Modeling the occurrence of events subject to a reporting delay via an EM algorithm

Roel Verbelen, Katrien Antonio, Gerda Claeskens, Jonas Crevecoeur

Modeling the occurrence of events subject to a reporting delay via an EM algorithm

Roel Verbelen[1,3,4], Katrien Antonio[1,2,3,4], Gerda Claeskens[1,3], and Jonas Crevecoeur[1,4]

[1] *Faculty of Economics and Business, KU Leuven, Belgium.*
[2] *Faculty of Economics and Business, University of Amsterdam, The Netherlands.*
[3] *LStat, Leuven Statistics Research Center, KU Leuven, Belgium.*
[4] *LRisk, Leuven Research Center on Insurance and Financial Risk Analysis, KU Leuven, Belgium.*

June 24, 2021

Abstract

A delay between the occurrence and the reporting of events often has practical implications such as for the amount of capital to hold for insurance companies, or for taking preventive actions in case of infectious diseases. The accurate estimation of the number of incurred but not (yet) reported events forms an essential part of properly dealing with this phenomenon. We review the current practice for analysing such data and we present a flexible regression framework to jointly estimate the occurrence and reporting of events. By linking this setting to an incomplete data problem, estimation is performed via an expectation-maximization algorithm. The resulting method is elegant, easy to understand and implement, and provides refined insights in the nowcasts. The proposed methodology is applied to a European general liability portfolio in insurance.

Keywords: EM algorithm, Nowcasting, Poisson regression model, Reporting delay

1 Introduction

This paper reviews and extends the literature on statistical models for problems where individuals (or objects) under study experience two events. The first, also called the initiating or primary, event occurs at time x, and the second, the so-called secondary or consequent, event only occurs at a later time $s \geq x$. The presence of the delay $u = s - x$ between the two events leads to statistical challenges, because the currently observed number of primary events is right-censored while observation delays are right-truncated, see e.g. the early contributions by Lagakos et al. (1988), Kalbfleisch and Lawless (1989), Harris (1990) and Kalbfleisch and Lawless (1991) for an introduction to the field. The term *back-calculation* (Brookmeyer and Gail, 1988; Bacchetti et al., 1993) refers to the reconstruction of the past history of first events that must have occurred to give rise to the observed pattern of second event cases, under the assumption of a known delay distribution. Nowadays, *nowcasting* is often used for estimating the current number of first events using only the available partial information on the reported or registered secondary events (see Höhle and an der Heiden, 2014; van de Kassteele et al., 2019; Bastos et al., 2019, for recent examples of nowcasting problems in epidemiology).

Such delays occur in different ways and in a variety of subject areas. In an insurance setting a claim is only reported some time after its occurrence, because the damage was not

immediately noticed or the insured needed some time to file the claim to the insurance company. A proper estimation of these unreported claims is important, since financial regulations force insurance companies to hold sufficient capital reserves to be able to fulfill their future liabilities with respect to such claims. Jewell (1989) sketches first contributions to the modelling of these occurred- or incurred-but-not-reported (OBNR or IBNR) claims in an insurance context. In disease modelling at least two examples of these delays are studied in the literature. In the first, x represents the time of diagnosis of a case (or another relevant event, like hospital admission in Donker et al., 2011) and s is the time of reporting of the case to the organization that coordinates the surveillance of the disease. This reporting delay may result from various processes including logistics, or the time to complete a test and to report a confirmed case in a health database. Statistical surveillance systems for the detection of outbreaks of infectious diseases have to properly adjust for reporting delays in order to take timely preventive action (see e.g. Farrington et al., 1996; Noufaily et al., 2015, 2016). In the second example x measures the time of infection (with a virus, for example) and s is the time of onset of disease symptoms. Insight in the distribution of the incubation time $u = s - x$ is important to estimate the current number of infections in a population. In reliability engineering and quality management, the statistical analysis of warranty data requires taking the time between the failure and its reporting into account in order to predict the number of future warranty claims from all units in service (see e.g. Kalbfleisch et al., 1991; Wu, 2013).

The contribution of our paper is threefold. First, Section 2 sketches (a selection of recent) contributions on nowcasting that appeared in actuarial, statistical or epidemiological literature. Our literature overview stretches across multiple disciplines and structures these papers along the time scale used in the modelling framework, where we distinguish between models for events in continuous time and models for data aggregated over a coarser discrete time grid. Second, we contribute to the literature by proposing in Section 3 a flexible yet practical modelling and estimation framework capable of dealing with any parametric structure for both the occurrence as well as the reporting process. We employ an expectation-maximisation (EM) method (Dempster et al., 1977) to jointly estimate the occurrence and reporting delay of the events in the presence of covariates, allowing us to acquire the necessary insights in the dynamics of both. Third, Section 4 presents the results of applying the model in a case study where out-of-time evaluations are used to asses the predictive performance of the method. Moreover, we benchmark our results against the findings obtained with a selection of modelling strategies from the literature overview. Section 5 concludes. The supplementary document contains some additional results regarding the case study.

2 Literature overview and notation

The event of interest (e.g. the failure of a product, the occurrence of an insured event or the diagnosis of disease) happens at time X, though it is only reported or observed at a later time S. The reporting delay is then $U = S - X$. Figure 1 represents the occurrence of multiple events over time. When the reporting occurs before τ, the observation is complete (events 1 and 3 in Figure 1), while reporting after τ corresponds with a currently unreported case (events 2 and 4 in Figure 1). At time τ the analyst has to predict or evaluate the number of events that incurred in the past, but will only be reported in the future. This is challenging because the analyst faces incomplete data, with no information available for the unreported cases. Leaving the continuous time setting, the dashed grid in Figure 1 pictures the aggregation of events when choosing a coarser discrete time scale. The objective is to use the observed, reported events in the upper triangle $\mathbf{\Delta}^{\text{r}} = \{(X, U) \mid X \leq \tau,\ U \leq \tau - x\}$ to predict unreported events in the lower triangle $\mathbf{\Delta}^{\text{nr}} = \{(X, U) \mid X \leq \tau,\ U > \tau - X\}$.

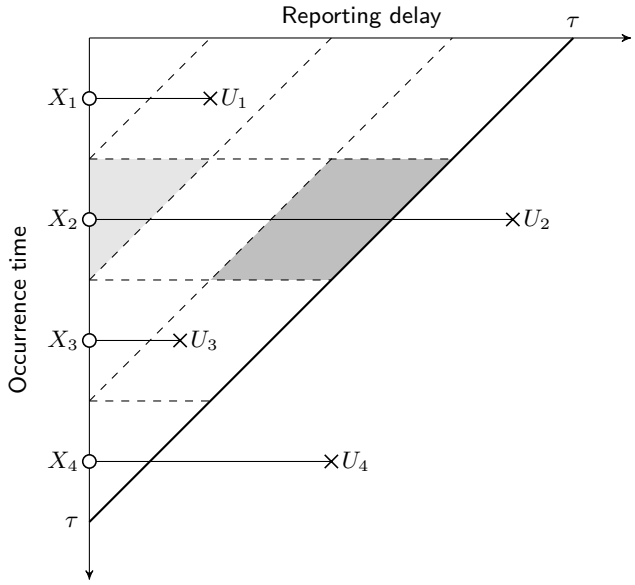

Figure 1: Occurrence and reporting of multiple events over time. Events 2 and 4 are unobserved as these events get reported after the evaluation date τ. The dashed grid indicates the aggregation of the continuous data towards a coarser discrete time scale. The events registered in continuous time in the shaded regions will be aggregated into a single cell or observation in the run-off triangle pictured in Table 1.

2.1 Continuous time models

When data are collected in (almost) real-time, predicting the number of unreported events reduces to estimating the two-dimensional density $f_{X,U}$ for the occurrence and reporting of events in $[0, \tau] \times [0, \tau]$. Hereto, only observations observed on the upper triangle in Figure 1 are available. Martínez Miranda et al. (2013) and more recently Hiabu et al. (2021) estimate this density with a two-dimensional kernel density estimator with support on the upper triangle. By assuming a multiplicative kernel, the local linear density estimate can be extrapolated to the lower triangle which then leads to a forecast for the occurred but not yet reported events. In many applications, we are interested in the marginal densities, i.e. the occurrence intensity of events on the one hand and the reporting delay distribution on the other hand. This interest in the marginal densities is commonly translated into the following assumptions

(A1) The event counting process $N(x) = \sum_{i \geq 1} \mathcal{I}\{X_i \leq x\}$ for $x \geq 0$ follows an inhomogeneous Poisson point process with an intensity $\lambda(x; \boldsymbol{\alpha})$, which depends on some parameter vector $\boldsymbol{\alpha}$.

(A2) Conditional on the occurrence time X, the reporting delay U follows a positive continuous distribution with density $f_{U|X}(\cdot; \boldsymbol{\theta})$ and parameter vector $\boldsymbol{\theta}$. The reporting delay U is independent of the event counting process $N(x)$.

