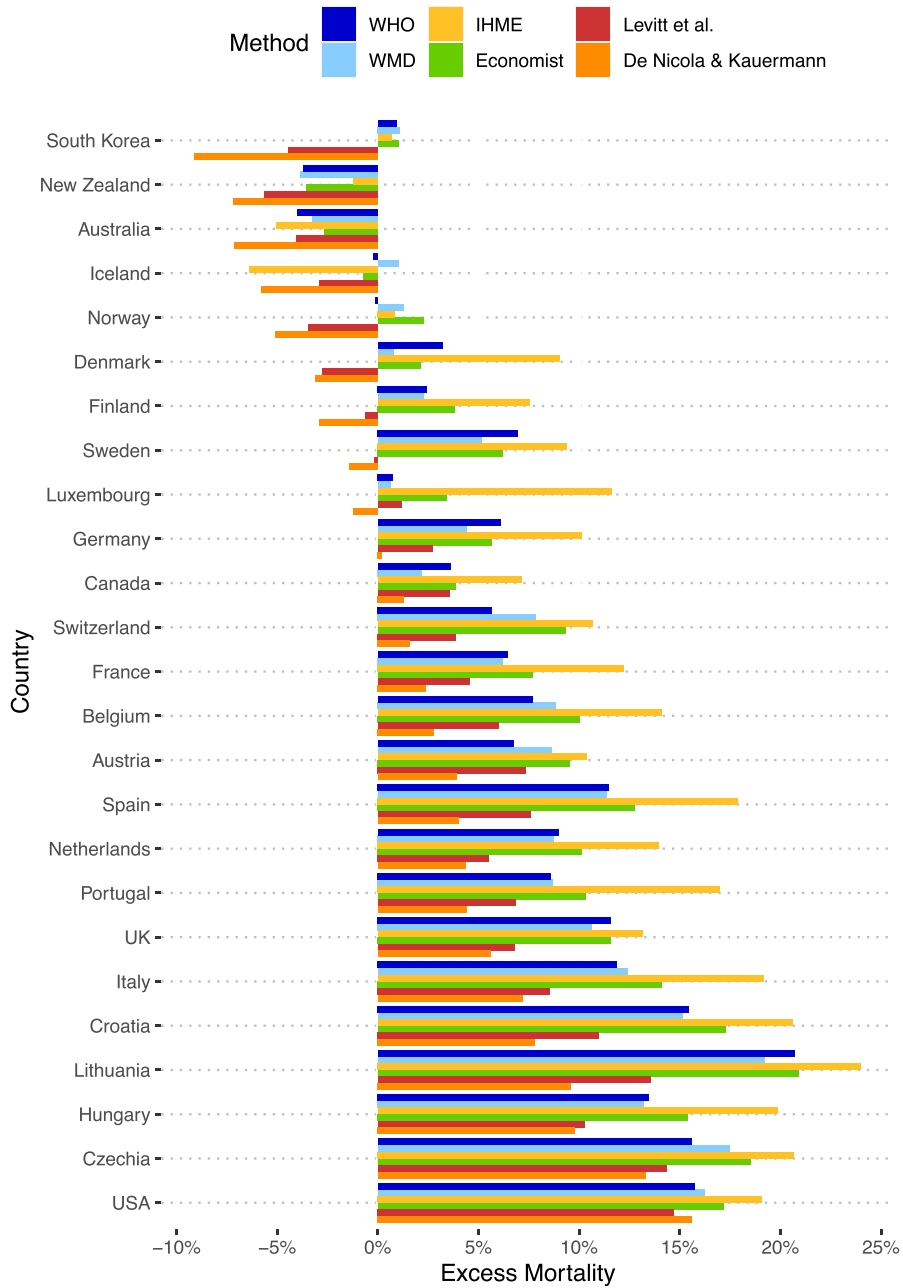


Figure 5. Comparison of country-specific excess mortality estimates obtained by six different studies, including ours. Methods not accounting for age are considerably overestimating excess mortality in countries with ageing populations.

