Martje Rave

Modelling ICU Occupancy and Patient Flow Dynamics during the COVID-19 Pandemic

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

Eingereicht am 22.09.2025

Martje Rave

Modelling ICU Occupancy and Patient Flow Dynamics during the COVID-19 Pandemic

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

Eingereicht am 22.09.2025

Erster Berichterstatter: Prof. Dr. Göran Kauermann (LMU München)
Zweiter Berichterstatter: Univ.-Prof. Dipl.-Math. Dr. Arne Bathke (Universität Salzburg)
Dritter Berichterstatter: Prof. Dr. Volker Schmid (LMU München)

Tag der Disputation: 12.12.2025

Acknowledgments

I am deeply grateful to my supervisor, Göran Kauermann, who guided me not only through my academic career but also through a difficult period in my personal life. You gave me the space to grow, the time to heal, and the encouragement I needed throughout. The Chair you lead has become a wonderfully diverse group of exceptionally kind, funny, and motivated individuals. Thank you all for accompanying me on this journey and for making it so special.

I would also like to thank Anne-Laure Boulesteix for her continued guidance through the mentoring programme, and Brigitte, Martina, and Elke for their patience, organisational support, and the joy their canine companions brought to the office.

My sincere thanks go to all my colleagues at the institute. From sport to coffee, it has been a privilege to be part of such a supportive environment. Special thanks to Nurzhan Sapargali and Hannah Blocher for our reading group — your feedback has been invaluable. I am also grateful to Daniel Racek and Oleg Vlasovets for our chats. To my office mates — Ben Sischka, Giacomo De Nicola, and Jan Anders — thank you for the friendship, the conversations, and the camaraderie.

I am indebted to the professors, postdocs, fellow PhD candidates, student assistants, and students with whom I had the pleasure of teaching. Your insight and trust made collaboration especially rewarding. I am particularly thankful to Helmut Küchenhoff and the StaBLab, whose collective commitment and teamwork are truly remarkable. Special gratitude also goes to Volker Schmid, whose calm and organised manner made teaching together such a delight.

I am grateful to Arne Bathke for kindly agreeing to serve as external reviewer on the examination committee.

I also thank Alexander Bauer for the cake we baked and the music we shared, and Edoardo Mosca, Sophia Althammer, and Jan Simson for our regular trips to Nordbad.

Further, I would like to acknowledge the linguistic support provided by ChatGPT.

Finally, I owe the deepest gratitude to my family — my mother, my father, and my brothers. Though my career choice has often seemed baffling to you, you have nurtured my passion for learning with your patience and debates throughout. Mor, I wish you could have been a part of this all the way. Hagen and Resa, you have become the most important pillars of my life, and I am profoundly thankful for your unconditional love and devotion.

Summary

In the wake of COVID-19, the world faced the mammoth task of understanding how this new disease functioned, its impact on humans, and, in particular, its implications for public health. The contributions of this thesis are part of the broader collaborative effort of understanding the COVID-19 pandemic by analysing the intensive care unit (ICU) dynamics in Germany. Since only data on ICU occupancy are publicly available at district level, and not on ICU admission or patient length of stay, ICU patient flow dynamics can not directly be analysed. This thus poses a missing data problem. The contributions of this thesis analyse the ICU occupancy and subsequently employ statistical methods in order to disentangle the unobserved patient inflow, length of stay and outflow from the observed occupancy data.

Part I of this thesis introduces metrics commonly employed to gain a comprehensive understanding of the impact of an infectious disease on public health, in the context of COVID-19. It also discusses the data available for public health surveillance in Germany and highlights the contextual motivation for the contributions of this thesis, namely the absence of district-level data on ICU patient flow dynamics. Chapter 3 introduces statistical modelling, with a particular focus on parametric models. It lays out the assumptions required for consistent estimation and discusses approaches to quantifying estimator uncertainty. Chapter 4 then links the statistical methodology to the previously introduced data, motivating the contributions of this thesis.

Part II analyses the ICU occupancy. The distribution of ICU beds among patients infected with COVID-19, patients not infected with COVID-19, and unoccupied status is modelled using a multinomial model. This approach allows estimation of associations between ICU bed distribution and infection rates in the previous week, spatial correlation (captured through a two-dimensional thin-plate spline), and district-level heterogeneity (modelled via a random intercept). The multinomial assumption is particularly valuable for prediction, as it reflects the mutually exclusive nature of ICU bed allocation.

Part III builds on the analysis of Part II by inferring patient inflow and outflow from ICU occupancy. Chapter 6 employs the stochastic Expectation-Maximisation (sEM) algorithm to iteratively simulate from a truncated Skellam distribution with incoming and outgoing intensity parameters, which are estimated using two independent Poisson models. The inflow model incorporates age-specific infection rates from the previous week, spatial correlation via a two-dimensional thin-plate spline, temporal correlation via a penalised B-spline over time, and categorical weekday effects. The outflow model is specified analogously, with an additional offset defined as the sum of previously incoming patients, weighted by the exit rate, which is taken to be fixed. Chapter 7 extends this methodology by treating exit rate, also referred to as length of stay, as a random parameter rather than a fixed input. Thus, the outflow model of the M-Step of the sEM is altered, such that the outflow is now solely a linear combination of inflow. This introduces a sum-to-one and non-negativity constraint. Thus, the parameters are estimated via constrained maximum likelihood, which introduces bias. This bias is corrected employing an additional simulation step in the sEM.

Zusammenfassung

Im Anfang COVID-19s stand die Welt vor der gewaltigen Aufgabe zu verstehen, wie diese neue Erkrankung funktioniert, welche Auswirkungen sie auf den Menschen hat und insbesondere, welche Implikationen sich für die öffentliche Gesundheit ergeben. Die Beiträge dieser Dissertation sind Teil des breiteren kooperativen Bemühens, die COVID-19-Pandemie durch die Analyse der Dynamik der Intensivstationen (Intensive Care Units, ICU) in Deutschland zu verstehen. Da auf Kreisebene lediglich Daten zur Auslastung der Intensivstationen öffentlich verfügbar sind, nicht jedoch zu Aufnahmen oder Verweildauern, können die Patientenflussdynamiken auf Intensivstationen nicht unmittelbar analysiert werden. Dies stellt somit ein Problem fehlender Daten dar. Die Beiträge dieser Arbeit analysieren die ICU-Auslastung und setzen anschließend statistische Methoden ein, um aus den beobachteten Auslastungsdaten die unbeobachteten Größen Zufluss, Verweildauer und Abfluss zu identifizieren.

Teil I dieser Arbeit führt einige der im Kontext von COVID-19 gebräuchlichen Metriken ein, mit denen die Auswirkungen einer Infektionskrankheit auf die öffentliche Gesundheit umfassend bewertet werden. Zudem werden die in Deutschland für die Gesundheitsüberwachung verfügbaren Daten diskutiert und die kontextuelle Motivation für die Beiträge dieser Arbeit herausgearbeitet, namentlich das Fehlen Daten zu Intensivpatientenflussdynamiken auf Kreisebene. Kapitel 3 führt in die statistische Modellierung ein, mit besonderem Fokus auf parametrische Modelle. Es legt die für konsistente Schätzung erforderlichen Annahmen dar und diskutiert Ansätze zur Quantifizierung der Schätzunsicherheit. Kapitel 4 verknüpft sodann die statistische Methodik mit den zuvor eingeführten Daten und motiviert die Beiträge dieser Dissertation.

Teil II analysiert die Auslastung der Intensivstationen. Die Verteilung der Intensivbetten auf COVID-19-infizierte Patientinnen und Patienten, nicht infizierte Patientinnen und Patienten sowie unbesetzte Betten wird durch ein multinomiales Modell beschrieben. Dieser Ansatz ermöglicht die Schätzung von Assoziationen zwischen der Bettenverteilung und den Infektionsraten der Vorwoche, räumlicher Korrelation (abgebildet durch einen zweidimensionalen Dünnplattenspline) sowie Heterogenität auf Kreisebene (modelliert über einen Random Intercept). Die Multinomialannahme ist für Prognosen besonders wertvoll, da sie die wechselseitig ausschließende Natur der Bettenallokation aufgegriffen wird.

Teil III baut auf der Analyse aus Teil II auf, indem aus der ICU-Auslastung auf Zu- und Abflüsse geschlossen wird. Kapitel 6 verwendet den stochastischen Expectation-Maximisation-Algorithmus (sEM), um iterativ aus einer trunkierten Skellam-Verteilung mit Zu- und Abfluss-Intensitätsparametern zu simulieren, die mithilfe zweier unabhängiger Poisson-Modelle geschätzt werden. Das Zuflussmodell umfasst alterspezifische Infektionsraten der Vorwoche, räumliche Korrelation über einen zweidimensionalen Thin-plate spline, sowie eine temporale Korrelation durch eine penalisierte B-spline, und einen kategorialen Wochentagseffekt. Das Abflussmodell wird analog spezifiziert, mit einem zusätzlichen Offset, definiert als Summe der zuvor eingeströmten Patientinnen und Patienten, gewichtet mit der als fix angenommenen Austrittsrate. Kapitel 7 erweitert diese Methodik, indem die Austrittsrate bzw. Verweildauer als zufälliger Parameter statt als fixe Eingabe behandelt wird. Der M-Schritt des sEM wird entsprechend so angepasst, dass der Abfluss nun ausschließlich als lineare Kombination des Zuflusses modelliert wird. Dies erfolgt unter Nebenbedingungen, da alle aufgenommenen Patientinnen und Patienten die Intensivstation auch wieder verlassen müssen und kein Zufluss einen negativen Effekt auf den Abfluss haben darf. Die

Parameterschätzung erfolgt daher mittels einer Constrained Maximum-Likelihood, was zu Verzerrungen führt. Diese Verzerrung wird mittels eines iterativen zusätzlichen Simulationsschritt korrigiert.

Contents

Part I.

Introduction and background

1. Introduction

> "A judicious man looks at statistics not to get knowledge, but to save himself from having ignorance foisted upon him."
>
> — Thomas Carlyle, 1840

The fields of statistics and epidemiology have long shared a symbiotic relationship. Statistics has gained much of its public relevance through applications in epidemiology, while epidemiological challenges have in turn motivated advances in statistical methodology. Although many contemporary epidemiologists, such as Frérot et al. (2018), argue that the field of epidemiology extends well beyond the realm of statistical analysis, it remains the case that, historical pioneers of epidemiology are equally regarded as pioneers of statistics.

For example, the systematic work of John Graunt, who recorded mortality data in seventeenth-century London (Stigler, 1986) and John Snow, who traced the 1854 cholera outbreak in Soho, London, back to a contaminated well (Anderson, 2018) are fundamental contributions both to epidemiology and statistics, and exemplifies the value of data in understanding public health. Despite the immense medical and technological progress achieved since, the outbreak of COVID-19 in late 2019 revealed how the core challenges which Graunt and Snow faced continue to persist, i.e. to describe and predict the spread of a disease, and to evaluate its impact on public health.

First reports of a novel pneumonia-like disease in Wuhan, China, were soon attributed to a new coronavirus strain, later named 'SARS-CoV-2' (Li et al., 2020). A notable characteristic of COVID-19, the disease caused by SARS-CoV-2, is its infectiousness (Peiris et al., 2004). Though first cases were only recorded in late 2019, COVID-19 had already been classified a global pandemic in early 2020 (Gallagher, 2020). In March, 2020, particularly affected regions, such as the Lombardy region in Italy, faced an extreme strain on their health care system which resulted in some hospitals having to resort to triage (Fagoni et al., 2020). Evidently, the impact of COVID-19 on public health and the health care system has been extremely severe. More than five years later, the World Health Organization reports over 777 million confirmed cases and approximately 7 million deaths attributed to COVID-19 worldwide (World Health Organization, 2025).

In an early response to the COVID-19 outbreak, health organisations, medical professionals, and researchers from a wide range of disciplines mobilised globally to provide some data-driven insight into the COVID-19 pandemic. One such initiative was the Covid-19 Data Analysis Group (CoDAG) at the Ludwig-Maximilians-Universität München (2020), which sought to deepen the understanding of COVID-19's impact on public health in Germany. CoDAG's work focused particularly on infection dynamics, hospitalisations, intensive care and mortality, and discussed the information needed to gain a holistic picture of COVID-19's impact on Germany's public health. For example, Fritz et al. (2023) give some commentary on the learnings from the COVID-19 pandemic by CoDAG. Where necessary, CoDAG would strive to develop statistical methodology to bridge the gap between the information needed to gain a complete picture of COVID-19 and the data available. Additionally, Jahn et al. (2022) and a subsequent discussion by Berger et al. (2022), or Spiegelhalter and Masters (2021) elaborate on the statistical modelling and reporting of COVID-19 beyond CoDAG.

The contributions of this thesis are part of CoDAG's collaborative effort. Specifically, this thesis is concerned with intensive care unit (ICU) occupancy and dynamics during the COVID-19 pandemic in Germany. The ICU occupancy is recorded and the number of ICU beds occupied by patients infected with COVID-19, the number of beds occupied by patients not infected with COVID-19 and number of unoccupied beds are published daily, aggregated over each district ('Landkreis' in German) in Germany. Though data on the occupancy are publicly available on district level, data on ICU patient inflow and outflow are not. Thus, as ICU patient dynamics are important to better understand the severity of COVID-19, statistical methodology is employed to disentangle inflow and outflow of ICU patients from the data available, i.e. ICU occupancy data. This thesis includes both analyses on the ICU occupancy and introduces methodology to disentangle the occupancy into inflow, length of stay and outflow. The thesis is structured as follows.

In **Part I**, Chapter 2 introduces commonly employed key metrics for comprehensively monitoring COVID-19, and discusses the publicly available data on COVID-19 in Germany. This Chapter aims to provide a holistic overview of public health surveillance with respect to COVID-19, with discussions on data available on infection, mortality, hospitalisation and intensive care in Germany. Chapter 3 presents the introduction into the statistical methodology underlying particularly the contributions of this thesis. It commences with an overview of parametric statistical models (Section 3.1), followed by approaches to parameter estimation (Section 3.2), and concluding with methods for quantifying uncertainty (Section 3.3). Chapter 4 then specifies methodology employed in the contributions of this thesis, tying in Chapter 2 and 3 to motivate the contributions and an outlook on directions for future research.

Part II comprises the first contribution, in Chapter 5, and focuses on the association between COVID-19 infection rates and the distribution of ICU occupancy in Germany. A central feature of this analysis is the assumption of a multinomial distribution when modelling the allocation of ICU beds to patients with COVID-19, patients without COVID-19, or unoccupied beds. This reflects the mutually exclusive nature of bed allocation. Specifically, the distribution of ICU beds is estimated in association with the infection rates of the 35–59, 60–79, and 80+ year-old age groups, as well as the previous week's occupancy, incorporated as an autoregressive covariate. Spatial correlation is modelled via a two-dimensional thin-plate spline across district centroids, and a random intercept is included to capture heterogeneous district effects. To account for dependencies in the data—specifically the incomplete reallocation of beds between observations—the sandwich estimator is employed to estimate variance.

In **Part III**, the ICU occupancy is taken to be a function of patient inflow, the length of stay, and outflow. Analysing occupancy alone may omit important information on COVID-19's pathogenesis. Yet, in Germany, data are not available on COVID-19 patient inflow, lengths of stay, or outflow on district level. If, for example, a constant number of beds are observed to be occupied by COVID-19 patients in a given district on two subsequent days, then between each observation either no patient may have been admitted and none have been discharged, or one may have been admitted and one discharged, and so forth. Thus, the difference in occupancy between two observations is equivalent to the difference in the number of incoming patients and the number of outgoing patients.

The second contribution, in Chapter 6, employs therefore the stochastic EM (sEM) algorithm to iteratively simulate ICU patient inflow and outflow from a truncated Skellam distribution in the E-step and to estimate the corresponding intensity parameters through two independent Poisson models in the M-step. The inflow model incorporates infection rates, spatial correlation via a two-dimensional thin-plate spline, temporal correlation through a penalised B-spline, and a weekday dummy effect. The outflow model is defined analogously and additionally includes an offset equal to the weighted sum of previously estimated inflow, weighted by the average probability of length of stay, taken from Tolksdorf et al. (2020). The variance is calculated using Rubin's rule.

Chapter 7 further extends this approach by treating the length of stay as a random parameter rather than fixed. This transforms the outflow model of the M-step from a standard Poisson model into a Poisson model where the length-of-stay parameters are contextually constrained to be non-negative and to sum to one, as all incoming patients must eventually leave the ICU and no inflow can have a negative effect on the outflow. Likelihood maximisation in this constrained parameter space produces biased results, which are corrected through an additional simulation step. The resulting methodology thus enables estimation in settings where only total net counts are observed, but where researchers are interested in the underlying inflow, outflow, and length-of-stay dynamics and is thus applicable to a plethora of data situations in which a total net count is periodically observed and the underlying flow dynamics are of interest.

2. Public health surveillance and COVID-19

Comprehensive reporting on the impact of an infectious disease on public health, such as by Spiegelhalter and Masters (2021) in the case of COVID-19, are typically comprised of both analyses on spread and severity. Assessing both aspects of an infectious disease through appropriate metrics is essential to allow for a holistic understanding of the disease. This chapter discusses some key metrics commonly employed in epidemiology to quantify both the spread and the severity of an infectious disease, such as COVID-19.

This chapter additionally elaborates on the publicly available data on COVID-19 in Germany and discusses the information necessary to calculate some of the introduced metrics necessary for measuring infection, mortality, hospitalisation and intensive care. This discussion provides the contextual motivation for the contributions of this thesis, which focuses on ICU occupancy and develop methodology to bridge the gap between available data and the metrics needed to understand COVID-19.

While the quantification of spread, or infectiousness, is arguably straight forward, usually being some function of either the number of new infections or viral load, the quantification of severity is a little more nuanced. Kelley and Bollens-Lund (2018) define a severe illness as a disease that poses a high risk of mortality or significantly impairs a patients' ability to manage daily life as they could prior to contracting the illness. By this definition, the severity of COVID-19 may be evaluated through several complementary factors. In a meta-analysis of clinical studies, Yuan et al. (2023) categorise the severity according to several outcome measures, which can broadly be summarised to metrics on mortality, hospitalisation, and intensive care. The Centers for Disease Control and Prevention (2025) also include the need for mechanical ventilation as an additional category of severity, which in most treatment centres is administered in the ICU. In addition, infection-associated chronic conditions such as long-COVID have emerged as an important dimension of severity Ely et al. (2024). Nonetheless, for the context of this thesis, metrics on mortality, hospitalisation, and intensive care are introduced, solely.

The validity of the aforementioned metrics naturally rely on high quality, transparent and available data. Systems for collection, storage, and publication of COVID-19 vary considerably between and even within countries. In Germany, the Robert Koch-Institut (2025e) (RKI) is responsible for centrally gathering, cleaning, processing, and publishing data on COVID-19 infections, mortality due to COVID-19, and hospitalisations of patients infected with COVID-19. Locally, the data collection of COVID-19 infections, hospitalisations and mortality is conducted by local health authorities. Local health authorities coincide with districts in Germany and collect the data from testing stations, doctors and hospitals. By calendar week 45 2020, 200 laboratories tested and reported test results on COVID-19 to doctors or patients and to the responsible local health authorities, (Willrich et al., 2021; Gross, 2024). The health authorities pass the number of recorded infections onto the RKI, as per the Infection Protection Act (IfSG) (Robert Koch-Institut, 2025c).

Data on ICU occupancy is also published by the RKI but collected by the Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin (2025) (DIVI), as an extension to a reporting system already established pre-pandemic by the German 'ARDS-Netzwerk', the German network for acute respiratory distress syndrome care (Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin, 2025). Treatment centres (usually hospitals), are liable to send on their ICU occupancy distribution to the DIVI by 12:00 (UTC+1:00) every day (Robert Koch-Institut, 2025d). The DIVI then publishes data on three levels of varying granularity, namely on district level, on county level ('Bundesland' in German) and on country level.

2.1. Infection

On its most rudimentary level, the number of new infections is a simple indicator on a diseases spread. However, arguably regions with a larger population, would likely exhibit a larger number of infections. Thus, the number of infections are commonly relativised by the size of the corresponding population. Though, if the number of infections are relatively low, compared to the population size, it is more conceptually intuitive to inspect the incidence, i.e. the infection rate per 100,000 inhabitants, defined by

$$\iota = \frac{\text{Total number of new COVID-19 cases}}{\text{Total population}} \times 100{,}000. \tag{2.1}$$

In Germany, non-pharmaceutical interventions, such as lockdowns and curfews, were directly tied to the 7-day incidence rate (Wunderlich, 2020), which is a common extension to (2.1) in which the total number of new COVID-19 cases is summed over seven days, mitigating daily fluctuation.

Naturally, the recording and registration of infections is subject to delay for systemic and random reasons. However, in an ongoing global pandemic, it is necessary having reliable real-time data. Therefore, nowcasting methods were employed. Günther et al. (2021), for example, develop nowcasting methods for COVID-19 infections in Bavaria, using previously observed dynamics in reporting delay, they estimate the number of infections which have not yet been reported at the given time point. De Nicola et al. (2022) give an intuitive elaboration on the delay of data registration of COVID-19 infections. Wolffram et al. (2023) compare different nowcasting methods in assessing hospitalised individuals infected with COVID-19. Schneble et al. (2021) further extend the nowcasting to reporting delay in fatalities due to COVID-19. In the first contribution of this thesis, Chapter 5, Maximilian Weigert applies nowcasting to the hospitalisations in Bavaria and shows its improved performance in forecasting the development of a COVID-19 pandemic.

A further shortcoming in all diagnostic data which relies on tests, is that the total infection counts are not only a function of prevalence of the disease but also of testing strategy, sensitivity and specificity of the respective tests used. As a simplified example, Figure 2.1, is presented. For fixed 20% of the population being tested, the probability of being unobserved and sick is plotted in dashed-turquoise over different values of prevalence. The probability of being falsely detected by being tested positive but healthy is plotted in solid blue. For low prevalence, we observe a higher probability of being falsely recorded than falsely unobserved, resulting in a higher probability of over-counting. For larger prevalence, the probability of being unobserved and sick increases, while the probability of being falsely recorded as sick decreases, eventually leading to a probable

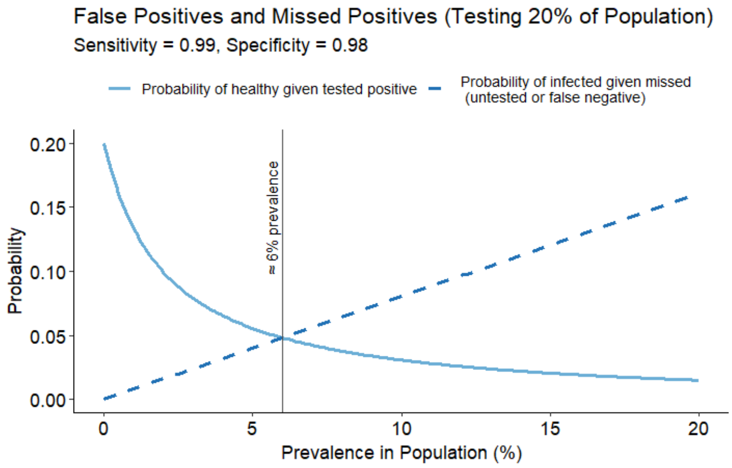


Figure 2.1.: Unobserved infected and falsely recorded cases over prevalence.

under-counting of cases. A similar result was shown empirically by Healy et al. (2021). In this simple example, this point is reached at a prevalence of 6%.

This further evokes the topic of systematically unreported cases. It may be that there has been a non-negligible number of undetected cases. Fiedler et al. (2021), for example, investigate the undetected spread of infection in a comparison between Italy and Germany through mathematical models.

Another key metric is the basic reproduction number, R_0, defined as the expected number of secondary infections caused by a single infected individual in a wholly susceptible population (Diekmann et al., 1990). A meta-analysis conducted in March 2020 estimated the mean R_0 of COVID-19 at 3.38, with a 95% confidence interval of (2.81, 3.82) (Alimohamadi et al., 2020). Over time, however, the proportion of susceptible individuals in a population changes due to interventions or immunity (acquired through infection or vaccination). Consequently, the effective reproduction number, R_t, became the more relevant measure, as it incorporates the proportion of immune individuals at a given time (Spiegelhalter and Masters, 2021).

A further perspective on infectiousness is provided by the viral load of infected individuals. Meyerowitz et al. (2021) estimated in 2021 that viral load peaks one to two days prior to symptom onset. This is of particular concern, since individuals are presumed to be more infectious when viral load is high, but may transmit the virus unknowingly while presymptomatic. Puhach et al. (2023) later find that the timing and magnitude of viral load vary depending on the variant with which an individual is infected.

2.2. Mortality

A key measure for assessing the mortality due to an infectious disease, such as COVID-19, the standardised mortality ratio (SMR), defined by

$$\text{SMR} = \frac{\text{Total number of observed events}}{\text{Expected number of events}}. \tag{2.2}$$

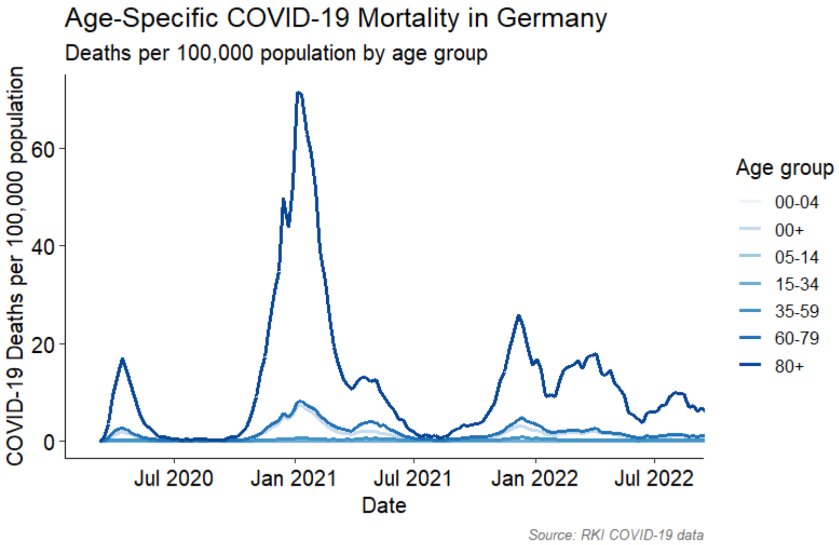


Figure 2.2.: Mortality rate per 100.000 inhabitants by age group in Germany from March 2020 to August 2022.

Farr (1852) introduced the SMR, which has since been widely accepted as a key metric for assessing incidences attributed to an adverse event (Armstrong, 1995). In terms of mortality, it is the number of deaths observed during an adverse event relative to the number of deaths expected in its absence.

However, the estimation of the counterfactual baseline, i.e. the expected number of deaths, is inherently uncertain and methodologically debated, particularly in outbreaks of novel diseases. In the COVID-19 pandemic, most analyses stress the importance of differentiating by age groups, while De Nicola et al. (2022) further argue in favour of accounting for demographic changes in order to produce appropriate baseline estimates. They thereby obtain more moderate estimates than earlier studies, such as Vestergaard et al. (2020).

Alternative measures include years of life lost, which quantify the difference between age at death and expected remaining life years, introduced conceptually by Gardner and Sanborn (1990). This measure has been used to place the severity of COVID-19 in comparative perspective with other causes of death, such as influenza or cardiovascular disease (Pifarré i Arolas et al., 2021).

For context, Figure 2.2 shows the recorded number of deceased infected with COVID-19 at time of death per 100.000 inhabitants given their recorded age group, calculated analogously to (2.1), as reported by the Robert Koch-Institut (2025b). In Germany, there are two data bases which comprise data on deaths due to COVID-19. The first being the Robert Koch-Institut (2025f) (RKI) and the other being collected and published by the Statistisches Bundesamt (2025) (Destatis). The RKI data set includes both patients who died 'with' and 'due to' COVID-19, without differentiating them, while the Destatis dataset includes only data on the people who died due to COVID-19. Additionally, the Destatis dataset differentiate between deaths which are proven to be COVID-19 cases and deaths which are suspected COVID-19 cases. In any case, Wollschläger et al. (2024) argues that assessing the number of deaths registered due to COVID-19 can be susceptible to bias, due to likely systematic differences in recording the cause of death across Germany. The discussion on the causality of adverse health events in patients during the COVID-19 pandemic also extends to data on hospitalisation and intensive care.

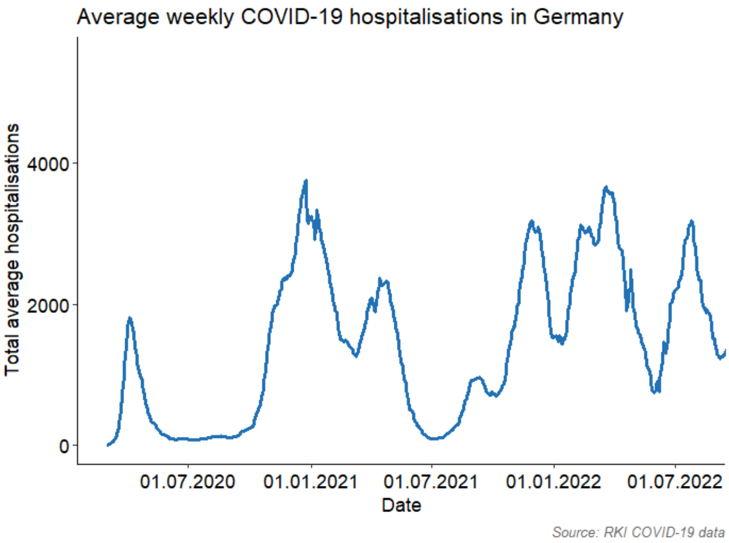
2.3. Hospitalisations

The key measure for hospitalisations is the hospitalisation rate per 100,000 inhabitants, analogously calculated to (2.1). This measure is more straightforward than the SMR but not without limitations, when assessing severity. For instance, some hospitalised patients tested positive for COVID-19 despite being admitted for unrelated reasons. Some studies account for comorbidities in assessing hospitalisation risk Mattey-Mora et al. (2022), though such adjustments are usually undertaken in clinical-level research or meta-analyses.

The previous discussion on fatality being 'due to' a COVID-19 infection or 'with' a COVID-19 infection also extends to the data on hospitalisations. In Germany, there are no publicly available health surveillance data which clearly differentiate between patients who are admitted due to COVID-19, directly or indirectly, or due to another cause and happened to be infected with COVID-19. While it is important to understand the strain on the health care system which infected individuals impose, regardless of the cause of their hospital admittance, it is equally important to understand the severity of COVID-19. Thus, one should ideally differentiate between hospitalisations due to COVID-19 and hospitalisations due to other causes.

In the beginning of the COVID-19 pandemic the conjectural belief of the physicians consulting CoDAG, was that the department to which the patient was admitted would inform on the cause of admission. For example, patients admitted to the ICU were thought to have been admitted 'due to' COVID-19. However, contrary to this initial intuition, Strobl et al. (2024) show that there is no significant difference in the patients admitted to the ICU or another hospital department, in terms of whether patients are admitted 'due to' or 'with' COVID-19. In this regard, the only significant difference was estimated between the ICU admittance and the surgical ward. Taking into account the testing strategy for non urgent surgical procedures, this finding is intuitive. However, this analysis comprises clinical level data, not publicly available.