Let $\boldsymbol{\Delta}^{\mathrm{r}} = \{(x_i, u_i),\ i = 1, \ldots, n | x_i \leq \tau,\ u_i \leq \tau - x_i\}$ denote the n observed events in the upper

triangle. The associated log-likelihood of these reported events becomes

$$\log \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\theta} | \boldsymbol{\Delta}^{\mathrm{r}}) = \sum_{(x,u) \in \boldsymbol{\Delta}^{\mathrm{r}}} \log \left\{ \lambda(x; \boldsymbol{\alpha}) \cdot F_{U|X}(\tau - x | x; \boldsymbol{\theta}) \right\} \tag{1}$$
$$- \int_0^{\tau} \lambda(v; \boldsymbol{\alpha}) \cdot F_{U|X}(\tau - v | v; \boldsymbol{\theta}) dv + \sum_{(x,u) \in \boldsymbol{\Delta}^{\mathrm{r}}} \log \left(\frac{f_{U|X}(u|x; \boldsymbol{\theta})}{F_{U|X}(\tau - x | x; \boldsymbol{\theta})} \right).$$

Optimization of this likelihood is complicated by the joint presence of $\lambda(\cdot; \boldsymbol{\alpha})$ from the occurrence process and $F_{U|X}(\cdot|\cdot; \boldsymbol{\theta})$ from the reporting process, prohibiting a split of the likelihood. The likelihood in (1) already appeared in Kalbfleisch and Lawless (1989), under the assumption of independence between the time of the initiating event and the duration of the delay. Two strategies for optimizing this likelihood commonly appear in the literature. First, a strand of research directly optimizes this likelihood, despite its complexity. For example, Kalbfleisch and Lawless (1989) use a parameterized Poisson process for infections and some parametric models for incubation time, allowing the joint estimation of the occurrence and reporting processes in continuous time. In actuarial science Haastrup and Arjas (1996) present a Bayesian and Wahl (2019) a frequentist approach with the likelihood in (1) as focus. The latter contributions build upon the fundamental ideas for the IBNR problem in actuarial science proposed by Jewell (1989), Norberg (1993) and Norberg (1999).

A second strategy opts for simplicity by proposing a (heuristic) two-stage approach. For example, Antonio and Plat (2014) fit a parametric distribution to the observed reporting delays in a first step and then plug in the estimated reporting delay distribution when estimating the parameters in the thinned Poisson process for the occurrence of reported events. The log-likelihood of the reporting process becomes

$$\log \mathcal{L}(\boldsymbol{\theta} | \boldsymbol{\Delta}^{\mathrm{r}}) = \sum_{(x,u) \in \boldsymbol{\Delta}^{\mathrm{r}}} \log \left(\frac{f_{U|X}(u|x; \boldsymbol{\theta})}{F_{U|X}(\tau - x | x; \boldsymbol{\theta})} \right).$$

Given the occurrence time $X = x$ of an observed event, its reporting delay is a right-truncated variable U with truncation point $\tau - x$. Lagakos et al. (1988), Kalbfleisch and Lawless (1991) and Lawless (1994) (among others) focus on estimating the reporting delay under right-truncation by inverting the direction of time, effectively transforming the right- into left-truncated data. Standard statistical methods for left-truncated data, such as the Cox proportional hazard model (Cox, 1972), can then be used to model the reporting of events. Badescu et al. (2016, 2019) and Avanzi et al. (2016) follow a strategy similar to Antonio and Plat (2014), but model the event occurrence process as a marked Cox process to allow for overdispersion and serial dependency. Along this strand Verrall and Wüthrich (2016) decouple the full likelihood in (1) by considering a plug-in estimate for the weekly periodic occurrence pattern of insurance claims, followed by estimating parametric distributions for the reporting delay.

2.2 Discrete time models

2.2.1 Aggregating events towards a coarser time scale.

Events are usually not registered in continuous, but in discrete time periods. Such discrete event counts result by aggregating the real time events towards a coarser time scale. Figure 1 sketches this aggregation from continuous to discrete time data. The day is the natural (discrete) time unit in many administrative systems. Daily data are often aggregated to weekly, monthly, quarterly or yearly event data, even though many granular data insights may get lost in this aggregation process. For example, Becker and Cui (1997) consider discrete time units

in quarterly periods and model the delay in entering AIDS diagnoses in a surveillance system. Insurance studies traditionally use quarterly or yearly claim counts (Wüthrich and Merz, 2008). In our notation and in the case study presented in Section 4 we consider event counts on a daily level, even though our notation is generic and may refer to any other time unit of interest (e.g., weeks, months or years). The number of events which occurred on day t is denoted by N_t, where the integer t indicates the occurrence date and ranges from 1 to τ. The number of these events which have been reported after precisely d days are denoted as N_{td} such that $N_t = \sum_{d=0}^{\infty} N_{td}$. Only those events that are reported before or at the evaluation date τ are observed, causing N_t to be right-censored and the reporting delay distribution of the observed events occurring on day t to be right-truncated at $\tau - t$. Throughout this paper we denote the reported, and therefore observed, number of events which occurred on day t by

$$N_t^{\mathrm{r}} = \sum_{d=0}^{\tau-t} N_{td}.$$

We denote the number of events that have happened on day t but are not yet reported by

$$N_t^{\mathrm{nr}} = \sum_{d=\tau-t+1}^{\infty} N_{td}\,,$$

and the total number of such unreported events over all occurrence days by

$$N^{\mathrm{nr}} = \sum_{t=1}^{\tau} N_t^{\mathrm{nr}} = \sum_{t=1}^{\tau} \sum_{d=\tau-t+1}^{\infty} N_{td}.$$

In discrete time nowcasting the overall objective is to use the observed events $\boldsymbol{N}^{\mathrm{r}} = \{N_{td} \mid 1 \leq t \leq \tau, d \geq 0, t+d \leq \tau\}$ to predict the events that are not yet reported $\boldsymbol{N}^{\mathrm{nr}} = \{N_{td} \mid 1 \leq t \leq \tau, d > 0, t+d > \tau\}$.

A convenient way to represent these event counts N_{td} is with a two-way contingency table where the rows correspond to occurrence time t and the columns indicate the reporting delay d. Since we can only observe what has been reported, an incomplete two-way contingency table results where the upper left-hand *triangle* is observed and the lower right-hand *triangle* is empty and has to be predicted. Table 1 depicts the structure of such a triangle with reported events, often called a run-off triangle in actuarial literature or a reporting triangle in epidemiology. The shaded regions in Figure 1 and Table 1 visualize the aggregation of events registered in continuous time into a single cell or observation in this triangle. The dimension of the triangle depends on the granularity of the discretization. Smaller bin sizes correspond to larger triangles, which usually result in more complex models to capture the heterogeneity present in the data.

2.2.2 A regression structure for the reporting intensity.

A first modelling strategy directly specifies a count regression model for the observed counts N_{td} in $\boldsymbol{N}^{\mathrm{r}}$. Most often these counts are assumed to be independent and Poisson distributed, i.e., $N_{td} \sim \mathrm{POI}(\lambda_{td})$, or, in case of overdispersion, a negative binomial distribution is used, i.e., $N_{td} \sim \mathrm{NB}(\lambda_{td}, \phi)$, where λ_{td} is the reporting intensity. In epidemiology, van de Kassteele et al. (2019) model the two-dimensional surface in Table 1 using bivariate P-splines with a smooth reporting intensity λ_{td} and day-of-the-week effects expressed as deviations from it. This intensity is modelled via a tensor product B-spline basis with pairwise products $B_i(t)B_j(d)$, where the basis functions $B_i(t)$ capture the effect of occurrence time and the $B_j(d)$ model the

Table 1: Triangular display with aggregated event counts. Only the event counts in the upper triangle are observed, whereas the occurred but not reported counts in the lower triangle have to be predicted. The coloured cells in the triangle represent the aggregation of events registered in continuous time in Figure 1 towards a coarser discrete time scale.

Occurrence period	Reporting delay				
	0	$\cdots$	$\tau - t$	$\cdots$	$\tau - 1$
1	N_{10}	$\cdots$	$N_{1,\tau-t}$	$\cdots$	$N_{1,\tau-1}$
$\vdots$					
t	N_{t0}	$\cdots$	$N_{t,\tau-t}$		
$\vdots$					
τ	$N_{\tau 0}$				

reporting delay. For the day-of-the-week effects dummy variables are included in the predictor, taking the value 1 if a certain combination of t and d corresponds to a specific weekday, and 0 otherwise. The smooth intensity and day of the week effects are estimated simultaneously using a penalized negative binomial regression model and the iterative reweighted least squares algorithm. Bastos et al. (2019) correct for reporting delays in disease surveillance data with a negative binomial regression model for the N_{td} that takes spatiotemporal variation into account. In their Bayesian framework the occurred-but-not-reported cases are estimated from the resulting posterior predictive distribution. A similar Bayesian modelling strategy is proposed in McGough et al. (2020), which is then applied in Greene et al. (2021) to nowcast COVID-19 cases in New York City.

In actuarial science the so-called chain ladder method is probably the most widely used technique to predict numbers of unreported claims. In this method the incremental event counts N_{td} are independently Poisson distributed (Hachemeister and Stanard, 1975; Renshaw and Verrall, 1998) with $\mathrm{E}[N_{td}] = \lambda_{td} = \lambda_t \cdot p_d$ for all $t = 1, \ldots, \tau$ and $d = 0, \ldots, \tau - 1$, where $\lambda_1, \ldots, \lambda_\tau > 0$ and $p_0, \ldots, p_{\tau-1} > 0$ with sum constraint $p_0 + \ldots + p_{\tau-1} = 1$. The latter assumes all claims to be reported by the end of the delay period $\tau - 1$. Beyond actuarial science, a similar modelling strategy was also proposed in (for instance) the disease modelling examples in Kalbfleisch and Lawless (1989) and Becker and Cui (1997). The log-likelihood corresponding to the Poisson model formulation of the chain ladder method (also referred to as the Poisson log-linear model in Sellero et al., 1996) is

$$\log \mathcal{L}(\boldsymbol{\Psi}; \boldsymbol{N}^{\mathrm{r}}) \quad = \quad \sum_{t=1}^{\tau} \sum_{d=0}^{\tau-t} \{-\lambda_t \cdot p_d + N_{td} \log(\lambda_t \cdot p_d)\} + c,$$

where we denote $\boldsymbol{\Psi} = \{\lambda_1, \ldots, \lambda_\tau, p_0, \ldots, p_{\tau-1}\}$ and c is a constant not depending on the model parameters. Maximizing the log-likelihood requires solving the following set of equations for $\boldsymbol{\Psi}$,

$$\sum_{d=0}^{\tau-t} \lambda_t \cdot p_d \quad = \quad \sum_{d=0}^{\tau-t} N_{td}\,, \qquad t = 1, \ldots, \tau\,, \tag{2}$$

$$\sum_{t=1}^{\tau-d} \lambda_t \cdot p_d \quad = \quad \sum_{t=1}^{\tau-d} N_{td}\,, \qquad d = 0, \ldots, \tau - 1\,, \tag{3}$$

subject to all elements of $\boldsymbol{\Psi}$ being positive and $\sum_{d=0}^{\tau-1} p_d = 1$. The Poisson maximum likelihood conditions equate the sums of the claim counts in each row and column of the observed upper

triangle $\boldsymbol{N}^{\mathrm{r}}$ to their expected value counterparts. Mack (1991, 1993) points out that this set of equations can be solved recursively due to the triangular structure. A more standard approach to estimate the parameters is to formulate the model as a generalized linear model and use a numerical optimizer (see Taylor, 2000; England and Verrall, 2002; Wüthrich and Merz, 2008).

