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.

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

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

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 $R_{w,r} = (R_{w,r,1}, ..., R_{w,r,A})$ 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 $\mathbb{E}(R_{r,a}) = \pi_{w,a}$ yields for model (3.1) under (3.3):

$$\mathbb{E}\left(\tilde{Y}_{w,r,a} \mid \tilde{Y}_{w-1,r}\right) = \mathbb{E}_{R_w, R_{w-1}}\left(\mathbb{E}_{Y_w}\left(R_{w,r,a} Y_{w,r,a} \mid \tilde{Y}_{w-1,r}, R_{w,r,a}, R_{w-1,r}\right)\right)$$

$$= \mathbb{E}_{R_w, R_{w-1}}\left(R_{w,r,a} \mathbb{E}_{Y_w}\left(Y_{w,r,a} \mid Y_{w-1,r}\right)\right)$$

$$= \pi_{w,a} \mathbb{E}_{R_{w-1}}\left(\exp\{\eta_{w,r,a}\}\right) \exp\{o_{r,a}\} \tag{3.4}$$

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

$$\mathbb{E}_{R_{w-1}}\left(\exp\{\eta_{w-1,r,a}\}\right) = \mathbb{E}_{R_{w-1}}\left(\exp\left\{\sum_{k=1}^{A}\log(R_{w-1,r,k}^{-1}\tilde{Y}_{w-1,r,k})\theta_{a,k} + \theta_w\right\}\right)$$

$$= \exp\left\{\tilde{\eta}_{w,r,a}\right\}\mathbb{E}_{R_{w-1}}\left(\exp\left\{\sum_{k=1}^{A}\log(R_{w-1,r,k}^{-1})\theta_{a,k} + \theta_w\right\}\right)$$

$$= \exp\left\{\tilde{\eta}_{w,r,a} + \tilde{\theta}_w\right\}, \tag{3.5}$$

where

$$\tilde{\eta}_{w,r,a} = \sum_{k=1}^{A}\log(\tilde{Y}_{w-1,r,k})\theta_{a,k}$$

and

$$\tilde{\theta}_w = \theta_w + \log\left(\mathbb{E}_{R_{w-1}}\left(\exp\left\{\sum_{k=1}^{A}\log(R_{w-1,r,k}^{-1})\theta_{a,k}\right\}\right)\right).$$

Hence, combining (3.4) and (3.5) shows that if we fit the model (3.2) to the observed data, which are affected by unreported cases, we obtain the same autoregressive coefficients $\theta_{a,k}$ for $k = 1, ..., A$ 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.
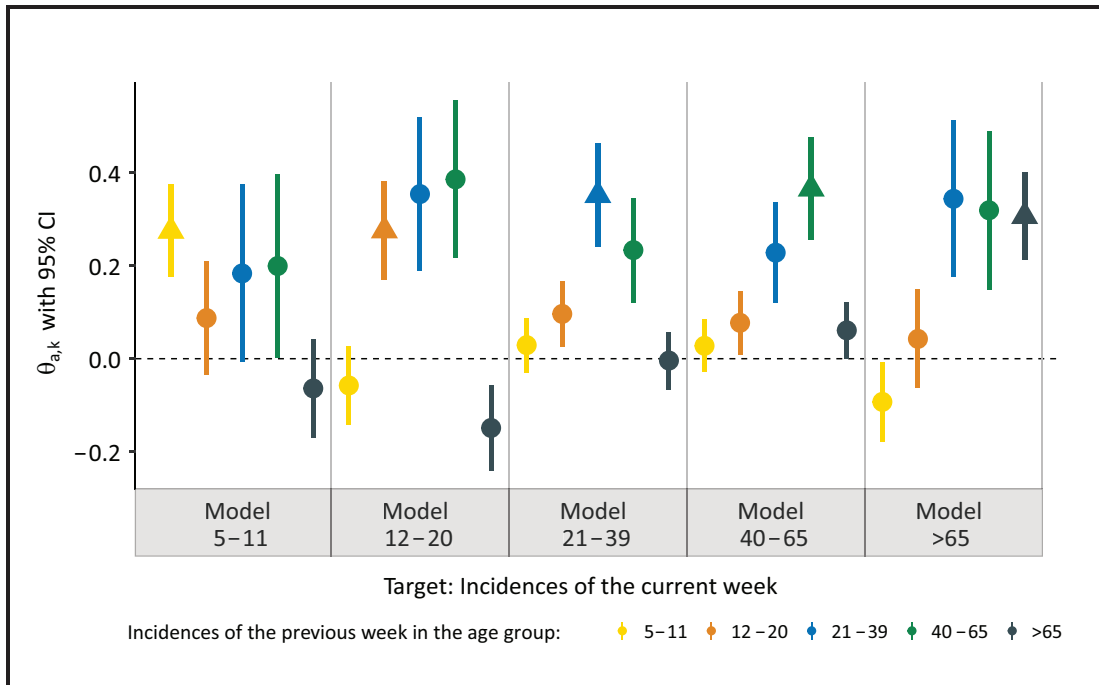
Figure 2   Association of previous week's incidences in different age groups (colour-coded) with the current-week incidences for calendar weeks 9–12 in 2021 stratified by age group (5 age groups correspond to 5 distinct Models)

