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).

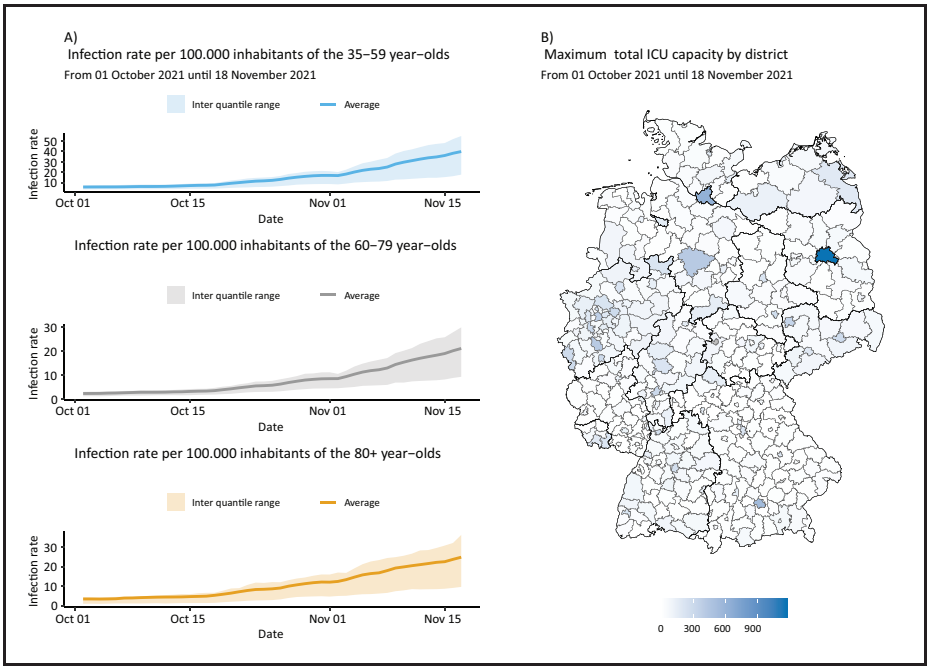

Figure 1 A: Summary over all districts of the infection rate per 100.000 inhabitants by age group, '45–59' year-olds, '60–79' year-olds and '80+' year-olds displayed by date, from the 1 October 2021 until the 18 November 2021, B: The maximum capacity of ICU beds per given district over the time span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

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

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

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:

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

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

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

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

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

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 200^{th} run to the 500^{th} 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

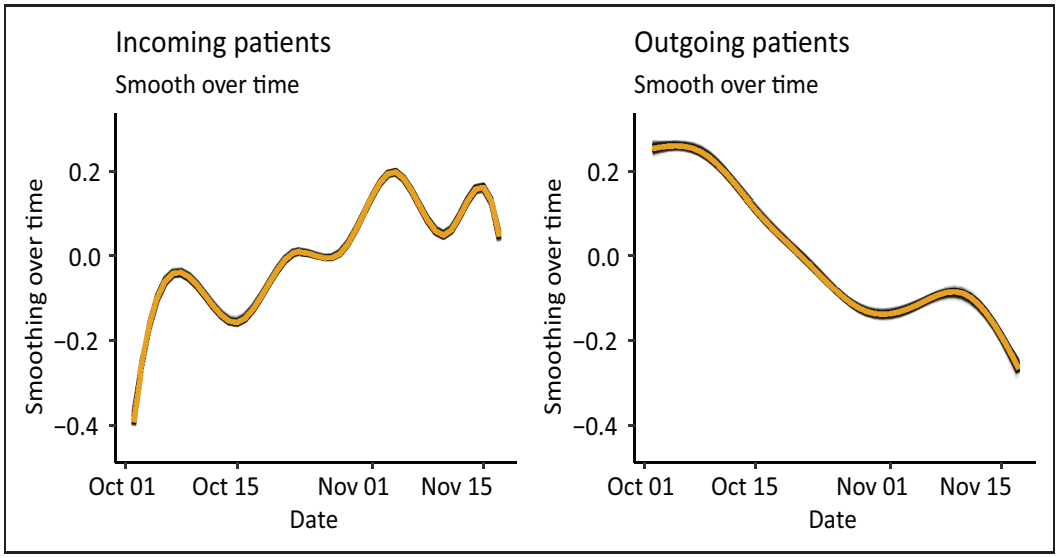

Figure 2 Estimated smooth functions of all runs, over time, rendered by the GAMs estimating the number of incoming patients (left hand side) and outgoing patients (right hand side) over the last 300 runs.