2.2.3 Modelling the occurrence and reporting processes.

An alternative approach explicitly models the underlying occurrence and reporting processes, in line with the strategy outlined in the continuous time setting. The framework is then hierarchical where the N_t follow an occurrence model and the $N_{td}|N_t$ are multinomially distributed. In discrete time the assumptions (A1)-(A2) become

(A1') The daily total event counts N_t for $t = 1, \ldots, \tau$ are independently Poisson distributed with intensity $\lambda_t = \exp(\boldsymbol{x}_t'\boldsymbol{\alpha})$, where $\boldsymbol{x}_t$ is the vector of covariate information corresponding to occurrence day t and $\boldsymbol{\alpha}$ is a parameter vector.

(A2') Conditional on N_t, the event counts N_{td} for $d = 0, 1, 2, \ldots$, are multinomially distributed with probabilities $p_{td} = p_{td}(\boldsymbol{\theta}, \boldsymbol{x}_{td})$. These reporting probabilities do not depend on the number of events that occurred on day t, sum to one and are modeled with parameter vector $\boldsymbol{\theta}$ and covariate information $\boldsymbol{x}_{td}$.

The discrete occurrence intensities and reporting probabilities can be retrieved from the continuous time assumptions in $(A1), (A2)$, when we assume λ_t to be piecewise constant between integer time points and properly integrate the reporting density $f_{U|X}(\cdot|\cdot;\boldsymbol{\theta})$. That is,

$$\lambda(x) = \lambda_t \quad \text{for } x \in [t-1, t),$$
$$p_{td} = \begin{cases} \int_{t-1}^{t} \left(\int_0^{t-x} f_{U|X}(u|x;\boldsymbol{\theta})\, du \right) dx & \text{for } d = 0 \\ \int_{t-1}^{t} \left(\int_{d-1+t-x}^{d+t-x} f_{U|X}(u|x;\boldsymbol{\theta})\, du \right) dx & \text{for } d > 0. \end{cases}$$

The split in cases with reporting delay zero and a strictly positive delay follows from the different shapes (i.e., the triangle in light gray versus the parallelogram in dark gray) when aggregating continuous data in Figure 1. Note that the chain-ladder method discussed in Section 2.2.2 fits in the framework outlined by (A1')-(A2') if we impose the sum constraint $p_0 + \ldots + p_{\tau-1} = 1$. The latter assumes all claims to be reported by the end of delay period $\tau - 1$, an assumption typically made in an insurance context. The chain-ladder method assumes a stationary reporting process since the probabilities p_d do not depend on the occurrence period t.

We bundle the parameters to be estimated in $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\theta})$. Based on the model assumptions (A1')-(A2') and the thinning property of Poisson random variables, the daily event counts N_{td} are independently Poisson distributed with intensities $\lambda_t(\boldsymbol{x}_t'\boldsymbol{\alpha}) \cdot p_{td}(\boldsymbol{\theta}, \boldsymbol{x}_{td})$ for $t = 1, \ldots, \tau$ and $d = 0, 1, 2, \ldots$. Below we (often) use the more compact notation $\lambda_t(\boldsymbol{x}_t'\boldsymbol{\alpha}) \cdot p_{td}(\boldsymbol{\theta}, \boldsymbol{x}_{td}) = \lambda_t \cdot p_{td}$ to make our writing more concise. In particular, the observed event count N_t^{r} on day t is Poisson distributed with intensity $\lambda_t \cdot p_t^{\mathrm{r}}$ where $p_t^{\mathrm{r}} = \sum_{d=0}^{\tau-t} p_{td}$ and the unreported event count N_t^{nr} is Poisson distributed with intensity $\lambda_t \cdot p_t^{\mathrm{nr}}$ where $p_t^{\mathrm{nr}} = 1 - p_t^{\mathrm{r}}$. Conditional on N_t^{r}, the observed daily event counts $\{N_{td} \mid d = 0, 1, \ldots, \tau - t\}$ are multinomially distributed with parameters N_t^{r} and $\{p_{td}/p_t^{\mathrm{r}} \mid d = 0, 1, \ldots, \tau - t\}$, since we have to account for the right-truncation of the reporting delay. The likelihood of the observed data can then be written as the product of a Poisson likelihood and a multinomial likelihood,

$$\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{r}}) = \prod_{t=1}^{\tau} \frac{\exp(-\lambda_t p_t^{\mathrm{r}}) \, (\lambda_t p_t^{\mathrm{r}})^{N_t^{\mathrm{r}}}}{N_t^{\mathrm{r}}!} \frac{N_t^{\mathrm{r}}!}{\prod_{d=0}^{\tau-t} N_{td}!} \prod_{d=0}^{\tau-t} \left(\frac{p_{td}}{p_t^{\mathrm{r}}} \right)^{N_{td}}. \tag{4}$$

Equivalently, the likelihood can also be constructed by treating the daily event counts as right-censored, since the number of unreported events is unknown,

$$\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{r}}) \;=\; \prod_{t=1}^{\tau} \sum_{n=N_t^{\mathrm{r}}}^{\infty} \frac{\exp(-\lambda_t)\lambda_t^n}{n!} \frac{n!}{(n-N_t^{\mathrm{r}})! \prod_{d=0}^{\tau-t} N_{td}!} \prod_{d=0}^{\tau-t} (p_{td})^{N_{td}} \, (p_t^{\mathrm{nr}})^{n-N_t^{\mathrm{r}}} \; .$$

Indeed, this expression reduces to (4) by rewriting the sum over n using the Taylor expansion for the exponential function. The corresponding log-likelihood equals

$$\log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{r}}) = \sum_{t=1}^{\tau} \Big[-\lambda_t p_t^{\mathrm{r}} + N_t^{\mathrm{r}} \log(\lambda_t) + \sum_{d=0}^{\tau-t} \{ N_{td} \log(p_{td}) - \log(N_{td}!) \} \Big]. \tag{5}$$

Similar to the continuous time setting, the log-likelihood in (5) contains terms which depend on the parameters of both the Poisson model for event occurrences (appearing in λ_t) and the reporting delay distribution (appearing in p_{td} and p_t^r). This complicates direct maximum likelihood estimation as it prevents separate optimization with respect to each of these parameter components. In line with the continuous time literature review in Section 2.1, two strategies for optimizing this likelihood are generally applied. A first strand puts focus on direct optimization of this likelihood using a standard numerical method such as Newton-Raphson. This is feasible, but we cannot rely on standard statistical modelling routines and we need to derive (analytically or numerically) the gradient and Hessian of the log-likelihood in (5). Höhle and an der Heiden (2014) propose a Bayesian approach for the joint modelling of the occurrence and delay process for discrete data collected in epidemiology. Günther et al. (2021) then apply this Bayesian model to construct nowcast estimates for the COVID-19 pandemic in Bavaria. In a second strand, a (heuristic) two-stage approach keeps the likelihood tractable by putting focus on the modelling of the reporting process, while using a nonparametric estimator for the occurrence process. Examples are Lawless (1994) on random temporal fluctuations in reporting delays and Crevecoeur et al. (2019) on the reporting of insurance claims in the presence of holiday effects.

2.3 The EM algorithm to tackle the missing data problem

Our literature overview for both continuous as well as discrete time models puts focus on either the direct optimization of the likelihood or the use of heuristic two-stage procedures. In the absence of right-truncation the complete likelihood analogue of (1) (continuous) or (5) (discrete) would split into an occurrence and a reporting likelihood and become tractable. This immediately suggests the use of the expectation-maximisation (EM) algorithm (Dempster et al., 1977) to handle the missing data problem, as is recognized by a series of contributions on nowcasting. Brookmeyer and Gail (1988) estimate a piecewise constant infection density while assuming a Weibull distribution for the incubation time, independent of the time of infection. With an iterative procedure similar to EM, Harris (1990) fits both categorical as well as continuous-time models for the incidence of infections and their incubation time. Pagano et al. (1994) use the EM algorithm but only focus on the estimation of the reporting delay distribution, while Bacchetti et al. (1993) and Bellocco and Marschner (2000) assume a known (incubation) distribution for the delays. These papers propose rather simple structures for the occurrence and reporting processes, e.g. a stationary reporting delay distribution, or categorical models with separate parameters for each t and d. In the next section we present a flexible yet practical estimation framework capable of dealing with any parametric structure for the occurrence and reporting processes. Because data are collected over discrete time units in administrative systems, we propose our general model for the intensities $\lambda_t(\boldsymbol{x}_t'\boldsymbol{\alpha})$ and the reporting delay probabilities $p_{td}(\boldsymbol{\theta}, \boldsymbol{x}_{td})$ in (A1')-(A2'). This framework not only allows for an easy calibration of each of

the models from Section 2.2.3, but also facilitates the estimation of new occurrence and reporting process specifications which were previously considered too difficult for joint optimization. Aiming for flexibility, we incorporate covariates in both the occurrence process of events and the reporting delay distribution. We capture evolutions over time or seasonal trends, by including fluctuations in both the event counts and their reporting delays by month, day of the month or day of the week of the occurrence date. Additionally, one can also model relationships with external covariates which might influence the events such as economic circumstances, business cycles and weather conditions. Day-specific particularities in the reporting delay (such as holiday effects) can be modeled using designated day probabilities. The case-study developed in Section 4 convincingly illustrates this, by letting the data set assist us in choosing an appropriate occurrence and reporting delay structure.