Figure 2.3 shows the total hospitalisations of 7-day-averages across Germany of patients which were hospitalized and infected with COVID-19. The data are provided by the Robert Koch-Institut (2025a).



Source: RKI COVID-19 data

Figure 2.3.: Total 7-day-average hospitalisations.

2.4. Intensive care

Admission to intensive care represents an additional dimension of severity of COVID-19's pathogenesis. COVID-19 commonly impairs respiratory function (Hosey and Needham, 2020) and the degree of respiratory support required can range from non-invasive assistance to highly invasive mechanical ventilation, and is usually administered in intensive care.

The most widely used public health metric, globally, on assessing the intensive care during the COVID-19 pandemic, is the ICU admission rate, defined analogously to (2.1) as the number of COVID-19 patients admitted to ICUs per 100,000 inhabitants.

The data on the ICU occupancy on district level are recorded and published daily and encompass information on the number of unoccupied beds, beds occupied by patients infected with COVID-19 and beds occupied by patients not infected with COVID-19. The data also contain the number of patients on respiratory aid and distinguish between adult and child care facility occupancy. Interestingly, on county and country level, there is more information comprised in the data. Namely, the county level data includes data on ICU admissions, while on country level, data on patient age groups are included (Robert Koch-Institut, 2025d). Thus, the ICU admission rate can only be provided on county level, or higher. Chapter 7 includes further illustration on county level ICU admission.

Figure 2.4 shows the maximal number of treatment centres, i.e. ICUs, reporting to the DIVI by district, between the 1^{st} of March, 2020, and 1^{st} of August, 2022. The ICU data during the COVID-19 pandemic are recorded for the purpose of understanding the strain on the ICU facilities. As such, solely the occupancy is recorded daily on district level. Data on the admittance, however, is only published on county level. Thus, while it is possible to analyse the ICU occupancy on district level, the ICU-admittance cannot be directly analysed by the publicly provided data.

Figure 2.5 shows the total occupancy in the intensive care units in Germany, as well as the ICU admittance which has been made available on county level from the 29^{th} of July 2021. The figure

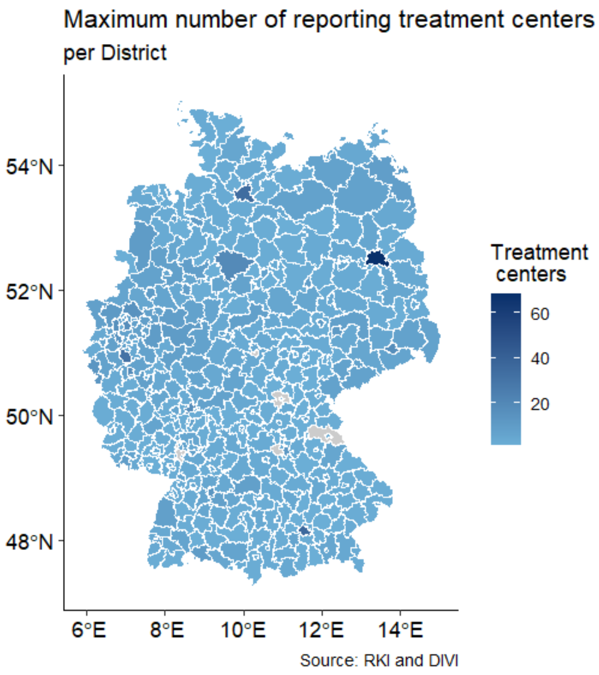


Figure 2.4.: Maximal number of treatment centres reporting to the DIVI by district, between the 1^{st} of March, 2020, and 1^{st} of August, 2022.

shows that occupancy numbers much higher and increase more drastically than the number of ICU patients admitted and infected with COVID-19. This evidences that the occupancy is not a direct function of admittance but also a function of length of stay. One should therefore also investigate the number of patients admitted to the ICU and infected with COVID-19, as well as the average length of stay of patients on the ICU.

The contributions of this thesis solely pertain to this aspect of the COVID-19 pandemic. In the first contribution the distribution of ICU occupancy is analysed, as seen in Chapter 5. In Chapter 6, the number of admitted and released patients are disentangled from the occupancy, thereby allowing for the analysis of an ICU admittance rate. In Chapter 7, this is further extended to allow for estimation on the average length of stay in the ICU.

To the best of our knowledge this has previously not been attempted. Karagiannidis et al. (2020) and Keller et al. (2023) explore the ICU admission, but do so by collecting further clinical level data. Karagiannidis et al. (2021) report the ICU admission on country level, rather than district level, which is the granularity in the contributions of this thesis.

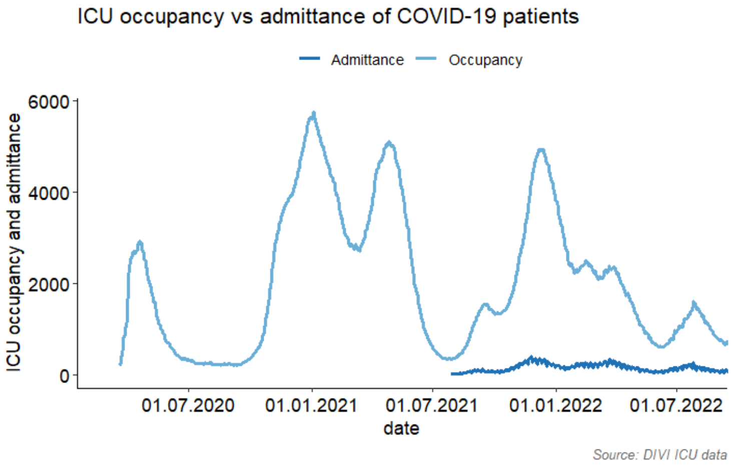


Figure 2.5.: ICU admittance and ICU occupancy. Data taken by Robert Koch-Institut (2025d).

3. Statistical Modelling

Following McCullagh (2002), statistical modelling is broadly understood as the methodology for reformulating real-world problems within a mathematical framework. In particular, parametric statistical modelling is defined by a parameter set Θ and a function P that assigns each $\theta \in \Theta$ to a probability distribution P_θ on the sample space $\mathcal{S}$, such that $P : \Theta \to \mathcal{P}(\mathcal{S})$, where $\mathcal{P}(\mathcal{S})$ denotes the set of all probability distributions on $\mathcal{S}$. Statistical modelling, and parametric modelling in particular, therefore form the cornerstone of empirical analysis.

The origins of statistical modelling lie in what has come to be known as regression, a concept introduced by Sir Francis Galton (1822–1911) in 1885. Galton analysed hereditary associations using the least squares method, which had been introduced by Gauss and Legendre approximately 80 years earlier (Stigler, 1986). These foundational ideas have since evolved into a comprehensive framework that underpins many of the parametric statistical models in widespread use today.

Following Fahrmeir et al. (2013) and Wood (2017), and Hastie et al. (2009), a distinction is commonly drawn between Normal Linear Models (LMs), Generalised Linear Models (GLMs), and Generalised Additive Models (GAMs), which reflects the historical chronology of their theoretical development. Moving from LMs to GLMs, the generalisation from the Gaussian distributional assumption to a broader class of distributions is credited to Nelder and Wedderburn (1972), who introduced GLMs for distributions within the exponential family, a concept initially attributed to Pitman (1936). Hastie and Tibshirani (1987) subsequently extended the linear predictor by allowing the linear combination of explanatory covariates to be substituted by smooth functions, thus extending GLMs to GAMs and moving from parametric to non-parametric statistical models. These models were later refined by Eilers and Marx (1996) and Wood (2015), among others, with further extensions outlined by Fahrmeir et al. (2013).

Fundamentally, these models are governed by three key assumptions: the distributional assumption, the structural assumption and assumption of independence. All three are necessary to quantify the relationship between explanatory covariates and the response, as outlined in Section 3.1.

In practical terms, statistical models aim to quantify the association between a response (or multiple responses) and explanatory variables assumed to contain relevant information. By assigning each observation in the data, X_i, a probability according to the assumed distribution, one can calculate the joint probability of the data given the distributional parameters. This forms the likelihood function, which can then be maximised with respect to the parameters $\boldsymbol{\theta}$. Estimation methods for GAMs are discussed in Section 3.2, while approaches to quantifying the uncertainty of these estimates are outlined in Section 3.3.

3.1. Model

To relate the response quantitatively to the explanatory variables, parametric statistical models rely on three fundamental assumptions: a *distributional* assumption, a *structural* assumption, and an *independence* assumption. The restrictiveness of these assumptions varies by context, but each must be specified to obtain a coherent framework for estimation. The independence assumption forms the foundation for likelihood based parametric modelling, where the distributional and structural assumption are further specified by the researcher. The distributional assumption supplies the general probabilistic form of the response; determining location, scale, and shape, while the distributional parameters are linked to explanatory variables. The structural assumption prescribes the functional form of this link between parameters and covariates. The following subsections introduce the general form of both assumptions.

3.1.1. Distributional assumptions

Following introductory treatments (e.g. Kauermann et al., 2021), let the response be a real-valued random variable Y,

$$Y : \Omega \to \mathbb{R}, \tag{3.1}$$

where Ω denotes the sample space and $\mathbb{R}$ the set of real numbers. Realisations are denoted $\boldsymbol{y} = [y_1, \dots, y_n]$. The probability law $P_{\boldsymbol{\theta}}$ assigns probabilities to events involving Y in accordance with Kolmogorov's axioms (Kolmogorov, 1933). Writing $F_Y(y \mid \boldsymbol{\theta})$ and $f_Y(y \mid \boldsymbol{\theta})$ for the cumulative distribution function (cdf) and probability density function (pdf), respectively,

$$\begin{aligned} Y &\sim F_Y(\,\cdot \mid \boldsymbol{\theta}), \\ P_{\boldsymbol{\theta}}(Y \leq y) = F_Y(y \mid \boldsymbol{\theta}) \; &= \; \int_{-\infty}^{y} f_Y(\tilde{y} \mid \boldsymbol{\theta}) \, \mathrm{d}\tilde{y}. \end{aligned} \tag{3.2}$$

For discrete responses the density is replaced by a probability mass function (pmf), and the integral by a sum.

The distribution depends on both the response and the parameter vector $\boldsymbol{\theta}$ that encodes location, scale and (where relevant) shape. A central task is to specify which parameters are to be estimated and to choose an appropriate distributional family. While most simply one-parameter models (e.g. estimating a mean) are encountered, many applications require *multi-parameter* families in which variance, skewness and kurtosis (or analogous characteristics) are also parameterised (Fahrmeir et al., 2013). Multivariate response models are, of course, inherently multi-parameter. A widely used class is the multi-parameter exponential family,

$$f_Y(\boldsymbol{y} \mid \boldsymbol{\theta}) = b(\boldsymbol{y}) \exp\Big\{\boldsymbol{\psi}(\boldsymbol{\theta})^\top T(\boldsymbol{y}) - A(\boldsymbol{\theta})\Big\}, \tag{3.3}$$

where $T(\boldsymbol{y})$, $b(\boldsymbol{y})$, $\boldsymbol{\psi}(\boldsymbol{\theta})$, and $A(\boldsymbol{\theta})$ are known functions of the data and parameters. For an intuitive overview, see DasGupta (2011). The choice of $\boldsymbol{\psi}(\boldsymbol{\theta})$ and parameterisation determines the link between covariates and parameters, and will interact with the structural assumption.

Extending (3.1), consider p real-valued responses on a common probability space,

$$Y_i : \Omega \to \mathbb{R} \quad, \forall \; i \in \{1, \dots, p\}. \tag{3.4}$$

The Multinomial distribution, Poisson distribution, and the Skellam distribution, which are relevant to this thesis, are outlined below. Thereafter, introduced models may be referred to as 'Multinomial model' or 'Poisson model', referring to regression models in which the response is assumed to follow the respectively titled distribution.

Multinomial distribution

Suppose an experiment with c categorical outcomes is repeated n times independently, with category probabilities $\boldsymbol{p} = (p_1, \ldots, p_c)$, such that $\sum_{i=1}^{c} p_i = 1$. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_c)$ denote the category counts; the sample space consists of vectors $\boldsymbol{y}$ satisfying

$$\sum_{i=1}^{c} y_i = n, \qquad y_i \in \mathbb{N}^0. \tag{3.5}$$

Then Y is Multinomially distributed, denoted by

$$\boldsymbol{Y} \sim \mathcal{M}\text{ult}(n, \boldsymbol{p}), \tag{3.6}$$

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \frac{n!}{\prod_{i=1}^{c} y_i!} \prod_{i=1}^{c} p_i^{y_i}. \tag{3.7}$$

Marginally, $Y_i \sim \mathcal{B}\text{in}(n, p_i)$, but the components are dependent (they sum to n) (Rudas, 2018).

Poisson distribution

Let the variable L follow a Poisson distribution, with pmf

$$L \sim \mathcal{P}\text{ois}(\lambda), \tag{3.8}$$

$$P(L = l) \;=\; \frac{\lambda^l e^{-\lambda}}{l!}., \tag{3.9}$$

where $\cdot!$ is the factorial function (Poisson, 1837).

Skellam distribution

Let $X \sim \mathcal{P}\text{ois}(\lambda_1)$ and $L \sim \mathcal{P}\text{ois}(\lambda_2)$ be independent. Then their difference $Z = X - L$ follows the Skellam distribution, with pmf

$$Z \sim \mathcal{S}\text{kellam}(\lambda_1, \lambda_2), \tag{3.10}$$

$$P(Z = z) \;=\; \exp\big(-(\lambda_1 + \lambda_2)\big) \left(\frac{\lambda_1}{\lambda_2}\right)^{z/2} I_{|z|}\Big(2\sqrt{\lambda_1 \lambda_2}\Big), \tag{3.11}$$

where $I_\nu(\cdot)$ is the modified Bessel function of the first kind, with $Z \in \mathbb{Z}$ (Skellam, 1948).

3.1.2. Structural assumptions

Given a distributional assumption, the structural assumption specifies how each distributional parameter depends on covariates. Parameters may be fixed (known) or modelled as functions of explanatory variables. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ denote the q-dimensional parameter vector. Each parameter is assumed to be a function of a set of covariates, thus the i^{th} observation of the j^{th} parameter is assumed to take the form

$$\theta_{i,j} \;=\; h_j(\eta_{i,j}), \tag{3.12}$$

where $h_j(\cdot)$ is the *response function* (its inverse $g_j(\cdot) = h_j^{-1}(\cdot)$ is the *link* function) and $\eta_{i,j}$ is referred to as the *(linear) predictor*. The response function is usually chosen so that its codomain matches the parameter's support and, where possible, to simplify estimation (e.g. using a canonical link for exponential-family models (Fahrmeir et al., 2013)). For readability, $\eta_{i,j}$ is going to be written as η_i hereafter.

Linear combination

The simplest specification is a linear predictor,

$$\eta_i \;=\; \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta}. \tag{3.13}$$

The linear combination of covariates may be restrictive for certain data situations. Transformations of the included covariates render the structural assumption of a linear combination a little less restrictive.

A common transformation of covariates is applied when the realisations of the a covariate, x_c, are non-numerical, e.g. categorical. If the support of x_c are categories, the variable needs to be transformed, such that likelihood can then be calculated. A common transformation is 'dummy' or 'reference' coding. On a high level, the variable is transformed into a binary matrix, X_c, in which the j^{th} entry of the k^{th} column takes the value of 1, if the j^{th} observation of the variable x_c takes the value of the k^{th} category, and 0 otherwise. To preserve identifiability a reference category is chosen, whose corresponding column is set to 0. For more detail see (Fahrmeir et al., 2013, p. 94).

Further, if prior knowledge suggests a specific functional form for a covariate's effect, the covariate may be transformed according to the functional form. For example, a quadratic term may be included as

$$\eta_i \;=\; \beta_0 + \boldsymbol{x}_{i,-k}^\top \boldsymbol{\beta}_{-k} + \beta_{k_1} x_{i,k} + \beta_{k_2} x_{i,k}^2, \tag{3.14}$$

with $\boldsymbol{x}_{i,-k}$ indicating the i^{th} observation of the included covariates, omitting the k^{th} covariate.

Splines and smooth functions

However, if the functional form of a given covariate is not known and cannot be represented by commonly well-known covariate transformations, effects may be modelled by regression splines, i.e. linear combinations of basis functions spanning the observed covariate range. Popular choices include truncated polynomial splines and B-splines (Fahrmeir et al., 2013; Green and Silverman, 1993).

In practice, the number, location and span of basis functions strongly influence the fit and are at the discretion of the analyst. Increasing the basis dimension can interpolate the data, risking numerical instability and overfitting (Wood, 2003).

Smooth splines address this by penalising roughness, balancing fit and smoothness. P-splines (Eilers and Marx, 1996; Eilers et al., 2015) combine a B-spline basis with a difference-penalty (typically on second differences or derivatives). Extending (3.13)–(3.14), a smooth effect of covariate $x_{i,k}$ is written

$$\eta_i \;=\; \beta_0 + \boldsymbol{x}_{i,-k}^\top \boldsymbol{\beta}_{-k} + f(x_{i,k}), \tag{3.15}$$

with $f(x_{i,k}) = \sum_{w=1}^{d} \gamma_w B_w(x_{i,k})$ built from $d = m + l - 1$ B-spline basis functions on $[\kappa_{1-l}, \kappa_{m+l}]$ with knots $\{\kappa_{1-l}, \ldots, \kappa_{m+l}\}$ and $\sum_{w=1}^{d} B_w(x_{i,k}) = 1$. For a B-spline basis of degree l,

$$B_w^{(l)}(x_{i,k}) \;=\; \frac{x - \kappa_{w-1}}{\kappa_w - \kappa_{w-1}}\, B_{w-1}^{(l-1)}(x_{i,k}) \;+\; \frac{\kappa_{w+l} - x}{\kappa_{w+l} - \kappa_w}\, B_w^{(l-1)}(x_{i,k}). \tag{3.16}$$

For multi-dimensional smoothing, thin-plate splines (a special case within the Duchon family of isotropic smoothers (Duchon, 2006)) are attractive because they avoid manual knot placement (Wood, 2017). A generic representation is

$$f(\boldsymbol{x}) \;=\; \sum_{i=1}^{n} \delta_i\, \mathcal{R}(\|\boldsymbol{x} - \boldsymbol{x}_i\|) \;+\; \sum_{j=1}^{M} \alpha_j\, \phi_j(\boldsymbol{x}), \tag{3.17}$$

with coefficients $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ subject to the constraint $\boldsymbol{T}^\top \boldsymbol{\delta} = \boldsymbol{0}$, where $T_{ij} = \phi_j(\boldsymbol{x}_i)$ and $\{\phi_j\}_{j=1}^{M}$ are linearly independent polynomials in $\mathbb{R}^d$ of degree less than m (with $M = \binom{m+d-1}{d}$). Additionally, $\boldsymbol{x}_i$ denotes the i^{th} observation of the covariates over which the smooth spline is calculated, $\boldsymbol{x}$. The radial basis $\mathcal{R}(r)$ is

$$\mathcal{R}(r) \;=\; \eta_{m,d}(r) \;=\; \begin{cases} \dfrac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-d/2)!}\, r^{\,2m-d} \log r, & d \text{ even}, \\ \dfrac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!}\, r^{\,2m-d}, & d \text{ odd}, \end{cases} \tag{3.18}$$

and computational burden can be reduced via an eigendecomposition of the associated kernel matrix (Wood, 2003).

Random effects and other extensions can be incorporated using the same penalised-spline approach (Eilers et al., 2015), e.g.

$$\eta_i \;=\; \beta_0 + \boldsymbol{x}_i^\top \boldsymbol{\beta} + u_i, \tag{3.19}$$

$$u_i \;\sim\; \mathcal{N}(0, \sigma_u^2), \tag{3.20}$$

with u_i represented through an appropriate (penalised) basis.

3.2. Estimation

Once the distributional and structural assumptions have been defined, a probability measure can be assigned to each observation given a set of distributional parameters. Together with the independence assumption, this permits the calculation of the joint conditional probability distribution corresponding to the likelihood, $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})$. The likelihood and its logarithm, the log-likelihood, are given by

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} f_Y(y_i \mid \boldsymbol{\theta}_i) \, , \tag{3.21}$$

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f_Y(y_i \mid \boldsymbol{\theta}_i) \, . \tag{3.22}$$

The likelihood is a function of the parameter vector $\boldsymbol{\theta}$ and the data $\boldsymbol{y}$, and can thus be used to identify plausible values for $\boldsymbol{\theta}$. Maximising $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y})$ with respect to $\boldsymbol{\theta}$ yields estimates for which the observed data are most probable, given the assumed model.

Under regularity conditions (Cramér, 1946), a maximum of the likelihood exists within the parameter space, denoted $\hat{\boldsymbol{\theta}}$, which is called the maximum likelihood estimate (MLE). When a model belongs to the exponential family but its variance assumption is overly restrictive, the quasi-likelihood provides a relaxation; see Fahrmeir et al. (2013, p. 309) for details.

In the frequentist framework, the objective is to approximate the true (but unknown) parameter vector, $\boldsymbol{\theta}_{\text{true}}$. By contrast, in Bayesian statistics $\boldsymbol{\theta}$ is treated as a random variable with posterior distribution

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \prod_{i=1}^{n} f_Y(y_i \mid \boldsymbol{\theta}) \, f(\boldsymbol{\theta}), \tag{3.23}$$

where $f(\boldsymbol{\theta})$ denotes the prior distribution. Although the two paradigms differ philosophically, they coincide mathematically in some settings (e.g. with certain non-informative priors). This thesis focuses on methods rooted primarily in the frequentist framework, and the following discussion emphasises likelihood maximisation.

For computational reasons, the log-likelihood is typically used. Estimates are obtained by solving the score equations, i.e. the system of first-order partial derivatives with respect to the parameters (Fahrmeir et al., 2013, pp. 653–659).

To ensure identifiability or to impose desirable properties, such as smoothness, which is previously referred to in Section 3.1, the maximisation of the likelihood can be penalised. The penalised maximum likelihood estimate is defined as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left\{ \ell(\boldsymbol{\theta}; \boldsymbol{y}) - \lambda \, \text{pen}(\boldsymbol{\theta}) \right\} . \tag{3.24}$$

The penalty term depends on the choice of model. For smooth functions, it is determined both by the choice of basis and by the properties one wishes to enforce. In P-splines (Eilers and Marx, 1996), the penalty takes the form

$$\text{pen}(\gamma_w) = \sum_{w=r+1}^{d} \left(\Delta^r \gamma_w\right)^2 = \int \left(f''(x_{i,k})\right)^2 \mathrm{d}x_{i,k}, \tag{3.25}$$

where Δ^r denotes the r^{th} order difference and $f''(x_{i,k})$ the second derivative of $f(x_{i,k})$.

For thin-plate splines, (3.17), the penalty is given by

$$\text{pen}(f) = J_{m,d}(f) = \int_{\mathbb{R}^d} \sum_{\nu_1+\cdots+\nu_d=m} \frac{m!}{\nu_1!\cdots\nu_d!} \left(\frac{\partial^m f}{\partial x_1^{\nu_1}\cdots \partial x_d^{\nu_d}} \right)^2 \mathrm{d}x_1 \cdots \mathrm{d}x_d, \tag{3.26}$$

with d indicating the number of covariates over which f is smoothed (Wood, 2003).

In some cases the MLE can be derived analytically, but this is rarely feasible in practice, so numerical approximations are used. A common class of iterative methods is based on Newton's method (Wood, 2017, p. 76). Newton–Raphson and Fisher scoring are widely applied variants (Fahrmeir et al., 2013, p. 660). These rely on successive quadratic approximations to the log-likelihood via second-order Taylor expansion around the current estimate. If the Hessian matrix is indefinite, modified Newton methods may be used, substituting a positive-definite approximation. Fisher scoring replaces the Hessian with the Fisher information matrix (Wood, 2017, pp. 76–77); see also Kauermann et al. (2021, p. 77) for an intuitive explanation. Quasi-Newton methods provide further flexibility by updating approximations to the Hessian iteratively (McLachlan and Krishnan, 2008, pp. 3–8).

The Expectation–Maximisation (EM) algorithm (Dempster et al., 1977) is another iterative estimation procedure, particularly valuable when data are incomplete. Let $\boldsymbol{x} \in \mathcal{X}$ denote covariates and $\boldsymbol{y} \in \mathcal{Y}$ the response. If part of $\boldsymbol{y}$ is unobserved, i.e. $\boldsymbol{y} = [\boldsymbol{y}_o, \boldsymbol{y}_{\neg o}]$, the likelihood must integrate over the missing data. This is often analytically intractable. The EM algorithm provides a practical solution and converges under mild conditions (Vaida, 2005). At iteration k, it alternates between:

1. **Expectation step (E-step):**

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}) = \sum_{i=1}^{n} \int \log f(y_i, \boldsymbol{\theta})\, f(y_{i,\neg o} \mid y_{i,o}, \boldsymbol{\theta}^{(k-1)})\, \mathrm{d}y_{i,\neg o} \tag{3.27}$$

$$= \mathbb{E}_{\boldsymbol{\theta}^{(k-1)}}[\ell(\boldsymbol{\theta}; \boldsymbol{y}) \mid \boldsymbol{y}_o]. \tag{3.28}$$

2. **Maximisation step (M-step):**

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}\in\Omega_{\boldsymbol{\theta}}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}). \tag{3.29}$$

Starting from initial values $\boldsymbol{\theta}^{(0)}$, these steps are repeated until convergence (McLachlan and Krishnan, 2008, p. 18). Cince its development, numerous variants of the EM algorithm have been proposed.

A relevant variant for this thesis is the stochastic EM (sEM) algorithm (Broniatowski et al., 1983; Celeux, 1985; Celeux and Diebolt, 1986b,a), a special case of the Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990a,b). In MCEM, the intractable expectation in the E-step is approximated by Monte Carlo integration using simulated missing data. Specifically,

$$\hat{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}) = \frac{1}{m} \sum_{j=1}^{m} \ell(\boldsymbol{\theta}; \boldsymbol{y}_o, \boldsymbol{y}_{\neg o}^{j}), \tag{3.30}$$

where $\boldsymbol{y}_{\neg o}^{j}$ are samples from $f(\boldsymbol{y}_{\neg o} \mid \boldsymbol{y}_o, \boldsymbol{\theta}^{(k-1)})$. The sEM algorithm corresponds to the special case $m = 1$. At each iteration it:

1. Simulates $\boldsymbol{y}_{\neg o}^{(k)}$ from $f_Y(y_i \mid \hat{\boldsymbol{\theta}}_i^{(k-1)})$,
2. Updates $\hat{\boldsymbol{\theta}}^{(k)}$ via MLE (or similar),

until convergence to a stationary distribution. For example, Häggström (2002, p. 28), give an intuitive introduction to- and definition of the stationary distribution of a stochastic process. Although convergence of sEM to a stationary distribution is guaranteed, the relationship to the true likelihood maximum is not always straightforward; see McLachlan and Krishnan (2008, p. 226).

Finally, depending on the likelihood, parameter space, or estimation method, estimates may be biased. Bias properties are critical for (weak) consistency (Wakefield et al., 2013, p. 39). The bias of an estimator $\hat{\boldsymbol{\theta}}$ is

$$\text{bias}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}. \tag{3.31}$$

If $\text{bias}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 0$ for all $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$, then $\hat{\boldsymbol{\theta}}$ is unbiased (Kauermann et al., 2021, p. 41). A wide range of bias-correction methods have been proposed; see Cordeiro and Cribari-Neto (2014) for an overview.

3.3. Uncertainty

Hüllermeier and Waegeman (2021) discuss uncertainty and its quantification in the context of supervised learning, to which regression and maximum likelihood estimation naturally belong. They distinguish between two forms of uncertainty: epistemic and aleatoric. Epistemic (or systematic) uncertainty arises from a lack of knowledge, for example about the correct model specification, while aleatoric uncertainty reflects inherent randomness in outcomes, i.e. variability in the data-generating process. Along similar lines, Begg et al. (2014) highlight the distinction between uncertainty and variability.

To reiterate, the likelihood (3.21) is maximised to obtain the MLE $\hat{\boldsymbol{\theta}}$. According to Hüllermeier and Waegeman (2021), the ‘peakedness’ of the likelihood function around $\hat{\boldsymbol{\theta}}$ reflects the certainty of the estimator. Thus, confidence regions and, by extension, the variance of the estimator provides a frequentist means of quantifying uncertainty—though not an infallible one. Namely, this approach does not allow the researcher to disentangle aleatoric and epistemic uncertainty. Nonetheless, in the context of this thesis, attention is restricted to variance-based measures.

In parametric modelling, as the number of independent observations n tends to infinity, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is asymptotically normal. More specifically,

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(0,\, I(\boldsymbol{\theta})^{-1}), \tag{3.32}$$

where

$$I(\boldsymbol{\theta}) \;=\; -\,\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j}\ell(\boldsymbol{\theta}; \boldsymbol{y})\right]_{1 \le i,j \le n} \tag{3.33}$$

is the Fisher information matrix (the negative expected Hessian of the log-likelihood at $\boldsymbol{\theta}$).

From this, approximate confidence regions for $\boldsymbol{\theta}$ around $\hat{\boldsymbol{\theta}}$ can be derived. A commonly used large-sample $(1 - \alpha)$ confidence region is

$$\mathcal{C}_{1-\alpha} \;=\; \left\{ \boldsymbol{\theta} \in \Omega_\theta \;:\; (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top I(\hat{\boldsymbol{\theta}})\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \;\le\; \chi^2_{p,\,1-\alpha} \right\}, \tag{3.34}$$

where $p = \dim(\boldsymbol{\theta})$ and $\chi^2_{p,\,1-\alpha}$ is the $(1-\alpha)$ quantile of the Chi-squared distribution with p degrees of freedom.