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.
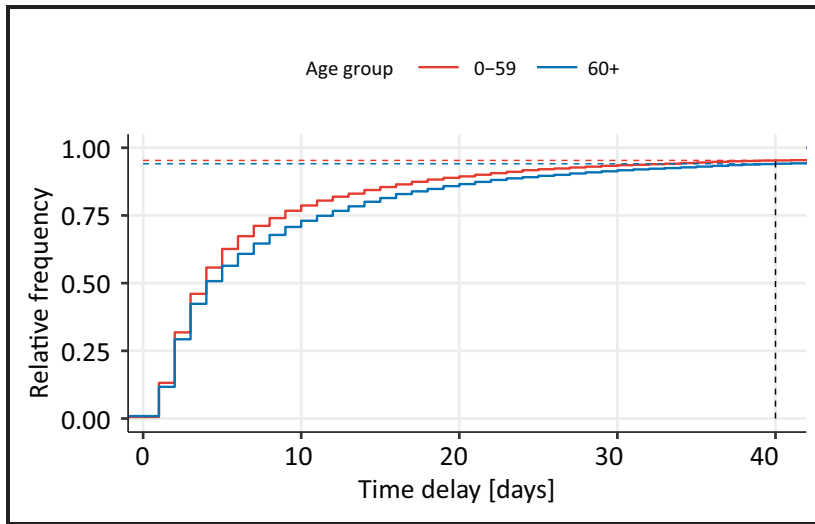
Figure 3 Cumulative distribution function of the time delay (in days) between hospitalization and its reporting, calculated with data from 1 January to 18 November of 2021, shown separately for the age groups 0–59 and 60+. The curves for both age groups are truncated at a delay of 40 days, when approximately 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 $R_{t,r,g}$
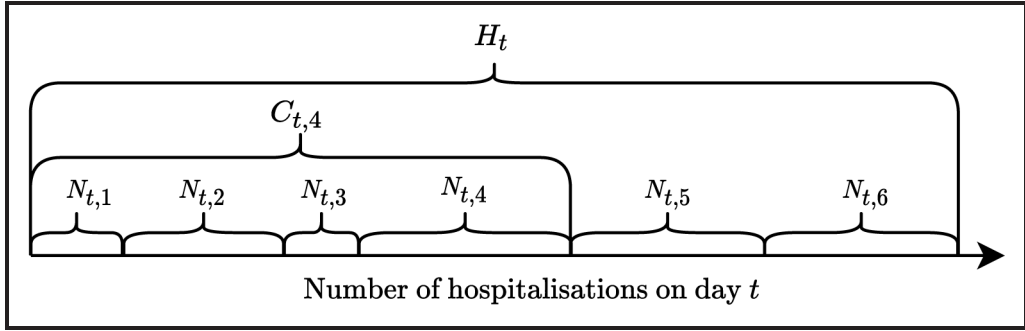
Figure 4 Illustration of the data setting for $d_{max} = 6$. $N_{t,d}$ indicates hospitalizations reported with a specific delay d, while $C_{t,d}$ denotes all those reported with delay up to d. H_t denotes the final number of hospitalized cases regardless of the delay with which they were reported, 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:

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

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

$$\mathbb{E}(R_{t,r,g}|x_{t,r,g}) = \exp\{\theta_0 + \theta_{\text{age}}x_{\text{age},g} + \theta_{\text{gender}}x_{\text{gender},g} + \theta_{\text{gender:age}}x_{\text{age},g}x_{\text{gender},g} +$$

$$\theta_{\text{weekday}}x_{\text{weekday},t} + s_1(t) + s_2(x_{\text{Lon},r}, x_{\text{Lat},r}) + u_r\}$$

$$= \exp\{\eta_{t,r,g}\}, \tag{4.8}$$

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:

$$C_{t,r,g,T-t}|x_{t,r,g} \sim \mathrm{NB}(\mu_{t,r,g}, \sigma^2),$$

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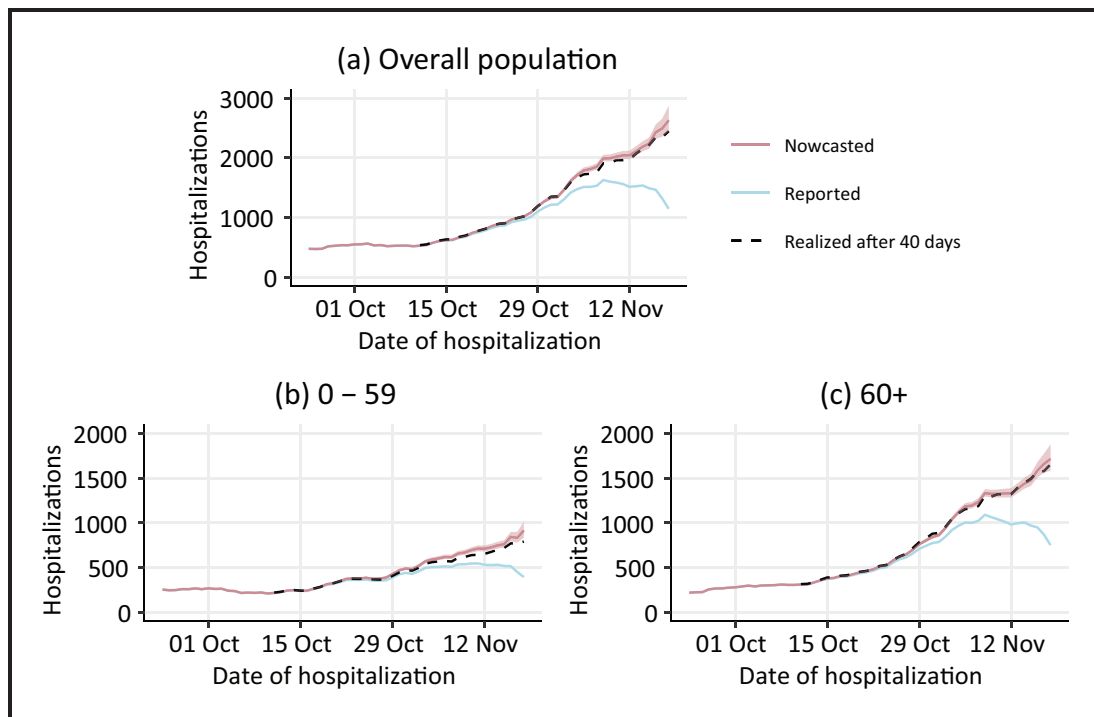
Figure 5 Comparison of nowcasted (red) and reported (blue) rolling weekly sums of hospitalization counts between 24 September and 18 November 2021, based on data reported as of 19 November 2021. Note: 95% confidence intervals of the nowcast estimates are indicated by the shaded areas. The dashed black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