slightly higher excess mortality estimates, while leaving country rankings largely unaffected (see also Levitt et al., 2023). More importantly, differences between the two sets of estimates persist even after aligning the reference periods, leading us to discuss the second main distinguishing factor between the two methods, namely the method used to perform age adjustment. Levitt et al.’s

procedure consists in dividing the population into five age strata (0–14, 15–64, 65–74, 75–85, and >85 years old) and calculating raw excess mortality for each strata, before summing them up to obtain the overall figure. While this is far better than no age adjustment at all, it still disregards age variation within the age groups. This can be especially problematic if the age classes are large, as for example the central age class of 15–64 used by the authors, or if variation within the age classes is large, as e.g. in the 75–85 group for Germany, depicted in Figure 1. The latter point may thus constitute a cause of bias in excess mortality estimation. In contrast, our method performs age standardization by partitioning the population pyramid at the finest available resolution of the data, i.e. 1-year age classes, which we argue to be preferable when data to do so is available.

6 Discussion

Accurately measuring excess deaths in times of COVID-19 is vital to assess the pandemic's impact on public health across different countries and regions of the world. Precise excess mortality estimation is also crucial for explaining the pandemic curve, for understanding the factors contributing to differences in the infection-fatality rate among populations, and for gauging the effectiveness of alternative policy options for managing future crises. In this article, we demonstrated the importance of taking age into account to obtain unbiased excess mortality estimates in high-income countries, and proposed a simple and effective method to do so when the necessary data is available. We applied our method to 30 different countries, and compared our results with other estimates obtained by five other high profile studies. The comparison sheds light on how different estimation methods result in sizeable differences in the estimates. Results are especially sensitive to the absence of age adjustment, which can lead to considerable downward bias, and is most relevant for countries with rapidly ageing populations. For example, estimates in South Korea, one of the countries with the lowest total fertility rates in the world, range from slightly positive excess mortality for methods without age adjustment to a 9.2% mortality deficit using our method. Large differences are also observable in most other high-income countries, while estimates for populations ageing at a slower rate, such as, e.g. the U.S., are more convergent. It is thus crucial to explicitly take age into account for excess mortality estimation in ageing high-income countries. Note that the focus of most of the studies included in the comparison does not lie specifically on high-income countries, but rather on obtaining global estimates of excess deaths, including situations in which high-quality data is not available. Nonetheless, even in the case of global estimates, we argue that it would be relevant to incorporate age adjustment in the estimates for countries in which data to do so is available, as the latter make up a non-negligible portion of the global population. Moreover, we are fortunate that countries where age adjustment tends to be most impactful are also the ones for which high-resolution data is available, making unbiased estimation possible. From the comparison, we further see that performing age adjustment at a fine level, as opposed to doing so by dividing the population into large age classes (as done by Levitt et al., 2022), also results in considerable differences in excess mortality estimates. Unsurprisingly, these differences are also more pronounced for countries in which the structure of the population pyramid is less stable over time. This speaks to the importance of utilizing high-resolution age-stratified data to perform the estimation.

Turning our attention from the method to the empirical results obtained, our estimates uncover large variation in pandemic outcomes among the 30 analysed countries. More specifically, over the 2020–2021 period, 10 countries showed excess mortality beyond what could be explained by standard variation, and 8 displayed a sizeable mortality deficit. In the remaining 12 countries mortality was neither markedly higher nor lower than during the 2015–2019 reference period. The countries with the worst outcomes relative to population were Bulgaria, the only non-high-income country in our sample, which had a 2 years excess of 22.8%, and the U.S., which have seen a 15.6% increase in mortality over the pandemic period. The latter increase is particularly notable given the size of the country's population, as it corresponds to more than 920.000 excess deaths over the 2 years. Other countries that experienced excess mortality (in the order of 5 to 10%) include most of the Eastern and Southern European states, as well as the UK. In contrast, considerable mortality deficits of similar magnitude were observed in some of the Nordic European countries, as well as in Australia, New Zealand, South Korea, Taiwan, and Japan. We here want to stress that, while the absence of excess mortality (or the presence of a mortality deficit)