3 An EM algorithm for the joint estimation of the occurrence and reporting delay of events

3.1 Parameter estimation using an EM algorithm

Starting from the log-likelihood in (5) we choose to treat the truncation as a missing data problem and employ an EM algorithm to simplify maximum likelihood parameter estimation. Consider the complete version of the data $\boldsymbol{N} = \boldsymbol{N}^{\mathrm{r}} \cup \boldsymbol{N}^{\mathrm{nr}} = \{N_{td} \mid 1 \leq t \leq \tau, d \geq 0\}$ which augments the observed daily event counts from the upper part of the triangular display in Table 1 with the unknown values of the future event counts in the lower triangle. Given the complete data $\boldsymbol{N}$, we construct the complete log-likelihood function

$$\log \mathcal{L}_c(\boldsymbol{\Theta}; \boldsymbol{N}) = \sum_{t=1}^{\tau} \Big[-\lambda_t + N_t \log(\lambda_t) + \sum_{d=0}^{\infty} \{N_{td} \log(p_{td}) - \log(N_{td}!)\} \Big], \tag{6}$$

which allows for a separate estimation of the parameters $\boldsymbol{\alpha}$ of the event occurrence model (appearing in $\lambda_t(\boldsymbol{x}_t'\boldsymbol{\alpha})$) and $\boldsymbol{\theta}$ of the reporting delay distribution (appearing in $p_{td}(\boldsymbol{\theta}, \boldsymbol{x}_{td})$).

From a numerical point of view, we propose to deal with the infinite sums over the reporting delay d in (6) by introducing

$$N_{t,\tau+} = \sum_{d \geq \tau} N_{t,d}$$

and

$$N^{\mathrm{nr}} = \{N_{t,d} \mid 1 \leq t \leq \tau, t + d > \tau, d < \tau\} \cup \{N_{t,\tau+} \mid 1 \leq t \leq \tau\}.$$

This implies grouping all events reported with a delay of at least τ days in a remainder category. The latter are the events reported beyond the right boundary of the reporting triangle in Table 1. The complete log-likelihood still factorizes into occurrence and reporting contributions and becomes

$$\begin{aligned} \log \mathcal{L}_c(\boldsymbol{\Theta}; \boldsymbol{N}) \quad = \quad & \sum_{t=1}^{\tau} \Big[-\lambda_t + N_t \log(\lambda_t) + \sum_{d=0}^{\tau-1} \{N_{td} \log(p_{td}) - \log(N_{td}!)\} \\ & + N_{t,\tau+} \cdot \log\Big(\sum_{d \geq \tau} p_{td} \Big) - \log(N_{t,\tau+}!) \Big]. \end{aligned} \tag{7}$$

The EM algorithm exploits the simpler form of $\mathcal{L}_c(\boldsymbol{\Theta}; \boldsymbol{N})$ by iterating between computing expectations in an E-step and maximization in an M-step. Applied to our setting, the numbers of unreported events are replaced by their expected values in the E-step and the log-likelihood of the augmented data is maximized in the M-step. While numerical optimization is required

in the M-step, the parameters of the event occurrence model can be estimated separately from the reporting delay parameters and standard software routines can be utilized in the absence of truncation. We discuss the steps of the EM algorithm in more detail for the kth iteration.

E-step. We take the conditional expectation of the complete log-likelihood (7) given the observed data $\boldsymbol{N}^{\mathrm{r}}$ and using the current estimate $\boldsymbol{\Theta}^{(k-1)}$ of the parameter vector $\boldsymbol{\Theta}$:

$$Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}) = \mathrm{E}_{\boldsymbol{\Theta}^{(k-1)}} \left[\log \mathcal{L}_c(\boldsymbol{\Theta}; \boldsymbol{N}) \mid \boldsymbol{N}^{\mathrm{r}} \right]. \tag{8}$$

This requires to compute the expected values of the future event counts

$$N_{td}^{(k)} = \mathrm{E}_{\boldsymbol{\Theta}^{(k-1)}} \left[N_{td} \mid \boldsymbol{N}^{\mathrm{r}} \right] = \begin{cases} N_{td} & \text{if } d \leq \tau - t \\ \lambda_t^{(k-1)} p_{td}^{(k-1)} & \text{otherwise,} \end{cases} \tag{9}$$

for $t = 1, \ldots, \tau$ and the total daily event counts are

$$N_t^{(k)} = \sum_{d=0}^{\tau-t} N_{td} + \sum_{d=\tau-t+1}^{\tau-1} N_{td}^{(k)} + N_{t,\tau+}^{(k)}.$$

The terms in (8) containing $\mathrm{E}_{\boldsymbol{\Theta}^{(k-1)}} \left[\log(N_{td}!) \mid \boldsymbol{N}^{\mathrm{r}} \right]$ depend on the observed event counts and the parameter values in iteration $k-1$ and these will therefore be treated as constants in the M-step where the maximization is with respect to the parameter vector $\boldsymbol{\Theta}$.

M-step. We maximize the expected value (8) of the complete data log-likelihood obtained in the E-step with respect to the parameter vector $\boldsymbol{\Theta}$. In order to optimize (8) with respect to $\boldsymbol{\alpha}$ as defined in model assumption (A1'), we have to maximize the terms related to the event occurrence model,

$$-\sum_{t=1}^{\tau} \lambda_t + \sum_{t=1}^{\tau} N_t^{(k)} \log(\lambda_t) = -\sum_{t=1}^{\tau} \exp(\boldsymbol{x}_t' \boldsymbol{\alpha}) + \sum_{t=1}^{\tau} N_t^{(k)} \boldsymbol{x}_t' \boldsymbol{\alpha}\,, \tag{10}$$

which is a weighted Poisson log-likelihood, possibly including an offset term (to include an exposure-to-risk measure, as illustrated in the case study covered in Section 4). The parameter values optimizing (10) are denoted by $\boldsymbol{\alpha}^{(k)}$. Updating the estimates for the parameters $\boldsymbol{\theta}$ of the reporting delay distribution requires the maximization of

$$\sum_{t=1}^{\tau} \sum_{d=0}^{\tau-1} N_{td}^{(k)} \log\{p_{td}(\boldsymbol{\theta}, \boldsymbol{x}_{td})\} + \sum_{t=1}^{\tau} N_{t,\tau+}^{(k)} \cdot \log\{\sum_{d \geq \tau} p_{td}(\boldsymbol{\theta}, \boldsymbol{x}_{td})\}. \tag{11}$$

We now have to optimize a right-censored likelihood with right-censoring at delay τ. Depending on the problem at hand, the likelihood can be simplified by omitting the censoring term when reporting events with a delay of more than $\tau - 1$ days is highly uncommon, see our discussion in the case study of Section 4.

Initial step. The first E-step of the EM algorithm with $k = 1$ requires a starting value $\boldsymbol{\Theta}^{(0)}$ for the parameter set. Our strategy is to first apply the chain ladder method (see Section 2.2) on the daily event counts to obtain initial predictions $N_{td}^{(0)}$ of the future event counts. Then, we initialize $\boldsymbol{\Theta}$ by applying an M-step based on these initial event count estimates. More specifically, we use the Wüthrich and Merz (2015) formulation of the chain ladder method in

terms of the chain ladder development factors, allowing for a fast and practical implementation. Therefore, we define the cumulative event counts as

$$C_{td} = \sum_{j=0}^{d} N_{tj} \quad \text{for } t = 1, \ldots, \tau, \text{and } d = 0, \ldots, \tau - t\,,$$

and estimate the development factors of the chain ladder method as

$$\widehat{f}_d = \frac{\sum_{t=1}^{\tau-d} C_{td}}{\sum_{t=1}^{\tau-d} C_{t,d-1}} \quad \text{for } d = 1, \ldots, \tau - 1\,. \tag{12}$$

The chain ladder method applies these development factors to the latest cumulative event count in each row to produce forecasts of future cumulative event counts:

$$\widehat{C}_{t,d} = C_{t,\tau-t}\widehat{f}_{\tau-t+1} \ldots \widehat{f}_d \quad \text{for } t = 2, \ldots, \tau, \text{and } d = \tau - t + 1, \ldots, \tau - 1\,.$$

We use these chain ladder estimates for the cumulative event counts to initialize the expected incremental event counts as

$$N_{td}^{(0)} = \begin{cases} N_{td} & \text{if } d \leq \tau - t \\ \widehat{C}_{t,\tau-t+1} - C_{t,\tau-t} & \text{if } d = \tau - t + 1 \\ \widehat{C}_{td} - \widehat{C}_{t,d-1} & \text{otherwise,} \end{cases}$$

for $t = 1, \ldots, \tau$ and apply an M-step, as outlined above with $k = 0$, to find decent starting values $\boldsymbol{\Theta}^{(0)}$. For delays beyond the maximum observed delay (i.e. $d > \tau - 1$) we put the initial values $N_{td}^{(0)}$ equal to zero.

Convergence. The log-likelihood (5) increases with each EM iteration (Dempster et al., 1977). Given proper starting values, the sequence $\boldsymbol{\Theta}^{(k)}$ converges to the maximum likelihood estimator (MLE) of $\boldsymbol{\Theta}$ corresponding to the incomplete data log-likelihood $\log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{r}})$ in (5). The stopping criterion we apply is based on the relative change in the log-likelihood. Namely, we iterate until the absolute value of $\{\log \mathcal{L}(\boldsymbol{\Theta}^{(k)}; \boldsymbol{N}^{\mathrm{r}}) - \log \mathcal{L}(\boldsymbol{\Theta}^{(k-1)}; \boldsymbol{N}^{\mathrm{r}})\}/\{0.1 + \log \mathcal{L}(\boldsymbol{\Theta}^{(k)}; \boldsymbol{N}^{\mathrm{r}})\}$ becomes smaller than 10^{-8}. The parameter vector estimate upon convergence is denoted by $\widehat{\boldsymbol{\Theta}}$.