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

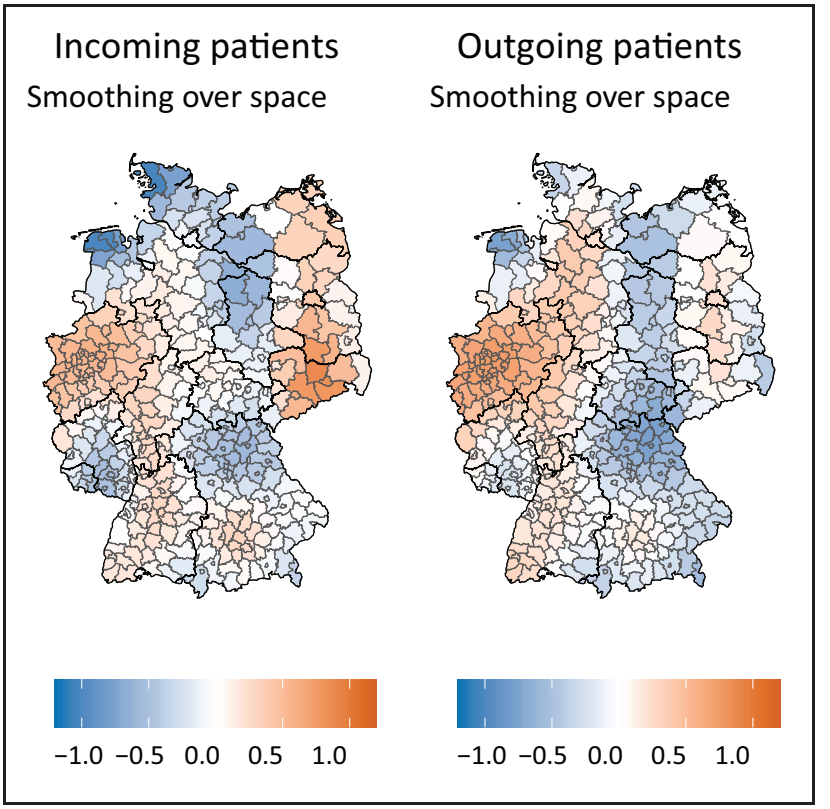

Figure 3 Estimated median smooth functions of the 200^{th} until the 500^{th} run over space rendered by the generalized additive models, estimating the number of incoming patients (left-hand side) and outgoing patients (right-hand side).

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

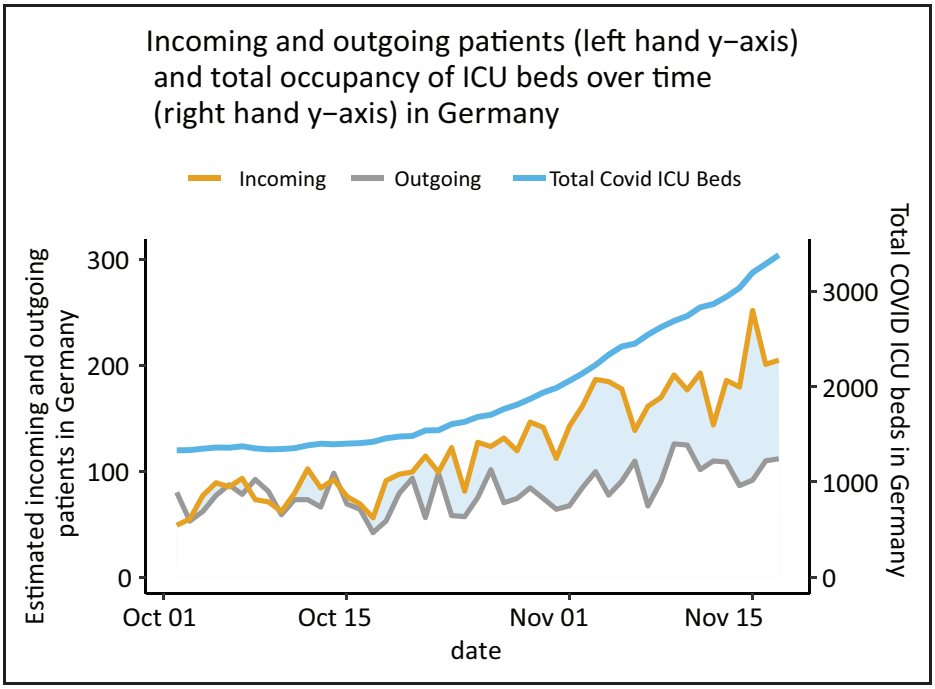

Figure 4 Estimated number of incoming and outgoing patients by date from the 1 October 2021 until the 18 November 2021, as well as the total number of COVID-19 patients in the ICUs of Germany.