in a given country does indicate that the country fared well in terms of life loss during the pandemic, it does not imply that no people died because of COVID-19. Instead, reduced mortality during pandemic years simply indicates that deaths from other causes which were prevented during the pandemic, e.g. through governmental NPIs and/or behavioural changes in the population, outnumbered COVID-19 victims. Likewise, for the same reasons, a value of overall excess mortality lower than the officially reported COVID-19 death toll in a given country does not imply over-reporting of COVID-related deaths. In fact, if the healthcare system keeps functioning normally (i.e. the same way as before the pandemic), one would actually expect there to be less overall excess mortality than total deaths caused by COVID-19, as lockdowns can prevent or postpone deaths from other causes, such as e.g. accidents or other respiratory infections (Calderon-Anyosa & Kaufman, 2021; Olsen et al., 2020). Assuming a perfectly functioning reporting system, this should also be reflected in the official death tolls (note that official COVID-19 death tolls for the studied period are given in the [supplementary material](#), for comparison purposes). Of course, even within our high-income sample, the latter assumption only holds to a certain degree in practice, as reporting systems can vary substantially between countries (Karanikolos & McKee, 2020). Likewise, the assumption on the functioning healthcare system is not always valid, as overwhelmed healthcare systems were reported at various stages of the crisis (Alderwick, 2022; Dorsett, 2020; Senni, 2020). Issues with reporting and healthcare systems pose even bigger problems in countries falling within low- and middle-income brackets (Chatterjee, 2020; Lone & Ahmad, 2020). It is therefore not a random occurrence that Bulgaria, being the sole non-high-income country in our sample, displays the highest relative excess mortality of all analyzed regions. Indeed, our selection of countries is not at all representative of the whole world, and outcomes are likely to be much worse in regions with fragile and less efficient healthcare systems (Bong et al., 2020). To obtain global estimates one would therefore need to use estimation techniques which work with deficient data, which our method does not do. Further, we note that our method performs retrospective excess mortality estimation using historical data, which is useful to understand the impact of a crisis and to learn policy lessons in preparation for future ones. On the other hand, the need to use full historical data limits its effectiveness for managing an ongoing crisis, for which a real-time monitoring tool would be needed. While it would be possible to use weekly data as soon as they are available, as done e.g. in Section 3 of De Nicola et al. (2022a), complete weekly data still usually arrive with a delay of at least several weeks, rendering it unfit for live monitoring. A possibility for developing a truly real-time tool would be to make use of so-called nowcasting techniques, i.e. methods bridging the delay between events and their reporting, to estimate the number of fatalities which already occurred based on the ones that were already reported. Examples of this are given by Schneble et al. (2021) and De Nicola et al. (2022b), who perform nowcasting for fatal and general COVID-19 infections, respectively. While this is beyond the scope of this article, nowcasting all-cause mortality data to create a real-time monitoring tool is certainly an interesting direction for future research, with great potential for real-world impact. Another thing to note is that our method for calculating expected mortality does not account for potential trends in life expectancy over time. This is equivalent to implicitly assuming constant age-specific hazards over the considered years (while we still, of course, account for the evolving age structure of the population). As discussed in Section 1, using a time trend to project changes in death rates observed during the reference period on the period of interest can lead to instability in the estimates, due to the large degree of natural variation that is present in all-cause deaths. Furthermore, there is no guarantee that mortality rates should continue to follow the same trend that was observed during the reference period, even in the absence of major perturbation events (Ioannidis et al., 2023; Levitt et al., 2023). For these reasons, we opted to not incorporate any trend, and instead simply use the age-specific average death rates over the 5-year reference period. This works particularly well in our case, as both the period of interest and the reference period are relatively short, and changes in life expectancy over the reference period were generally moderate in high-income countries (Aburto et al., 2022). However, if one would aim at estimating excess mortality over a longer period following the pandemic, accounting for (expected) changes in life expectancy would be recommendable. While certainly not straightforward, as it requires several further assumptions, such an adjustment could be attempted by adapting projection techniques (see e.g. Lee, 2000) while still keeping the estimators' variance in check by, e.g. incorporating cross-country time trends instead of country-specific ones.