3.2 Asymptotic variance-covariance matrix

It is expected that when the number of parameters is finite, asymptotic normality of the estimators holds under quite general conditions, though they depend on the specific problem settings. However, if the number of parameters increases with τ, such as in a chain ladder model, the asymptotic properties of the maximum likelihood estimators are difficult to obtain and require a more detailed and rigorous study.

While an EM algorithm by itself does not output an estimated covariance matrix of the parameter estimators, with some additional effort, such matrix may be constructed. Louis (1982) showed how the observed information matrix can be expressed in terms of the gradient and second derivative of the complete data log-likelihood function. This approach avoids working with the incomplete data log-likelihood function and leads to analytic expressions which can be computed using the model specifications. The supplemented EM algorithm (Meng and Rubin, 1991) provides a computational, iterative approach using the negative second Hessian matrix of the expected complete-data log likelihood. Oakes (1999) explains how the matrix of second derivatives of the observed data log-likelihood can be obtained using the derivatives of the function Q, see (8), and provides expressions for exponential family models. For models for which analytical results are too hard to obtain, numerical derivatives could be used instead.

3.3 Variable selection and model diagnostics

If one works with the log-likelihood from (5), in principle all known methods for likelihood estimation are applicable, in particular the use of information criteria such as AIC (Akaike, 1973) and BIC (Schwarz, 1978) for variable selection, see Claeskens and Hjort (2008, Chap. 2, 3) for more explanation about their use. However, when an EM algorithm is used for estimation, care needs to be taken since at convergence the function Q of (8) is not the maximized log-likelihood. Cavanaugh and Shumway (1998) developed AICcd which allows to directly work with Q though it requires an adjustment to the 'penalty' part of the AIC. Such an adjustment is used by Claeskens and Consentino (2008) for variable selection with missing covariate data, and by Dirick et al. (2015) for selection in multiple event mixture cure models.

The complete data AIC value for use with an EM algorithm is defined by

$$\text{AICcd} = -2Q(\widehat{\boldsymbol{\Theta}};\widehat{\boldsymbol{\Theta}}) + 2\text{trace}\{I_c(\widehat{\boldsymbol{\Theta}};\boldsymbol{N}^{\text{r}})I_o^{-1}(\widehat{\boldsymbol{\Theta}};\boldsymbol{N}^{\text{r}})\},$$

with the square matrices

$$I_c(\widehat{\boldsymbol{\Theta}};\boldsymbol{N}^{\text{r}}) = -\frac{\partial^2}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}'}Q(\widehat{\boldsymbol{\Theta}};\widehat{\boldsymbol{\Theta}}); \quad I_o(\widehat{\boldsymbol{\Theta}};\boldsymbol{N}^{\text{r}}) = -\frac{\partial^2}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}'}\log\mathcal{L}(\widehat{\boldsymbol{\Theta}};\boldsymbol{N}^{\text{r}}),$$

where I_o estimates the limiting covariance matrix of $\widehat{\boldsymbol{\Theta}}$, for which methods such as described in Section 3.2 may be used.

For model selection purposes, the AICcd values are computed for each considered model and the model with the smallest value of AICcd gets selected.

The matrix $I_c(\widehat{\boldsymbol{\Theta}};\boldsymbol{N}^{\text{r}})$ is also the main ingredient for model diagnostic measures after using an EM algorithm, see Zhu et al. (2001) and Barreto-Souza and Simas (2017). To define the generalized Cook's distance to investigate the influence of a single observation, thus of a single N_{td} with $t = 1,\ldots,\tau$ and $d = 0,\ldots,\tau-1$, we use a one-step approximation as in Zhu et al. (2001) to avoid refitting the model. Define $\dot{Q}_{-[t,d]}(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}}) = E[\partial\log\mathcal{L}_{c,-[t,d]}(\boldsymbol{\Theta};\boldsymbol{N})/(\partial\boldsymbol{\Theta})]$ with observation t,d removed from the summation, which is estimated by replacing $\boldsymbol{\Theta}$ by $\widehat{\boldsymbol{\Theta}}$. The generalized Cook's distance for use with the output of an EM algorithm now reads, for $t = 1,\ldots,\tau$ and $d = 0,\ldots,\tau-1$,

$$\text{GD}_{t,d} = \dot{Q}_{[t,d]}(\widehat{\boldsymbol{\Theta}};\widehat{\boldsymbol{\Theta}})'I_c^{-1}(\widehat{\boldsymbol{\Theta}};\boldsymbol{N}^{\text{r}})\dot{Q}_{[t,d]}(\widehat{\boldsymbol{\Theta}};\widehat{\boldsymbol{\Theta}}).$$

Large values should encourage further investigation of those observations. We refer to the mentioned references for other influence measures that can be used with an EM algorithm.

3.4 Prediction and out-of-time evaluation tools

Using the estimated parameter vector $\widehat{\boldsymbol{\Theta}}$, we predict the number of daily unreported events in the lower triangle of Table 1. Point estimators for all $N_{td} \in \boldsymbol{N}^{\text{nr}}$ can be obtained using the expected values $\widehat{N}_{td} = \widehat{\lambda}_t\,\widehat{p}_{td}$. Similarly, the total number of unreported events per day is estimated by $\widehat{N}_t^{\text{nr}} = \sum_{d=\tau-t+1}^{\infty}\widehat{N}_{td} = \widehat{\lambda}_t\,\widehat{p}_t^{\text{nr}}$ and the total number of unreported events over all occurrence days by $\widehat{N}^{\text{nr}} = \sum_{t=1}^{\tau}\widehat{N}_t^{\text{nr}}$. Moreover, under the model assumptions (A1')-(A2'), the future daily number of events N_{td} ($t \le \tau, t+d > \tau$) are independently Poisson distributed and we thus have that $N_{td} \sim \text{Poisson}(\lambda_t\,p_{td})$, $N_t^{\text{nr}} \sim \text{Poisson}(\lambda_t\,p_t^{\text{nr}})$ and $N^{\text{nr}} \sim \text{Poisson}\,(\sum_{t=1}^{\tau}\lambda_t\,p_t^{\text{nr}})$. This allows us to construct prediction intervals and to make probabilistic statements concerning the number of unreported events after replacing the intensities by their maximum likelihood estimates. With a model defined on a daily level (by means of example) these unreported events can be divided into daily nowcasts by occurrence date or by reporting date, as we demonstrate in the case study in Section 4. The latter is of particular interest when using our model in practice

as it gives the analyst a refined view on the future reporting times (at daily level or aggregated e.g. by future reporting weeks or months). An out-of-time evaluation is then a valuable tool to assess the predictive performance of a proposed nowcasting model. Hereby, we restrict the time window, say to events with occurrences $1 \leq t \leq \tau^\star < \tau$ and reporting delays $d \geq 0$ and $t + d \leq \tau^\star$, calibrate the model with the adjusted evaluation date $\tau^\star$ and use it to nowcast the events with $1 \leq t \leq \tau^\star$ and $\tau \geq t + d > \tau^\star$. These nowcasts can then be compared to the actual observed reporting of events. We are particularly in favor of performing a moving window out-of-time evaluation, where we repeat the single out-of-time evaluation multiple times with different evaluation dates $\tau^\star$. This not only allows to assess the stable predictive performance of a model across multiple evaluation dates, but also allows to verify possible changes in the parameters when calibrated over different subsets of data. When such (substantial) changes are detected, the use of change-points in the modelling of the occurrence and reporting processes may be an interesting direction to explore, see Tabnak et al. (2000) for an example in modelling reporting delays in AIDS surveillance data.

3.5 Example: revisiting the chain ladder method with an EM algorithm

The chain ladder introduced in Section 2.2 imposes a multiplicative structure on the reporting intensity, with a stationary reporting delay distribution. Classic ways to estimate this model either rely on a log-linear Poisson regression model or use the development factors discussed in the initialization step of an EM algorithm (see Section 3.1). We now revisit the estimation of the chain ladder method and discuss how its parameters can equivalently be estimated using our proposed estimation framework. The complete log-likelihood related to the chain ladder method is similar to (6) with the difference that the reporting delay probabilities p_d do not depend on the occurrence period t and the sum over d runs until $\tau - 1$. The latter is linked to the fact that the chain ladder method does not allow for extrapolation beyond the range of data (cfr. England and Verrall, 2002, for extensions of the classical chain ladder method which involve the estimation of tail factors). The E-step is the same as (9) with, for $t = 1, \ldots, \tau$ and $d = 0, \ldots, \tau - 1$, $p_{td}^{(k-1)}$ replaced by $p_d^{(k-1)}$, whereas the M-step simplifies to

$$\lambda_t^{(k)} = N_t^{(k)} = \sum_{d=0}^{\tau-1} N_{td}^{(k)} \text{ and } \quad p_d^{(k)} = \frac{\sum_{t=1}^{\tau} N_{td}^{(k)}}{\sum_{t=1}^{\tau}\sum_{d=0}^{\tau-1} N_{td}^{(k)}} = \frac{\sum_{t=1}^{\tau} N_{td}^{(k)}}{\sum_{t=1}^{\tau} N_t^{(k)}}.$$

In case the chain ladder factors (12) are used in the initial step (see Section 3.1), such that in fact $N_{td}^{(0)} = \widehat{\lambda}_t \, \widehat{p}_d$ for $t = 2, \ldots, \tau$ and $d = \tau - t + 1, \ldots, \tau - 1$ due to the equivalence with the Poisson model, convergence is reached immediately the first time we apply the M-step above. Indeed, we then obtain

$$\lambda_t^{(0)} = \sum_{d=0}^{\tau-1} N_{td}^{(0)} = \sum_{d=0}^{\tau-t} N_{td} + \sum_{d=\tau-t+1}^{\tau-1} N_{td}^{(0)} = \widehat{\lambda}_t \sum_{d=0}^{\tau-1} \widehat{p}_d = \widehat{\lambda}_t,$$

where we used (2). Similarly, using (3), we find

$$p_d^{(0)} = \frac{\sum_{t=1}^{\tau} N_{td}^{(0)}}{\sum_{t=1}^{\tau}\sum_{d=0}^{\tau-1} N_{td}^{(0)}} = \frac{\sum_{t=1}^{\tau-d} N_{td} + \sum_{t=\tau-d+1}^{\tau} N_{td}^{(0)}}{\sum_{t=1}^{\tau} \widehat{\lambda}_t} = \frac{\sum_{t=1}^{\tau} \widehat{\lambda}_t \, \widehat{p}_d}{\sum_{t=1}^{\tau} \widehat{\lambda}_t} = \widehat{p}_d.$$

Using an EM algorithm, the iterative steps are easy and intuitive and, upon convergence, the same parameter estimators for λ_t and p_d are obtained compared to direct maximum likelihood optimization. When structuring the occurrence and reporting delay parameters (as we propose in Section 2.2.3), the model can no longer be solved analytically nor formulated as a generalized linear model. An EM algorithm then offers an elegant solution.