It follows from (3.32) that the Fisher information provides a lower bound—known as the Cramér–Rao bound—for the variance of an unbiased estimator $\hat{\boldsymbol{\theta}}$, under regularity conditions (Nielsen, 2013). Asymptotically, the variance estimate approaches this bound (Cordeiro and Cribari-Neto, 2014, p. 4). However, if assumptions, such as independence are violated, or if the variance specification is too restrictive, and one may under or overestimate the certainty of the estimators. Thus, adjustments are required. A common adjustment is the *sandwich estimator* (Wakefield et al., 2013, p. 33), which estimates the variance as

$$\hat{\mathbb{V}}(\boldsymbol{\theta}) \;=\; \boldsymbol{A}^{-1}(\boldsymbol{\theta})\,\boldsymbol{B}(\boldsymbol{\theta})\,\boldsymbol{A}^{-1}(\boldsymbol{\theta}), \tag{3.35}$$

where, in the context of likelihood estimation,

$$\boldsymbol{A}(\boldsymbol{\theta}) = \mathbb{E}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\frac{\partial}{\partial\boldsymbol{\theta}^\top}\ell(\boldsymbol{\theta};\boldsymbol{y})\right), \tag{3.36}$$

$$\boldsymbol{B}(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\boldsymbol{y})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\boldsymbol{y})\right)^\top\right]. \tag{3.37}$$

In the context of missing data problems, the variance of estimates obtained by the stochastic EM (sEM) algorithm must be adjusted to account for imputation during the E-step. Rubin's rule provides an appropriate correction (Kauermann et al., 2021, p. 301). Suppose the sEM runs for K iterations after convergence. At each iteration $k \in \{1,\ldots,K\}$, both $\hat{\boldsymbol{\theta}}^{(k)}$ and $\hat{V}^{(k)} = I^{-1}(\hat{\boldsymbol{\theta}}^{(k)})$ are calculated, with $I(\cdot)$ denoting the Fisher information matrix of the complete data. The pooled parameter estimate is

$$\hat{\boldsymbol{\theta}}^{*} \;=\; \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\theta}}^{(k)}. \tag{3.38}$$

Rubin's rule then defines the variance estimate as

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\theta}}^{*}) \;=\; \frac{1}{K}\sum_{k=1}^{K}\hat{V}^{(k)} \;+\; \frac{1+K^{-1}}{K-1}\sum_{k=1}^{K}(\hat{\boldsymbol{\theta}}^{(k)}-\hat{\boldsymbol{\theta}}^{*})(\hat{\boldsymbol{\theta}}^{(k)}-\hat{\boldsymbol{\theta}}^{*})^\top, \tag{3.39}$$

which simplifies to

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\theta}}^{*}) \;=\; \frac{1}{K}\sum_{k=1}^{K}\hat{V}^{(k)} \;+\; \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\boldsymbol{\theta}}^{(k)}-\hat{\boldsymbol{\theta}}^{*})(\hat{\boldsymbol{\theta}}^{(k)}-\hat{\boldsymbol{\theta}}^{*})^\top, \tag{3.40}$$

in the case where all data are imputed in the E-step.

4. Concluding Remarks

During the outbreak and spread of COVID-19, the world faced a multitude of unforeseen challenges, one of which was assessing the extent of its impact on public health. The contributions of this thesis form part of the effort to understand the impact of COVID-19 on ICUs in Germany. The methodology developed herein is not limited to COVID-19 but is applicable to a wide range of contexts. In essence, the contributions concern the estimation of the association between ICU bed occupancy and COVID-19 infection rates, the inference of patient inflow and outflow from occupancy data, and the extension of this analysis to estimate the average length of stay.

The following sections provide an overview of the key findings of the thesis' contributions, together with an outlook on possible applications and extensions of the developed methodology.

4.1. Contributions

Part II analyses the distribution of the ICU occupancy. The Multinomial distribution is assumed when analysing the ICU occupancy with respect to beds either being occupied by patients infected with COVID-19, patients not infected with COVID-19, or unoccupied. This approach allows estimation of associations between the ICU bed distribution and lagged infection rates by age group, the bed distribution of the previous week, spatial correlation (captured through a two-dimensional thin-plate spline), and district-level heterogeneity (modelled via a random intercept). The multinomial assumption is particularly valuable for prediction, as it reflects the mutually exclusive nature of ICU bed allocation.

Part III extends the analysis of Part II by inferring patient inflow and outflow from ICU occupancy. Chapter 6 employs the stochastic Expectation-Maximisation (sEM) algorithm to iteratively simulate from a truncated Skellam distribution with incoming and outgoing intensity parameters, which are estimated using two independent Poisson models. The covariates included in the inflow model extend analogously from Part II, further to temporal correlation via a thin-plate spline over time, and a categorical weekday effect. The outflow model is specified analogously, with an additional offset defined as the weighted sum of previously estimated incoming patients. While Chapter 6 takes the weight of previously incoming patients as fixed input, Chapter 7 extends this by treating length of stay as a random parameter. These parameters are estimated via constrained maximum likelihood, which introduces bias. This bias is corrected employing an additional simulation step, rendering a non-standard application of the sEM.

4.2. Outlook

In the contributions of this thesis, intensive care has been analysed from two main perspectives. The first contribution examines ICU occupancy, providing insight into the distribution of ICU beds and their association with infection rates. This model preserves interpretability and explainability while offering predictions that can support facility planning. Building on this, the methodology is extended to disentangle the previously unobserved patient inflow, length of stay, and outflow from occupancy data. In this particular use case, it provides valuable insight into a pandemic for which publicly available data are limited, thereby enabling a more complete understanding of the pandemic than would otherwise have been possible.

As noted above, the methodology developed in this thesis, particularly in Chapter 7, has potential applications well beyond the context of ICU occupancy of COVID-19 patients. It can be applied in any setting where underlying inflows and outflows are of interest but only net counts are observed, such as herd dynamics, for example.

Nonetheless, a number of limitations remain.

A limitation arising from the specific context is that there should ideally be a distinction between patients who are discharged to intermediate care units (or similar) and those who die. These are evidently two very different outcomes, which may systematically affect the length of stay. Extending the model to capture such distinctions—similar in spirit to multi-state models such as those of Johnston and Hay (2006)—could enhance the framework, though possibly at the cost of reduced interpretability.

Particularly in Chapter 7, the identifiability of the model poses a considerable challenge. The longer the estimated length of stay, the more parameters the model must estimate, which can quickly render the model non-identifiable and necessitate larger data sets. Similarly, the more complex the model in the M-step, the less likely it is to be identifiable.

Further, the bias correction applied in Chapter 7 does not necessarily eliminate all bias, and refining this approach could yield deeper insights into the consistency of the estimators. Finally, while the extension of the sEM algorithm in Chapter 7 is conceptually straightforward, its computational efficiency could be improved, offering another avenue for future research. Its stochastic nature prohibits parallelisation; however, Lu and Li (2024), for example, suggest methods to improve the runtime of the MCEM, which may represent a promising extension to the algorithm.

References

Alimohamadi, Y., Taghdir, M., and Sepandi, M. (2020). Estimate of the basic reproduction number for covid-19: A systematic review and meta-analysis. *J Prev Med Public Health*, 53(3): 151–157.

Anderson, W. (2018). The history in epidemiology. *International Journal of Epidemiology*, 48(3): 672–674.

Armstrong, B. G. (1995). Comparing standardized mortality ratios. *Annals of epidemiology*, 5(1): 60–64.

Pifarré i Arolas, H., Acosta, E., López-Casasnovas, G., Lo, A., Nicodemo, C., Riffe, T., and Myrskylä, M. (2021). Years of life lost to covid-19 in 81 countries. *Scientific reports*, 11(1): 3504.

Begg, S., Bratvold, R., and Welsh, M. (2014). Uncertainty vs. variability: What's the difference and why is it important?

Berger, U., Kauermann, G., and Küchenhoff, H. (2022). Discussion on on the role of data, statistics and decisions in a pandemic. *AStA Adv Stat Anal 106, 387–390 (2022).*

Broniatowski, M., Celeux, G., and Diebolt, J. (1983). Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. *Data analysis and informatics*, 3: 359–373.

Celeux, G. and Diebolt, J. (1986a). The sem and em algorithms for mixtures: numerical and statistical aspects. In *Procedings of the 7th Franco Belgian Meeting of Statistics. Bruxelles: Publication Des Facultes Universitaries St. Louis.*

Celeux, G. (1985). The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational statistics quarterly*, 2: 73–82.

Celeux, G. and Diebolt, J. (1986b). L'algorithme sem: un algorithme d'apprentissage probabiliste pour la reconnaissance de mélange de densités. *Revue de statistique appliquée*, 34(2): 35–52.

Centers for Disease Control and Prevention. (2025). Underlying conditions and the higher risk for severe COVID-19. Accessed: 2025-06-11.

Cordeiro, G. M. and Cribari-Neto, F. (2014). *An introduction to Bartlett correction and bias reduction.* Springer.

Cramér, H. (1946). *Mathematical methods of statistics*, volume 1. Princeton university press.

DasGupta, A. (2011). The exponential family and statistical applications. *Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics.*

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(1): 5–20.

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(3): 407–426.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1): 1–22.

Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin. (2025). DIVI-Intensivregister. Accessed: 2025-06-24.

Diekmann, O., Heesterbeek, J. A. P., and Metz, J. A. J. (1990). The impact of false positive COVID-19 results in an area of low prevalence. *Journal of Mathematical Biology*, 28(4): 365–382.

Duchon, J. (2006). Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables: proceedings of a conference held at oberwolfach April 25–May 1, 1976*, pages 85–100. Springer.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2): 89–121.

Eilers, P. H., Marx, B. D., and Durbán, M. (2015). Twenty years of p-splines. *SORT: statistics and operations research transactions*, 39(2): 0149–186.

Ely, E. W., Brown, L. M., and Fineberg, H. V. (2024). Long covid defined.

Fagoni, N., Perone, G., Villa, G. F., Celi, S., Bera, P., Sechi, G. M., Mare, C., Zoli, A., and Botteri, M. (2020). The lombardy emergency medical system faced with covid-19: the impact of out-of-hospital outbreak. *Prehospital Emergency Care*, 25(1): 1–7.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. (2013). Regression models. In *Regression: Models, methods and applications*. Springer.

Farr, W. (1852). *Report on the Mortality of Cholera in England, 1848–49*. Printed by W. Clowes and Sons, Stamford Street, for Her Majesty's Stationery Office, London.

Fiedler, J., Moritz, C. P., Feth, S., Speckert, M., Dreßler, K., and Schöbel, A. (2021). Ein mathematisches modell zur schätzung der dunkelziffer von sars-cov-2-infektionen in der frühphase der pandemie am beispiel deutschland und italien. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz*, 64(9): 1067.

Frérot, M., Lefebvre, A., Aho, S., Callier, P., Astruc, K., and Aho Glélé, L. S. (2018). What is epidemiology? changing definitions of epidemiology 1978-2017. *PloS one*, 13(12): e0208442.

Fritz, C., Nicola, G. D., Günther, F., Rügamer, D., Rave, M., Schneble, M., Bender, A., Weigert, M., Brinks, R., Hoyer, A., Berger, U., Küchenhoff, H., and Kauermann, G. (2023). Challenges in interpreting epidemiological surveillance data – experiences from germany. *Journal of Computational and Graphical Statistics*, 32(3): 765–766.

Gallagher, B. J. (2020). Analysis: How close are we to a pandemic? Online; accessed 28 May 2025.

Gardner, J. W. and Sanborn, J. S. (1990). Years of potential life lost (ypll)—what does it measure? *Epidemiology*, 1(4): 322–329.

Green, P. and Silverman, B. (1993). Nonparametric regression and generalized linear models.

Gross, G. (2024). Gesundheitsämter: Ärztliche Unabhängigkeit sichern. *Deutsches Ärzteblatt*, 121(23): –8504–.

Günther, F., Bender, A., Katz, K., Küchenhoff, H., and Höhle, M. (2021). Nowcasting the covid-19 pandemic in bavaria. *Biometrical Journal*, 63(3): 490–502.

Häggström, O. (2002). *Finite Markov chains and algorithmic applications*, volume 52. Cambridge University Press.

Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398): 371–386.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Healy, B., Khan, A., Metezai, H., Blyth, I., and Asad, H. (2021). The impact of false positive covid-19 results in an area of low prevalence. *Clinical Medicine*, 21(1): e54–e56.

Hosey, M. M. and Needham, D. M. (2020). Survivorship after covid-19 icu stay. *Nature reviews Disease primers*, 6(1): 60.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn*, 110: 457–506.

Jahn, B., Friedrich, S., Behnke, J., Engel, J., Garczarek, U., Münnich, R., Pauly, M., Wilhelm, A., Wolkenhauer, O., Zwick, M., et al. (2022). On the role of data, statistics and decisions in a pandemic. *AStA Advances in Statistical Analysis*, 106(3): 349–382.

Johnston, R. and Hay, A. (2006). Voter transition probability estimates: An entropy-maximizing approach*. *European Journal of Political Research*, 11: 93 – 98.

Karagiannidis, C., Mostert, C., Hentschker, C., Voshaar, T., Malzahn, J., Schillinger, G., Klauber, J., Janssens, U., Marx, G., Weber-Carstens, S., et al. (2020). Case characteristics, resource use, and outcomes of 10 021 patients with covid-19 admitted to 920 german hospitals: an observational study. *The Lancet Respiratory Medicine*, 8(9): 853–862.

Karagiannidis, C., Windisch, W., McAuley, D. F., Welte, T., and Busse, R. (2021). Major differences in icu admissions during the first and second covid-19 wave in germany. *The Lancet Respiratory Medicine*, 9(5): e47–e48.

Kauermann, G., Küchenhoff, H., and Heumann, C. (2021). *Statistical foundations, reasoning and inference*. Springer.

Keller, K., Farmakis, I. T., Valerio, L., Koelmel, S., Wild, J., Barco, S., Schmidt, F. P., Espinola-Klein, C., Konstantinides, S., Muenzel, T., et al. (2023). Predisposing factors for admission to intensive care units of patients with covid-19 infection—results of the german nationwide inpatient sample. *Frontiers in Public Health*, 11: 1113793.

Kelley, A. S. and Bollens-Lund, E. (2018). Identifying the population with serious illness: The "denominator" challenge. *Journal of Palliative Medicine*, 21(S2): S–7–S–16. PMID: 29125784.

Kolmogorov, A. N. (1933). *Grundbegriffe Der Wahrscheinlichkeitsrechnung*. Berlin: Springer.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., et al. (2020). Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England journal of medicine*, 382(13): 1199–1207.

Lu, Y. and Li, Q. (2024). An improved monte carlo em acceleration algorithm. *International Core Journal of Engineering*, 10(5): 493–499.

Ludwig-Maximilians-Universität München. (2020). CODAG (COVID-19 Data Analysis Group). Accessed: 2025-09-05.

Mattey-Mora, P. P., Begle, C. A., Owusu, C. K., Chen, C., and Parker, M. A. (2022). Hospitalised versus outpatient covid-19 patients' background characteristics and comorbidities: a systematic review and meta-analysis. *Reviews in Medical Virology*, 32(3): e2306.

McCullagh, P. (2002). What is a statistical model? *The Annals of Statistics*, 30(5): 1225–1310.

McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*. John Wiley & Sons.

Meyerowitz, E. A., Richterman, A., Gandhi, R. T., and Sax, P. E. (2021). Transmission of sars-cov-2: A review of viral, host, and environmental factors. *Annals of internal medicine*, 174(1): 69–79.

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.

Nielsen, F. (2013). *Cramér-Rao Lower Bound and Information Geometry*, pages 18–37. Hindustan Book Agency, Gurgaon.

Peiris, J. S., Guan, Y., and Yuen, K. (2004). Severe acute respiratory syndrome. *Nature medicine*, 10(Suppl 12): S88–S97.

Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. *Mathematical Proceedings of the Cambridge Philosophical Society*, 32(4): 567–579.

Poisson, S.-D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile: précédées des règles générales du calcul des probabilités*. Bachelier.

Puhach, O., Meyer, B., and Eckerle, I. (2023). Sars-cov-2 viral load and shedding kinetics. *Nat Rev Microbiol*, 21: 147–161.

Robert Koch-Institut. (2025a). COVID-19-Hospitalisierungen in Deutschland. Accessed: 2025-05-26.

Robert Koch-Institut. (2025b). COVID-19-Todesfälle in Deutschland. Accessed: 2025-05-26.

Robert Koch-Institut. (2025c). German infection protection act. Accessed: 2025-09-12.

Robert Koch-Institut. (2025d). Intensivkapazitäten und COVID-19-Intensivbettenbelegung in Deutschland. Accessed: 2025-05-26.

Robert Koch-Institut. (2025e). Public health - health for all. Accessed: 2025-09-12.

Robert Koch-Institut. (2025f). SARS-CoV-2 Infektionen in Deutschland. Accessed: 2025-06-11.

Rudas, T. (2018). *Lectures on categorical data analysis*. Springer.

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.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2): 257–261.

Spiegelhalter, D. and Masters, A. (2021). *Covid by numbers: making sense of the pandemic with data*. Penguin UK.

Statistisches Bundesamt. (2025). Gestorbene: Deutschland, Jahre, Todesursachen, Geschlecht, Altersgruppen. Accessed: 2025-06-11.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Strobl, R., Misailovski, M., Blaschke, S., Berens, M., Beste, A., Krone, M., Eisenmann, M., Ebert, S., Hoehn, A., Mees, J., et al. (2024). Differentiating patients admitted primarily due to coronavirus disease 2019 (covid-19) from those admitted with incidentally detected severe acute respiratory syndrome corona-virus type 2 (sars-cov-2) at hospital admission: A cohort analysis of german hospital records. *Infection Control & Hospital Epidemiology*, 45(6): 746–753.

Thomas Carlyle. (1840). *Chartism*, page 11. Chicago, New York, Belford, Clarke & co.

Tolksdorf, K., Buda, S., Schuler, E., Wieler, L., and Haas, W. (2020). Eine hoehere letalitaet und lange beatmungsdauer unterscheiden covid-19 von schwer verlaufenden atemwegsinfektionen in grippewellen. *Epidemiologisches Bulletin*, (41): 3–10.

Vaida, F. (2005). Parameter convergence for em and mm algorithms. *Statistica Sinica*, 15(3): 831–840.

Vestergaard, L. S., 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. *Eurosurveillance*, 25(26): 2001214.

Wakefield, J. et al. (2013). *Bayesian and frequentist regression methods*, volume 23. Springer.

Wei, G. C. G. and Tanner, M. A. (1990a). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411): 699–704.

Wei, G. C. G. and Tanner, M. A. (1990b). Posterior computations for censored regression data. *Journal of the American Statistical Association*, 85(411): 829–839.

Willrich, N., Böttcher, S., Stern, D., Biegala, W., Albrecht, S., Oh, D.-Y., Feig, M., Schneider, M., Noll, I., Abu Sin, M., et al. (2021). Update: Erfassung der sars-cov-2-pcr-testzahlen in deutschland und die entwicklung der testzahlen in ärztlichen praxen. *Epidemiologisches Bulletin*.

Wolffram, D., Abbott, S., An der Heiden, M., Funk, S., Günther, F., Hailer, D., Heyder, S., Hotz, T., van de Kassteele, J., Küchenhoff, H., et al. (2023). Collaborative nowcasting of covid-19 hospitalization incidences in germany. *PLOS Computational Biology*, 19(8): e1011394.

Wollschläger, D., Fückel, S., Blettner, M., and Gianicolo, E. (2024). Übersterblichkeit im kontext der covid-19-pandemie in deutschland. *Die Kardiologie*, 18(2): 101–108.

Wood, S. (2015). Package 'mgcv'. *R package version*, 1(29): 729.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(1): 95–114.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. chapman and hall/CRC.

World Health Organization. (2025). Number of covid-19 deaths reported to who (cumulative total). Online; accessed 28 May 2025.

Wunderlich, C. (2020). Diese landkreise reißen die lockdown-linie. Online; accessed 28 May 2025.

Yuan, Z., Shao, Z., Ma, L., and Guo, R. (2023). Clinical severity of sars-cov-2 variants during covid-19 vaccination: A systematic review and meta-analysis. *Viruses*, 15(10).

Part II.

ICU Occupancy during the COVID-19 pandemic

5. 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:, 24(4):344–367.* `https://doi.org/10.1177/1471082X221124628`.

Data and code

Available at https://github.com/corneliusfritz/Statistical-modelling-of-COVID-19-data.

Copyright information

Author contributions

This article presents a comprehensive statistical modelling framework for COVID-19 data in Germany, structured into three main components: infections, hospitalisations, and intensive care unit (ICU) occupancy. These components were led by Dr Cornelius Fritz, Maximilian Weigert, Dr Giacomo De Nicola, and Martje Rave, respectively.

Martje Rave was responsible for the ICU occupancy analysis. This work comprised the independent collection, cleaning, and preprocessing of the data, as well as the complete modelling and interpretation of the results, undertaken in collaboration with and under the supervision of Prof. Dr Göran Kauermann. In addition, Rave implemented a sandwich estimator to adjust standard errors in light of potential violations of multinomial assumptions—specifically, the fact that ICU beds are not fully reallocated at each time point. This approach was developed in discussion with Dr Cornelius Fritz.

Rave also co-authored the introduction, data section, and discussion of the article, and provided detailed feedback on sections drafted by the other authors. In particular, sections three and four were authored by Fritz, Weigert, and De Nicola. Rave was actively involved in the final revision process, editing her own contributions independently and coordinating with the group during the preparation of the final manuscript.

The project was supervised by Prof. Dr Helmut Küchenhoff, Prof. Dr Göran Kauermann, and Prof. Dr Ursula Berger.

Statistical Modelling 2024; **24(4)**: 344–367

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 Küchenhoff[1] and Göran Kauermann[1]

[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

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 interdepency 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

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

$$\mathbb{E}(Y_{w,r,a} | Y_{w-1,r}) = \exp\{\eta_{w,r,a} + o_{r,a}\} \tag{3.1}$$

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

$$\eta_{w,r,a} = \theta_w + \sum_{k=1}^{A} \log(Y_{w-1,r,k} + \delta)\theta_{a,k}, \tag{3.2}$$

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

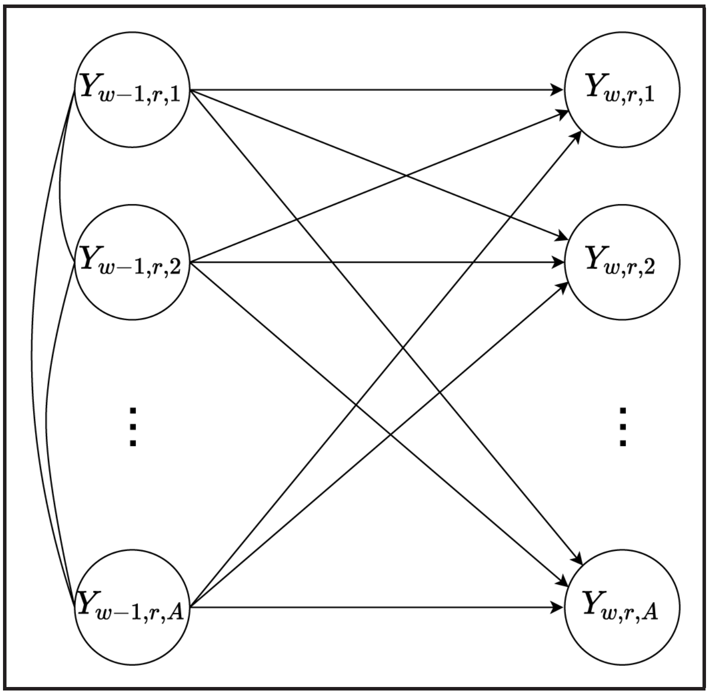


Figure 1 Assumed temporal dependence structure visualized as a directed acyclic graph (DAG)

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 $R_{w,r} = (R_{w,r,1}, \ldots, R_{w,r,A})$ 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):

$$\begin{aligned}\mathbb{E}\left(\tilde{Y}_{w,r,a} | \tilde{Y}_{w-1,r}\right) &= \mathbb{E}_{R_w, R_{w-1}}\left(\mathbb{E}_{Y_w}\left(R_{w,r,a} Y_{w,r,a} | \tilde{Y}_{w-1,r},\, R_{w,r,a},\, R_{w-1,r}\right)\right) \\ &= \mathbb{E}_{R_w, R_{w-1}}\Big(R_{w,r,a}\, \mathbb{E}_{Y_w}\left(Y_{w,r,a} | Y_{w-1,r}\right)\Big) \\ &= \pi_{w,a}\, \mathbb{E}_{R_{w-1}}\Big(\exp\{\eta_{w,r,a}\}\Big)\, \exp\{o_{r,a}\} \qquad (3.4)\end{aligned}$$

where for clarity we include the random variable as an index in the notation of the expectation. Note that

$$\begin{aligned}\mathbb{E}_{R_{w-1}}\Big(\exp\{\eta_{w-1,r,a}\}\Big) &= \mathbb{E}_{R_{w-1}}\left(\exp\left\{\sum_{k=1}^{A}\log(R_{w-1,r,k}^{-1}\tilde{Y}_{w-1,r,k})\theta_{a,k} + \theta_w\right\}\right) \\ &= \exp\left\{\tilde{\eta}_{w,r,a}\right\}\mathbb{E}_{R_{w-1}}\left(\exp\left\{\sum_{k=1}^{A}\log(R_{w-1,r,k}^{-1})\theta_{a,k} + \theta_w\right\}\right) \\ &= \exp\left\{\tilde{\eta}_{w,r,a} + \tilde{\theta}_w\right\}, \qquad (3.5)\end{aligned}$$

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 $k = 1, ..., A$ 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.

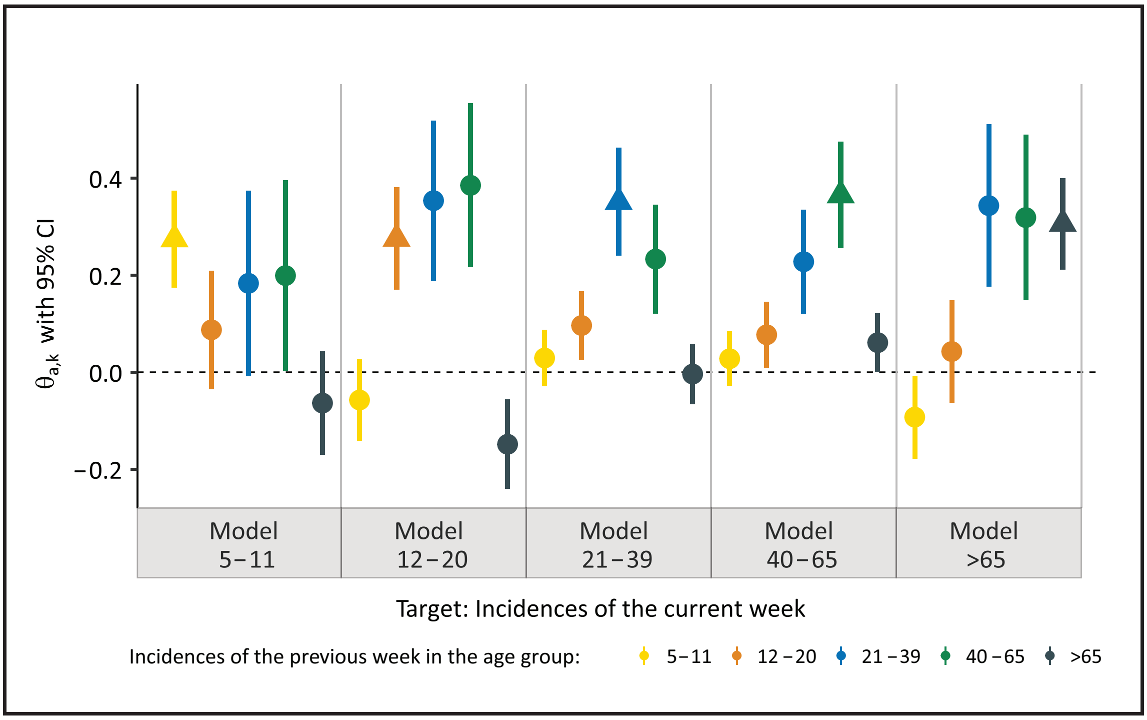


Figure 2 Association of previous week's incidences in different age groups (colour-coded) with the current-week incidences for calendar weeks 9–12 in 2021 stratified by age group (5 age groups correspond to 5 distinct Models)

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.

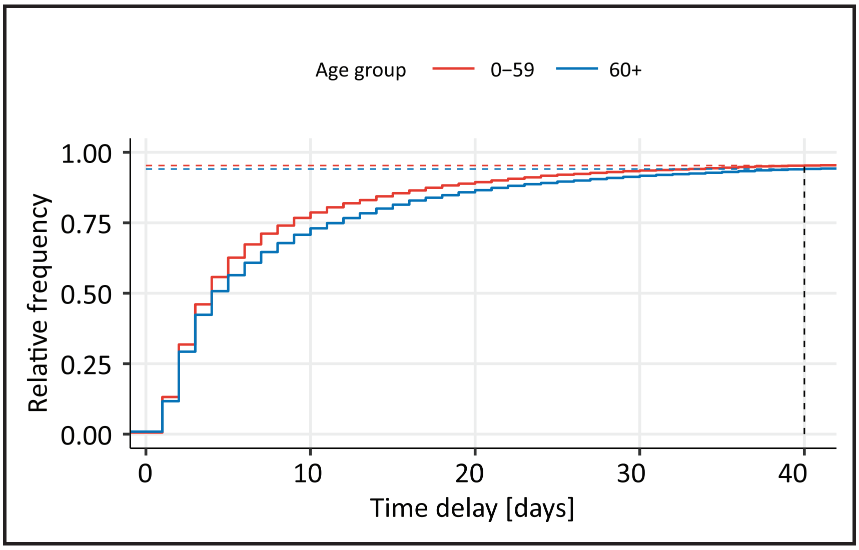


Figure 3 Cumulative distribution function of the time delay (in days) between hospitalization and its reporting, calculated with data from 1 January to 18 November of 2021, shown separately for the age groups 0–59 and 60+. The curves for both age groups are truncated at a delay of 40 days, when approximately 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

$$R_{t,r,g} = \frac{H_{t,r,g}}{x_{\text{pop},r,g}}. \tag{4.1}$$

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 $R_{t,r,g}$

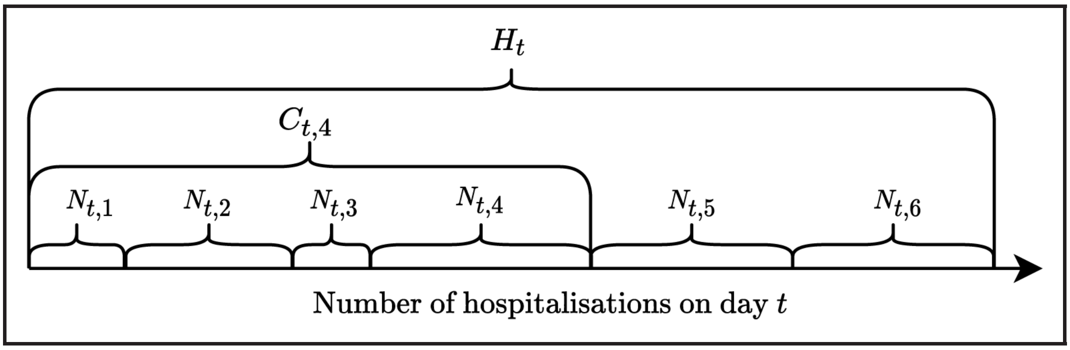


Figure 4 Illustration of the data setting for $d_{\max} = 6$. $N_{t,d}$ indicates hospitalizations reported with a specific delay d, while $C_{t,d}$ denotes all those reported with delay up to d. H_t denotes the final number of hospitalized cases regardless of the delay with which they were reported, 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 $t \in \{1, ..., T\}$ and $d \in \{1, ..., T - t\}$. 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 $N_t = (N_{t,1}, ..., N_{t,d_{\max}}) \in \mathbb{N}^{d_{\max}}$ 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:

$$N_t \sim \text{Multinomial}(H_t, \pi_t), \tag{4.2}$$

where $\pi_t = (\mathbb{P}(D_t = 1; t), ..., \mathbb{P}(D_t = d_{\max}; t))$ 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

$$p_t(d|x_t) := \mathbb{P}(D_t = d | D_t \leq d; x_t), \tag{4.3}$$

conditional on covariates x_t. It follows that the cumulative distribution function of D can be written as:

$$\begin{aligned} F_t(d|x_t) &= \mathbb{P}(D_t \leq d; x_{t,a}) \\ &= \mathbb{P}(D_t \leq d | D_t \leq d+1; x_t)\mathbb{P}(D_t \leq d+1; x_t) \\ &= \prod_{k=d}^{d_{\max}-1} \mathbb{P}(D_t \leq k | D_t \leq k+1; x_t) \\ &= \prod_{k=d}^{d_{\max}-1} (1 - \mathbb{P}(D_t = k+1 | D_t \leq k+1; x_t)) \\ &= \prod_{k=d+1}^{d_{\max}} (1 - \mathbb{P}(D_t = k | D_t \leq k; x_t)) \\ &= \prod_{k=d+1}^{d_{\max}} (1 - p_t(k|x_t)). \end{aligned} \tag{4.4}$$

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

$$\begin{aligned}\mathbb{E}(R_{t,r,g}|x_{t,r,g}) = \exp\{&\theta_0 + \theta_{\text{age}} x_{\text{age},g} + \theta_{\text{gender}} x_{\text{gender},g} + \theta_{\text{gender:age}} x_{\text{age},g} x_{\text{gender},g} + \\ &\theta_{\text{weekday}} x_{\text{weekday},t} + s_1(t) + s_2(x_{\text{Lon},r}, x_{\text{Lat},r}) + u_r\} \\ = \exp\{&\eta_{t,r,g}\},\end{aligned} \tag{4.8}$$

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 $t > T - d_{\max}$, 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

$$\mathbb{E}(R_{t,r,g}|x_{t,r,g}) = \frac{\mathbb{E}(C_{t,r,g,T-t}|x_{t,r,g})}{x_{\text{pop},r,g} F_{t,g}(T - t|x_{t,g})}, \tag{4.9}$$

where we set $C_{t,r,g,T-t} = H_{t,r,g}$ if $T - t \geq d_{\max}$. Rearranging (4.9) shows that modelling the count variable $C_{t,r,g,T-d}$ 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

$$\mathbb{E}(C_{t,r,g,T-t}|x_{t,r,g}) = \exp\left\{\eta_{t,r,g} + \log(x_{\text{pop},r,g} F_{t,g}(T - t|x_{t,g}))\right\} = \mu_{t,r,g} \tag{4.10}$$

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.