One of the features of our method is that, in addition to point estimates, it also allows to produce excess mortality ranges, providing us with best- and worst-case mortality scenarios under conditions observed during the reference period. We stress that these are not classical confidence intervals, and as such they do not give us a probabilistic measure of uncertainty. As detailed in Section 2, however, calculating standard confidence intervals would require imposing unconvincing distributional assumptions on both the reference population and the mortality process, thus injecting a large amount of model-related uncertainty in the figures. We, therefore, opted for the data-driven, multiverse-style approach described, which is considerably more robust and allows us to make clear statements on whether or not mortality was substantially different in the period of interest than in the reference period.

Given the primary role of age in both overall and COVID-related mortality, incorporating it into the estimation in some way is essential to obtain unbiased estimates. With this article, we hope to make explicit age adjustment a standard practice for excess mortality estimation in cases for which age-stratified data is available. To this avail, we have publicly shared all data and code relevant to this study, to facilitate researchers in reproducing our findings as well as to enable them to utilize and build on our methods for future applications.

Acknowledgments

The authors would like to thank Juan Camilo Rosas Romero as well as all members of the COVID-19 Data Analysis Group (CODAG@LMU) for invaluable comments and fruitful discussions.

Conflicts of interest: The authors declare no competing interests.

Data Availability

We provide full replication code and materials, including data, in our GitHub repository, available at <https://github.com/gdenicola/excess-mortality-world>. All data used were gathered from publicly available sources, with details on the specific sources given in Section 3 of the manuscript, the [supplementary material](#), and the repository itself.

Supplementary material

[Supplementary material](#) is available online at *Journal of the Royal Statistical Society: Series A*.

References

- Aburto J. M., Schöley J., Kashnitsky I., Zhang L., Rahal C., Missov T. I., Mills M. C., Dowd J. B., & Kashyap R. (2022). Quantifying impacts of the COVID-19 pandemic through life-expectancy losses: A population-level study of 29 countries. *International Journal of Epidemiology*, 51(1), 63–74. <https://doi.org/10.1093/ije/dyab207>
- Acosta E. (2023). Global estimates of excess deaths from COVID-19. *Nature*, 613(7942), 31–33. <https://doi.org/10.1038/d41586-022-04138-w>
- Alderwick H. (2022). Is the NHS overwhelmed? *BMJ*, 376, o51. <https://doi.org/10.1136/bmj.o51>
- Barnes S. R., Beland L.-P., Huh J., & Kim D. (2022). COVID-19 lockdown and traffic accidents: Lessons from the pandemic. *Contemporary Economic Policy*, 40(2), 349–368. <https://doi.org/10.1111/coep.12562>
- Beaney T., Clarke J. M., Jain V., Golestaneh A. K., Lyons G., Salman D., & Majeed A. (2020). Excess mortality: The gold standard in measuring the impact of COVID-19 worldwide? *Journal of the Royal Society of Medicine*, 113(9), 329–334. <https://doi.org/10.1177/0141076820956802>
- Bergeron-Boucher M.-P., & Kjærgaard S. (2022). Mortality forecasting at age 65 and above: An age-specific evaluation of the Lee-Carter model. *Scandinavian Actuarial Journal*, 2022(1), 64–79. <https://doi.org/10.1080/03461238.2021.1928542>
- Bloom D. E., Canning D., & Lubet A. (2015). Global population aging: Facts, challenges, solutions & perspectives. *Daedalus*, 144(2), 80–92. https://doi.org/10.1162/DAED_a_00332
- Boka D. M., & Wainer H. (2020). How can we estimate the death toll from COVID-19? *Chance*, 33(3), 67–72. <https://doi.org/10.1080/09332480.2020.1787743>
- Bong C.-L., Brasher C., Chikumba E., McDougall R., Mellin-Olsen J., & Enright A. (2020). The COVID-19 pandemic: Effects on low-and middle-income countries. *Anesthesia and Analgesia*, 131(1), 86–92. <https://doi.org/10.1213/ANE.0000000000004846>