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.

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

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

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

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

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

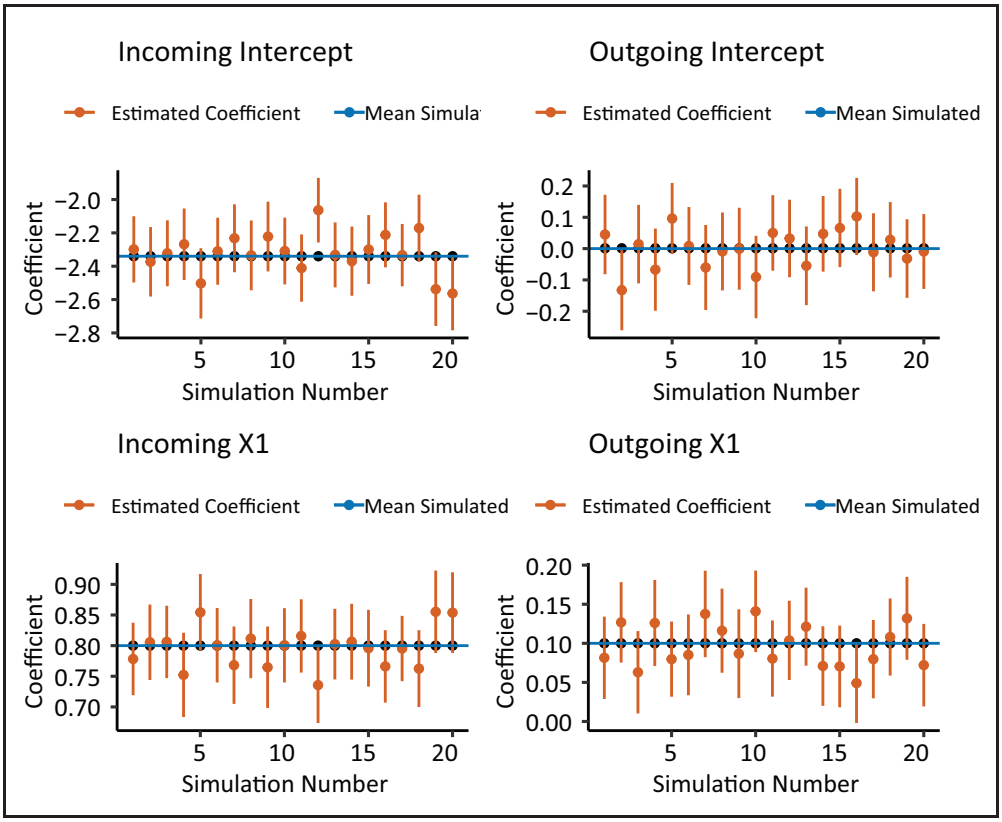

Figure 5 The estimated coefficients for twenty simulated data sets.

the twenty runs include the real coefficient, except for the 'Incoming Intercept' coefficient in the 12^{th} 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.

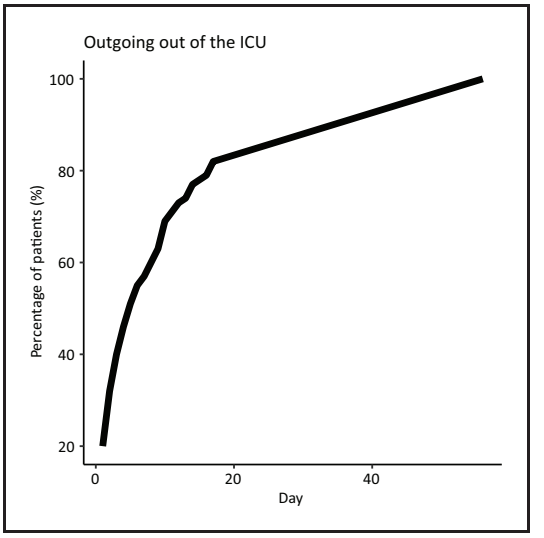

Figure A1 Percentage of outgoing ICU patients after the day of admission.

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.