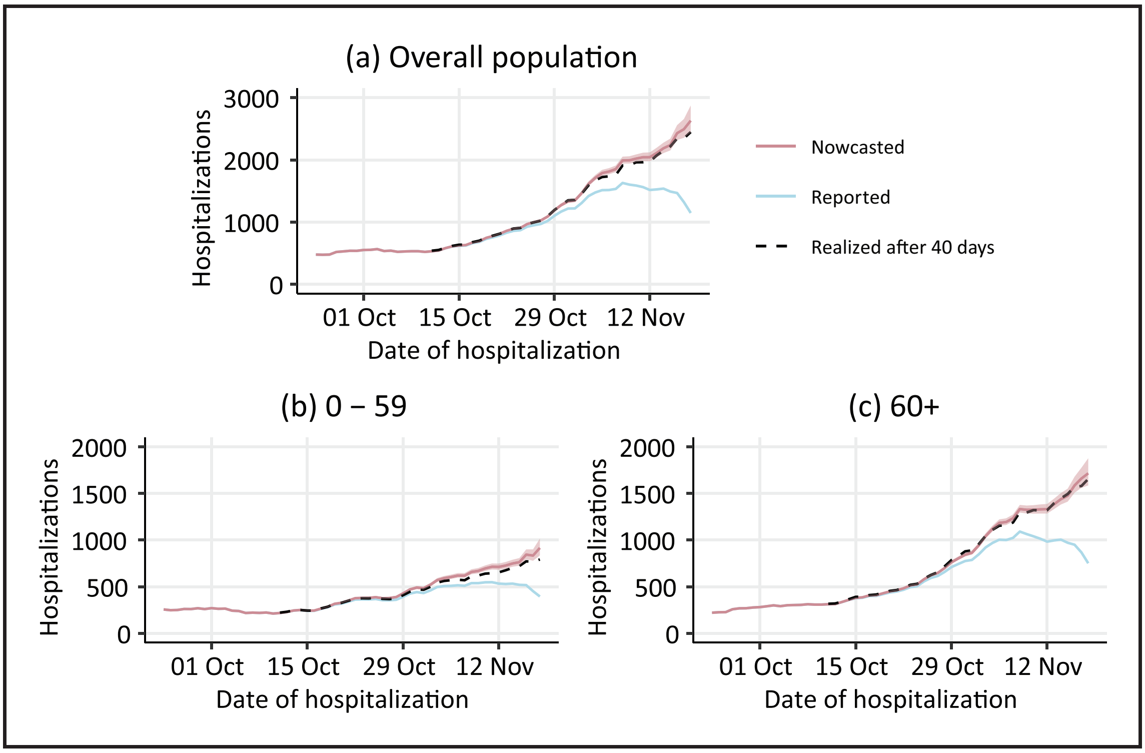
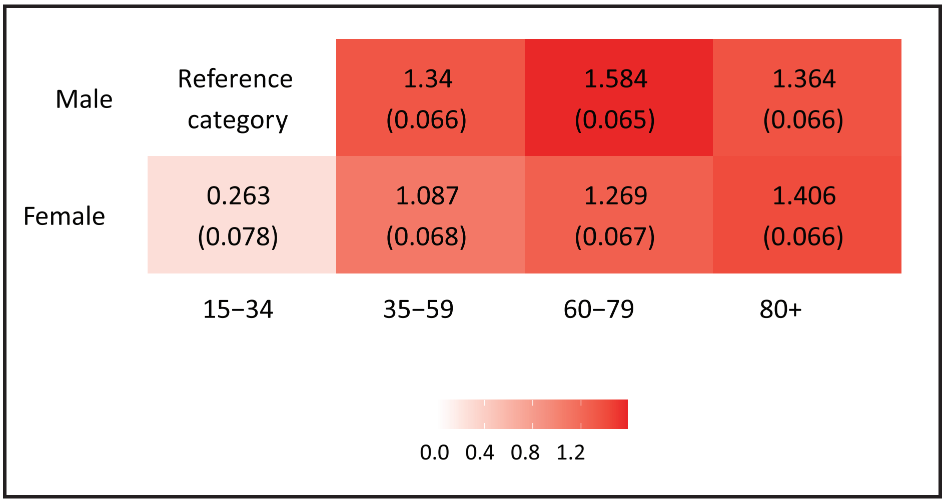


Figure 5 Comparison of nowcasted (red) and reported (blue) rolling weekly sums of hospitalization counts between 24 September and 18 November 2021, based on data reported as of 19 November 2021. Note: 95% confidence intervals of the nowcast estimates are indicated by the shaded areas. The dashed black 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



	15–34	35–59	60–79	80+
Male	Reference category	1.34 (0.066)	1.584 (0.065)	1.364 (0.066)
Female	0.263 (0.078)	1.087 (0.068)	1.269 (0.067)	1.406 (0.066)

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

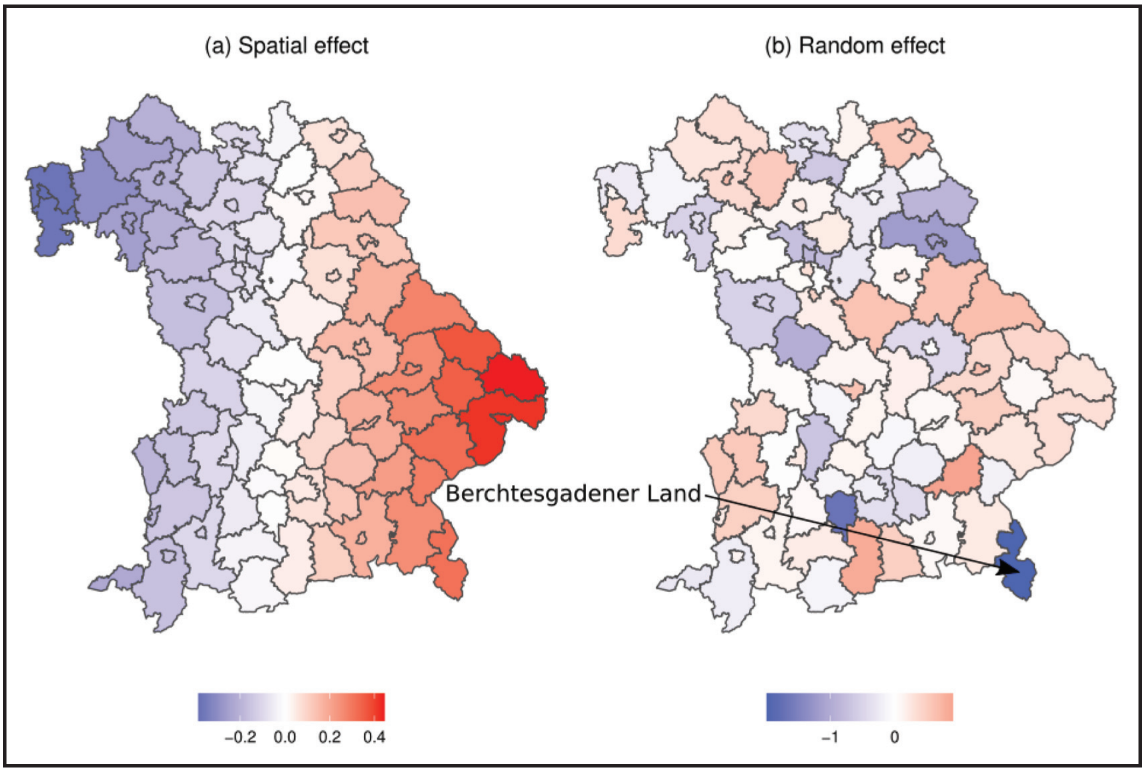


Figure 7 Estimated smooth spatial effect (a) and district-specific random effect (b) in the hospitalization model

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:

$$Z_{w,r} \sim \text{Multinomial}\big(N_{w,r}, \pi_{w,r}\big), \tag{5.1}$$

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 $-N_{w,d}\pi_{w,r,k}\pi_{w,r,l}\ \forall\, k, l \in \{1, 2, 3\}, k \neq l$.

Taking the number of beds occupied by patients infected with COVID-19 as the reference category, we effectively parametrize pairwise comparisons via

$$\log\left(\frac{\pi_{w,r,j}}{\pi_{w,r,3}}\right) = \eta_{w,r,j}\ \forall\ j = 1, 2, \tag{5.2}$$

where the linear predictors $\eta_{w,r,j}$ are functions of covariates labeled as $x_{w,r}$ and defined by:

$$\begin{aligned}\eta_{w,r,j} =&\theta_{0,j} + \theta_{AR(1),j}^{\top}(\tilde{Z}_{w-1,r,1}, \tilde{Z}_{w-1,r,2})^{\top} + \theta_{I,j}^{\top} \log(Y_{w-1,r} + \delta)+ \\ &s_j(x_{\text{Lon},r}, x_{\text{Lat},r}) + u_{r,j} \; \forall \; j = 1, 2, \end{aligned} \tag{5.3}$$

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 jth 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 $\theta_{I,j} \; \forall \; j = 1, 2$. To control for district-specific heterogeneity, we include Gaussian random effects, that is, $u_{r,j} \sim N(0, \tau^2) \; \forall \; r \in \{1, \dots, R\} \; \forall \; j = 1, 2$. For smooth spatial deviations from these random effects, we add a bivariate function $s_j(\cdot, \cdot) \; \forall j = 1, 2$ 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 patents 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:

$$A_{w,r} \sim \text{Multinomial}\big(\alpha N_{w,r}, \pi_{w,r}\big), \tag{5.4}$$

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,

$$f_\pi(\pi_{w,r}) \propto \prod_{j=1}^{3} \pi_{w,r,j}^{(1-\alpha)Z_{w-1,r,j}}. \tag{5.5}$$

Combining the prior (5.5) with the likelihood from (5.4), leads to the posterior

$$f_\pi(\pi_{w,r}|A_{w,d}) \propto \prod_{j=1}^{3} \pi_{w,r,j}^{A_{w,r,j}+(1-\alpha)Z_{w-1,r,j}} = \prod_{j=1}^{3} \pi_{w,r,j}^{Z_{w,r,j}} \tag{5.6}$$

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:

$$\mathbf{V}(\theta) = \mathbf{A}(\theta)^{-1}\mathbf{B}(\theta)\mathbf{A}(\theta)^{-1}, \tag{5.7}$$

where we set $\mathbf{A}(\theta) = \mathbb{E}\left(-\frac{\partial}{\partial\theta\partial^\top\theta}\ell(\theta)\right)$, $\mathbf{B}(\theta) = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\ell(\theta)\right)$, 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:

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\sqrt{\frac{1}{n}\sum_j^n (x_j - \bar{x})^2}} \quad \text{with } \bar{x} = \frac{\sum_j^n x_j}{n}. \tag{5.8}$$

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

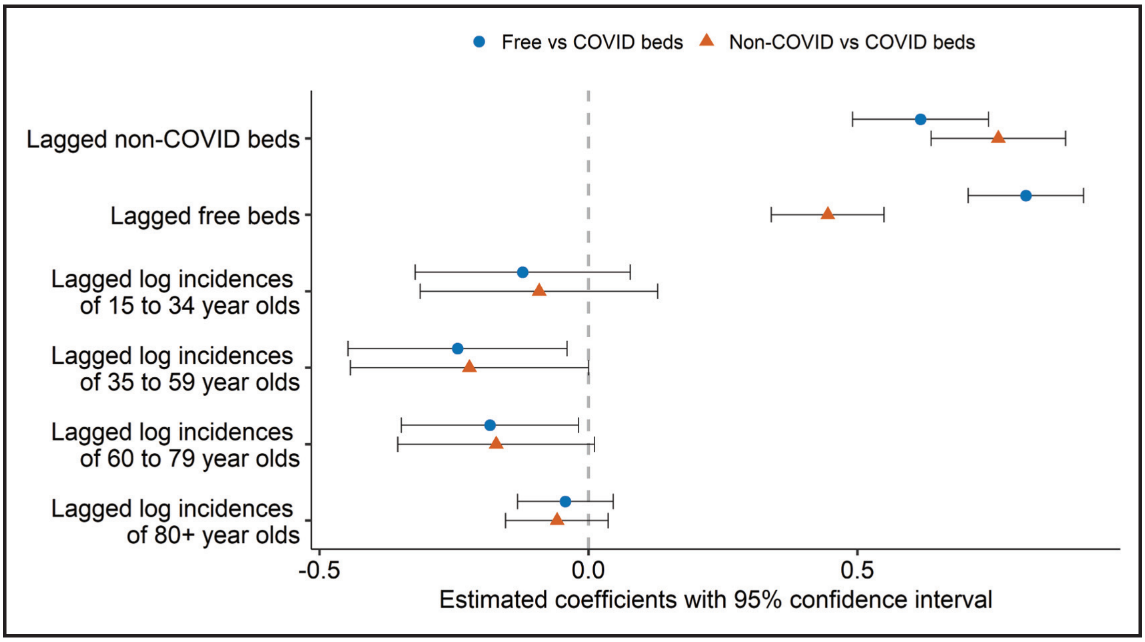


Figure 8 Estimated coefficients with confidence interval of the associations between normalized linear covariates included in the multinomial model and the logarithmic odds of a bed being free vs occupied by a patient infected with COVID-19 (blue dots) and the logarithmic odds of a bed being occupied 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 information 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 Kassteele 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).

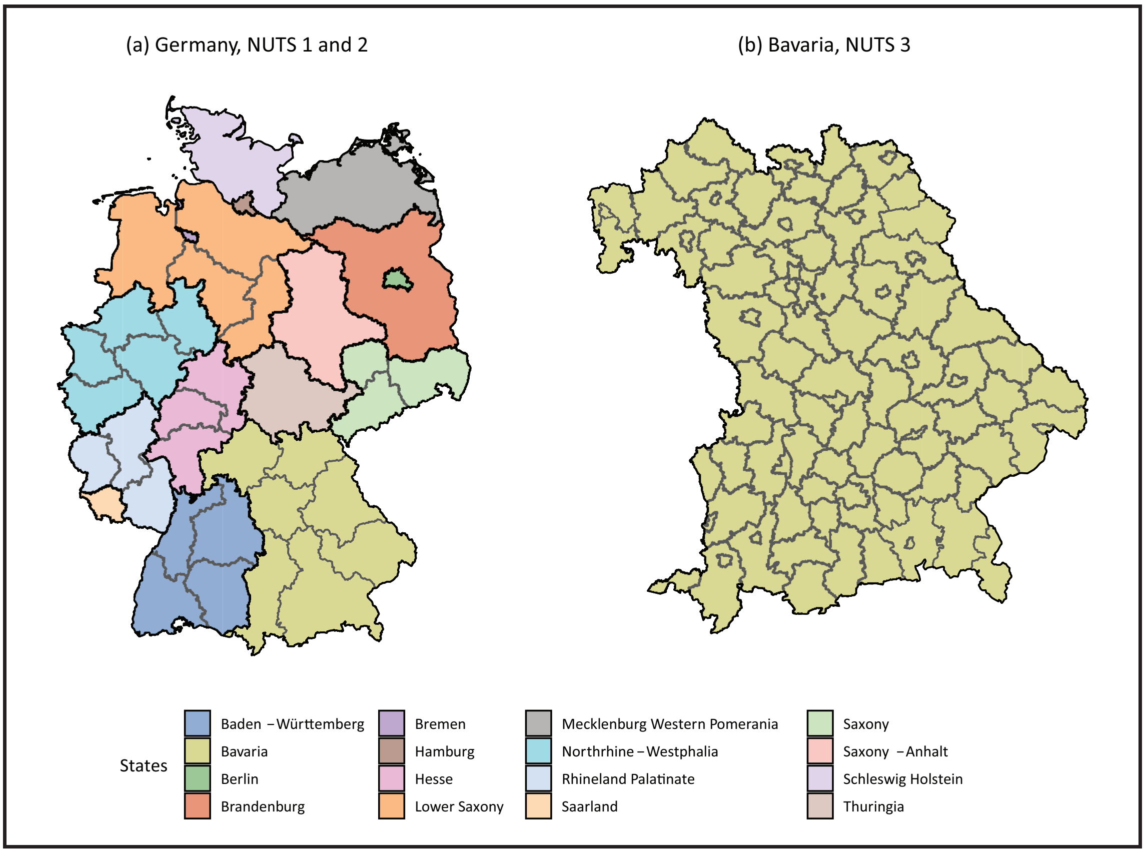


Figure A.1 (a): Map of Germany, where the NUTS 1 regions are indicated by the black borders and the different colours. The NUTS 2 regions, on the other hand, are drawn in grey. Note that all NUTS 1 region borders are also NUTS 2 region borders. (b): Map of Bavaria where also the NUTS 3 regions are marked. 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.

Bańbura 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 connectedness, 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, Sifft 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 Kassteele 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, Eddleston 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 III.

Unobserved inflow and outflow of ICU patients with COVID-19

6. The Skellam distribution revisited: Estimating the unobserved incoming and outgoing ICU COVID-19 patients on a regional level in Germany

Contributing article

Martje Rave and Göran Kauermann (2024). The Skellam distribution revisited: Estimating the unobserved incoming and outgoing ICU COVID-19 patients on a regional level in Germany. *Statistical Modelling: 2024;25(3):270-280.* `https://doi.org/10.1177/1471082X241235024`.

Data and code

Available at https://github.com/MartjeRave/The-Skellam-Distribution-Revisited.

Copyright information

Author contributions

Prof. Dr Göran Kauermann conceived the research approach. The statistical methodology was jointly designed by Kauermann and Martje Rave, and subsequently implemented by Rave. All results were computed and validated by Rave, who also prepared the figures, tables, and supplementary materials, and authored the full manuscript. Input and feedback from Kauermann were continuously incorporated throughout the process. The review process was led and coordinated by Rave, with valuable contributions from Kauermann.

Statistical Modelling 2025; **25(3)**: 270–288

The Skellam distribution revisited: Estimating the unobserved incoming and outgoing ICU COVID-19 patients on a regional level in Germany

Martje Rave[1] and Göran Kauermann[1]

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

Abstract: With the beginning of the COVID-19 pandemic, we became aware of the need for comprehensive data collection and its provision to scientists and experts for proper data analyses. In Germany, the Robert Koch Institute (RKI) has tried to keep up with this demand for data on COVID-19, but there were (and still are) relevant data missing that are needed to understand the whole picture of the pandemic. In this article, we take a closer look at the severity of the course of COVID-19 in Germany, for which ideal information would be the number of incoming patients to ICU units. This information was (and still is) not available. Instead, the current occupancy of ICU units on the district level was reported daily. We demonstrate how this information can be used to predict the number of incoming as well as released COVID-19 patients using a stochastic version of the Expectation Maximization algorithm (SEM). This, in turn, allows for estimating the influence of district-specific and age-specific infection rates as well as further covariates, including spatial effects, on the number of incoming patients. The article demonstrates that even if relevant data are not recorded or provided officially, statistical modelling allows for reconstructing them. This also includes the quantification of uncertainty which naturally results from the application of the SEM algorithm.

Key words: EM, Skellam distribution, stochastic EM, imputation, COVID-19, ICU patients

Received April 2023; **revised** October 2023; **accepted** January 2024

1 Introduction

Albeit its atrocity, in its aftermath, the COVID-19 pandemic has taught Germany, among many other countries, the shortcomings of inadequate data availability in its healthcare system. In fact, in Germany, while intensive care unit (ICU) occupancy was provided by the DIVI e.V. (2021), the numbers of newly hospitalized patients (incoming) and released patients (outgoing), either cured or deceased, has (until now) not been included in the database. This can be criticized since a relevant number, which measures the pressure of the disease on the healthcare system—the number of incoming patients—is not available to the public. We show, in this article, how to disentangle incoming and

Address for correspondence: Martje Rave, Chair of Applied Statistics in Social Sciences, Economics and Business, Department of Statistics, Faculty of Mathematics, Informatics and Statistics, Ludwig-Maximilians-Universität München, Ludwigstr. 33 80539 München Germany
E-mail: martje.rave@stat.uni-muenchen.de

10.1177/1471082X241235024

outgoing patients from pure occupancy data using statistical models. This, in particular, allows us to investigate how hospitalizations depend on time, age, and spatial factors.

We assume that admission to- and release of the ICU units follow Poisson distributions with inhomogeneous intensities. Consequently, the changes in ICU occupancy result from the difference between incoming and outgoing patients. This in turn gives the framework of the Skellam distribution, originally introduced by Skellam (1948). The distribution is described as resulting from the difference of two independent Poisson distributed random variables. This distributional approach has been used in different settings. For instance, in sports statistics Karlis and Ntzoufras (2009) apply the distribution for modelling the goal difference in football games. In network analysis, Gan and Kolaczyk (2018) and Schneble and Kauermann (2022) look at network flows while Koopman et al. (2014) utilize the idea to model financial trades. Further application areas include image analysis when comparing intensity differences of pixels, see for example, Hwang et al. (2007), Hwang et al. (2011) or Hirakawa et al. (2009). Extensions towards bivariate Skellam processes are provided for example, in Genest and Mesfioui (2014), see also Aissaoui et al. (2017). A general discussion on the Skellam distribution and its application fields is provided in Tomy and Veena (2022). In this article, we provide an application of the Skellam distribution for disentangling incoming and outgoing patients in ICUs.

The occupancy of ICU units was a central component of the COVID-19 pandemic. Numerous tools have been developed for forecasting the number of patients who require ICU admission, see for example, Grasselli et al. (2020), Goic et al. (2021), Murray (2020) or Farcomeni et al. (2021) to just mention a few. Our focus in this article is not primarily on prediction but on investigating the risk of admission and how this depends on the infection rates and further covariates, including spatial components. To do so we assume that the number of incoming and released patients comes from an inhomogeneous Poisson process, but we only observe the difference between incoming and released patients, leading to a Skellam distribution. Treating incoming and released patients as missing data, allows us to simulate the patient flows (stochastic E step) and refit the model (M step). Parameter estimation in the Skellam distribution is cumbersome due to its numerically complex form of the likelihood function, which requires the use of the Bessel function. Even though these are implemented in standard software packages, we refer to Schneble and Kauermann (2022), who report some numerical instabilities in the case of parameters at the boundary of the parameter space. We also refer to Lewis et al. (2016) or Aissaoui et al. (2017) who pursue moment-based estimation. In this article, we aim to use implemented routines to achieve stability. In fact, the data can be rewritten as a missing data constellation, which itself suggests the use of an EM algorithm. We here use the Stochastic Expectation Maximization (SEM) algorithm and present it as a suitable and numerically stable alternative to available estimation routines. Originally proposed by Celeux et al. (1996), the stochastic version of the EM algorithm gained interest in recent years, in particular in mixture models, see for example, Noghrehchi et al. (2021). We also refer to Nielsen (2000) for asymptotic results on the algorithm. The EM algorithm relates the estimation to a missing data problem, which is easily described. We assume that instead of the complete data with incoming and outgoing patients, we only observe the changes in occupancy of ICUs. In other words, the exact number of incoming and outgoing is missing. Replacing these missing numbers iteratively with simulated numbers, based on the current estimates of the model, provides the stochastic version of the 'E'-step. This, in turn, leads to full data, which allows for standard maximum likelihood estimation of two Poisson processes—the M step. The algorithm is easily implemented, and Rubin's rule, Rubin (1976), provides inference statements.

A particularly interesting attribute that this approach provides is the simplification of the initial complexity of the problem. We are able to break our problem down from a fairly complex distributional assumption, with respect to deriving an association between the infection rates and the number of incoming and outgoing patients, to land at essentially two iteratively updated generalized additive models (GAMs) with simulated responses, each response simultaneously sampled from a joint distribution, comprised of the product of two Poisson distributions. This allows us to not only circumvent rather cumbersome calculations and modifications of the first and second derivative of the Skellam distribution, as, for example, shown by Schneble and Kauermann (2022) but also almost effortlessly interpret the association between the number of incoming and outgoing patients and the infection rate.

The article is structured as follows; in Section 2, we give a detailed data description. In Section 3, we elaborate on the model approach to our problem, while in Section 4, we will provide the results of our model approach. A simple simulation exercise to validate our findings can be found in Section 5, and in Section 6, we conclude our article which also includes a discussion of the shortcomings of our approach.

2 Data description

The database for our analyses consists of two main components; data on COVID-19 infections and data on the ICU occupancy of COVID-19 patients. The infections and the ICU occupancy are collected by the German health care departments, recorded by the Robert Koch Institute (2021) (RKI), the German federal government agency and scientific institute responsible for health reporting and disease control, and published by the RKI and DIVI e.V. (2021), respectively. We here focus on data during the fourth infection wave in Germany, that is, from the 2nd October 2021 until the 17th November 2021, though the method is readily extendable to other time frames, so long that the data included are subject to homogeneous testing or lock down strategies. We visualize the average infection rates over all districts in Figure 1 (left-hand side).

The RKI collects and publishes data on infections on a daily basis. Due to privacy protection, the RKI aggregates the number of COVID-19 patients, ICU occupancy and general hospital admission of patients infected with COVID-19 over NUTS3 districts, European Commission (2021), but separates by demographic groups. These namely are the age categories; ‘0–4’ year-olds, ‘5–14’ year-olds, ‘15–34’ year-olds, ‘35–59’ year-olds, ‘60–79’ year-olds and ‘80+’ year-olds and the sex; ‘male’, ‘female’ and ‘not disclosed’. For the purpose of this analysis, the infections are aggregated over the age groups. The data were directly downloaded through the ArcGIS website, Robert Koch Institute (2021). The infection rates per 100.000 inhabitants are then calculated as a weekly average for each age group. For each district, the infection rate is averaged over the seven days immediately preceding the respective observed day change in ICU occupancy.

The data on ICU occupancy is also collected by the RKI and published by DIVI. These data are also on a district level, however, the occupancy can only be differentiated by the number of beds occupied by patients infected with COVID-19, by the number of beds occupied by patients not infected with COVID-19 and the number of empty beds, the sum of which is the overall ICU capacity in a given district on a given date. We solely take the COVID-19 ICU-patients into account and visualize the ICU data for one day in Figure 1 (right-hand side).

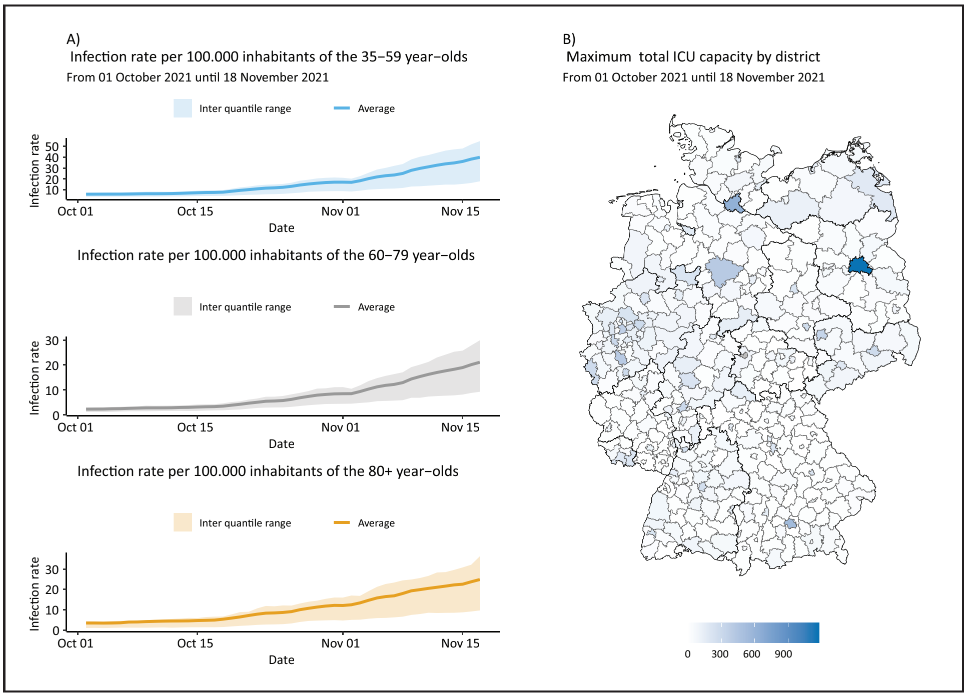


Figure 1 A: Summary over all districts of the infection rate per 100.000 inhabitants by age group, '45–59' year-olds, '60–79' year-olds and '80+' year-olds displayed by date, from the 1 October 2021 until the 18 November 2021, B: The maximum capacity of ICU beds per given district over the time span from the 1 October 2021 until the 18 November 2021 by district.