0.0 0.4 0.8 1.2

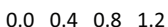**Figure 6** Estimated linear effects for different age and gender groups in the hospitalization model, where males aged 15–34 are the reference category. Note: Estimated standard deviations are written in brackets

(a) Spatial effect

(b) Random effect

Berchtesgadener Land

−0.2 0.0 0.2 0.4

−1 0

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:

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

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:

$$\eta_{w,r,j} = \theta_{0,j} + \theta_{AR(1),j}^\top (\check{Z}_{w-1,r,1}, \check{Z}_{w-1,r,2})^\top + \theta_{I,j}^\top \log(Y_{w-1,r} + \delta) +$$

$$s_j(x_{\text{Lon},r}, x_{\text{Lat},r}) + u_{r,j} \ \forall \ j = 1, 2, \quad (5.3)$$

where $\theta_{0,j}$ is the intercept term. Further, we incorporate an autoregressive component in (5.3) by including the relative ICU occupancy observed in the previous week as a regressor. We denote the distribution of the different occupancies of the previous week as $\check{Z}_{w-1,d} = (Z_{w-1,r,1}, Z_{w-1,r,2})/(\sum_{j=1}^{3} Z_{w-1,r,j})$, 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, $u_{r,j} \sim N(0, \tau^2) \ \forall \ r \in \{1, \ldots, R\} \ \forall \ j = 1, 2$. 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:

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

with $\pi_{w,r}$ specified in (5.3). Note, however, that we do not know α and that no information is provided in the data concerning the length of stay or the number of beds changing their status. To account for that data deficiency, we impose a Dirichlet distribution on the vector $\pi_{w,r}$, where the

prior information is determined by the available beds, that is,

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

(5.5)

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

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

(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:

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

(5.7)

where we set $\mathbf{A}(\theta) = \mathbb{E}\left(-\frac{\partial}{\partial\theta\partial^\top\theta}\ell(\theta)\right)$, $\mathbf{B}(\theta) = \text{Var}\left(\frac{\partial}{\partial\theta}\ell(\theta)\right)$, and $\ell(\theta)$ is the logarithmic likelihood resulting from (5.1) or equivalently the logarithm of the posterior of (5.3). See also Stefanski and Boos (2002) and Zeileis (2006).

5.3 Application to the third wave

We now employ the multinomial logistic regression (5.1) to ICU data recorded during the third wave between March and June 2021. For the incidence data used in the covariates, we employ the RKI data; hence we set $A = 4$ and the age groups are: 15–34, 35–59, 60–79 and 80+. Further, we normalize all non-binary covariates:

$$\tilde{x}_i = \frac{x_i - \bar{x}}{\sqrt{\frac{1}{n}\sum_{j}^{n}(x_j - \bar{x})^2}} \text{ with } \bar{x} = \frac{\sum_{j}^{n} x_j}{n}.$$

(5.8)

This way, we facilitate the interpretation of associations and guarantee a meaningful comparison between the covariates. Due to space restrictions, we here only present the linear effects from (5.3) and refer to the Supplementary Material for the random and smooth estimates.