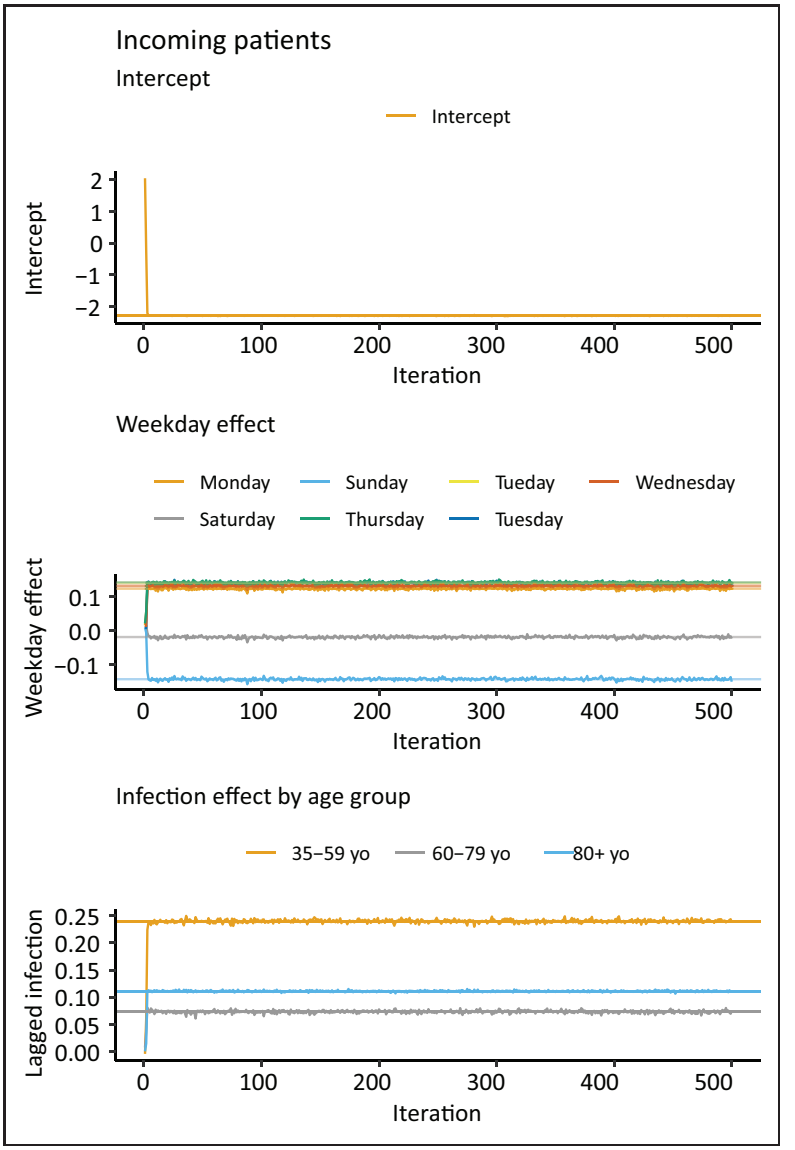

Figure A2 Coefficients estimated by the generalized additive models of the last three hundred runs of the EM algorithm of the incoming patients.