Conveniently, both data sets can also be found in the daily updated GitHub repository maintained by the RKI, Robert Koch Institute (2023). We take a closer look at the infection rates by age group in the Supplemental Material.

3 Model

3.1 Assumption

Let $Y_{(t,d)}$ be the number of COVID-19 ICU patients in a given district d at day t. This is the official number issued by DIVI, described above and freely accessible from the given sources. We define with $I_{(t,d)}$ the number of incoming patients in district d at day t, which is the number of newly admitted COVID-19 patients in the ICUs located in district d. Accordingly, we denote with $R_{(t,d)}$ the number of released patients, meaning that they are discharged, deceased or transferred to a non-ICU. We assume both to come from an inhomogeneous Poisson process such that

$$I_{(t,d)} \sim \text{Poisson}\left(\lambda^I_{(t,d)}\right) \tag{3.1}$$

$$R_{(t,d)} \sim \text{Poisson}\left(\lambda^R_{(t,d)}\right). \tag{3.2}$$

The explicit modelling of the intensities $\lambda^{I}_{(t,d)}$ and $\lambda^{R}_{(t,d)}$ is of primary interest and discussed in depth later in this section. For now, note that Equation (3.2) is an approximation, and formally we have a right censored Poisson distribution with $R_{(t,d)} \leq Y_{(t-1,d)}$ since no more patients can be released than are currently in the ICU. We can omit this point, though, since, based on the disease, we know that not all patients are discharged at a time, so the formal censoring does not play any practical relevance due to a generally small discharge intensity $\lambda_{(t,d)}$.

With these definitions, we can now define the difference $\Delta_{(t,d)}$ in occupancy of COVID-19 ICU patients per district d and day t to the previous day $t-1$.

$$\Delta_{(t,d)} = Y_{(t,d)} - Y_{(t-1,d)} = I_{(t,d)} - R_{(t,d)}. \tag{3.3}$$

Assuming independence for the number of incoming and outgoing ICU COVID-19 patients together with (3.1) and (3.2) leads to a Skellam distribution Skellam (1948).

$$\Delta_{(t,d)} \sim \text{Skellam}(\lambda^{I}_{(t,d)}, \lambda^{R}_{(t,d)}). \tag{3.4}$$

Before we derive how to estimate the two intensities in (3.4) we want to discuss the suitability of the distributional assumptions. Note that the approach relies on independence of $I_{(t,d)}$ and $R_{(t,d)}$. This would be violated if discharges of the ICU in t depend on the number of incoming patients in t. A conceivable scenario where $I_{(t,d)}$ and $R_{(t,d)}$ are dependent results if the ICUs get to their limit capacity and triage of patients is inevitable. This situation has not been observed in Germany—over the entire course of the pandemic—so we can argue that assuming independence between incoming and outgoing patients is reasonable.

There was, however, relocation of patients if local hospitals reached the edge of capacity. This followed a national plan, called 'Kleeblattkonzept', literally translated as clover-leaf-concept, see Pfenninger et al. (2022). This also implies that some ICU patients are not local.

We also want to add a comment given by the referee, in that a Skellam distribution also results in a more general setup. Assume that we have noisy data in that incoming and released patients have an additive shift. That is instead of $I_{(t,d)}$ we have $\tilde{I}_{(t,d)} = I_{(t,d)} + Z_{(t,d)}$ and analogously $R_{(t,d)}$ becomes $\tilde{R}_{(t,d)} = R_{(t,d)} + Z_{(t,d)}$ where $Z_{(t,d)}$ is some discrete noise. Apparently, now $\tilde{I}_{(t,d)}$ and $\tilde{R}_{(t,d)}$ are not any longer independent, but their difference like in (3.3) is again Skellam distributed. Hence, we can slightly weaken the independence assumption if we assume additive noise on incoming and released patient counts.

Finally, the intensities $\lambda^{I}_{(t,d)}$ and $\lambda^{R}_{(t,d)}$ are modelled to depend on a set of covariates denoted by $\mathbf{x}_{(t,d)}$ as well as previous data. To be specific, we set

$$\lambda^{I}_{(t,d)} = \exp\left(\eta^{I}_{(t,d)} + s^{I}(t) + h^{I}(\text{longitude}_{d}, \text{latitude}_{d})\right), \tag{3.5}$$

$$\lambda^{R}_{(t,d)} = \exp\left(\eta^{R}_{(t,d)} + s^{R}(t) + h^{R}(\text{longitude}_{d}, \text{latitude}_{d}) + \underbrace{\log(\sum_{j=t-56}^{t} \omega_j \hat{I}_{(j,d)})}_{= \text{offset}}\right), \tag{3.6}$$

where $\eta^I_{(t,d)}$ and $\eta^R_{(t,d)}$ are the linear combinations of the covariates included in the models. Namely, the logged infection rates of the age groups '35–59' year-olds, '60–79' year-olds and '80+' year-olds, as well as the weekday, included as a categorical variable, with Friday as its reference category. $s^I(t)$ and $s^R(t)$ are smooth functions in time, and $h^I(\text{longitude}_d, \text{latitude}_d)$ and $h^R(\text{longitude}_d, \text{latitude}_d)$ are two-dimensional thin-plate smooth functions over the coordinates of the centroids of the respective districts, Wood (2003). Note that $\hat{I}_{(j,d)}$ is not observed, and we, therefore, replace it with its simulated value from the 'E'-step. Moreover, the weights ω_j are fixed and not estimated but instead obtained from duration time models for COVID-19 patients in ICU units. We make use of the epidemiological bulletin published by the RKI in 2020, Tolksdorf et al. (2020), see Figure A1 in the Appendix. The maximum length of stay is set to 56 days, which explains the number in the formula above.

Finally, we impose the standard identifiability constraints, that is, that both $s^I(t)$ and $s^R(t)$ as well as the spatial effects $h^I(\text{longitude}_d, \text{latitude}_d)$ and $h^R(\text{longitude}_d, \text{latitude}_d)$ integrate out to zero. We refer to Wood (2017) for more details.

3.2 SEM algorithm

Instead of maximizing the Skellam likelihood, as done for instance in Schneble and Kauermann (2022), we pursue an EM algorithm, with the E-step replaced by a simulation step, leading to the stochastic EM algorithm, as discussed in Celeux et al. (1996). The approach has the advantage, that estimation can be carried out iteratively using implemented procedures and, even more importantly, we directly obtain predicted values for the incoming and released patients, which are the quantities of interest. Note that we observe $\Delta_{(t,d)}$ from which we can 'calculate' $I_{(t,d)}$ and $R_{(t,d)}$. Given the data we have

$$I_{(t,d)} = \Delta_{(t,d)} + R_{(t,d)} \tag{3.7}$$

with the additional constraints that both, $I_{(t,d)} \geq 0$ and $R_{(t,d)} \geq 0$. Hence, based on the data, we have the joint probability model for incoming and released ICU patients:

$$\begin{aligned} & P(I_{(t,d)} = k,\, R_{(t,d)} = j \mid \Delta_{(t,d)} = \delta) \\ & \propto \begin{cases} P(I_{(t,d)} = k) \times P(R_{(t,d)} = k - \delta) & \text{for } j = k - \delta \text{ and } k \geq \max(\delta, 0) \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{3.8}$$

with $P(I_{(t,d)} = k)$ and $P(R_{(t,d)} = k - \delta)$ resulting from the Poisson model (3.5) and (3.6), respectively. While model (3.8) is a clumsy convolution model which does not simplify to an analytic form. Simulation from the model is simple by just replacing the infinite pairs k for $I_{(t,d)}$ and $k-\delta$ for $R_{(t,d)}$ by a set of finite pairs, such that the resulting cumulative probabilities are approximately equal to one. To be specific, we have

$$\begin{aligned} & P(I_{(t,d)} = k,\, R_{(t,d)} = j \mid \Delta_{(t,d)} = \delta, \lambda^I, \lambda^R) \\ & \qquad = \lim_{K\to\infty} \frac{P_{\lambda^I}(I_{(t,d)} = k) P_{\lambda^R}(R_{(t,d)} = k - \delta)}{\sum_{i=1}^K P_{\lambda^I}(I_{(t,d)} = i) P_{\lambda^R}(R_{(t,d)} = i - \delta)}. \end{aligned} \tag{3.9}$$

We approximate this numerically by assuming that either $P_{\lambda^I}((I_{(t,d)}) = k)$, or $P_{\lambda^R}(R_{(t,d)} = k - \delta)$ is sufficiently close to zero at $k > 1000$ making the product of the two distributions sufficiently close to zero, such that the sum of probabilities for events $k > 1000$ may be negligible. This results in the finite approximation

$$\begin{aligned} P(I_{(t,d)} = k, R_{(t,d)} = j | \Delta_{(t,d)} = \delta, \lambda^I, \lambda^R) \\ \approx \frac{P_{\lambda^I}(I_{(t,d)} = k) P_{\lambda^R}(R_{(t,d)} = k - \delta)}{\sum_{i=1}^{1000} P_{\lambda^I}(I_{(t,d)} = i) P_{\lambda^R}(R_{(t,d)} = i - \delta)}. \end{aligned} \tag{3.10}$$

Numerically this is easily carried out and allows to simulate data pairs $(I^*_{(t,d)}, R^*_{(t,d)})$ based on the current estimates of the intensities using (3.10) as an approximate version of (3.8). This provides a stochastic ‘E’-step and leads to a full data set with (simulated) incoming and (simulated) released patients for all districts and all time points. With the resulting (simulated) full data set, we can now directly estimate the intensities in models (3.5) and (3.6), which in turn is conducted in the ‘M’ step. The ‘M’ step can be carried out by fitting two generalized additive Poisson models using standard software, see Wood (2017).

Iterating between the two steps gives a stochastic version of the EM algorithm. Each simulation step provides an estimate, and following the classical EM algorithm, we can easily see that on average, we increase the (marginal) likelihood in each step. The outline of which is sketched in Figure A4, in the Appendix.

The results of the model which simulates from the joint probability distribution with $K = 2.000$, instead of $K = 1.000$, are shown in the Supplemental Material.

3.3 Inference based on SEM

Unlike the EM algorithm, where calculating the variance of the estimates is not straightforward, and one typically relies on Louis’ formula Louis (1982), the stochastic version allows to take the uncertainty due to the missing data into account. The derivation shows similarities to Rubin’s formula for imputation, see Rubin (1976). Let the parameter vector of linear and smooth functions, $\hat{\boldsymbol{\beta}}^{(k)} = (\hat{\boldsymbol{\beta}}^{I(k)^T}, \hat{\boldsymbol{\beta}}^{R(k)^T})^T$, be the resulting estimate in the k^{th} step of the SEM algorithm. We assume $k > k_0$, where k_0 refers to the step when convergence seems to be achieved. The final estimate results through

$$\hat{\boldsymbol{\beta}} = \frac{1}{K - k_0} \sum_{k=k_0+1}^{K} \hat{\boldsymbol{\beta}}^{(k)}. \tag{3.11}$$

The variance is estimated via

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \frac{1}{K - k_0} \sum_{k=k_0+1}^{K} \widehat{Var}(\hat{\boldsymbol{\beta}}^{(k)}) + \frac{1 + (K - k_0)^{-1}}{(K - k_0) - 1} \sum_{j=k_0+1}^{K} (\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^{(k)} - \hat{\boldsymbol{\beta}})^T \tag{3.12}$$

where $\widehat{Var}(\hat{\boldsymbol{\beta}}^{(k)})$ is the variance estimate in the k iteration step, based on the imputed data set. The latter directly results through the applied fitting algorithm.

4 Results

A great advantage of our approach is that we can directly interpret the estimated association between the included covariates and the incoming patients and outgoing patients separately. To do so, we look at covariates containing information on the infection rates for each of the three age groups and the weekday effects. The estimated coefficients and their standard deviation, calculated based on Rubin's formula, see Equation (3.12), are provided in Table 1. We use the last 300 runs to determine the coefficient estimates through their median, as well as their variance through the Equation (3.12). The estimates over the last 300 runs are shown through line plots in Figures A2 and A3 in the Appendix for the incoming and outgoing patients, respectively. We include extensions to the runs included in the analysis in the Supplemental Material. We find, however, that the inclusion of more runs will not result in a change in the estimated coefficients.

First, we look at the association between our covariates and the number of incoming and outgoing patients, as seen in the middle and right column of the output table, Table 1. Recall that the weekday effect is included in the model through a categorical variable, with Friday as its reference category. For the model estimating the number of incoming patients, keeping respectively all other variables constant, we can observe that there is an increased number of incoming patients on other weekdays, compared to Friday, whereas on the weekend, there is a decreased number of patients, compared to Friday. For outgoing patients, the behaviour is slightly different. On Monday, Thursday, Saturday, and especially Sunday, fewer patients are released compared to Friday. Conversely, Tuesday and Wednesday seem slightly increased.

The number of incoming and outgoing patients is positively associated with the infection rates of all age groups. Notably, the strongest effect exists for the infection rate of '35–59' year-olds. This is interesting, bearing in mind that '60–79' year-olds are the predominant age group DIV. We should,

Table 1 Estimated coefficients and standard deviations presented on the level of incoming and outgoing patients. The estimates are the exponential of the median of the coefficient estimates from the 200th run to the 500th run of the EM algorithm.

	Incoming		Outgoing	
	Estimates	Std. Err.	Estimates	Std. Err.
Intercept	−2.28	0.10	−6.41	0.12
Monday effect	0.12	0.05	−0.21	0.06
Tuesday effect	0.14	0.05	0.03	0.06
Wednesday effect	0.13	0.05	0.02	0.06
Thursday effect	0.14	0.05	−0.10	0.06
Saturday effect	−0.02	0.05	−0.09	0.06
Sunday effect	−0.14	0.05	−0.39	0.06
Infection 35–59 yo	0.24	0.05	0.28	0.06
Infection 60–79 yo	0.07	0.05	0.07	0.05
Infection 80+ yo	0.11	0.02	0.10	0.02

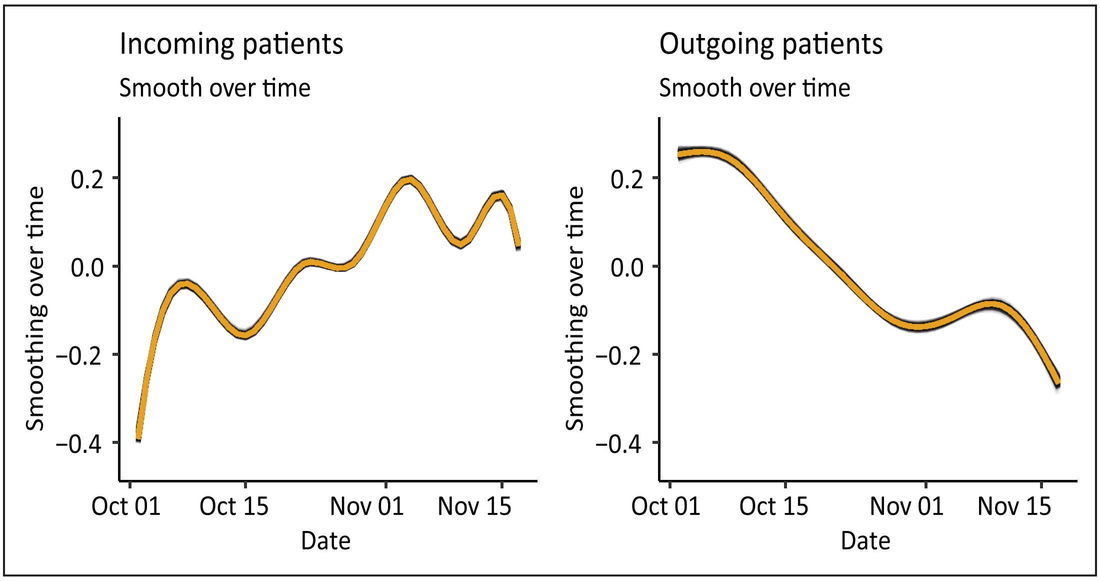


Figure 2 Estimated smooth functions of all runs, over time, rendered by the GAMs estimating the number of incoming patients (left hand side) and outgoing patients (right hand side) over the last 300 runs.

however, not omit that there is strong collinearity between the infection rates themselves which could affect our interpretability of the coefficients. More on the change of coefficients, when we look at different time frames over which the data is observed is discussed in the Supplemental Material.

Recall further, that we included smooth functions to estimate both the spatial-, and the temporal effects. They are included to pick up on additional spatial and temporal structural dependencies. Let us first look at the smooth effects over time, as seen in Figure 2. The averaged smooth function over time for incoming patients (left-hand side) is generally increasing. Evidently, we can see some fluctuation and there seems to be a fortnightly rhythm within the overall trend. Here we observe an increase in the number of incoming patients for the first seven days, then a decrease in the following seven days, followed by a subsequent increase, and so forth. In contrast, as shown on the right-hand side of Figure 2, we see a general decrease in the number of outgoing patients without a biweekly rhythm.

Finally, we look at the spatial effects for the incoming patients, see the left-hand side of Figure 3, and for the outgoing patients, shown with the right-hand side of Figure 3. There seems to be an increased level of incoming patients in Saxony (east Germany) and North Rhine-Westphalia (west Germany) and a slight increase around the larger cities of Germany (Frankfurt, Stuttgart, and Munich, south and southwest of Germany). We observe a similar structure in the spatial smooth function in the model estimating the outgoing patients, except for the strong increase around Saxony. Overall, we see clear spatial heterogeneity.

At last, we visualize in Figure 4 the estimated number of incoming and outgoing patients, summed up over the entirety of Germany, for the observed time frame. The left-hand axis scales the number of incoming and outgoing patients, whereas the right-hand axis scales the number of overall ICU patients with COVID-19. We see that the model picks up the somewhat constant occupancy, from the 1 October 2021, until the 17 October 2021, in Germany's ICUs rather well, where

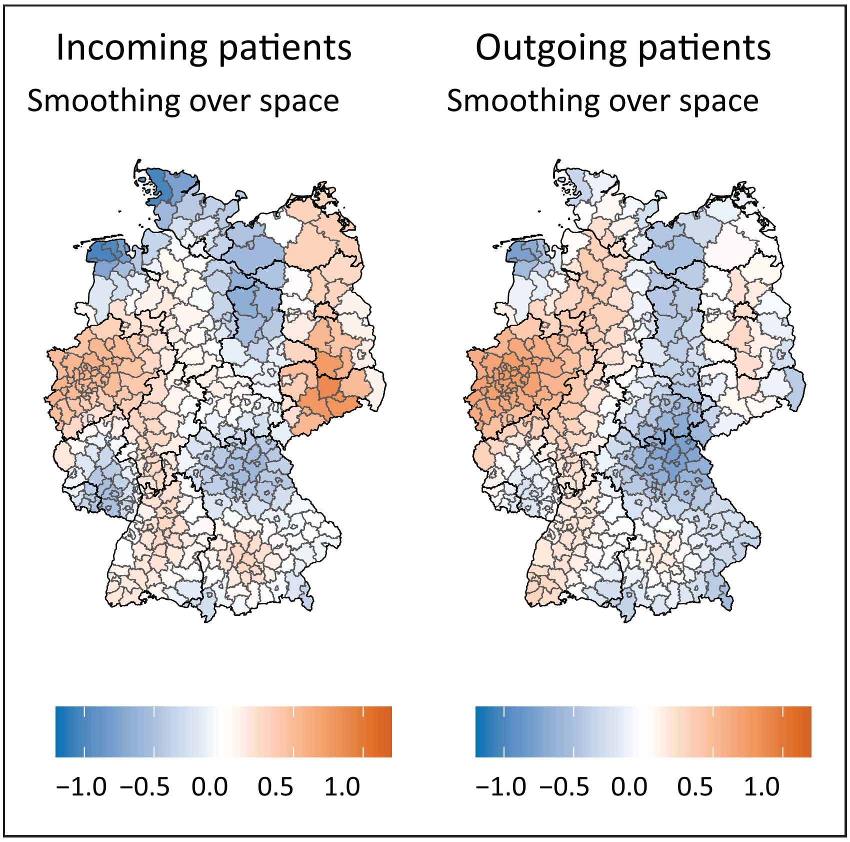


Figure 3 Estimated median smooth functions of the 200th until the 500th run over space rendered by the generalized additive models, estimating the number of incoming patients (left-hand side) and outgoing patients (right-hand side).

the number of incoming and outgoing patients are estimated to be similar, if not equal. Thereafter, the number of ICU patients in the ICU increases, around this time, we also observe a higher estimated number of incoming patients than outgoing patients. It is not unusual for patients, especially the critically ill, to stay in the ICU for more than four weeks, making the divergence in estimation for the number of incoming patients and outgoing patients entirely plausible.

With respect to model validation, we provide some additional analyses in the Supplemental Material of the article. In particular, we look at serial correlation and show that due to the autoregressive component in the model, the Pearson residuals show no autocorrelated structure.

5 Simulation

This section is aimed to investigate the goodness of fit of the modelling approach we chose a simple version to emulate the data used above. We use one covariate, randomly drawn from a normal distribution, whose mean and variance are taken from the observed mean and variance of the logged

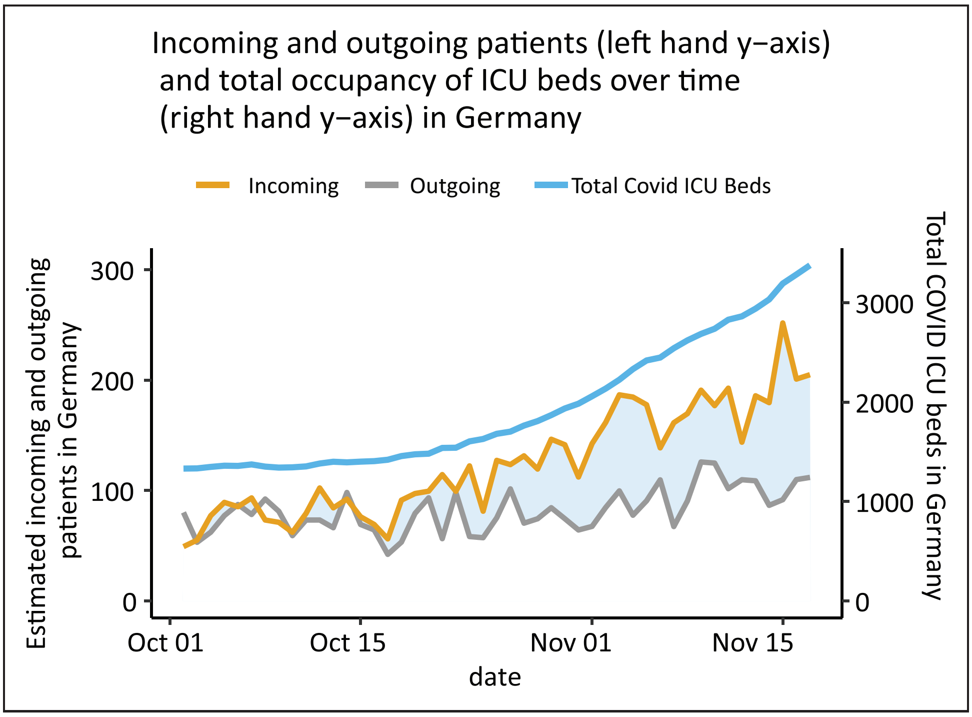


Figure 4 Estimated number of incoming and outgoing patients by date from the 1 October 2021 until the 18 November 2021, as well as the total number of COVID-19 patients in the ICUs of Germany.

infection rates of ‘60–79’ year-olds. We choose this age group, as ‘60–79’ year-olds are the predominant group in the German ICUs during the fourth wave, see Robert Koch-Institut (2023). The coefficients for the simulation are chosen in a way such that the difference in the simulated incoming and outgoing patients is somewhat similar to the range of the difference in the observed incoming and outgoing patients, namely (−24, 20) in the observed data. The incoming and outgoing number of patients are then simulated, outlined in Equation 5.1.

$$I_i \sim Poi(exp(\beta_0^{in} + \beta\beta_1^{in} X_i)), \tag{5.1}$$

$$R_i \sim Poi(exp(\beta_0^{out} + \beta\beta_1^{out} X_i + log(I_{i-1}))), \tag{5.2}$$

$$X_i \sim N(1.978, 1.397), \tag{5.3}$$

$$\forall i \in (1, \dots, 1000). \tag{5.4}$$

Here, β_0^{in} is taken to be −2.340, β_1^{in} is 0.800, β_0^{out} is 0.001 and β_1^{out} is taken to be 0.100. Here, $N(\mu, \sigma)$ refers to the Gaussian distribution with mean μ and standard deviation σ, and $Poi(\lambda)$, refers to the Poisson distribution with intensity parameter λ. . The simulation algorithm is sketched out in Figure A5 and the resulting estimated coefficients of twenty independent runs are shown in Figure 5, where we see that the confidence intervals of each of the coefficient estimates of each of

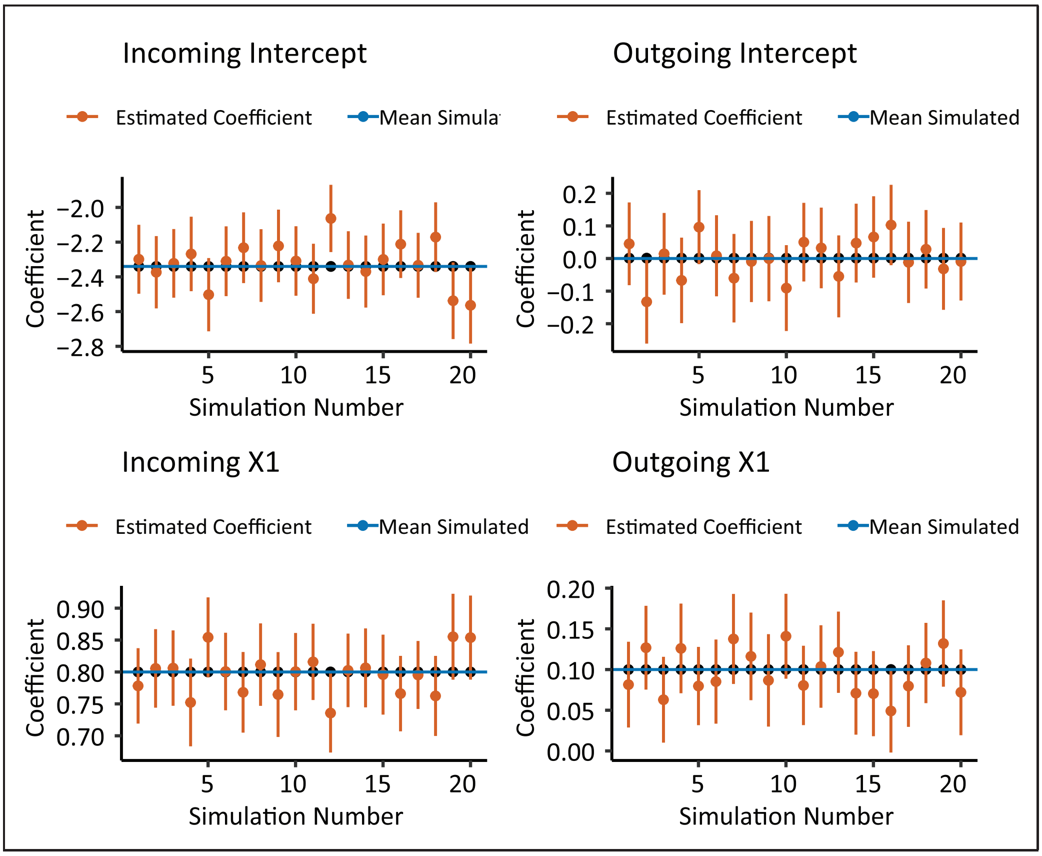


Figure 5 The estimated coefficients for twenty simulated data sets.

the twenty runs include the real coefficient, except for the 'Incoming Intercept' coefficient in the 12th simulated data set.

Overall, the simulation confirms that we are able to uncover incoming and outgoing patients from pure hospitalizations.

6 Conclusion

Overall, in this application of the SEM, we are not only able to simulate unobserved data but also estimate the association between the weekday effect and the infection rates and the number of incoming and outgoing patients in a simple and intuitive manner. We achieve some insight into the estimated association between the infection rates and the number of incoming and outgoing patients. Namely, the driving force of the estimated number of incoming and outgoing patients seems to be the infection rates of '35–59' year-olds. Although we are not able to validate the predictions against the actual number of incoming and outgoing ICU patients, our findings seem to be mostly reasonable. Additionally, the SEM estimates the association of the simulated number of incoming and outgoing ICU patients and the simulated covariate well. In this situation, the

SEM seems to be an appropriate application and allows us to gain a more complete picture of the COVID-19 pandemic, even when dealing with incomplete information.

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

We acknowledge the support of the Deutsche Forschungsgemeinschaft (DFG, Project KA 1188/13-1).

Supplementary material

Supplementary material for this article is available online.

Appendix

1 Maximum length of stay in the ICU

Figure A1 illustrates the information provided by the RKI on how long COVID-19-infected patients stayed in the ICU in Germany in 2020, see Tolksdorf et al. (2020). The maximum number of days is here 56.

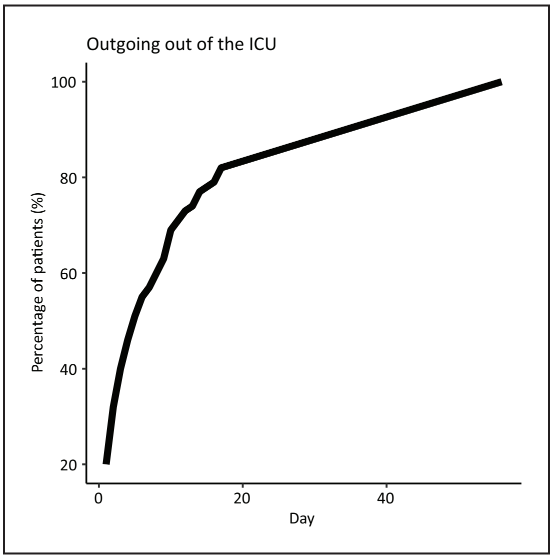


Figure A1 Percentage of outgoing ICU patients after the day of admission.

2 Convergence of the algorithm

Figure A2 and Figure A3 show the estimated coefficients in the 'M'-Step of the SEM, at each of the 500 total iterations. We see that convergence seems to have been achieved at around fifty runs and then oscillates around respective constants, just as we expect the SEM to perform.