4 Case study: insurance nowcasting

We analyze a data set with the occurrence and reporting dates of claims in a portfolio of general liability insurance policies for private individuals from a European insurance company. The goal is to predict (or: nowcast) the number of claims that already incurred in the past, but are not yet reported to the insurance company.

4.1 Description of the insurance dataset

Detailed claim and policy information is available from January 2000 until August 2009. This includes the occurrence date of a claim, the time between occurrence and reporting of the claim to the insurance company and an exposure-to-risk measure. The online supplement further details the exposure measure that is available in this dataset. To enable out-of-time prediction, we restrict our analysis to claims that have occurred between January 1, 2000 and August 31, 2004. We set the new evaluation date, say $\tau^\star$, at the end of this time window, on August 31, 2004 and want to estimate the total incurred but not reported (IBNR) claim count, as well as the reporting dates of these IBNR claims. Based on the full data set until August 2009, 176 671 claims have occurred during this time window. Due to a reporting delay, only 174 624 of these have been reported by the evaluation date, as visualized in Figure 6 in the online supplement. The remaining 2047 claims are IBNR claims, i.e. claims which have occurred between January 2000 and August 2004 but have only been reported after the evaluation date and before the end of the observation period.

4.2 Parametric models for the occurrence and reporting processes

In line with Section 2.2.3, we work in discrete time and explicitly model the occurrence and reporting processes.

Occurrence model. Let N_t be the total insurance claim counts that occur on day t, for $t = 1, \dots, \tau^\star$ where $t = 1$ is Jan 1, 2000 and $t = \tau^\star$ refers to August 31, 2004. We model the N_t as independent Poisson distributed random variables with intensity λ_t, see assumption (A1'), structured as

$$\lambda_t \quad = \quad e_t \cdot \exp\left(\alpha_0 + \alpha_{\texttt{jan1}(t)} + \alpha_{\texttt{dec31}(t)} + \alpha_{\texttt{month}(t)} + \alpha_{\texttt{dow}(t)} + \alpha_{\texttt{dom}(t)}\right), \tag{13}$$

where e_t is the exposure on day t, $\texttt{month}(t)$ indicates the month, $\texttt{dow}(t)$ the day of the week and $\texttt{dom}(t)$ the day of the month to which t belongs. We also include parameters for occurrence days t on January 1 and December 31, because on these days many claims occur.

Reporting model. Figures 7a and 7b in the online supplement illustrate how reporting probabilities decrease over time and are low on Saturdays and Sundays. Supported by these empirical findings we structure reporting delay as the product of week probabilities and day (or intra-week) probabilities $p_{td}(\boldsymbol{\theta}, \boldsymbol{x}_{td}) = p^{\mathrm{W}}_{tw}(\boldsymbol{\theta}, \boldsymbol{x}_t) \cdot p^{\mathrm{intra}}_{td}$. Here, p^{W}_{tw} denotes the probability of reporting an event from occurrence day t in the wth week after occurrence. The intra-week reporting probabilities are denoted with p^{intra}_{td} and sum to 1 over the days within the reporting week.

The reporting week probabilities

$$p^{\mathrm{W}}_{tw} = \frac{\Gamma(\phi + w)}{w!\Gamma(\phi)} \frac{\phi^\phi \mu_t^w}{(\phi + \mu_t)^{\phi + w}} \quad \text{for } w = 0, 1, 2, \dots \ ,$$

are modeled using the probability mass function of a negative binomial distribution with expected value $\mu_t = \exp(\boldsymbol{x}_t'\boldsymbol{\theta})$ and variance $\mu_t + \mu_t^2/\phi$, where ϕ is the dispersion parameter and $\boldsymbol{x}_t$ is the covariate vector corresponding to occurrence day t. Figure 7b in the supplement indicates that the negative binomial distribution is a good fit. We structure the μ_t as follows,

$$\mu_t \quad = \quad \exp\left(\theta_0 + \theta_{\texttt{jan1}(t)} + \theta_{\texttt{dec31}(t)} + \theta_{\texttt{month}(t)} + \theta_{\texttt{dow}(t)} + \theta_{\texttt{dom}(t)}\right), \tag{14}$$

including parameters for occurrence day t on January 1 and December 31.

A first modelling strategy for the reporting delay day probabilities symbolically writes these probabilities as

$$p_{td}^{\text{intra}} = \text{P}(\mathsf{dow}(t), \mathsf{wday}(t, t+d)), \tag{15}$$

where $\mathsf{wday}(t, t+d)$ orders the working days within the reporting week with separate levels for Saturday and Sunday. For example, when the claim occurred on a Thursday, Friday is $\mathsf{wday2} = \mathsf{wday}(t, t+1)$ and Monday is $\mathsf{wday3} = \mathsf{wday}(t, t+2)$. The 7×7-matrix P then contains the day probabilities related to the first week. Each element in P is between 0 and 1 and all row sums equal 1.

Alternatively, we project the seven intra-week reporting probabilities p_{td}^{intra} to six probabilities, inspired by the reverse time strategy of Kalbfleisch and Lawless (1991). This allows to leave the sum-to-one restriction on the intra-week reporting probabilities. Moreover, we can now include covariate information related to both the day of occurrence t and the actual reporting date $t + d$. We define

$$\begin{aligned} q_{t,7w+j} &= P(\text{delay} = 7w + j \,|\, 7w \leq \text{delay} \leq 7w + j) \\ &= \frac{p_{t,7w+j}^{\text{intra}}}{\sum_{k=0}^{j} p_{t,7w+k}^{\text{intra}}} \quad \text{for } j = 1, \ldots, 6, \end{aligned} \tag{16}$$

with $w = \left\lfloor \frac{d}{7} \right\rfloor$ (the reporting week after occurrence) and $j = d - 7w$ the intra-week day of reporting. Expression (16) takes the form of a discrete time hazard rate, but due to the reverse time strategy we now condition on failure before day $7w + j$ instead of survival. We structure

$$\texttt{logit}(q_{t,d}) = \boldsymbol{x}_{td}'\boldsymbol{\gamma} = \gamma_0 + \gamma_{\texttt{workdays}(t,t+d)} + \gamma_{\texttt{dow}(t+d)} + \gamma_{\texttt{holiday}(t+d)}, \tag{17}$$

where $\texttt{workdays}(t, t+d)$ counts the number of elapsed working days in the current reporting week, i.e. excluding the weekend and holidays, and $\texttt{holiday}(t+d)$ indicates whether $t + d$ is a holiday. Reporting on holidays is exceptional in our data set. The inclusion of dummy variables for holidays allows to capture this empirical fact in the intra-day reporting probabilities, while distinguishing between national holidays on which all companies are closed and two unofficial holidays (New Year's Eve and Good Friday).

4.3 Competing strategies: directly modelling the reporting intensity and the chain ladder method

We also study some competing modelling strategies proposed in the literature. We consider two versions of the models discussed in Section 2.2.2 where the N_{td} are independent Poisson distributed random variables with mean λ_{td} structured as

$$\begin{aligned} \lambda_{td} \quad = e_t \cdot \exp\big(&\alpha_{\texttt{jan1}(t)} + \alpha_{\texttt{dec31}(t)} + \alpha_{\texttt{month}(t)} + \alpha_{\texttt{dow}(t)} + \alpha_{\texttt{dom}(t)} + \\ &\beta_{\texttt{holiday}(t+d)} + \beta_{\texttt{holiday}(t+d+1)} + \beta_{\texttt{dow}(t+d)} + \beta_{\texttt{delay}(d)}\big), \end{aligned} \tag{18}$$

or

$$\begin{aligned} \lambda_{td} \quad = e_t \cdot \exp\big(&\alpha_t + \beta_{\texttt{holiday}(t+d)} + \beta_{\texttt{holiday}(t+d+1)} + \\ &\beta_{\texttt{dow}(t+d)} + \beta_{\texttt{delay}(d)}\big). \end{aligned} \tag{19}$$

While (18) uses covariates capturing relevant information from t, (19) includes a parameter α_t for each t. Both model specifications structure the information on the reporting delay with a parameter $\beta_{\texttt{delay}(d)}$ for each observed delay (see e.g. Bastos et al., 2019) as well as $\beta_{\texttt{dow}(t+d)}$ referring to the day of the week of the effective reporting date $t+d$. The $\beta_{\texttt{holiday}(t+d)}$ and $\beta_{\texttt{holiday}(t+d+1)}$ correspond to reporting on a holiday, or the day after. No explicit reporting delay distribution is proposed, but instead a large number of parameters is used.

We also calibrate the chain ladder method using a yearly grid. That is,

$$\lambda_{td} \quad = \quad \lambda_{\texttt{year(t)}} \cdot p_{\texttt{year}(t+d)-\texttt{year}(t)}, \tag{20}$$

where $\lambda_{\texttt{year(t)}}$ is the effect of the year of occurrence to which day t belongs. Reporting delay is modelled independently from the occurrence, with a parameter per year of reporting, thus $p_{\texttt{year}(t+d)-\texttt{year}(t)}$ describes the effect of the reporting year of the claims that occurred on day t and were reported with d days of delay.