In Figure 8, we visualize the estimated coefficients, including their confidence intervals. The reference category in both pairwise comparisons is COVID-beds; thus, we refer to the two models as free vs COVID beds and non-COVID vs COVID beds. In particular, the coefficients relate to the association between the covariates and the logarithmic odds of a bed not being occupied compared to being occupied by a patient with COVID-19, shown with blue dots in Figure 8. Analogously, the orange triangles in Figure 8 illustrate the estimated association between the covariates and the logarithmic odds of a bed being occupied by a patient not infected with COVID-19 in comparison to a
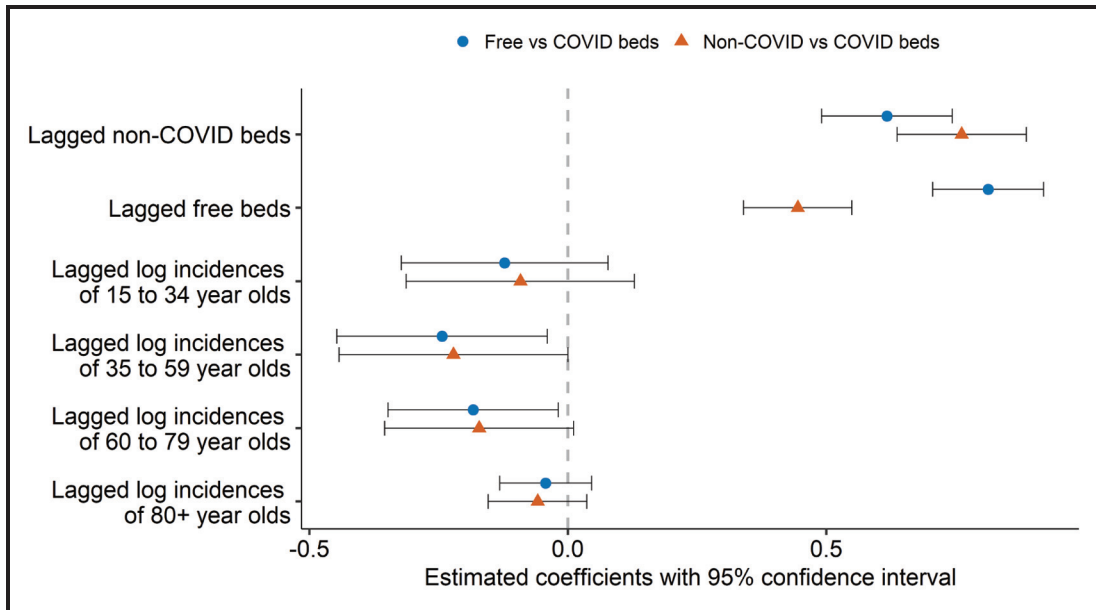
Figure 8 Estimated coefficients with confidence interval of the associations between normalized linear covariates included in the multinomial model and the logarithmic odds of a bed being free vs occupied by a patient infected with COVID-19 (blue dots) and the logarithmic odds of a bed being occupied by a patient not infected with COVID-19 vs a patient infected with COVID-19 (orange triangles)

bed being occupied by a patient infected with COVID-19. To demonstrate the uncertainty of each estimate, a 95% confidence interval is added. Keeping the other variables constant, the normalized lagged log-incidences of all age groups generally have a negative effect on the logarithmic odds of both pairwise comparisons. This translates to the finding that an increase in the incidences leads to a decrease in the proportion of non-COVID and free-beds in when compared to COVID beds. The lagged normalized proportion of free and non-COVID beds is estimated to have a stronger, positive association with the logarithmic odds of both pairwise comparisons. We, therefore, expect a higher number of non-COVID beds in the previous week to be followed by a higher number of non-COVID beds in the next week.

The model can be extended to a forecasting model, as shown in the supplementary material. In particular, we demonstrate how forecasting performance changes over the different waves of the pandemic. In principle, we could also incorporate further covariates like district-specific proportions of vaccinated people. Unfortunately, these numbers are not very reliable and require sophisticated cleaning, so we prefer not to present results in this direction here.

6 Discussion

The COVID-19 pandemic poses numerous complex challenges to scientists from different disciplines. Statisticians and epidemiologists, in particular, face the problem of extracting meaningful in-

formation from imperfect, incomplete and rapidly changing data. Generalized additive models are a powerful tool that, if used correctly, can help solving some of these challenges. In this work, we have addressed three such challenges where the utilization of GAMs provided meaningful insight.

1. We investigated whether children are the main drivers of the pandemic under a time-varying case-detection ratio.
2. We modelled hospitalization incidences controlling for delayed registrations, thereby providing both up to dates estimates of current hospitalization numbers as well as insight on the demographic and spatio-temporal drivers of COVID-19.
3. We developed an interpretable predictive tool for ICU bed occupancy that is actively used by the Bavarian government.