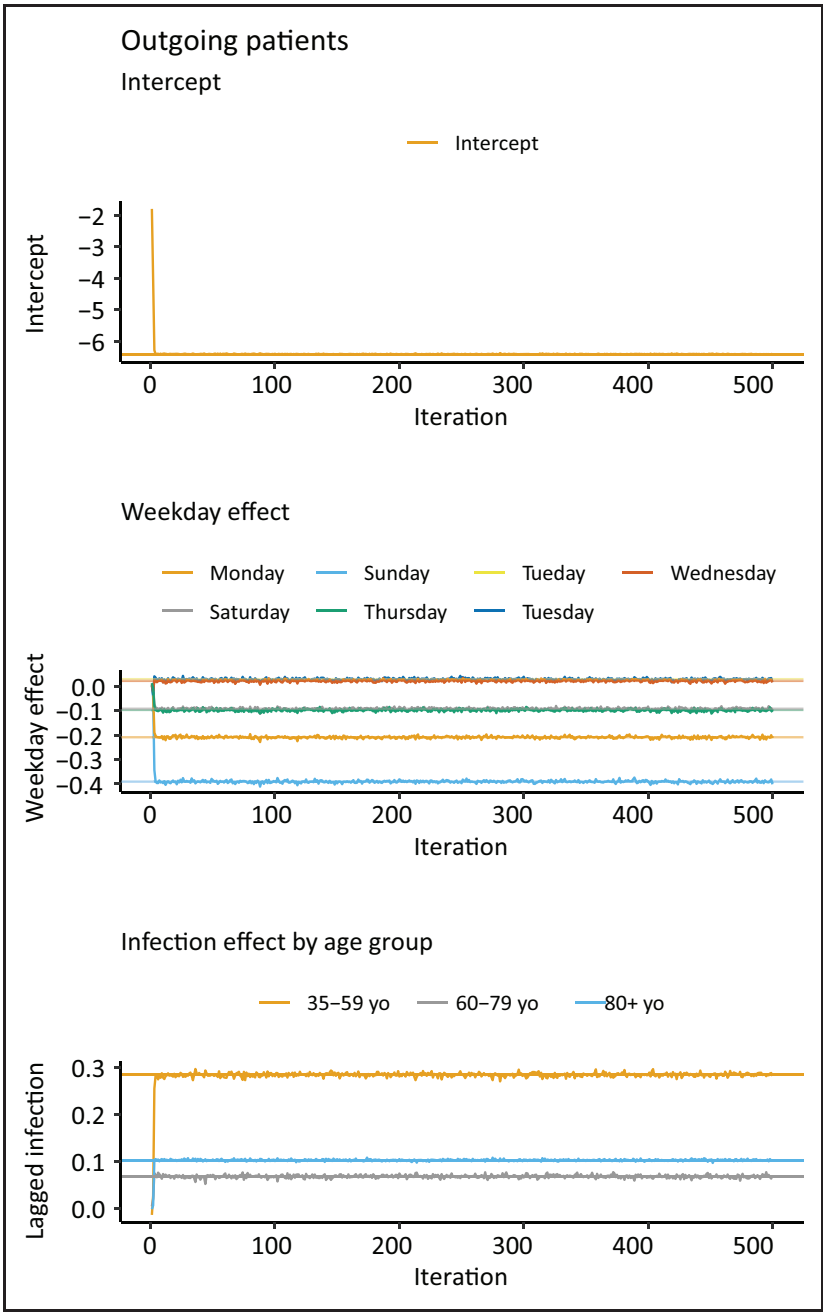

Figure A3 Coefficients estimated by the generalized additive models of the last three hundred runs of the EM algorithm of the outgoing patients.

3 Pseudoalgorithms

Algorithm 1 Pseudo algorithm of the SEM

Require: $\boldsymbol{\delta}$ $\rightarrow$ $\boldsymbol{\delta}$ is the observed difference.

· $\boldsymbol{\lambda}^0_{In} \in \{1,\ldots,10\}$, $\boldsymbol{\lambda}^0_{Out} \in \{11,\ldots,20\}$ $\rightarrow$ Randomly chosen starting values for the Poisson parameters.

· $\mathbb{P}(\boldsymbol{I}^0 = \boldsymbol{k}, \boldsymbol{R}^0 = \boldsymbol{k} - \boldsymbol{\delta} | \boldsymbol{\lambda}^0_{In}, \boldsymbol{\lambda}^0_{Out}) = Poi_{\boldsymbol{\lambda}^0_{In}}(\boldsymbol{k}) Poi_{\boldsymbol{\lambda}^0_{Out}}(\boldsymbol{k} - \boldsymbol{\delta})$ $\rightarrow$ Calculate the probability density function for $\boldsymbol{k} = [1,\ldots,1000]$. *

· $[\boldsymbol{I}^0_{sim}, \boldsymbol{R}^0_{sim}] \sim \mathbb{P}(\boldsymbol{I}^0, \boldsymbol{R}^0)$ $\rightarrow$ Simulate $\boldsymbol{I}^0_{sim}$ and $\boldsymbol{R}^0_{sim}$ from the joint probability distribution.

for $i \in \{1,\ldots,500\}$ **do**

'M'-Step Estimate $\hat{\boldsymbol{\lambda}}^i_{In}$ and $\hat{\boldsymbol{\lambda}}^i_{Out}$ by using two generalized additive models.

'E'-step Simulate the number of incoming and outgoing patients from $[\boldsymbol{I}^i_{sim}, \boldsymbol{R}^i_{sim}] \sim \mathbb{P}(\boldsymbol{I} = \boldsymbol{k}, \boldsymbol{R} = \boldsymbol{k} - \boldsymbol{\delta} | \boldsymbol{\lambda}^i_{In}, \boldsymbol{\lambda}^i_{Out}) = Poi_{\boldsymbol{\lambda}^i_{In}}(\boldsymbol{k}) Poi_{\boldsymbol{\lambda}^i_{Out}}(\boldsymbol{k} - \boldsymbol{\delta})$.

end for

Return $\rightarrow$ 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.