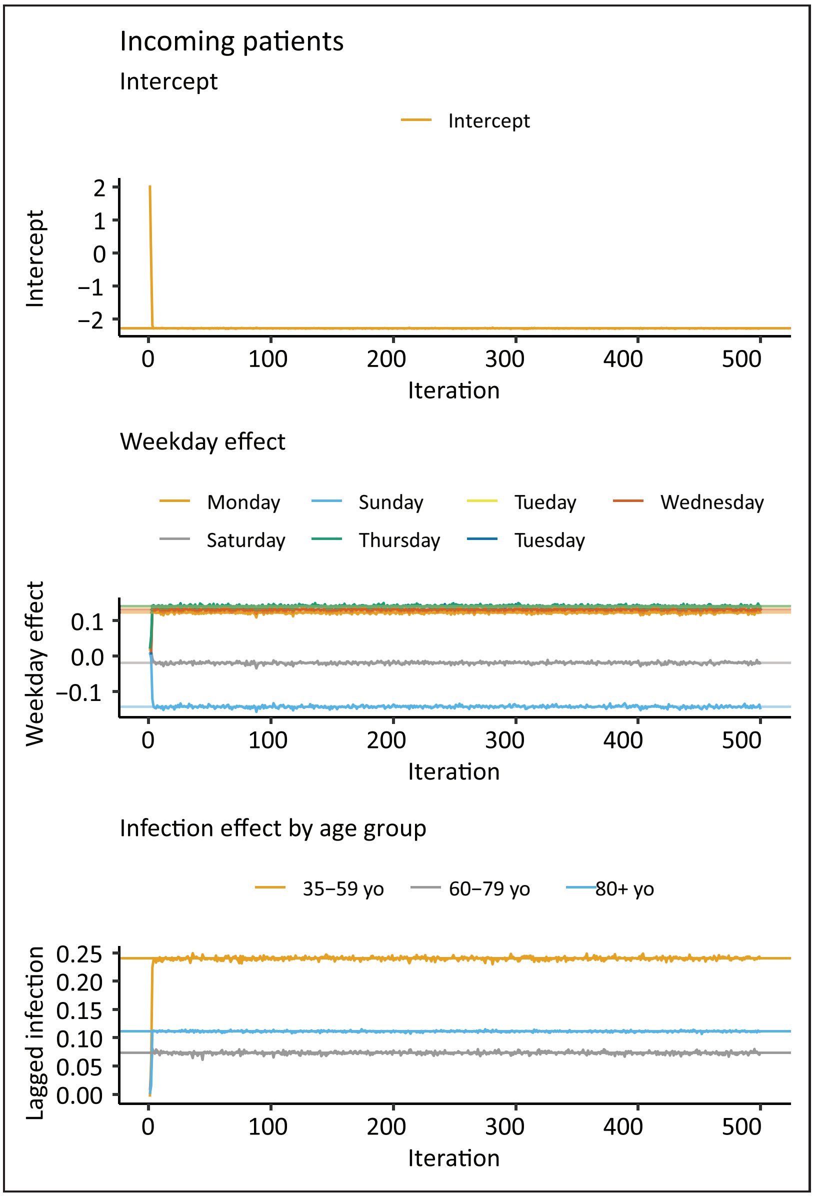


Figure A2 Coefficients estimated by the generalized additive models of the last three hundred runs of the EM algorithm of the incoming patients.

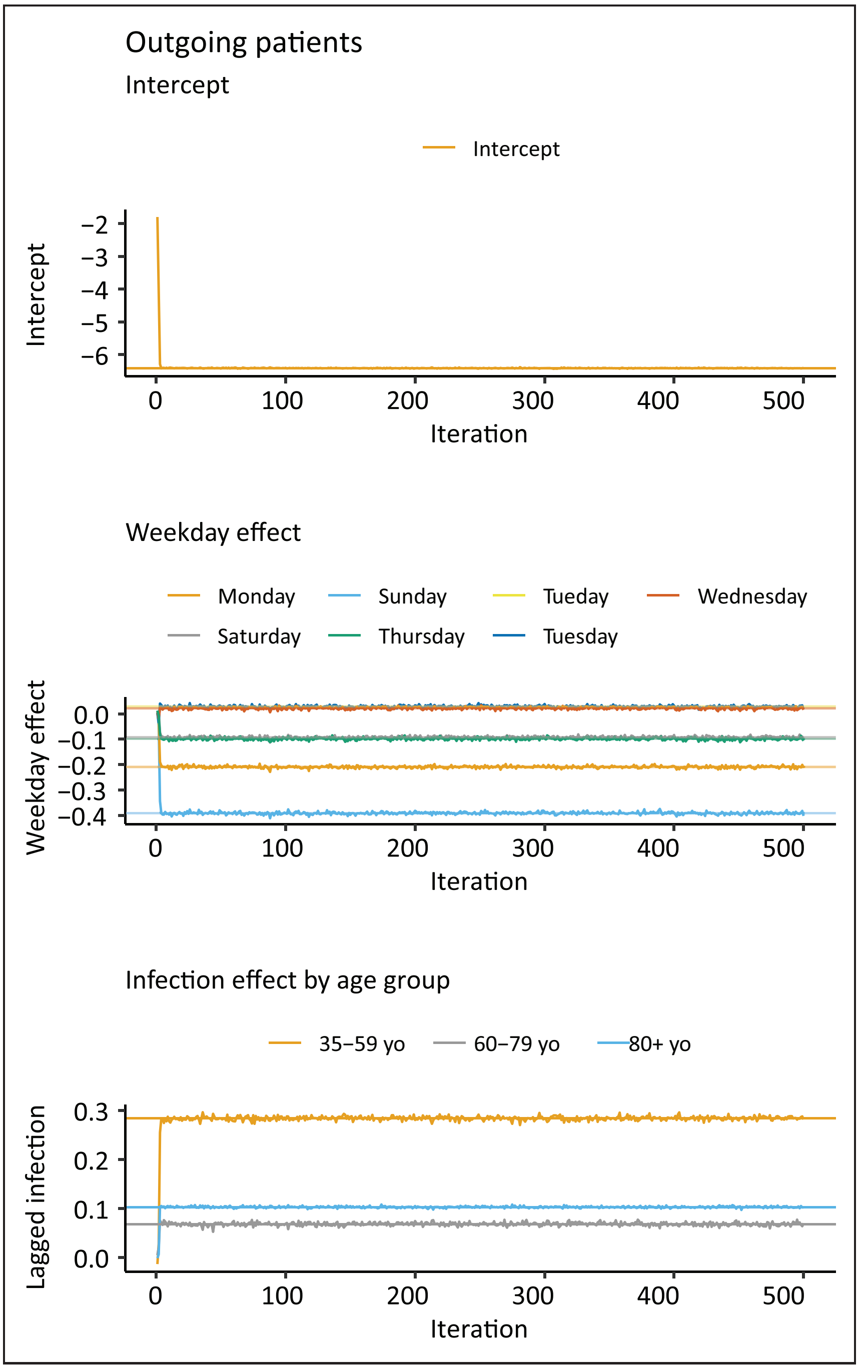
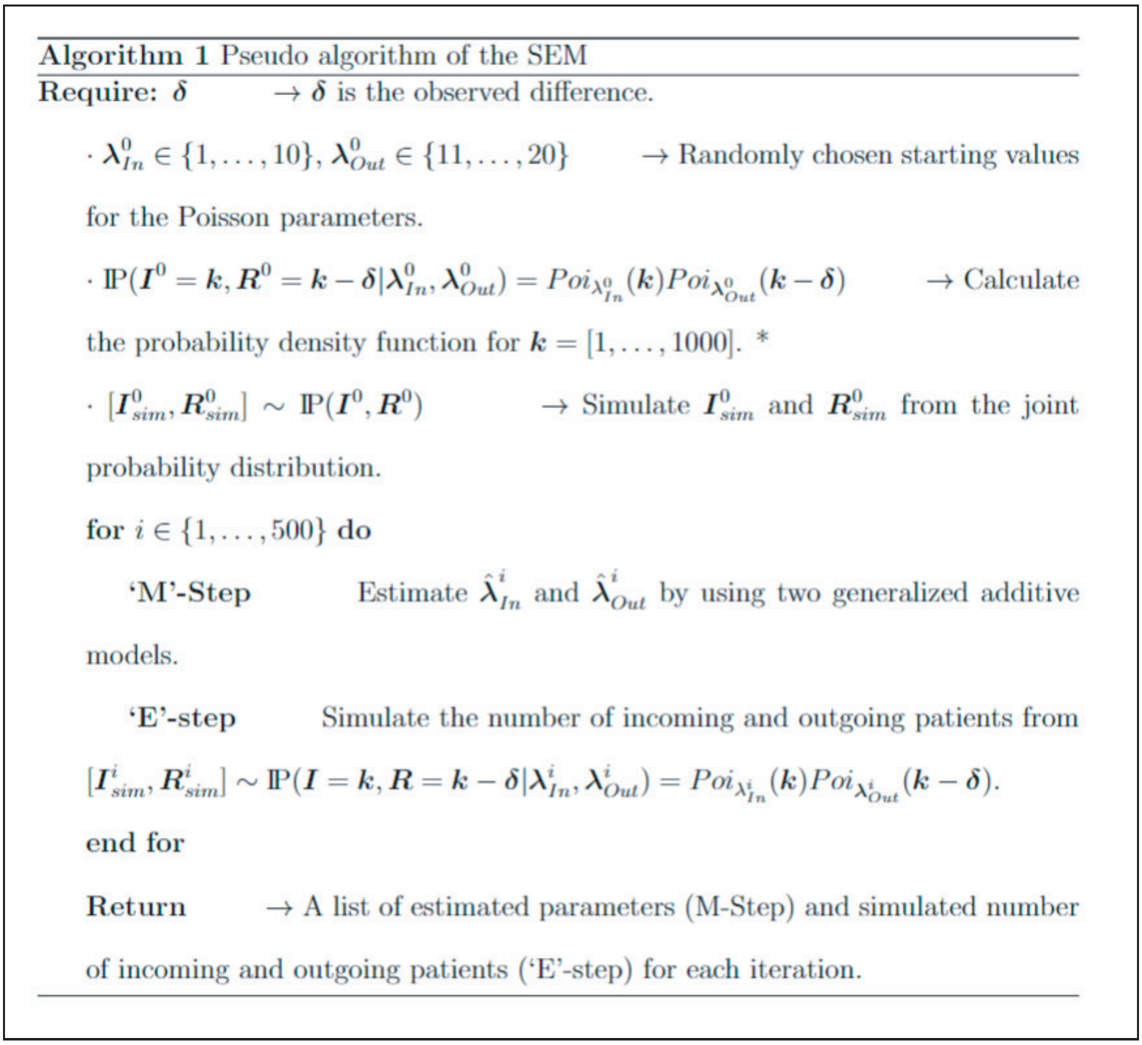


Figure A3 Coefficients estimated by the generalized additive models of the last three hundred runs of the EM algorithm of the outgoing patients.

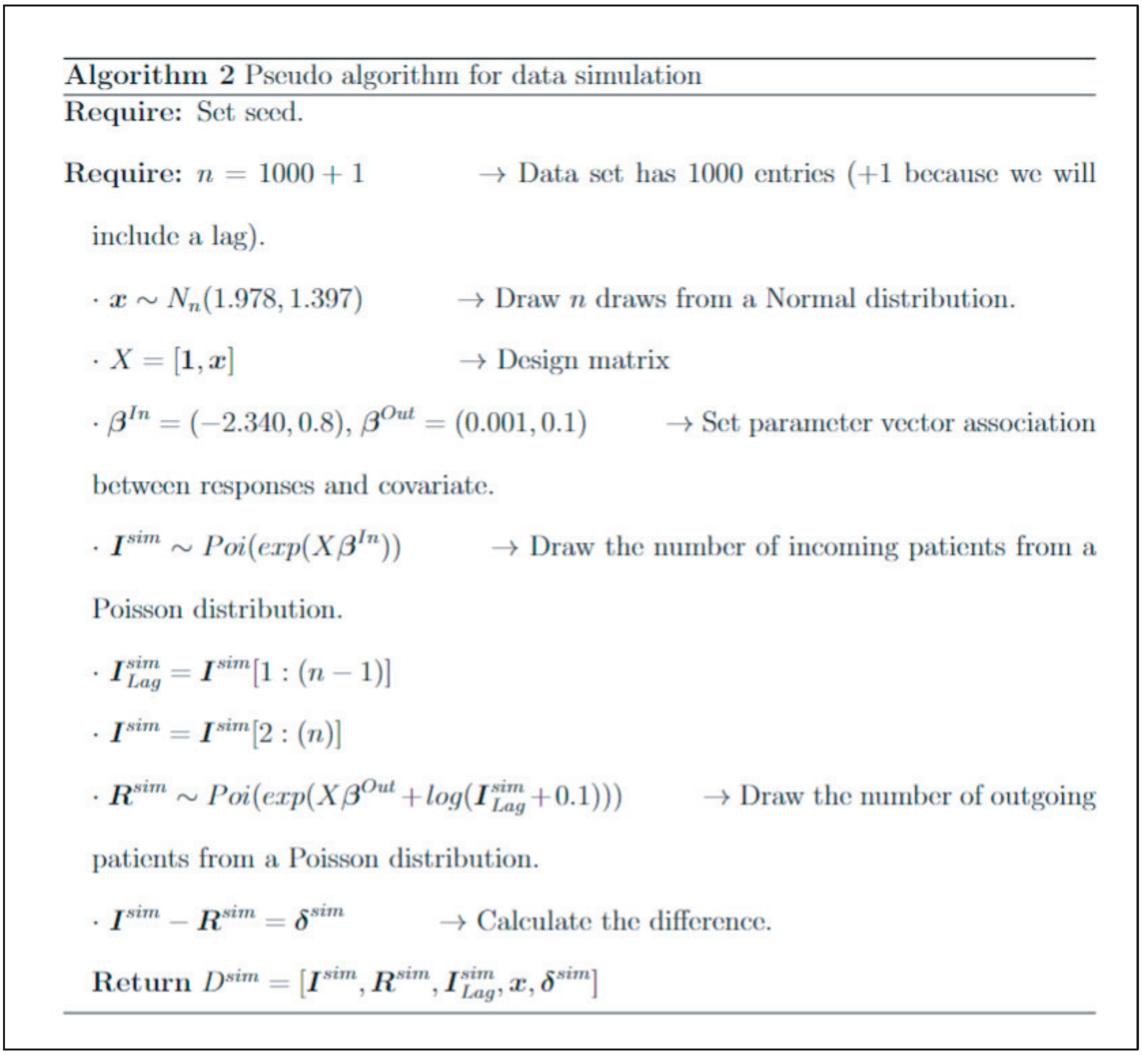
3 Pseudoalgorithms



Algorithm 1 Pseudo algorithm of the SEM

```
Require: $\delta$        → $\delta$ is the observed difference.
    · $\lambda^0_{In} \in \{1, \ldots, 10\}$, $\lambda^0_{Out} \in \{11, \ldots, 20\}$        → Randomly chosen starting values for the Poisson parameters.
    · $\mathbb{P}(I^0 = k, R^0 = k - \delta | \lambda^0_{In}, \lambda^0_{Out}) = Poi_{\lambda^0_{In}}(k) Poi_{\lambda^0_{Out}}(k - \delta)$        → Calculate the probability density function for $k = [1, \ldots, 1000]$. *
    · $[I^0_{sim}, R^0_{sim}] \sim \mathbb{P}(I^0, R^0)$        → Simulate $I^0_{sim}$ and $R^0_{sim}$ from the joint probability distribution.
    for $i \in \{1, \ldots, 500\}$ do
        'M'-Step        Estimate $\hat{\lambda}^i_{In}$ and $\hat{\lambda}^i_{Out}$ by using two generalized additive models.
        'E'-step        Simulate the number of incoming and outgoing patients from $[I^i_{sim}, R^i_{sim}] \sim \mathbb{P}(I = k, R = k - \delta | \lambda^i_{In}, \lambda^i_{Out}) = Poi_{\lambda^i_{In}}(k) Poi_{\lambda^i_{Out}}(k - \delta)$.
    end for
    Return        → A list of estimated parameters (M-Step) and simulated number of incoming and outgoing patients ('E'-step) for each iteration.
```

Figure A4 The algorithm describes the SEM which simulates the number of incoming and outgoing patients and their association with the infection rates of COVID-19 and other covariates. * 1000 is a semi-arbitrary value, but during the time span analysed the maximum number of beds per district in the data set is 1300, so reasonable.



```
Algorithm 2 Pseudo algorithm for data simulation
Require: Set seed.
Require: $n = 1000 + 1$        → Data set has 1000 entries (+1 because we will include a lag).
    · $\boldsymbol{x} \sim N_n(1.978, 1.397)$        → Draw $n$ draws from a Normal distribution.
    · $X = [\mathbf{1}, \boldsymbol{x}]$        → Design matrix
    · $\boldsymbol{\beta}^{In} = (-2.340, 0.8)$, $\boldsymbol{\beta}^{Out} = (0.001, 0.1)$        → Set parameter vector association between responses and covariate.
    · $\boldsymbol{I}^{sim} \sim Poi(exp(X\boldsymbol{\beta}^{In}))$        → Draw the number of incoming patients from a Poisson distribution.
    · $\boldsymbol{I}^{sim}_{Lag} = \boldsymbol{I}^{sim}[1 : (n-1)]$
    · $\boldsymbol{I}^{sim} = \boldsymbol{I}^{sim}[2 : (n)]$
    · $\boldsymbol{R}^{sim} \sim Poi(exp(X\boldsymbol{\beta}^{Out} + log(\boldsymbol{I}^{sim}_{Lag} + 0.1)))$        → Draw the number of outgoing patients from a Poisson distribution.
    · $\boldsymbol{I}^{sim} - \boldsymbol{R}^{sim} = \boldsymbol{\delta}^{sim}$        → Calculate the difference.
    Return $D^{sim} = [\boldsymbol{I}^{sim}, \boldsymbol{R}^{sim}, \boldsymbol{I}^{sim}_{Lag}, \boldsymbol{x}, \boldsymbol{\delta}^{sim}]$
```

Figure A5 The algorithm describes the data simulation process of the number of incoming and outgoing patients. Here we use 1000 observations, one covariate for both incoming and outgoing patients, while for the outgoing patients, we additionally take the logged lag of incoming patients of the previous 'day'.

References

Aissaoui SA, Genest C and Mesfioui M (2017) A second look at inference for bivariate Skellam distributions. *Stat*, **6**, 79–87. doi: 10.1002/sta4.136.

Celeux G, Chauveau D and Diebolt J (1996) Stochastic versions of the em algorithm: an experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, **55**, 287–314. doi: 10.1080/00949659608811772.

DIVI e.V. (2021) Daily ICU occupancy data for COVID-19 and non-COVID-19 patients. URL https://www.divi.de/register/tagesreport.

European Commission (2021) Eurostat Europe NUTS maps. URL https://ec.europa.eu/eurostat/web/nuts/nuts-maps.

Farcomeni A, Maruotti A, Divino F, Jona-Lasinio G and Lovison G (2021) An ensemble approach to short-term forecast of COVID-19 intensive care occupancy in Italian regions. *Biometrical Journal*, **63**, 503–513. doi:10.1002/bimj.202000189.

Gan HL and Kolaczyk ED (2018) Approximation of the difference of two Poisson-like counts by Skellam. *Journal of Applied Probability*, **55**, 416–430. doi:10.48550/arXiv.1708.04018.

Genest C and Mesfioui M (2014) Bivariate extensions of Skellam's distribution. *Probability in the Engineering and Informational Sciences*, **28**, 401–417. doi:10.1017/S0269964814000072.

Goic M, Bozanic-Leal MS, Badal M and Basso LJ (2021) COVID-19: Short-term forecast of ICU beds in times of crisis. *PLOS One*, **16**, e0245272. doi:10.1371/journal.pone.0245272.

Grasselli G, Pesenti A and Cecconi M (2020) Critical care utilization for theCOVID-19 outbreak in Lombardy, Italy: early experience and forecast during an emergency response. *Jama*, **323**, 1545–1546. doi:10.1001/jama.2020.4031.

Hirakawa K, Baqai F and Wolfe PJ (2009) Wavelet-based Poisson rate estimation using the Skellam distribution. In *Computational Imaging VII*, volume 7246, pages 215–226. SPIE. doi: 10.1117/12.815487.

Hwang Y, Kim JS and Kweon IS (2007) Sensor noise modeling using the Skellam distribution: Application to the color edge detection. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE. doi:10.1109/CVPR.2007.383004.

Hwang Y, Kim JS and Kweon IS (2011) Difference-based image noise modeling using skellam distribution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**, 1329–1341. doi:10.1109/TPAMI.2011.224.

Karlis D and Ntzoufras I (2009) Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, **20**, 133–145. doi:10.1093/imaman/dpn026

Koopman SJ, Lit R and Lucas A (2014) The dynamic Skellam model with applications. doi: 10.2139/ssrn.2406867.

Lewis JW, Brown PE, Tsagris M and Brown MPE (2016) Package 'skellam'.

Louis TA (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **44**, 226–233. URL http://www.jstor.org/stable/2345828.

Murray C (2020) Forecasting COVID-19 impact on hospital beddays, ICU-days, ventilator-days and deaths by US state in the next 4 months. medRxiv. doi:10.1101/2020.03.27.20043752. URL https://www.medrxiv.org/content/early/2020/03/30/2020.03.27.20043752.full.pdf.

Nielsen SF (2000) The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli*, **6**, 457–489. doi: 10.2307/3318671.

Noghrehchi F, Stoklosa J, Penev S and Warton DI (2021) Selecting the model for multiple imputation of missing data: Just use an IC! *Statistics in Medicine*, **40**, 2467–2497. doi:10.1002/sim.8915.

Pfenninger EG, Faust JO, Klingler W, Fessel W, Schindler S and Kaisers UX (2022) Eskalations-/Deeskalationskonzept zur

COVID-19-bedingten Freihal tung von Intensivkapazitäten an Kliniken. *Der Anaesthesist*, **71**, 12–20. doi:10.1007/s00101-021-00982-z.

Robert Koch-Institut (2023) Altersstruktur. URL https://www.intensivregister.de/#/aktuelle-lage/altersstruktur.

Robert Koch Institut (2021) Daily COVID-19 cases data. URL https://www.arcgis.com/home/item.html?id=f10774f1c63e40168479a1feb6c7ca74.

Robert Koch Institut (2023) Daily COVID-19 cases data. URL https://github.com/robert-koch-institut.

Rubin DB (1976) Inference and missing data. *Biometrika*, **63**, 581–592. doi:10.2307/2335739.

Schneble M and Kauermann G (2022). Estimation of latent network flows in bike-sharing systems. *Statistical Modelling*, **22**, 349–378. doi:10.1177/1471082X20971911.

Skellam JG (1948) A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological)*, **10**, 257–261.

Tolksdorf K, Buda S, Schuler E, Wieler LH and Haas W (2020) Eine höhere Letalität und lange Beatmungsdauer unterscheiden COVID-19 von schwer verlaufenden Atemwegsinfektionen in Grippewellen. *Epidemiologisches Bulletin*, **41**, 3–10. doi:10.25646/7111.

Tomy L and Veena G (2022) A retrospective study on Skellam and Related Distributions. *Austrian Journal of Statistics*, **51**, 102–111. doi:10.17713/ajs.v51i1.1224.

Wood SN (2003) Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **65**, 95–114. doi:10.1111/1467-9868.00374.

Wood SN (2017) *Generalized additive models: An introduction with R*. Boca Raton: CRC press.

7. Deriving Duration Time from Occupancy Data – A Case Study in the Length of Stay in Intensive Care Units for COVID-19 Patients

Contributing article

Martje Rave and Göran Kauermann (2025). Deriving Duration Time from Occupancy Data – A Case Study in the Length of Stay in Intensive Care Units for COVID-19 Patients. Accepted at *Biometrical journal: submitted* 5th *of May, 2025, accepted* 15th *of October, 2025. Preprint: arXiv preprint.* https://doi.org/10.48550/arXiv.2505.02587.

Data and code

Available at https://github.com/MartjeRave/OccupancyDuration.

Copyright information

Author contributions

The research approach was initiated by Prof. Dr Göran Kauermann. In developing the estimation procedure, Kauermann and Martje Rave collaborated closely. Together, they extended the existing sEM framework to incorporate the estimation of a discrete hazard function, constrained to be non-negative and to sum to one. This extension introduced a systematic shrinkage towards the uniform distribution, an effect identified and further investigated by Rave. The conceptual explanation of this effect in the paper was contributed by Kauermann.

To correct for this shrinkage bias, a simulation-based method was developed as a second-stage estimation step, designed by Kauermann and implemented by Rave.

All methodological innovations—including constraint implementation, bias diagnostics, and correction procedures—were implemented by Rave. The manuscript was written and revised jointly by Rave and Kauermann.

0 (0) 0, zzz–zzz / DOI: 10.1002/

Deriving Duration Time from Occupancy Data – A case study in the length of stay in Intensive Care Units for COVID-19 patients

Martje Rave*[1] and **Göran Kauermann** [1, 2]

[1] Chair of Applied Statistics in Social Sciences, Economics and Business, Department of Statistics, Faculty of Mathematics, Informatics and Statistics, Ludwig-Maximilians-Universität München, Ludwigstr. 33, 80539 München, Germany

[2] Munich Center for Machine Learning (MCML)

This paper focuses on drawing information on underlying processes, which are not directly observed in the data. In particular, we work with data in which only the total count of units in a system at a given time point is observed, but the underlying process of inflows, length of stay and outflows is not. The particular data example looked at in this paper is the occupancy of intensive care units (ICU) during the COVID-19 pandemic, where the aggregated numbers of occupied beds in ICUs on the district level ('Landkreis') are recorded, but not the number of incoming and outgoing patients. The Skellam distribution allows us to infer the number of incoming and outgoing patients from the occupancy in the ICUs. This paper goes a step beyond and approaches the question of whether we can also estimate the average length of stay of ICU patients. Hence, the task is to derive not only the number of incoming and outgoing units from a total net count but also to gain information on the duration time of patients on ICUs. We make use of a stochastic Expectation-Maximisation algorithm and additionally include exogenous information which are assumed to explain the intensity of inflow.

Key words: Stochastic EM Algorithm, Skellam distribution, Survival, COVID-19, ICU-patients

The GitHub repository accompanying this paper can be found under `https://github.com/MartjeRave/OccupancyDuration.git`

1 Introduction

In this paper, we introduce a method which enables one to estimate an underlying, unobserved inflow, length of stay, and consequent outflow of units, using only sporadically observed net count data of said units. While we look at intensive care units (ICU) in the paper, we want to make clear right at the start that the approach is transferable to similar data constellations. Consider, for instance, the research of an ornithologist who is investigating the hatches and deaths in a given penguin colony. The scientist sporadically observes the number of penguins at given time points. Between each observation, some penguins will have hatched and some will have died. The methodology developed in this paper allows us to estimate the number of incoming units (hatched penguins), the length of stay (average life span) and the number of outgoing units (penguins which have died). The same question is posed on our data example. We observe data on the occupancy of ICUs during the COVID-19 pandemic, but the real focus of interest is on obtaining information of incoming and outgoing patients as well as on the (average) length of stay in the ICU.

Throughout the COVID-19 pandemic, there were arguably a good amount of data published in Germany, foremost by the Robert Koch Institute (RKI), on COVID-19 infections, and the Deutsche Interdisziplinäre Vereinigung für Intensiv- und Notfallmedizin (DIVI), on hospital and ICU occupancy. However, in the beginning of the pandemic there were no data published on the number of incoming and outgoing ICU patients infected with COVID-19, only the ICU occupancy. While these data were made available on the

*Corresponding author: e-mail: martje.rave@stat.uni-muenchen.de, Phone: (+49)-89-2180-2248, Fax: (+49)-89-2180-5040 state level ('Bundeslandebene') from 2021 onwards, data on district level- which is what we consider in this paper- have not been published.

Data on ICU admissions for the whole of Germany were analysed by, for example, [Karagiannidis et al., 2021] to evidence the difference in the initial pandemic waves. Others, particularly medical practitioners, conducted studies on individual treatment centre level, to assess the treatment strategy and the success thereof, see e.g. [Rieg et al., 2020].

In our earlier work, [Fritz et al., 2024], we analyse the occupancy in relation to the infection rates in order to understand the strain on the healthcare system. Clearly, the occupancy is a function of admission and length of stay. This is the core assumption in our subsequent work [Rave and Kauermann, 2024], in which we take the length of stay as fixed, relying on results of [Tolksdorf et al., 2020]. Here, we extend our previous work and demonstrate, that the length of stay can also be estimated from occupancy data, besides obtaining information on incoming and outgoing patients. By doing so, our approach allows us to gain more understanding of the epidemiological dynamics of the disease.

The key component of our statistical model looks at the difference in two independent counting processes, each assumed to be Poisson distributed. This leads to a Skellam distribution [Skellam, 1948] with parameters equivalent to the intensity parameters of the respective underlying in- and outgoing Poisson processes. We model the unobserved number of incoming and outgoing units to depend on a set of covariates, as well as spatio-temporal information. The required independence of the two Poisson processes is achieved by conditioning on the history of the process, i.e., we assume some Markov structure.

Fitting is pursued by applying the stochastic Expectation-Maximisation (sEM) algorithm as introduced by [Celeux et al., 1996] and further discussed among others in [Chen et al., 2018] concerning running time or [Figueroa-Zúñiga et al., 2023] for estimation of complex or uncommon distributions; see also [Yang et al., 2016] for latent variable estimation in survival models. In our application, we iteratively and sequentially simulate the number of incoming and outgoing units, using the Skellam distribution. This replaces the unobserved values by simulated values (stochastic E step), and the sequential simulation allows us to condition on the past, so that we can utilise the Markov structure in the simulations. The E-step provides a complete data set which is used to estimate the incoming and outgoing intensity parameter (M-Step) employing two independent Generalised Additive Models (GAMs), [Wood, 2017]. The outgoing intensity is modelled to depend on the (unobserved) number of incoming patients, which allows to model the average length of stay of COVID-19 patients on ICUs. This part of the model is non-standard and requires specially tailored estimation routines. In simulations, we demonstrate the usability of our estimates and apply the routine to real data, as described above.

The paper is organised as follows. In Section 2, we describe the COVID-19 ICU data in more detail. In Section 3, we go into further detail of our estimation process, by describing the sEM algorithm, first through our initial approach, then by our extension thereof. We then show the application to simulated data in Section 4 and the application to COVID-19 ICU data in Section 5. We discuss the method in Section 6.

2 COVID-19 ICU Data

We define with $Y_{(t,d)}$ the observed COVID-19-related occupancy of the ICUs at time point t in the administrative district d. For time, we take the interval 1^{st} of August 2021 to the 31^{st} of December 2021 with $t = 1, 2, \ldots$ denoting the days. For the districts, we have a total of 400 different administrative regions, districts, in Germany. Data on the ICU occupancy are provided by DIVI[1] [Robert Koch-Institut, 2025a], and additionally, we take the daily infection rates as provided by the RKI[2] [Robert Koch-Institut, 2025b].