- Calderon-Anyosa R. J., & Kaufman J. S. (2021). Impact of COVID-19 lockdown policy on homicide, suicide, and motor vehicle deaths in Peru. *Preventive Medicine*, 143, 106331. <https://doi.org/10.1016/j.ypmed.2020.106331>
- Carballo A., Durban M., Kauermann G., & Lee D.-J. (2021). A general framework for prediction in penalized regression. *Statistical Modelling*, 21(4), 293–312. <https://doi.org/10.1177/1471082X19896867>
- Chatterjee P. (2020). Is India missing COVID-19 deaths? *The Lancet*, 396(10252), 657. [https://doi.org/10.1016/S0140-6736\(20\)31857-2](https://doi.org/10.1016/S0140-6736(20)31857-2)
- Checchi F., & Roberts L. (2005). HPN network paper 52: Interpreting and using mortality data in humanitarian emergencies: A primer for non-epidemiologists. <https://odihpn.org/publication/interpreting-and-using-mortality-data-in-humanitarian-emergencies/>.
- Dale W. (1772). *Calculations deduced from first principles, in the most familiar manner, by plain arithmetic, for the use of the societies instituted for the benefit of old age: Intended as an introduction to the study of the doctrine of annuities. By a member of one of the societies.* J. Ridley.
- De Nicola G. (2023). Github repository. Accessed May 30, 2023. <https://github.com/gdenicola/excess-mortality-world>.
- De Nicola G., Kauermann G., & Höhle M. (2022a). On assessing excess mortality in Germany during the COVID-19 pandemic. *AStA Wirtschafts-Und Sozialstatistisches Archiv*, 16(1), 5–20. <https://doi.org/10.1007/s11943-021-00297-w>
- De Nicola G., Schneble M., Kauermann G., & Berger U. (2022b). Regional now-and forecasting for data reported with delay: Toward surveillance of covid-19 infections. *AStA Advances in Statistical Analysis*, 106(3), 407–426. <https://doi.org/10.1007/s10182-021-00433-5>
- Destatis (2021). Press release No. 563 of 9 December 2021. Accessed May 30, 2023. https://www.destatis.de/DE/Presse/Pressemitteilungen/2021/12/PD21_563_12.html.
- Dorsett M. (2020). Point of no return: COVID-19 and the US healthcare system: An emergency physician's perspective. *Science Advances*, 6(26), eabc5354. <https://doi.org/10.1126/sciadv.abc5354>
- Dowd J. B., Andriano L., Brazel D. M., Rotondi V., Block P., Ding X., Liu Y., & Mills M. C. (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. *Proceedings of the National Academy of Sciences of the United States of America*, 117(18), 9696–9698. <https://doi.org/10.1073/pnas.2004911117>
- Eurostat (2023). Population change - demographic balance and crude rates at national level, 2023. Accessed May 30, 2023. https://ec.europa.eu/eurostat/databrowser/view/DEMO_GIND__custom_2089936.
- Gianicolo E. A., Russo A., Büchler B., Taylor K., Stang A., & Blettner M. (2021). Gender specific excess mortality in Italy during the COVID-19 pandemic accounting for age. *European Journal of Epidemiology*, 36(2), 213–218. <https://doi.org/10.1007/s10654-021-00717-9>
- Giattino C., Ritchie H., Roser M., Ortiz-Ospina E., & Hasell J. (2023). Excess mortality during the Coronavirus pandemic (COVID-19), 2023. Accessed May 30, 2023. <https://ourworldindata.org/excess-mortality-covid>.
- HMD (2023). Human Mortality Database. Max planck institute for demographic research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France), 2023. Accessed May 30, 2023. Available at www.mortality.org.
- Institute for Health Metrics and Evaluation (2022). COVID-19 Excess Mortality Estimates 2020–2021. Accessed May 30, 2023. https://ghdx.healthdata.org/record/ihme-data/covid_19_excess_mortality.
- Ioannidis J. P., Zonta F., & Levitt M. (2023). Flaws and uncertainties in pandemic global excess death calculations. *European Journal of Clinical Investigation*, 23(8), e14008. <https://doi.org/10.1111/eci.14008>
- Islam N., Shkolnikov V. M., Acosta R. J., Klimkin I., Kawachi I., Irizarry R. A., Alicandro G., Khunti K., Yates T., Jdanov D. A., White M., Lewington S., & Lacey B. (2021). Excess deaths associated with COVID-19 pandemic in 2020: Age and sex disaggregated time series analysis in 29 high income countries. *BMJ*, 373, n1137. <https://doi.org/10.1136/bmj.n1137>
- Johnson N. P., & Mueller J. (2002). Updating the accounts: Global mortality of the 1918–1920 “Spanish” influenza pandemic. *Bulletin of the History of Medicine*, 76(1), 105–115. <https://doi.org/10.1353/bhm.2002.0022>
- Karanikolos M., & McKee M. (2020). How comparable is COVID-19 mortality across countries? *Eurohealth*, 26(2), 45–50. <https://eurohealthobservatory.who.int/monitors/hstrm/analyses/hstrm/how-comparable-is-covid-19-mortality-across-countries>
- Karlinsky A., & Kobak D. (2021). Tracking excess mortality across countries during the COVID-19 pandemic with the world mortality dataset. *eLife*, 10, e69336. ISSN 2050-084X. <https://doi.org/10.7554/eLife.69336>
- Keiding N. (1987). The method of expected number of deaths, 1786–1886–1986. *International Statistical Review*, 55(1), 1–20. <https://doi.org/10.2307/1403267>
- Knutson V., Aleshin-Guendel S., Karlinsky A., Msemburi W., & Wakefield J. (2023). Estimating global and country-specific excess mortality during the COVID-19 pandemic. *The Annals of Applied Statistics*, 17(2), 1353–1374. <https://doi.org/10.1214/22-AOAS1673>