4.4 Parameter estimates

For the models proposed in Section 4.2, we use the EM algorithm of Section 3 to estimate the parameters in the occurrence and reporting processes. For the models in Section 4.3, maximum likelihood estimates are obtained with the glm4 routine in the R package MatrixModels (Bates and Maechler, 2015).

To grasp the insights obtained via an explicit specification of the occurrence and reporting processes, we focus on the maximum likelihood estimates of the parameters in the occurrence model (13) combined with the reporting model with reporting week probabilities in (14) and intra-week reporting probabilities in (15). We evaluated the impact of the censoring term in the reporting delay likelihood used in the M-step, see (11). We obtained similar parameter estimates when omitting the reporting of events with a delay of more than $\tau^\star - 1$ days compared to the estimates obtained under the simplified likelihood that drops the censoring term in (11) (comparison not shown). Therefore, the results below are obtained with the simplified reporting likelihood. The parameter estimates of some alternative specifications are deferred to Section A.3 in the online supplement.

The effects related to the categorical predictors in the Poisson occurrence model specified in (13) are visualized in Figure 2. Figure 8 in the online supplement shows the parameter estimates for the negative binomial regression model of the reporting delay in weeks, while Table 2 in the online supplement collects the estimated intra-week probabilities from (15). The maximum likelihood estimates are shown along with Bonferroni adjusted simultaneous 95% confidence intervals based on the inverse of the expected information matrix. The intercept in the Poisson model is estimated as -5.965 with 95% confidence interval $[-6.020, -5.910]$.

Figure 2 reveals a seasonal pattern in which the number of claims rises in the middle of the year and falls around the year end. Claims most likely occur in June and least likely in December with an estimated difference in expected value of 42%. The calibrated day of the week effect shows an increase in the expected number of claims on Saturdays and a slight decrease on Tuesdays and Thursdays. The categorical effect of the day of the month shows a remarkable pattern which is similar in both the claim occurrence and the reporting delay model. On the 1st and 15th, the number of claims as well as the reporting delays have significantly higher expected values. To a lesser extent, this is also present for the 5th, 10th, 20th, 25th, and 30th or 31st day of each month. This pattern can most likely be explained by rounding errors of the occurrence date when insureds have to report a claim which took place several weeks or months ago. Since this misreporting of dates is more likely to occur for claims which are only reported after a longer time period, we simultaneously see an increase in the expected reporting delay

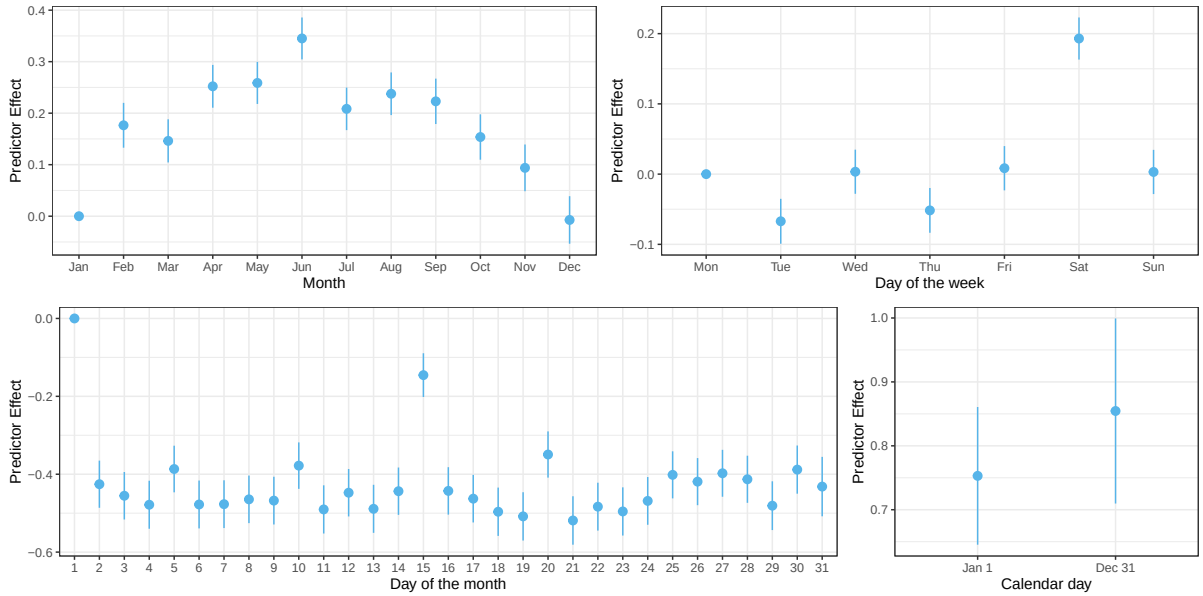

Figure 2: Maximum likelihood estimates and 95% simultaneous confidence intervals for $\boldsymbol{\alpha}$ in the Poisson claim occurrence model.

for claims occurring on these rounded month days. There is a high number of occurrences on December 31 and January 1 with on average 112%, respectively 135%, more claims. However, the reporting delay of these claims is not significantly different from the other ones.

4.5 Prediction of unreported claim counts

The main goal of our proposed models is to estimate the number of unreported claims. Using an out-of-time evaluation with $\tau^\star =$ August 31, 2004, we know there are 2047 IBNR claims in the full data set, which runs until August 2009, of which we know the corresponding occurrence date and reporting delay.

We first demonstrate the insights that can be derived from a nowcasting model using the occurrence specification in (13) and the reporting model in (14) combined with (15). Calibrated on the data from January 1, 2000 to August 31, 2004 this model estimates the total number of IBNR claims as $\widehat{N}^{\text{nr}} = 2055.8$, which is close to the actual count. Moreover, the distributional assumptions of our model can be used to provide a 95% prediction interval given by $[1697, 2145]$, see Section 3.4. Furthermore, since the model is defined on a daily level, the total IBNR prediction can be divided into daily forecasts by occurrence date and by reporting date. To illustrate this strong point of our model, we predict the IBNR claim counts by reporting date in Figure 3. Figure 9 in the online supplement shows a similar split of the IBNR claims by occurrence dates.

In Figure 3 we disperse the total predicted IBNR claim count according to the date on which the claims will be reported to the insurer. It means we now focus on estimating $\sum_{t=1}^{\tau^\star} N_{t,\rho-t}$ for $\rho = \tau^\star + 1, \tau^\star + 2, \ldots$, i.e. the number of unreported claims reported on day 1, 2, ... of the out-of-time period. This forms an appealing way to use our model in practice as it gives the insurer a refined view on the reporting times. The predictions on a daily level in Figure 3a are accompanied by 95% simultaneous prediction intervals and range from September 1, 2004, until November 7, 2004, i.e. the first two months following our training period. When compared to the out-of-time actual values, the forecasts clearly capture the downward trend in the reporting of IBNR claims and the nearly absence of reporting in weekends. This is primarily the case

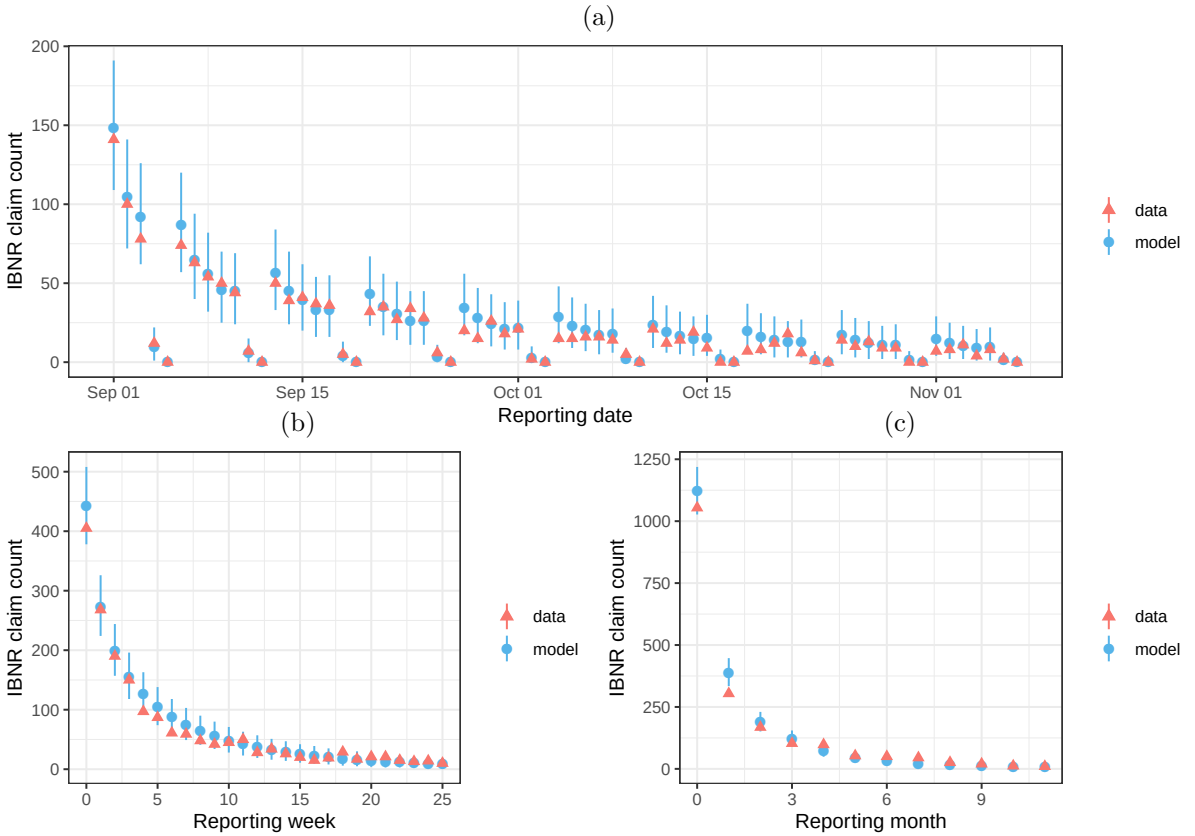

Figure 3: Predictions of the IBNR claim counts and simultaneous 95% prediction intervals by reporting date. (a) Daily, for reporting dates in between September 1 and November 7, 2004; (b) weekly (7 days) for the next 26 weeks; (c) monthly (30 days) for the next 12 months.

due to the day probabilities in our model which reflect the day-specific aspects of the reporting delay. In Figure 3b (resp. 3c) the reporting dates are grouped by weeks (resp. months) after the evaluation date and the IBNR claim counts are predicted for the next 26 weeks (resp. 12 months). We notice how, also over longer time spans, the predictions by reporting week or month follow the pattern observed in the actual unreported counts.