To the best of our knowledge, there are no data on the incoming, length of stay or mortality of ICU patients infected with COVID-19, publicly available in Germany. So in order to later link the inflow and outflow of patients to data, which are observed, we take the ICU occupancy and we can calculate its

[1] https://robert-koch-institut.github.io/Intensivkapazitaeten_und_COVID-19-Intensivbettenbelegung_in_Deutschland/

[2] https://robert-koch-institut.github.io/COVID-19_7-Tage-Inzidenz_in_Deutschland/

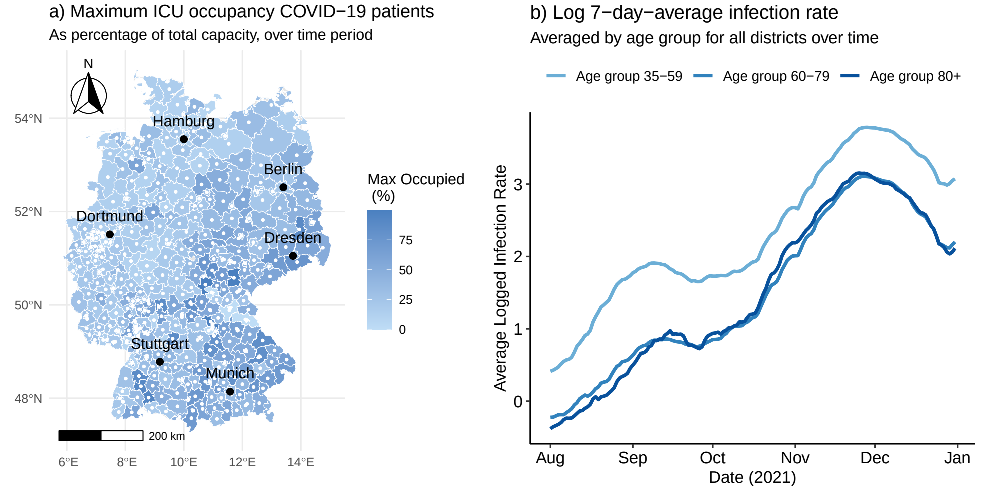


Figure 1 Introduction to COVID-19 Data with; a) maximum ICU occupancy, as a percentage of the total ICU beds, per district ('Landkreis') and b) average of the logged 7-day-average infection rate per 100.000 inhabitants, per age group from the 1^{st} of August until the 31^{st} of December, 2021.

difference

$$\Delta_{(t,d)} = Y_{(t,d)} - Y_{(t-1,d)}. \tag{1}$$

The infection rates are provided for each district, day and age group, namely '35-59' year olds, '60-79' year olds and '80+' year olds. We calculate the 7-day-average of the infection rate per 100.000 inhabitants and take the natural logarithm thereof.

For data exploration, we plot the maximum ICU occupancy in Figure 1 a). We show the percentage COVID-19 occupation of the total ICU beds, per district. The logged 7-day-average infection rates per 100.000 inhabitants of the three age groups included in the analysis, plotted over time and averaged for all districts in Germany are visualised in Figure 1 b).

Figure 1 a) shows a somewhat constant maximum occupancy rate over space, with some rural districts exhibiting a larger occupancy rate than cities. One might add that some districts report to have as little as 6 ICU beds available for patients. We would therefore expect to see these filled up more quickly than others. The cities Hamburg and Berlin are observed to have a maximum occupancy of COVID-19 patients of around 12.6% and 8.2%, respectively. Dresden, München and Stuttgart are observed to have a maximum occupancy of around 22% to 26%. Dortmund's maximum occupancy is observed at around 38%. More information on the ICU dynamics in Germany are published by the [Bundesministerium für Gesundheit, 2025]. The centroids of the given districts are marked by the respective white dots seen in Figure 1 a).

Figure 1 b) shows two spikes in the average of the logged infection rate per 100.000 inhabitants, per age group; one in mid September and another, larger spike, at the end of November, 2021. While there were non-pharmaceutical interventions in place, such as curfews and testing strategies, some readers might remember a dramatic infection wave towards the end of the second half of 2021. We also observe this in Figure 1 b).

3 Modelling Incoming and Outgoing

3.1 Skellam Modell

We are interested in the underlying process of incoming, length of stay and outgoing units, which is not observed. We therefore define with $I_{(t,d)}$ the incoming and with $R_{(t,d)}$ the outgoing (released) units of the ICUs in district d at time point t. We use the equivalence between $\Delta_{(t,d)}$, as given in (1), and the difference between the incoming units and outgoing units, i.e.

$$\Delta_{(t,d)} = Y_{(t,d)} - Y_{(t-1,d)} = I_{(t,d)} - R_{(t,d)}. \tag{2}$$

As $I_{(t,d)}$ and $R_{(t,d)}$ are both counting processes, it is reasonable to assume that the two random variables follow Poisson distributions, with intensity parameters $\lambda^I_{(t,d)}$ and $\lambda^R_{(t,d)}$, respectively, i.e.

$$I_{(t,d)} \sim \text{Poisson}\left(\lambda^I_{(t,d)}\right) \tag{3}$$

$$R_{(t,d)} \sim \text{Poisson}\left(\lambda^R_{(t,d)}\right). \tag{4}$$

We define with $H_{t,d}$ the history of the incoming process, that is $H_{t,d} = \{I_{\tilde{t}} : \tilde{t} < t\}$. Given the history of the incoming, we assume that $I_{t,d}$ and $R_{t,d}$ are conditionally independent, so that the difference of the two Poisson distributions follows the Skellam distribution, [Skellam, 1948],

$$\Delta_{(t,d)} | H_{t,d} \sim \text{Skellam}(\lambda^I_{(t,d)}, \lambda^R_{(t,d)}). \tag{5}$$

For the incoming intensity we set

$$\lambda^I_{(t,d)} = \exp\left(\eta^I_{(t,d)}\right) \tag{6}$$

where the linear predictor $\eta^I_{(t,d)}$ is modelled to depend on explanatory variables denoted by $\boldsymbol{x}^I_{(t,d)}$.

The linear predictor in estimating the number of incoming ICU patients with COVID-19 is defined as

$$\begin{aligned} \lambda^I{}_{(t,d)} = \exp(&\beta_0 + \beta_1 \text{Infec}_{35-59\,(t,d)} + \beta_2 \text{Infec}_{60-79\,(t,d)} + \\ &\beta_3 \text{Infec}_{80\,(t,d)} + \beta_4 \text{Monday}_{(t,d)} + \\ &\beta_5 \text{Tuesday}_{(t,d)} + \beta_6 \text{Wednesday}_{(t,d)} + \\ &\beta_7 \text{Thursday}_{(t,d)} + \beta_8 \text{Saturday}_{(t,d)} + \beta_9 \text{Sunday}_{(t,d)} + \\ &f_1(\text{long}_{(d)}, \text{lat}_{(d)}) + f_2(t)). \end{aligned} \tag{7}$$

The variables included are the logged 7-day-average infection rate of the week prior to t for the age groups, '35-59' year olds, '60-79' year olds and '80+' year olds. We further include a weekday effect through a dummy-coded categorical variable, with 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Saturday', 'Sunday' denoting dummy indicator variables for respective weekdays and 'Friday' being the reference category. Information on space is included by $f_1(\text{long}_{(d)}, \text{lat}_{(d)})$, a thin plate spline over the longitude and latitude of the districts' centroids. The function $f_2(t)$ denotes a thin plate spline across the date of observation, t.

For the outgoing units, $R_{(t,d)}$, we come to the understanding that this number depends on the count of incoming patients. This is modelled multiplicative as follows. Let l denote the time delay, i.e. the time between admission to the ICU and the current day t. We define with parameters ω_l for $l = 1, \ldots, L$ the exit rates, comparable to the hazard of leaving the ICU, where L is the maximum length of stay which is taken sufficiently large. One may also take the intensity of the number of outgoing units to be a function of external information, defined by a linear predictor $\eta^R_{(t,d)}$ which can depend on covariates $\boldsymbol{x}^R_{(t,d)}$. This leads to the model

$$\lambda^R_{(t,d)} = \exp\{\eta^R_{(t,d)} + \log(\sum_{l=1}^{L} \omega_l I_{(t-l,d)})\}. \tag{8}$$

In our example we will simplify the setup and set $\eta^R_{(t,d)} \equiv 0$. Moreover, as argued before, $\lambda^R_{(t,d)}$ is assumed to be a function of the incoming units and the length of stay. We thus need to postulate constraints on the parameters ω_l, namely

$$\sum_{l=1}^{L} \omega_l = 1 \text{ with } \omega_l \geq 0 \,\forall\, l \,\in \{1, \ldots, L\}. \tag{9}$$

for a sufficiently large L.

Assuming $I_{(t,d)}$ and $R_{(t,d)}$ to be known, the estimation of the parameters of $\lambda^I_{(t,d)}$ and $\lambda^R_{(t,d)}$ would be conceptionally simple. Following the distributional assumption of (3), we would be able to maximise the likelihood, given the incoming intensity parameter using a Generalized Linear Model [Wood, 2017]. The maximization of the likelihood of the outgoing number of units is, however, a little more intricate. We again assume a Poisson distribution leading to the (partial) log-likelihood

$$l^R_P(\boldsymbol{\omega}) = \sum_{t=1}^{T}\sum_{d=1}^{D} \left(R_{(t,d)} \log(\sum_{l=1}^{L} \omega_l I_{(t-l,d)}) - \sum_{l=1}^{L} \omega_l I_{(t-l,d)} \right). \tag{10}$$

Maximization of the log-likelihood in (10) needs to be done under linear constraints (9). This can be done iteratively through quadratic optimisation, see e.g. [Goldfarb and Idnani, 1983]. Second-order approximation yields

$$\begin{aligned} l^R_P(\boldsymbol{\omega}) &\approx l^R_P(\hat{\boldsymbol{\omega}}^{(k)}) + s^T(\hat{\boldsymbol{\omega}}^{(k)})(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}^{(k)}) - \frac{1}{2}(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}^{(k)})^T \mathcal{I}(\hat{\boldsymbol{\omega}}^{(k)})(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}^{(k)}) \qquad (11) \\ &\approx [s^T(\hat{\boldsymbol{\omega}}^{(k)}) + \hat{\boldsymbol{\omega}}^{(k)T} \mathcal{I}(\hat{\boldsymbol{\omega}}^{(k)})]\boldsymbol{\omega} - \frac{1}{2}(\boldsymbol{\omega}^T \mathcal{I}(\hat{\boldsymbol{\omega}}^{(k)})\boldsymbol{\omega}) + K, \end{aligned}$$

with $\hat{\boldsymbol{\omega}}^{(k)}$ denoting the estimate for the length of stay at the k^{th} iteration. Quadratic optimization allows to maximize (11) with respect to the linear constraints given above. More details are provided in Appendix A.

3.2 Estimation Approach

Since the number of incoming units and outgoing units are not observed (or observable), we can not directly estimate both, the incoming intensity (6) and outgoing intensity (8), respectively. We therefore pursue a sEM-algorithm, where the E-step is replaced by a simulation step to iteratively obtain simulations of incoming, $\boldsymbol{I}^{(k)}$, and outgoing, $\boldsymbol{R}^{(k)}$, at the k-th iteration. We then use the procedures outlined in the previous subsection to estimate $\hat{\boldsymbol{\lambda}}^{I(k+1)}$ and $\hat{\boldsymbol{\lambda}}^{R(k+1)}$, given the simulated incoming and outgoing units. To be more specific, we set the parameters to some (reasonable) starting values and then simulate incoming and outgoing patients, which builds the stochastic E-step (see Celeux et al., 1996). This leads to a complete dataset, which easily allows for (re-) estimating the parameters following the results from above. This, in turn, gives the M-step. Formally, the algorithm proceeds as follows.

1. Simulation E-Step
   Naturally, the first observation for all districts $d = 1, \ldots, D$ is at $t = 1$. However, since we assume

$\hat{\lambda}^{R\ (k)}_{(t=u,d)} = \sum_{l=1}^{L} \hat{\omega}_l^{(k)} I^{(k)}_{(t=u-l,d)}$, for all $u = \{1, \ldots, L\}$ we need the number of incoming patients before the first day of the observation period. We thus simulate $I^{(k)}_{(\tilde{t},d)} \sim \text{Poisson}(\hat{\lambda}^{I\ (k)}_{(t=1,d)})$ as 'burn-in', for $\tilde{t} = \{-L+1, \ldots, 0\}$. These 'burn-in' values are utilised in the E-Step simulations but not used for estimation of the incoming intensity. For $t = 1, \ldots, T$ we proceed to simulate both incoming and outgoing units conditional on the observed values $\Delta_{(t,d)}$. To be specific, we assume

$$I^{(k)}_{(t,d)} \sim \text{Poisson}(\hat{\lambda}^{I\ (k)}_{(t,d)}) \tag{12}$$

$$R^{(k)}_{(t,d)} \sim \text{Poisson}(\exp(\log(\sum_{l=1}^{L} \hat{\omega}_l^{(k)} I^{(k)}_{(t-l,d)}))), \tag{13}$$

subject to

$$I^{(k)}_{(t,d)} - R^{(k)}_{(t,d)} = Y_{(t,d)} - Y_{(t-1,d)} = \Delta_{(t,d)}.$$

Note that $I^{(k)}_{(t,d)}$ and $R^{(k)}_{(t,d)}$ are dependent and can be simulated as shown in [Rave and Kauermann, 2024]. We reiterate the general idea, ignoring for the moment the iteration index k. First, we define a reasonable range $[0, I_{max}]$ of probable income values $I_{(t,d)}$. Then we calculate the truncated conditional probability

$$\begin{aligned} &p(I_{(t,d)} = i, R_{(t,d)} = i - \Delta_{(t,d)} | I_{(t,d)} \leq I_{max}; \lambda^I_{(t,d)}, \lambda^R_{(t,d)}) = \\ &\frac{\exp(-\lambda^I_{(t,d)})[\lambda^I_{(t,d)}]^i \ \exp(-\lambda^R_{(t,d)})[\lambda^R_{(t,d)}]^{i-\Delta_{(t,d)}}(i!(i-\Delta_{(t,d)})!)^{-1}}{\sum_{j=0}^{I_{max}} [\exp(-\lambda^I_{(t,d)})[\lambda^I_{(t,d)}]^j \exp(-\lambda^R_{(t,d)})[\lambda^R_{(t,d)}]^{j-\Delta_{(t,d)}}(j!(j-\Delta_{(t,d)})!)^{-1}]}. \end{aligned} \tag{14}$$

The derivation is given in the Appendix B. We then sample from this normalised truncated joint probability mass function to obtain $I^{(k)}_{(t,d)}$ and $R^{(k)}_{(t,d)}$.

2. M-step
With the simulated values, we can now update the estimates for the linear predictor of the incoming intensity, $\hat{\lambda}^{I(k+1)}_{(t,d)}$, as well as the outgoing intensity $\hat{\lambda}^{R\ (k+1)}_{(t,d)}$.

3.3 Bias Correction

By defining the constraints in (9) in the estimation of the outgoing intensity (8), we obtain a prior structure on the coefficient vector $\boldsymbol{\omega}$, which induces a systematic bias. Namely, we find a pull towards a discrete uniform distribution. To accentuate this, suppose $Y_{(t,d)}$ is constant over time $Y_{(t,d)} = Y_{(t-1,d)} = Y_{(t-2,d)} = \cdots = Y_{(t-n,d)}$, which may occur, for instance, when we encounter an utterly closed system with no incoming nor outgoing units. In this case the vector $\boldsymbol{\omega}$ consists of zero entries, which violates the assumption $\sum_{l=1}^{L} \omega_l = 1$. The likelihood for $\boldsymbol{\omega}$ is thus flat and the constraints would lead to the estimate $\hat{\omega}_l = \frac{1}{L}$, which is evidently biased. To illustrate the bias problem empirically, we refer to simulated data, which are described in more depth in Section 4. We apply the sEM outlined in Section 3.2, above. We thus estimate the exit rate without adjustment, for which a pull towards the uniform distribution can be observed. We visualize this in Figure 2 a), top left-hand side plot. The light blue squares give the final estimates for ω_l. The dark blue dots indicate the ground truth exit rate. The horizontal dashed line is $1/L = 1/12$, which indicates the probability of a discrete uniform distribution with maximum length of stay equal to $L = 12$. We observe an evident pull towards the $1/12$ line.

To correct this bias, we propose bias-corrected estimates of the exit rate. The basic idea relies on the pull towards the $1/L$ line. In Figure 2 c) we plot the squared difference between $\hat{\omega}_l$ as well as ω_L and $1/L$,

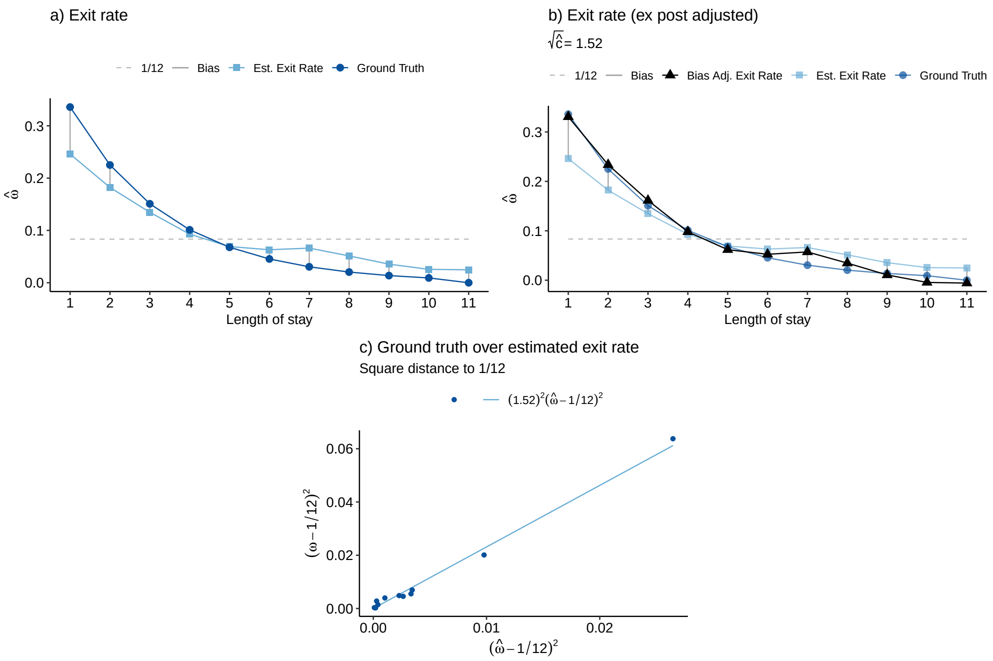


Figure 2 a) Estimated exit rate, $\hat{\boldsymbol{\omega}}$ (light blue squares) and ground truth (dark blue dots) plotted over the length of stay, l. The estimated exit rate is not bias adjusted, thus the pull in the estimates from the ground truth towards $1/12$ is shown by the grey vertical lines. (Nota bene: $\hat{\omega}_{12} = 1 - \sum_{l=1}^{11} \hat{\omega}_l$.) b) Illustrated bias adjustment with adjusted estimates $(\hat{\hat{\boldsymbol{\omega}}})$) (black triangles) with an estimated pull $\sqrt{\hat{c}} = 1.52$ using the ground truth and the unadjusted exit rate estimate, illustrated in a). c) Square difference between the ground truth exit rate and $1/12$, $(\omega_l - 1/12)^2$, and square difference between the estimated exit rate and $1/12$, $(\hat{\omega}_l - 1/12)^2$, with line $y = \hat{c}(\hat{\omega} - 1/12)^2$

.

respectively. This suggests the approximate proportionality

$$\left(\hat{\omega}_l - \frac{1}{L}\right)^2 \approx c\left(\omega_l - \frac{1}{L}\right)^2 \tag{15}$$

for some value c. The ‘best’ value of c could be estimated through least squares

$$\hat{c} = \text{argmin}_c \sum_{l=1}^{L} \left[\left(\hat{\omega}_l - \frac{1}{L}\right)^2 - c\left(\omega_l - \frac{1}{L}\right)^2\right]^2 . \tag{16}$$

A bias-corrected version of the estimate is then available by replacing c in (15) by $\hat{c}$ and reversing the pull towards $1/L$. To be precise, we define a bias-corrected version through

$$\hat{\hat{\omega}}_l = \begin{cases} \frac{1}{L} + \sqrt{\hat{c}(\hat{\omega}_l - \frac{1}{L})^2} & \text{for } \hat{\omega}_l \geq \frac{1}{L} \\ \frac{1}{L} - \sqrt{\hat{c}(\hat{\omega}_l - \frac{1}{L})^2} & \text{for } \hat{\omega}_l \leq \frac{1}{L}. \end{cases} \tag{17}$$

The resulting bias-corrected estimate is shown in Figure 2 b) on the top right-hand side as black triangles, in addition to the true values and the raw, biased estimates. We see a close concordance with the true values, demonstrating that the bias correction works in the right direction.

Apparently, looking at formula (16), it becomes obvious that the idea is not directly applicable in practice, since we would need the true values ω_l for $l = 1, \ldots, L$. However, we will utilise the idea and insert an extension to the sEM loop, where we simulate from the k-th estimated model and refit the model subesequently. By doing so, we can take the current estimates $\hat{\omega}$ as ground truth and are thereby able to estimate c, as described above. The idea is laid out as follows.

A bias correction is indeed necessary in each iteration step of the sEM algorithm, because a biased estimate of the exit rate ω_l will induce biased simulations of the incoming patients (sE-step), which in turn will lead to biased estimates of the incoming intensity (M-step). Hence, ignoring the bias creates a chain of problems. To avoid these problems, we propose to extend the sEM-steps 1 and 2 in Section 3.2 with a bias correction.

3. Simulate data from fitted model
   Let $\hat{\boldsymbol{\lambda}}^{I(k+1)}$ and $\hat{\boldsymbol{\lambda}}^{R(k+1)}$ be the estimates resulting after step 1 and 2 in the k-th step of the sEM algorithm described in Section 3.2. These estimates are biased and need to be corrected. For the bias correction, the estimates are taken as (current) ground truth. Therefore, simulate $\tilde{I}^{(k)}_{(t,d)}$ and $\tilde{R}^{(k)}_{(t,d)}$ using the current estimates and do **not** impose $\tilde{I}^{(k)}_{(t,d)} - \tilde{R}^{(k)}_{(t,d)} = \Delta_{(t,d)}$. Instead calculate

   $$\tilde{\Delta}^{(k)}_{(t,d)} = \tilde{I}^{(k)}_{(t,d)} - \tilde{R}^{(k)}_{(t,d)}$$

   and use these numbers as ‘simulated observations’ from a model, where the parameters are known.

4. Inner E-Step *(on simulated data)*
   Conditional on the ‘simulated observations’, simulate $\check{I}_{(t,d)}$ and $\check{R}_{(t,d)}$ using the current estimates from a Skellam distribution under the condition

   $$\tilde{\Delta}^{(k)}_{(t,d)} \equiv \check{I}_{(t,d)} - \check{R}_{(t,d)}.$$

   This can be done as described in Section 3.2. Note, $\tilde{\Delta}^{(k)}_{(t,d)}$ here are the simulated differences from step 3 and not the observed data.

5. Inner M-Step: Outgoing
   Use the simulated data from step 4 to obtain estimates $\tilde{\boldsymbol{\omega}}_l$ for $l = 1, \ldots, L$. This can be done as described in Section 3.2.

6. Bias Correction for Outgoing (ω)
   Based on the ‘raw’ estimates $\hat{\omega}^{(k+1)}$ from step 2 and the derived estimates $\tilde{\omega}$ from step 5, calculate the optimal $\hat{c}$ using (16), with ω_l in (16) replaced by $\hat{\omega}^{(k+1)}$ and $\hat{\omega}$ replaced by $\tilde{\omega}$. This yields a bias corrected version for $\hat{\omega}^{(k+1)}$, which is available through (17), that is

   $$\hat{\hat{\omega}}_l^{(k+1)} = \begin{cases} \frac{1}{L} + \sqrt{\hat{c}(\hat{\omega}_l^{(k+1)} - \frac{1}{L})^2}, & \hat{\omega}_l^{(k+1)} \geq \frac{1}{L} \\ \frac{1}{L} - \sqrt{\hat{c}(\hat{\omega}_l^{(k+1)} - \frac{1}{L})^2}, & \hat{\omega}_l^{(k+1)} < \frac{1}{L} \end{cases}$$

7. Bias Correction for Incoming (λ^I)
   Simulate incoming and outgoing patients again, like in step 1, but now using the current (raw) estimates $\hat{\boldsymbol{\lambda}}^{I\ (k+1)}$ and the bias-corrected estimates $\hat{\hat{\omega}}^{(k+1)}$ and conditional on the observed data

   $$\Delta_{(t,d)} \equiv \tilde{I}^{(k)}_{(t,d)} - \tilde{R}^{(k)}_{(t,d)},$$

Note, this is like the original step 1 in the sEM algorithm, but a bias-corrected version replaces the exit rates.
Use the simulated incoming patients to obtain a bias-corrected version $\hat{\hat{\boldsymbol{\lambda}}}^{I\,(k+1)}$.

8. Concluding the loop
   Replace $\hat{\boldsymbol{\omega}}^{(k+1)}$ by $\hat{\hat{\boldsymbol{\omega}}}^{(k+1)}$ and $\hat{\boldsymbol{\lambda}}^{I\,(k+1)}$ by $\hat{\hat{\boldsymbol{\lambda}}}^{I\,(k+1)}$ and proceed with step 1 in the sEM algorithm.

In the application, we suggest extending steps 1 and 2 of the sEM loop with the extra steps 3 to 8 not immediately, but only after some 'burn-in' phase. This accelerates the estimation process.

3.4 Inference

Given the application of the sEM we can use the variability of the estimates within the sEM chain to adjust for the underestimated variance, as given by [Rubin, 1976]. Let therefore $\boldsymbol{\beta}$ denote the parameter vector with all model parameters stacked together. We use the variance estimation

$$\hat{\Sigma}_{\boldsymbol{\beta}} = \frac{\sum_{k=k'}^{K} \Sigma_{\boldsymbol{\beta}}^{(k)}}{K-k'} + \frac{\sum_{k=k'}^{K} (\boldsymbol{\beta}^{(k)} - \bar{\boldsymbol{\beta}})(\boldsymbol{\beta}^{(k)} - \bar{\boldsymbol{\beta}})^T}{K-k'-1}, \tag{18}$$

with $\Sigma^{(k)}$ being the covariance matrix estimated at the k^{th} iteration, $\bar{\boldsymbol{\beta}}$ being the mean (or median in case of outliers) estimate of the last $K-k'$ runs of the column coefficient vector $\boldsymbol{\beta}$, with k' being a starting point at which convergence is assumed to have occurred. The estimated covariance matrix for the model on incoming units, (12), is a standard estimation. For the model on outgoing units, (13), we take the inverse of (27) as an estimate for the covariance matrix. For simplicity, we assume the incoming and outgoing units to be independent.

4 Simulation

We simulate a data example, which is aimed to emulate the real data closely. We simulate 200 districts, d, for which we observe data at 200 time points, t. This results in 40.000 observations. We then simulate two covariates from which the incoming units are simulated, as seen in (19).

$$\begin{aligned}
x1_{(d)} &\sim \text{Gamma}(0.1, 0.5), \text{ (Nota bene: varying over districts, constant over time)}\\
x2_{(t,d)} &\sim \text{Gamma}(1, 3),\\
\lambda^{I}_{(t,d)} &= \exp(0.5 + \text{x}1_{(d)} + 0.2\,\text{x}2_{(t,d)})\\
I_{(t,d)} &\sim \text{Poisson}(\lambda^{I}_{(t,d)})\\
&\forall\, t \in \{1, \dots, 200\},\ d \in \{1, \dots, 200\}.
\end{aligned} \tag{19}$$

For the simulation setup, we choose the maximum length of stay to be $L = 10$. The probability mass function is

$$\pi_l = P(L = l) = \frac{\exp(-0.4l)}{\sum_{s=1}^{10} \exp(-0.4s)}. \tag{20}$$

From this, we now simulate the outgoing number of units in a slightly different way to the estimation procedure. Namely, let $(\pi_1, \dots, \pi_{10})$ and for each incoming patient $i_{(t,d)} \in \{1, \dots, I_{(t,d)}\}$ at time t and district d we simulate a length of stay, $l_{i_{(t,d)}}$, from $(\pi_1, \dots, \pi_{10})$. Then

$$R_{(t,d)} = \sum_{l=1}^{L} \sum_{i=1}^{I_{(t-l,d)}} 1(l_{i_{(t,d)}} = l), \tag{21}$$

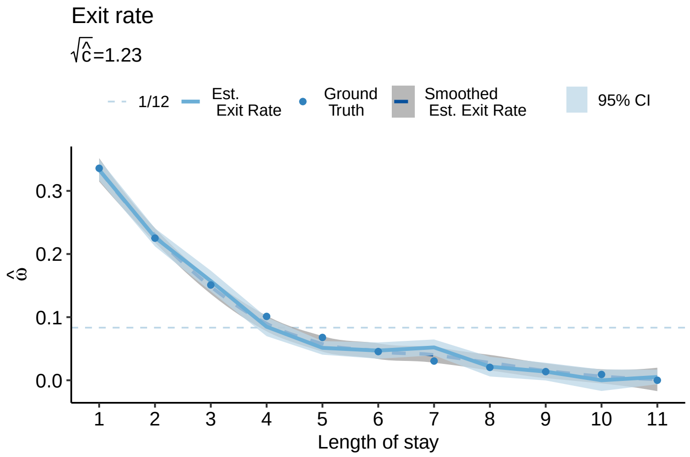


Figure 3 Estimated exit rate over the length of stay (denoted lag) with 95% confidence interval (of the last 200 runs of the sEM) against ground truth, with $\hat{\omega}_{12} = 1 - \sum_{l=1}^{11} \hat{\omega}_l$.

with $1(.)$ denoting the indicator function. To summarise, the number of outgoing units, at time point t and district d, is the sum of units which have previously come in l days before.

From (19) and (21) we obtain the difference,

$$\Delta_{(t,d)} = I_{(t,d)} - R_{(t,d)}. \tag{22}$$

Once the data are generated, the sEM is applied for 400 iterations. For different starting values, the sEM would take a different number of iterations until convergence is observed. However, we conjecturally observe a convergence rather quickly, maximally after 100 iterations. In Appendix C, we observe that the likelihood has reached some convergence after around 50 iterations of the sEM. We summarise the results for the last 200 runs of the applied sEM, by the median of respective point estimates and the estimated standard deviation. The M-Step comprises the estimation of the incoming intensity parameter and the exit rates. For the exit rate, we select a maximum lag L that exceeds the true lag used in the simulation. This should mirror a plausible estimation strategy, where, for estimation, one sets the maximum lag large enough, potentially larger than needed. To be specific, we set the maximum lag for fitting to be 12. The estimate of the incoming intensity parameter is given by

$$\hat{\lambda}^I_{(t,d)} = \exp(\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x1}_{(d)} + \hat{\beta}_2 \mathrm{x2}_{(t,d)}). \tag{23}$$

In Figure 3 and Table 1 the median of the point estimates and the standard deviation, the square root of the variance estimate as given in (18), are displayed, where the median of the simulated incoming and outgoing units are displayed in Figure 4.

Table 1 shows the estimated and true effects of the covariates on the incoming units. We observe that the estimates approximate the ground truth for both coefficient estimates. However, we observe a somewhat larger deviation for the estimated intercept. We will get back to this point in a second simulation setup below. The true and estimated exit rates, shown in Figure 3, evidence an estimation close to the ground truth for all estimates of the exit rate, with some slight deviation from the 95% confidence interval at lag 5 and lag 7. Note this is just one simulation and overinterpretation should be avoided. Therefore, we additionally fit a smooth fit to estimated exit rates, which mitigates the random deviations from the true exit rates.