We achieved all of those results by using GAMs with different methodological extensions. Nevertheless, the use of our proposed models to extract novel information from the data provided is still subject to both data-related and methodological limitations. In general, our data sources are subject to exogenous shocks (e.g., policy changes) that lead to sudden changes in population behaviour and pose a danger to the validity of our results. Regarding the study of infection dynamics of school kids, revised testing policies hinder the long-range comparability of our findings. In the hospitalization data, the exact date of hospitalization is missing for about 10% of the hospitalized cases, which we impute by the given registration date of the infection. Furthermore, the records on the ICU-bed occupancy do not include intrinsic constraints, as the capacity of beds available to COVID-19 patients does not equate to the capacity of beds available to patients not infected with COVID-19. There are also methodological limitations. First of all, note that the data is observational, not experimental. Additionally, the set of covariates in our model can easily be extended to control for other factors, such as meteorological and socioeconomic ones.

We close this work by emphasizing that the nowcasting model can also be used as a stand-alone model. In the German COVID-19 Nowcast Hub (KIT), the described model is used among other nowcasting methods, including the work of Günther et al. (2020) and van de 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

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.
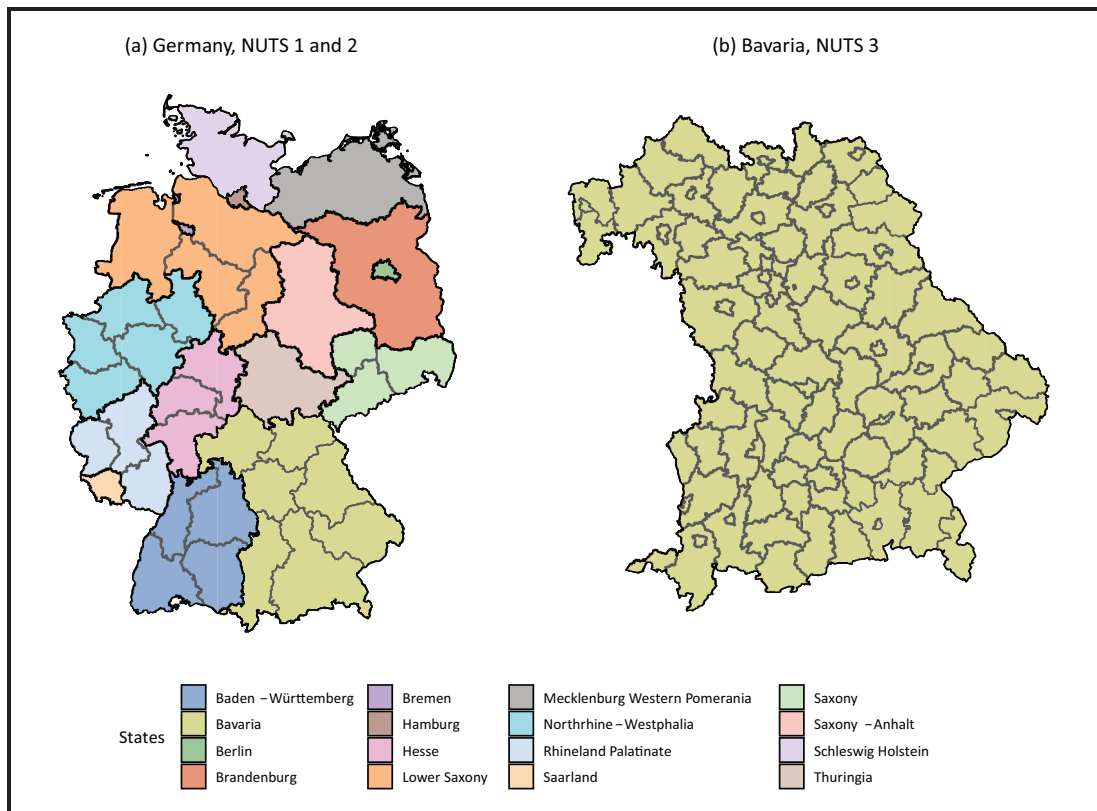
Funding

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.