15. Estimating excess mortality in high-income countries during the COVID-19 pandemic

- Konstantinou G., Cameletti M., Gómez-Rubio V., Gómez I. L., Pirani M., Baio G., Larrauri A., Riou J., Egger M., Vineis P., & Blangiardo M. (2022). Regional excess mortality during the 2020 COVID-19 pandemic in five European countries. *Nature Communications*, 13(1), 482. <https://doi.org/10.1038/s41467-022-28157-3>
- Kung S., Doppin M., Black M., Hills T., & Kearns N. (2021). Reduced mortality in New Zealand during the COVID-19 pandemic. *The Lancet*, 397(10268), 25. [https://doi.org/10.1016/S0140-6736\(20\)32647-7](https://doi.org/10.1016/S0140-6736(20)32647-7)
- Lee R. (2000). The Lee-Carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal*, 4(1), 80–91. <https://doi.org/10.1080/10920277.2000.10595882>
- Leon D. A., Shkolnikov V. M., Smeeth L., Magnus P., Pechholdová M., & Jarvis C. I. (2020). COVID-19: A need for real-time monitoring of weekly excess deaths. *The Lancet*, 395(10234), e81. [https://doi.org/10.1016/S0140-6736\(20\)30933-8](https://doi.org/10.1016/S0140-6736(20)30933-8)
- Levin A. T., Hanage W. P., Owusu-Boaitey N., Cochran K. B., Walsh S. P., & Meyerowitz-Katz G. (2020). Assessing the age specificity of infection fatality rates for COVID-19: Systematic review, meta-analysis, and public policy implications. *European Journal of Epidemiology*, 35(12), 1123–1138. <https://doi.org/10.1007/s10654-020-00698-1>
- Levitt M., Zonta F., & Ioannidis J. P. (2022). Comparison of pandemic excess mortality in 2020–2021 across different empirical calculations. *Environmental Research*, 213, 113754. <https://doi.org/10.1016/j.envres.2022.113754>
- Levitt M., Zonta F., & Ioannidis J. P. (2023). Excess death estimates from multiverse analysis in 2009–2021. *European Journal of Epidemiology*, 38(11), 1129–1139. <https://doi.org/10.1007/s10654-023-00998-2>
- Lone S. A., & Ahmad A. (2020). COVID-19 pandemic – An African perspective. *Emerging Microbes & Infections*, 9(1), 1300–1308. <https://doi.org/10.1080/22221751.2020.1775132>
- Mazick A. (2007). Monitoring excess mortality for public health action: Potential for a future European network. *Eurosurveillance, Weekly Releases (1997–2007)*, 12(1), 3107. <https://doi.org/10.2807/esw.12.01.03107-en>
- Msemburi W., Karlinsky A., Knutson V., Aleshin-Guendel S., Chatterji S., & Wakefield J. (2023). The WHO estimates of excess mortality associated with the COVID-19 pandemic. *Nature*, 613(7942), 130–137. <https://doi.org/10.1038/s41586-022-05522-2>
- Nepomuceno M. R., Klimkin I., Jdanov D. A., Alustiza-Galarza A., & Shkolnikov V. M. (2022). Sensitivity analysis of excess mortality due to the COVID-19 pandemic. *Population and Development Review*, 48(2), 279–302. <https://doi.org/10.1111/padr.12475>