Parameter	Estimate	Std. Dev.	Ground Truth
β_0	0.3788	0.0089	0.5
β_1	0.9828	0.0167	1
β_2	0.2149	0.0036	0.2

Table 1 Results of coefficients against ground truth.

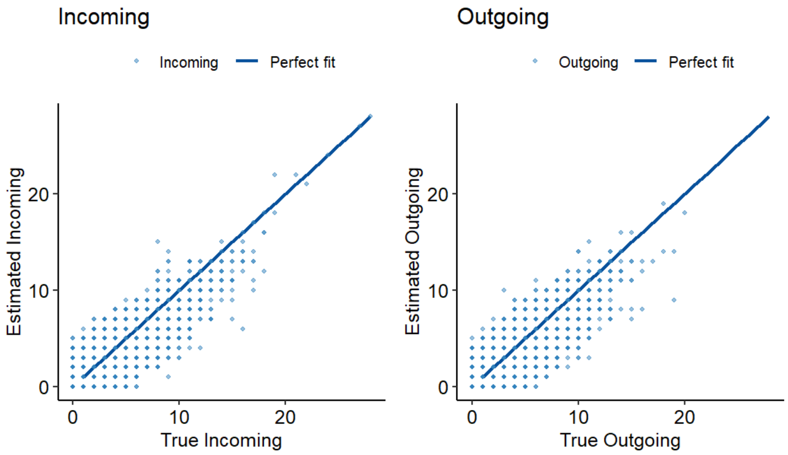


Figure 4 Estimated incoming and outgoing number of units.

Though this simulation shows that the model is able to estimate the true underlying incoming and outgoing units, as well as the true coefficients, in practice, there is likely to be overdispersion in inflow and outflow relative to the Poisson model assumption. We therefore extend the simulation process by generating inflow data from a Negative Binomial distribution (24), instead of a Poisson distribution. In doing so, we retain the true incoming coefficients for the expected value of incoming units, as specified in (19):

$$I_{(t,d)} \sim \text{Negative-Binomial}(\lambda^I_{(t,d)}, \theta). \tag{24}$$

In (24) the variance assumption extends from that of the Poisson distribution to

$$\mathbb{V}_{NB}(I_{(t,d)}) = \lambda^I_{(t,d)} + \frac{\left(\lambda^I_{(t,d)}\right)^2}{\theta}. \tag{25}$$

We simulate data for different values of θ, with $\theta_1 = 0.5, \theta_2 = 1, \theta_3 = 5,\ \theta_4 = 10$. The simulated overdispersion decreases with increasing θ. For each θ, we again simulate 200 districts and 200 time points, resulting in $40{,}000$ observations per data set, and estimate inflow and outflow analogously to the previous setup.

Table 2 reports the estimated and true covariate effects for each simulated data set, including the Poisson-based simulation for comparison. We observe that the estimates approach the ground truth as overdispersion decreases. We also refer to Appendix D, where we show simulated incomings, from one of the last stochastic E-steps, plotted against the true incomings, based on the simulations. Overall, the models' estimates tend to approximate the true coefficients more closely as overdispersion diminishes.

Table 2 Estimated coefficients from the 200^{th} to the 400^{th} of misspecified models compared to the true values.

	β_0	β_M	β_N
True	0.500	1.000	0.200
$\theta = 0.5$	1.269	1.925	0.099
$\theta = 1$	0.937	1.662	0.137
$\theta = 5$	0.499	1.406	0.178
$\theta = 10$	0.487	1.025	0.219
Poisson	0.379	0.983	0.215

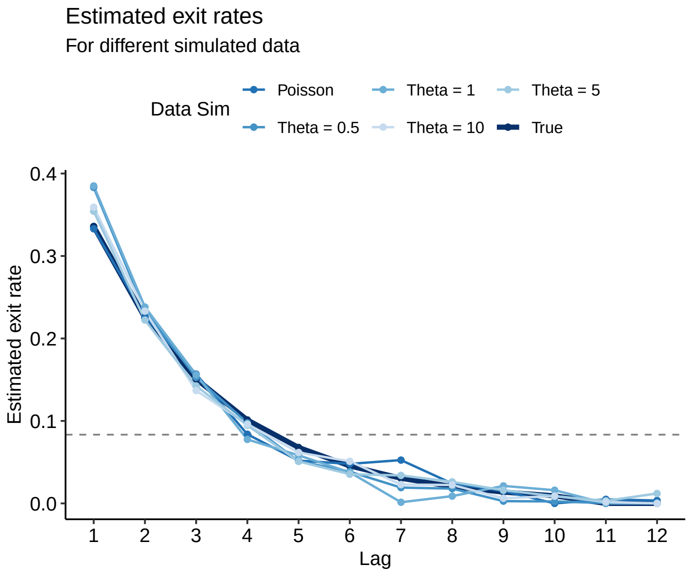


Figure 5 Estimated exit rate over the length of stay (denoted lag) for models applied to data simulated, with incoming units simulated from a Poisson distribution and Negative-Binomial distributions, with $\theta_1 = 0.5,\ \theta_2 = 1,\ \theta_3 = 5,\ \theta_4 = 10$.

Looking at the exit rate, which is the primary focus of interest, we see from Figure 5 that overdispersion does not disturb roughly consistent estimation. The estimates of the exit rate all show a similar performance.

5 Results

With the above prerequisites, we are now able to apply our model to the ICU data. For stability in our estimation, we first apply the sEM, as a 'pre-run', to the data for a total 200 iterations, without conducting any bias adjustment. Said 'pre-run' renders results which are assumed to be in a reasonable range for starting values of the sEM with bias adjustment, i.e. actually used in estimation results. The sEM with bias adjustment runs for another 150 iterations. The final results are summarized over the last 100 runs. The log-likelihood over the initial 200- unadjusted- iterations and the subsequent 150- adjusted- iterations are shown in Appendix C. For reference, the linear predictor for the incoming intensity is given in (7).

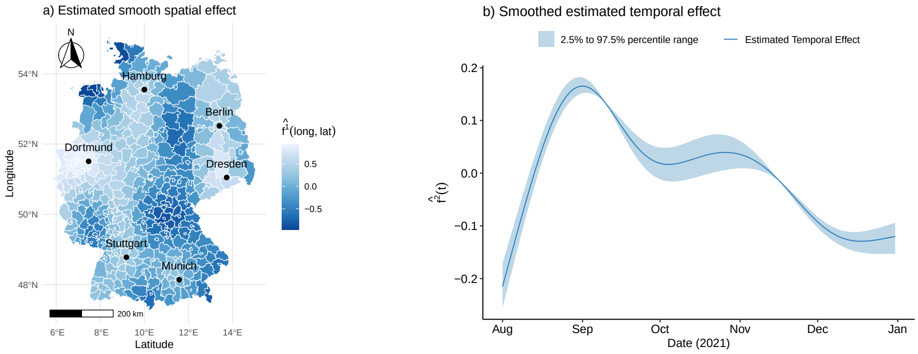


Figure 6 a) Estimated smooth function over space b) Estimated smooth function over time.

In Table 3 we see the results of the estimated effects of the infection rates and the weekday effects. We see that the estimated effect of the lagged infection rates of the ‘35-59’ year olds is largest, which agrees with the findings of our first paper, see [Rave and Kauermann, 2024]. The estimated weekday effects further agree with our initial findings, where we estimate to see less incoming patients into the ICU on weekends, compared to Fridays, and more during weekdays, again, compared to Fridays. Contextually, one might argue that the severity of a disease might not care about the day of the week. However, this might be explained by internal movements within a treatment centre, where severe cases might first be treated in an Emergency Room (ER), and only be moved to the ICU, once the appropriate personnel has authorised it.

In Figure 6 a), we see the estimated spatial effects, were we observe an increase, to varying degrees, in and around large cities, such as Dortmund, Hamburg, Dresden, Berlin, Stuttgart and Munich. This also agrees with the findings of our earlier work. Contextually, in the centralised health care system of Germany, we tend to have more ICU capabilities in the cities, which leads to ICU patients from surrounding rural areas typically being treated in near cities, rather than in their district. Particularly, during the COVID-19 pandemic, city hospitals were usually the treatment centres with treatment capabilities for isolation and respiration of COVID-19 patients. So rather than directly inferring that the severity of the disease being stronger in urban environments, the factor of the hospitalisation logistics may also be a driving factor here.

In Figure 6 b), we observe the estimated smooth function over time. It is wrapped by the 2.5^{th} and 97.5^{th} percentile of the estimated smooth functions of the 100 included sEM iterations. We observe an initial increase in the estimated smooth function until September, 2021, with a subsequent sharp decrease, a slight pick up from October until November and a following decrease, which seems to pick up again in the end of December, 2021. The interpretation of the estimated temporal effect is, as all other interpretations, conducted ceteris paribus. Thus, we estimate an increasing admittance to the ICU until September, which cannot be entirely explained by the other covariates included in our estimation. This is followed by a subsequent fall in ICU admittance, likewise not explained by the other estimated effects.

In Figure 8, we show the estimated incoming patients aggregated to Bundesland level, plotted against the “Erstaufnahmen” (incoming) patients, reported by the [Robert Koch-Institut, 2025a]. We see that our model underestimates the number of incoming patients in Berlin, which would fit intuition, following our centralised health care system interpretation of the estimated spatial effects. For a better visual impression, we included a smooth estimate of the exit rates.

Figure 7 shows the estimated exit rates up until a maximum of a 30 day lag. We estimate a sharp decline in the estimated exit rate along the initial 16 days, and a subsequent slough off thereafter. More specifically, we see an estimate of around 13% of ICU patients with COVID-19 leaving after one day, 50% of patients

Covariates	Estimates	Std. Dev.
Intercept	-2.811	0.040
Infection Rate 35-59	0.545	0.023
Infection Rate 60-79	0.098	0.024
Infection Rate 80+	0.112	0.011
Monday	0.105	0.022
Tuesday	0.045	0.023
Wednesday	0.033	0.023
Thursday	0.071	0.022
Saturday	-0.018	0.023
Sunday	-0.086	0.023

Table 3 Estimated coefficients on inflow of ICU patients.

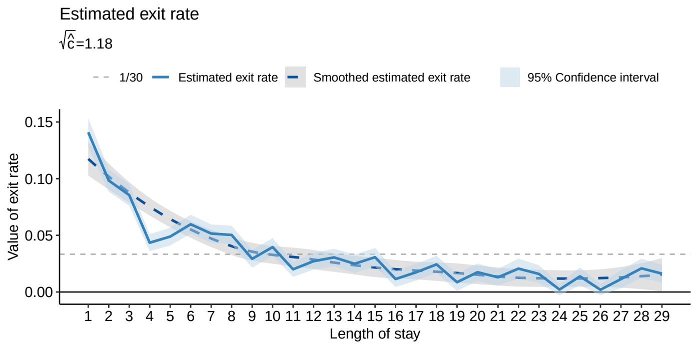


Figure 7 Estimated exit rate with 95% confidence interval (of the last 100 runs of the sEM) with smoothed estimate over exit rates.

are estimated to have left by their 6^{th} day in the ICU and 80% of COVID-19 patients are estimated to have left the ICU by the 17^{th} day. Finally, 90% of COVID-19 patients are estimated to have left after 22 days.

Inspecting the [Robert Koch-Institut, 2025a] repository, one may discover, that since 2021 data on the number of admitted ICU patients with COVID-19 have been published. However, the most granular these data are published, are on state level (there are 16 states in Germany), while our data are on the district level, which make up each of the respective counties to which they belong. We may therefore aggregate our estimated admitted ICU patients and compare them with the data reported by the RKI. In Figure 8, we plot our aggregated estimation against the RKI reported data. Specifically in Berlin and Brandenburg (titles marked by the blue outline), we observe a clear deviance. This may be due to hospital logistics, which we have not included in our analysis. The health care system in Germany induces that treatment facilities in cities tend to be more equipped to treat patients in need of specialised care, such as isolation for patients infected with COVID-19. A short outline of this principle during the COVID-19 pandemic and the planned cooperation between counties is given by [Gräsner et al., 2020]. We thus underestimate the number of admitted ICU patients with COVID-19 in Berlin, as we suspect that many of which will have been moved from surrounding counties, such as Brandenburg, where we overestimate the number of admitted ICU patients.

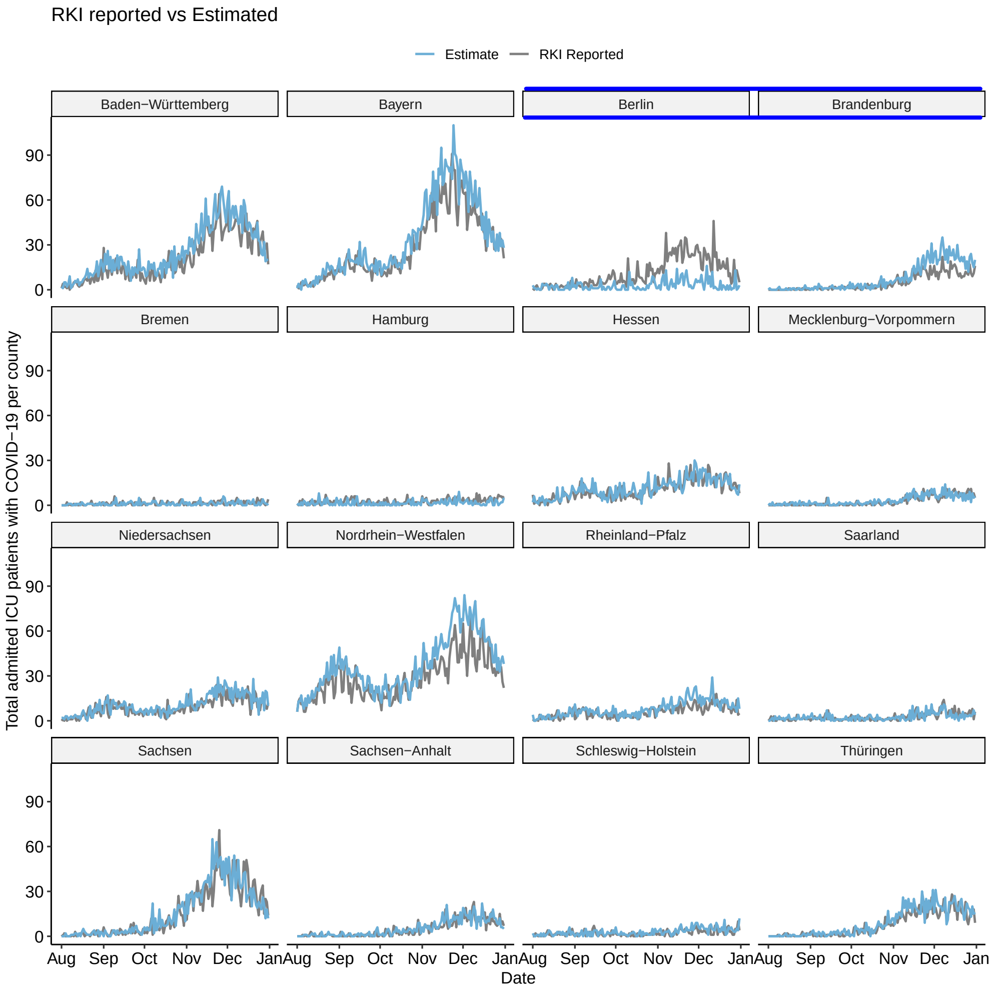


Figure 8 RKI reported admitted ICU patients with COVID-19 over time plotted against the estimated admitted ICU patients.

6 Discussion

Our approach demonstrates that we can extract information on underlying inflow and outflow processes by observing current snapshots of a system only. We also show how to include further covariates which influence the incoming intensity. As remarked in the introduction, the idea can be extended to similar data constellations. For example, the field of population dynamics would benefit from our approach in that herd inflow and outflow are often expensive to record continuously over a long period of time. Our model is able to circumvent this predicament elegantly by including information on the inflow.

In the estimation of the length of stay, we draw on a ‘non-standard’ estimation process through the bias adjustment. There is a possibility that the bias is merely mitigated, but still present, thus implying we underestimate the variance. In further work, one could refine the approach to adjust for bias in the variance estimation and thereby achieve better coverage of the estimates.

Despite the advantages of our approach, we do encounter some challenges when fitting the sEM. We have a clear disadvantage in the running time of the algorithm. This is likely optimisable in our particular model, however, only to a certain degree, with a clear limitation being the stochastic nature of the algorithm. All in all, the estimation requires iterative simulations due to the sequential pattern of the model. This leads inevitably to heavy computation.

A further possible extension to the model arises from the context of the COVID-19 ICU data. We do not differentiate between patients who were moved to Intermediate Care Units (IMCU) or other units within the hospital and patients who die during their stay at the ICU. We also do not take the movement of patients between counties into account. It is therefore likely that our model predicts the number of admitted ICU patients by district of origin well, but does not take patients’ placement between districts into account and therefore deviates from the RKI-reported data.

Our approach allows us to obtain information about data which were not made public at the time of the analysis. Apparently, for practical purposes, it is certainly better to record the original data and omit the modelling exercise pursued in this paper. Meaning that in our view it seems advisable to enable any data system to incorporate the true data on incoming and outgoing patients in ICU units.

Declaration of Conflicting Interests

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

Data Availability Statement

The data that support the findings of this study are openly available at `https://robert-koch-institut.github.io/Intensivkapazitaeten_und_COVID-19-Intensivbettenbelegung_in_Deutschland/` and `https://robert-koch-institut.github.io/COVID-19_7-Tage-Inzidenz_in_Deutschland/`.

References

Bundesministerium für Gesundheit. Intensive Care Unit Utilization – Infektionsradar. `https://infektionsradar.gesund.bund.de/en/covid/intensivecare`, 2025. Accessed: 2025-03-02.

G. Celeux, D. Chauveau, and J. Diebolt. Stochastic versions of the em algorithm: an experimental study in the mixture case. *Journal of statistical computation and simulation*, 55(4):287–314, 1996.

J. Chen, J. Zhu, Y. W. Teh, and T. Zhang. Stochastic expectation maximization with variance reduction. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/aba22f748b1a6dff75bda4fd1ee9fe07-Paper.pdf`.

J. Figueroa-Zúñiga, J. G. Toledo, B. Lagos-Alvarez, V. Leiva, and J. P. Navarrete. Inference based on the stochastic expectation maximization algorithm in a kumaraswamy model with an application to covid-19 cases in chile. *Mathematics*, 11(13), 2023. ISSN 2227-7390. doi: 10.3390/math11132894. URL `https://www.mdpi.com/2227-7390/11/13/2894`.

C. Fritz, G. De Nicola, M. Rave, M. Weigert, Y. Khazaei, U. Berger, H. Küchenhoff, and G. Kauermann. Statistical modelling of covid-19 data: Putting generalized additive models to work. *Statistical Modelling*, 24(4):344–367, 2024. doi: 10.1177/1471082X221124628. URL `https://doi.org/10.1177/1471082X221124628`.

D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27(1):1–33, September 1983. ISSN 1436-4646. doi: 10.1007/BF02591962. URL `https://doi.org/10.1007/BF02591962`.

J. Gräsner, L. Hannappel, M. Zill, B. Alpers, S. Weber-Carstens, and C. Karagiannidis. COVID-19-Intensivpatienten: Innerdeutsche Verlegungen, 2020. URL `https://www.aerzteblatt.de/archiv/216919/COVID-19-Intensivpatienten-Innerdeutsche-Verlegungen`. Dtsch Arztebl 2020; 117(48): A-2321 / B-1959.

C. Karagiannidis, W. Windisch, D. McAuley, T. Welte, and R. Busse. Major differences in icu admissions during the first and second covid-19 wave in germany. *The Lancet Respiratory Medicine*, 9:47 – 48, 2021. URL `https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(21)00101-6/fulltext`.

M. Rave and G. Kauermann. The skellam distribution revisited: Estimating the unobserved incoming and outgoing icu covid-19 patients on a regional level in germany. *Statistical Modelling*, page (to appear), 2024. doi: 10.1177/1471082X241235024. URL `https://doi.org/10.1177/1471082X241235024`.

S. Rieg, M. von Cube, J. Kalbhenn, S. Utzolino, K. Pernice, L. Bechet, J. Baur, C. Lang, D. Wagner, M. Wolkewitz, W. Kern, and P. Biever. Covid-19 in-hospital mortality and mode of death in a dynamic and non-restricted tertiary care model in germany. *PLOS ONE*, 15(11):1–16, 11 2020. doi: 10.1371/journal.pone.0242127. URL `https://doi.org/10.1371/journal.pone.0242127`.

Robert Koch-Institut. Intensivkapazitäten und covid-19-intensivbettenbelegung in deutschland, March 2025a. URL `https://robert-koch-institut.github.io/Intensivkapazitaeten_und_COVID-19-Intensivbettenbelegung_in_Deutschland/`. Accessed: 2025-03-02.

Robert Koch-Institut. 7-tage-inzidenz der covid-19-fälle in deutschland, March 2025b. URL `https://robert-koch-institut.github.io/COVID-19_7-Tage-Inzidenz_in_Deutschland/`. Accessed: 2025-03-02.

D. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

J. Skellam. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):257–261, 1948.

K. Tolksdorf, S. Buda, E. Schuler, L. Wieler, and W. Haas. Eine hoehere letalitaet und lange beatmungsdauer unterscheiden covid-19 von schwer verlaufenden atemwegsinfektionen in grippewellen. *Epidemiologisches Bulletin*, (41):3–10, 2020. doi: http://dx.doi.org/10.25646/7111.

Simon N. Wood. *Generalized additive models: An introduction with R*. Boca Raton: CRC press, 2017.

Y. Yang, H. Ng, and N. Balakrishnan. A stochastic expectation-maximization algorithm for the analysis of system lifetime data with known signature. *Computational Statistics*, 31:609–641, 2016.

A Score function and Fisher Information

We derive the approximate score function from (10),

$$s(\hat{\omega}_l^{(k)}) = \frac{\partial l_P^R(\boldsymbol{\omega})}{\partial \omega_l}\Big|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}^{(k)}} = \sum_{t=1}^{T}\sum_{d=1}^{D}\Big((I_{(t-l,d)} - I_{(t-L,d)})\Big(\frac{R_{(t,d)}}{\sum_{l=1}^{L}\hat{\omega}_l^{(k)} I_{(t-l,d)}} - 1\Big)\Big) \tag{26}$$

and the second-order derivative

$$\mathcal{I}_{jk}(\hat{\boldsymbol{\omega}}^{(k)}) = \frac{\partial l_P^R(\boldsymbol{\omega})^2}{\partial \omega_j \partial \omega_k}\Big|_{\boldsymbol{\omega}=\hat{\boldsymbol{\omega}}^{(k)}} = -\sum_{t=1}^{T}\sum_{d=1}^{D} R_{(t,d)} \frac{(I_{(t-l,d)} - I_{(t-L,d)})(I_{(t-k,d)} - I_{(t-L,d)})}{(\sum_{l=1}^{L}\hat{\omega}_l^{(k)} I_{(t-l,d)})^2}, \tag{27}$$

for $l = \{1, \dots, L-1\}$, $j = \{1, \dots, L-1\}$ and $k = \{1, \dots, L-1\}$. These terms are derived to determine the second-order approximation (11).

B Truncated joint probability mass function

First, we define a reasonable range $[0, I_{max}]$ of probable income values $I_{(t,d)}$, such that p$(I_{(t,d)} \geq I_{max},\ R_{(t,d)} \geq I_{max} - \Delta_{(t,d)} | \lambda^I_{(t,d)}, \lambda^R_{(t,d)}) \approx 0$. Then we calculate the conditional probability

$$\begin{aligned} p(I_{(t,d)} = i,\, R_{(t,d)} = i - \Delta_{(t,d)} | I_{(t,d)} \leq I_{max}; \lambda^I_{(t,d)}, \lambda^R_{(t,d)}) \quad (28) \\ = \lim_{Q\to\infty} \frac{\exp(-\lambda^I_{(t,d)})[\lambda^I_{(t,d)}]^i \exp(-\lambda^R_{(t,d)})[\lambda^R_{(t,d)}]^{i-\Delta_{(t,d)}}(i!(i-\Delta_{(t,d)})!)^{-1}}{\sum_{j=0}^{Q}[\exp(-\lambda^I_{(t,d)})[\lambda^I_{(t,d)}]^j \exp(-\lambda^R_{(t,d)})[\lambda^R_{(t,d)}]^{j-\Delta_{(t,d)}}(j!(j-\Delta_{(t,d)})!)^{-1}]} \\ \approx \frac{\exp(-\lambda^I_{(t,d)})[\lambda^I_{(t,d)}]^i \exp(-\lambda^R_{(t,d)})[\lambda^R_{(t,d)}]^{i-\Delta_{(t,d)}}(i!(i-\Delta_{(t,d)})!)^{-1}}{\sum_{j=0}^{I_{max}}[\exp(-\lambda^I_{(t,d)})[\lambda^I_{(t,d)}]^j \exp(-\lambda^R_{(t,d)})[\lambda^R_{(t,d)}]^{j-\Delta_{(t,d)}}(j!(j-\Delta_{(t,d)})!)^{-1}]}, \end{aligned}$$

$\forall\ i\ \in \{0, \dots, I_{max}\}$. For conciseness, we omitted the indicator for sampling at the k-th iteration.

Nota bene: The bias correction need not be conducted at every iteration of the sEM. Particularly, the estimation of parameters where the likelihood is multimodal, or the assumed model is highly complex. A suggested solution is to conduct a sEM, without the bias correction until convergence is reached, and then use the obtained estimates as starting values for conducting an sEM with bias correction.

C Convergence

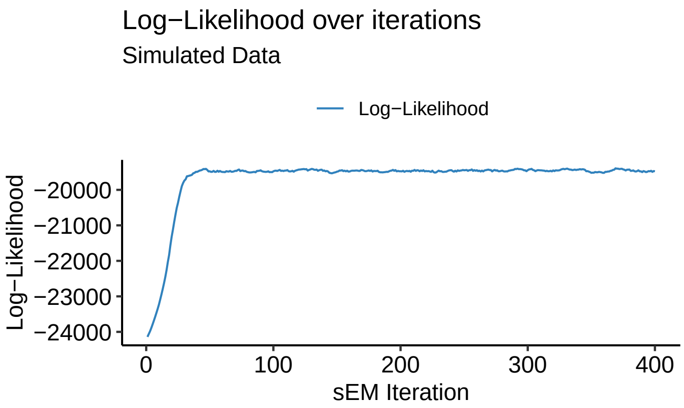


Figure 9 Log-Likelihood over 400 iterations of the sEM applied to simulated data.

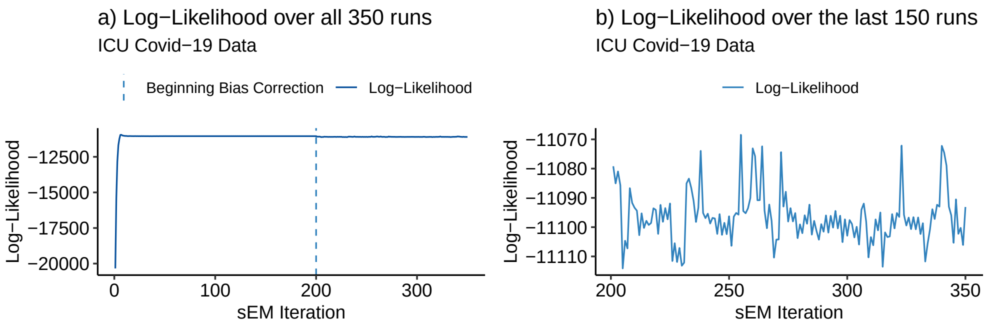


Figure 10 a) Log-Likelihood over 350 iterations of the sEM applied to ICU COVID-19 data (initial 200 iterations 'burn-in' without bias correction, subsequent 150 iterations are implemented using bias correction). b) Log-Likelihood zoomed in over the last 150 iterations of the sEM applied to ICU COVID-19 data.
NB: The y-axes in a) and b) are on different scales.

D Simulation-Predicted Incoming and Outgoing

Inspecting Figure 4, and Figure 11 to Figure 14, we observe an overestimation of inflow and outflow, which diminishes as the overdispersion reduces.

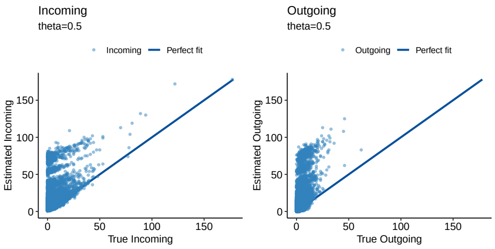


Figure 11 The estimated inflow and outflow for data with chosen data with Negative Binomial- $\theta = 0.5$.

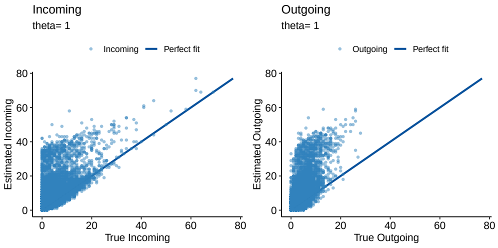


Figure 12 The estimated inflow and outflow for data with chosen data with Negative Binomial- $\theta = 1$.

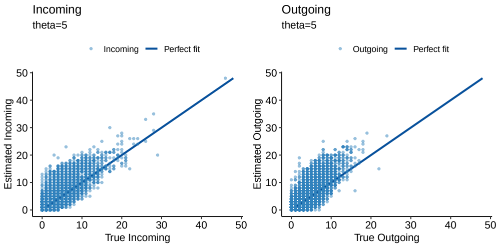


Figure 13 The estimated inflow and outflow for data with chosen data with Negative Binomial- $\theta = 5$.

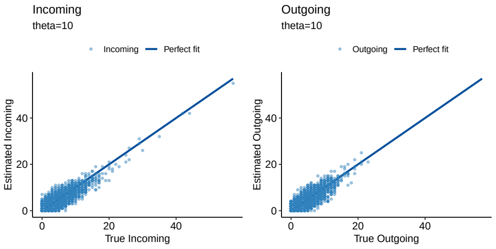


Figure 14 The estimated inflow and outflow for data with chosen data with Negative Binomial- $\theta = 10$.

Contributing Publications

Martje Rave and Göran Kauermann (2025). Deriving Duration Time from Occupancy Data – A Case Study in the Length of Stay in Intensive Care Units for COVID-19 Patients. Accepted at *Biometrical journal: submitted* 5th *of May, 2025, accepted* 15th *of October, 2025. Preprint: arXiv preprint.* *https://doi.org/10.48550/arXiv.2505.02587*.

Martje Rave and Göran Kauermann (2024). The Skellam distribution revisited: Estimating the unobserved incoming and outgoing ICU COVID-19 patients on a regional level in Germany. *Statistical Modelling: 2024;25(3):270-280.* https://doi.org/10.1177/1471082X241235024.

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:, 24(4):344–367.* https://doi.org/10.1177/1471082X221124628.

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.

Flensburg, den 22.09.2025

Martje Rave