- Olsen S. J., Azziz-Baumgartner E., Budd A. P., Brammer L., Sullivan S., Pineda R. F., Cohen C., & Fry A. M. (2020). Decreased influenza activity during the COVID-19 pandemic—United States, Australia, Chile, and South Africa, 2020. *American Journal of Transplantation*, 20(12), 3681–3685. <https://doi.org/10.1111/ajt.16381>
- Price R. (1771). *Observations on reversionary payments; on schemes for providing annuities for widows, and for persons in old age; on the method of calculating the values of assurances on lives, and on the national debt*. Cadell.
- Raths J. (1909). Die Sterblichkeitsmessung in der allgemeinen Bevölkerung. In *Denkschriften und Verhandlungen des 6. Internationalen Kongressess für Versicherungswissenschaften* (pp. 115–129).
- Rocklöv J., & Sjödin H. (2020). High population densities catalyse the spread of COVID-19. *Journal of Travel Medicine*, 27(3), taaa038. <https://doi.org/10.1093/jtm/taaa038>
- Schneble M., De Nicola G., Kauermann G., & Berger U. (2021). Nowcasting fatal COVID-19 infections on a regional level in Germany. *Biometrical Journal*, 63(3), 471–489. <https://doi.org/10.1002/bimj.202000143>
- Schöley J., Aburto J. M., Kashnitsky I., Kniffka M. S., Zhang L., Jaadla H., Dowd J. B., & Kashyap R. (2022). Life expectancy changes since COVID-19. *Nature Human Behaviour*, 6(12), 1649–1659. <https://doi.org/10.1038/s41562-022-01450-3>
- Senni M. (2020). COVID-19 experience in Bergamo, Italy. *European Heart Journal*, 41(19), 1783–1784. <https://doi.org/10.1093/eurheartj/ehaa279>
- Simonsen L., Spreeuwenberg P., Lustig R., Taylor R. J., Fleming D. M., Kroneman M., Van Kerkhove M. D., Mounst A. W., Paget W. J., & Teams G. C. (2013). Global mortality estimates for the 2009 influenza pandemic from the GLaMOR project: A modeling study. *PLoS Medicine*, 10(11), e1001558. <https://doi.org/10.1371/journal.pmed.1001558>
- Stang A., Standl F., Kowall B., Brune B., Böttcher J., Brinkmann M., Dittmer U., & Jöckel K.-H. (2020). Excess mortality due to COVID-19 in Germany. *Journal of Infection*, 81(5), 797–801. <https://doi.org/10.1016/j.jinf.2020.09.012>
- Suthar A. B., Wang J., Seffren V., Wiegand R. E., Griffing S., & Zell E. (2022). Public health impact of COVID-19 vaccines in the US: Observational study. *BMJ*, 377, e069317. <https://doi.org/10.1136/bmj-2021-069317>
- The Economist (2023). Tracking COVID-19 excess deaths. Accessed May 5, 2023. <https://www.economist.com/graphic-detail/coronavirus-excess-deaths-tracker>
- Van Noorden R. (2022). COVID death tolls: Scientists acknowledge errors in WHO estimates. *Nature*, 606(7913), 242–244. <https://doi.org/10.1038/d41586-022-01526-0>

- Wang H., Paulson K. R., Pease S. A., Watson S., Comfort H., Zheng P., Aravkin A. Y., Bisignano C., Barber R. M., Alam T., & Fuller J. E. (2022). Estimating excess mortality due to the COVID-19 pandemic: A systematic analysis of COVID-19-related mortality, 2020–21. *The Lancet*, 399(10334), 1513–1536. [https://doi.org/10.1016/S0140-6736\(21\)02796-3](https://doi.org/10.1016/S0140-6736(21)02796-3)
- Wilmoth J., Andreev K., Jdanov D., Gleit D., & Riffe T. (2021). Methods protocol for the human mortality database (Version 6). Accessed May 30, 2023. Available at www.mortality.org/File/GetDocument/Public/Docs/MethodsProtocolV6.pdf.
- Woolf S. H., Chapman D. A., Sabo R. T., & Zimmerman E. B. (2021). Excess deaths from COVID-19 and other causes in the US, March 1, 2020, to January 2, 2021. *Jama*, 325(17), 1786–1789. <https://doi.org/10.1001/jama.2021.5199>
- World Bank (2022). New World Bank country classifications by income level: 2022–2023. Accessed May 30, 2023. Available at blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2022-2023.
- World Health Organization (2022). Global excess deaths associated with COVID-19, January 2020 - December 2021, Accessed May 30, 2023. <https://www.who.int/data/stories/global-excess-deaths-associated-with-covid-19-january-2020-december-2021>.

Contributing Publications

- De Nicola, G., Sischka, B., and Kauermann, G. (2022). Mixture models and networks: The stochastic blockmodel. *Statistical Modelling*, 22(1-2):67–94. <https://doi.org/10.1177/1471082X211033169>.
- De Nicola, G., Fritz, C., Mehrl, M., and Kauermann, G. (2023). Dependence matters: Statistical models to identify the drivers of tie formation in economic networks. *Journal of Economic Behavior & Organization*, 215:351–363. <https://doi.org/10.1016/j.jebo.2023.09.021>.
- Fritz, C., De Nicola, G., Kevork, S., Harhoff, D., and Kauermann, G. (2023). Modelling the large and dynamically growing bipartite network of German patents and inventors. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(3):557–576. <https://doi.org/10.1093/jrssa/qnad009>.
- De Nicola, G., Tuekam Mambou, V.H., and Kauermann, G. (2023). COVID-19 and social media: Beyond polarization. *PNAS Nexus*, 2(8):pgad246. <https://doi.org/10.1093/pnasnexus/pgad246>.
- Schneble, M., De Nicola, G., Kauermann, G., and Berger, U. (2021). Nowcasting fatal COVID-19 infections on a regional level in Germany. *Biometrical Journal*, 63(3):471–489. <https://doi.org/10.1002/bimj.202000143>.
- De Nicola, G., Schneble, M., Kauermann, G., and Berger, U. (2022). Regional now- and forecasting for data reported with delay: Towards surveillance of COVID-19 infections. *AStA Advances in Statistical Analysis*, 106:407–426. <https://doi.org/10.1007/s10182-021-00433-5>.
- Schneble, M., De Nicola, G., Kauermann, G., and Berger, U. (2021). A statistical model for the dynamics of COVID-19 infections and their case detection ratio in 2020. *Biometrical Journal*, 63(8):1623–1632. <https://doi.org/10.1002/bimj.202100125>.
- Fritz, C., De Nicola, G., Rave, M., Weigert, M., Khazaei, Y., Berger, U., Küchenhoff, H. and Kauermann, G. (2022). Statistical modelling of COVID-19 data: Putting generalized additive models to work. *Statistical Modelling (OnlineFirst)*. <https://doi.org/10.1177/1471082X221124628>.
- De Nicola, G., Kauermann, G., and Höhle, M. (2022). On assessing excess mortality in Germany during the COVID-19 pandemic. *AStA Wirtschafts-und Sozialstatistisches Archiv*, 16:5–20. <https://doi.org/10.1007/s11943-021-00297-w>.
- De Nicola, G., and Kauermann, G. (2022). An update on excess mortality in the second year of the COVID-19 pandemic in Germany. *AStA Wirtschafts-und Sozialstatistisches Archiv*, 16:21–24. <https://doi.org/10.1007/s11943-022-00303-9>.

De Nicola, G., and Kauermann, G. (2024). Estimating excess mortality in high-income countries during the COVID-19 pandemic. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *qnae031*. <https://doi.org/10.1093/jrsssa/qnae031>.

Eidesstattliche Versicherung (Affidavit)

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

München, den 25.05.2024

Giacomo De Nicola