4.6 Comparing nowcasting models with a moving window evaluation

We compare the performance of different model specifications via a moving window out-of-time evaluation. We adjust the evaluation date $\tau^\star$, which was previously chosen to be August 31, 2004, to any date in between August 31, 2003, and August 31, 2004. For each such $\tau^\star$, we refit the model based on the observed data by that date, $\boldsymbol{N}^{\mathrm{r}} = \{N_{td} \mid 1 \leq t \leq \tau^\star, d \geq 0, t+d \leq \tau^\star\}$, and produce an estimate of the total unreported count $N^{\mathrm{nr}} = \sum_{t=1}^{\tau^\star} N_t^{\mathrm{nr}}$ corresponding to claims that occurred before or at $\tau^\star$.

Figures 4a and (zoom in) 4b show the predictions obtained with the occurrence specification in (13) and the reporting model in (14) and (15), as well as the actual total IBNR claim counts based on the full data set. Overall, the estimates follow the seasonal pattern quite well. The deviations from the observed trend in the first months of 2014 might be explained by variations from year to year in the effect of the month on the occurrence and reporting of claims. In our model, we assume the seasonal monthly pattern to be the same over the different years. As a result, the parameter estimates are averaged values. The model could potentially be refined

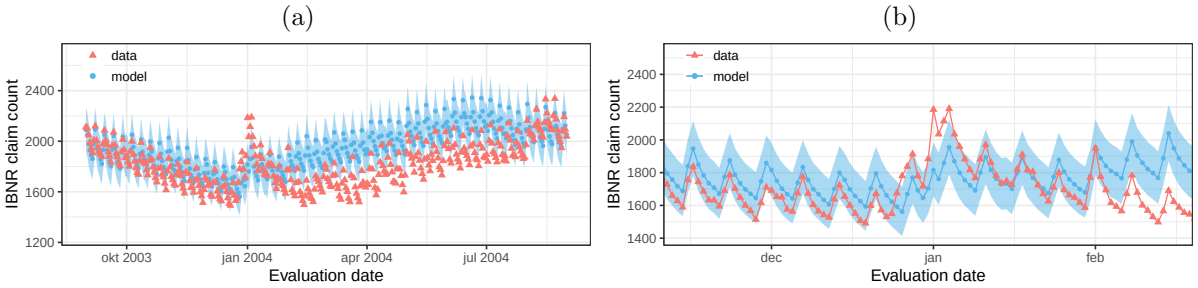

Figure 4: Predictions and 95% simultaneous prediction intervals of the total IBNR claim counts for varying evaluation dates $\tau^\star$: (a) in between August 31, 2003, and August 31, 2004 and (b) zoom in on the estimates from November 15 to February 15. Red triangles: data, blue circles: model.

(by including extra covariates) if expert-knowledge is available on how changes in e.g. product design and conditions, the claims handling process, the business environment or legislation, may impact the claim arrival process or the reporting delay. We notice a seven day pattern in the number of unreported claims in Figure 4b, which results from the delayed reporting during the weekend. This aspect is incorporated in our model through the intra-week reporting probabilities in (15). Moreover, we observe an increase in the number of IBNR claim counts around the end of the year. This is due to the fact that the insurance company is closed around the holidays, preventing any claims from being reported at that time. This effect is only partially captured by the dummy variables for the occurrence dates `jan1` and `dec31` in (13) and (14).

We compare these predictions to two other model specifications that benefit from an EM algorithm for calibration. Figure 5 zooms in on evaluation dates between November 15 and February 15, 2004. The corresponding plots with evaluation dates between August 31, 2003 and August 31, 2004 are available in the online supplement, Figure 15. Figure 5a shows predictions obtained with the occurrence specification in (13) and the reporting model in (14) and (17). In the specification of the intra-week probabilities this model designates dummy indicators for reporting on national and unofficial holidays in (17), which leads to clear improvements around the end-of-year holidays as compared to Figure 4. The yearly chain ladder method is used in Figure 5b. While this method overall performs well, detailed insights on the number of reportings in e.g. the next days, weeks or months can not be deduced from this model calibrated on data aggregated into yearly grids. Moreover, each year ends on a different day of the week. Since the chain ladder averages the reporting probability over past years, the estimated 7-day week pattern in the predictions is slightly misaligned with the actual pattern in the data. Crevecoeur et al. (2019) show that when more years of data are available the 7-day pattern completely disappears from the yearly chain ladder predictions, which then results in a systematic underestimation of the IBNR claim count on Saturdays and Sundays. Experiments to capture the week pattern with a more granular chain ladder method resulted in a sharp loss in the overall performance (not shown).

The nowcasting models illustrated in Figure 5c and 5d directly specify a structure for the reporting intensity. In (18) (shown in panel 5c) the information on the occurrence date t is structured via a set of covariates, whereas panel 5d calibrates a parameter for each occurrence date t (as in (19)). With this type of models it is more difficult to capture insights from the calibrated parameters, as there is no explicit distinction between the dynamics in the occurrence and the reporting process. On the one hand, with a structured representation of the influence of the occurrence dates t does not sufficiently capture the actual pattern in the data. The model

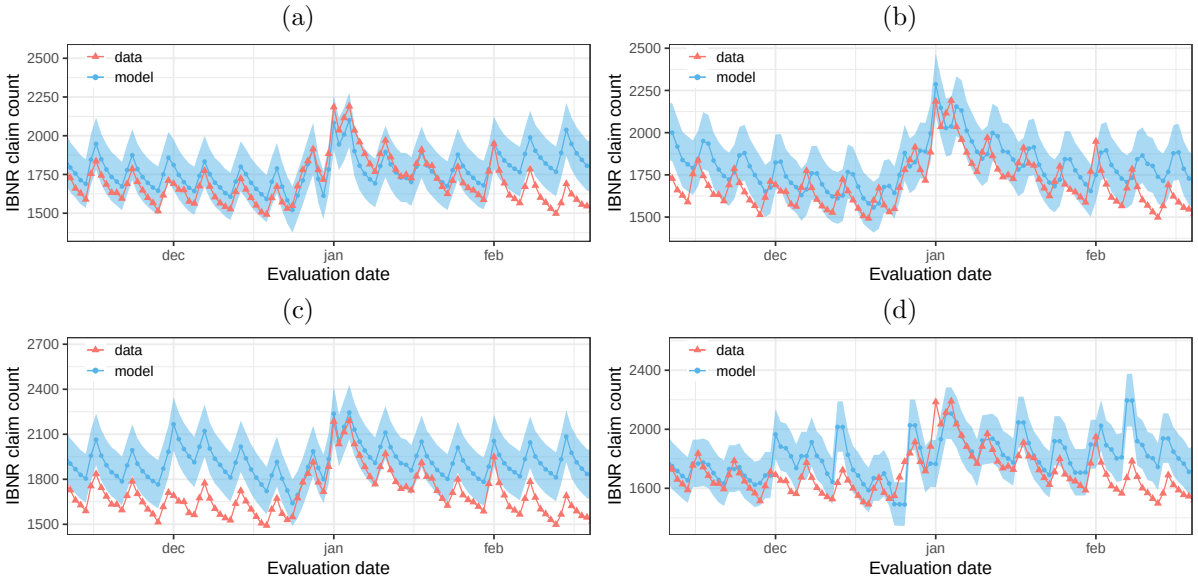

Figure 5: Predictions and 95% simultaneous prediction intervals of the total IBNR claim counts for varying evaluation dates from November 15 to February 15, 2004. (a) uses (13), (14) and (17), (b) yearly chain ladder method specified in (20), (c) direct specification of reporting intensity in (18), (d) direct specification of reporting intensity in (19). Red triangles: data, blue circles: model.

specification with a parameter for each occurrence date t on the other hand leads to volatile nowcasts, e.g., one prediction was removed (96 995 predicted IBNR claims on Sunday, January 11 2004). Since reporting on a Sunday is extremely rare in our data set the occurrence and immediate reporting of a claim on this date resulted in an unreasonably high λ_t estimate.

5 Conclusions and outlook

We review and structure the literature on nowcasting the occurrence of events subject to a reporting or observation delay. Our literature overview bridges multiple disciplines and incorporates papers from the actuarial, statistical and epidemiological literature. We propose a general modelling and estimation framework capable of dealing with any parametric structure for both the occurrence as well as the reporting process. The framework uses regression models for count data and treats the right truncation of the reporting delays as a type of missing data. Applying an EM algorithm strongly simplifies maximum likelihood estimation as it allows for the use of standard statistical software to fit the regression models. As an example, we demonstrate how the parameter estimators of a classical method for nowcasting with aggregated data, also known as the chain ladder method, can be obtained using an EM approach. We investigate the performance of our proposed framework on an insurance case study where the focus is on predicting the number of incurred but not (yet) reported claims in a European portfolio of general liability insurance policies for private individuals. We benchmark the predictions obtained with our proposed strategy against the results from other nowcasting models that directly structure the reporting intensity and can be calibrated with standard statistical routines. The presented model provides a better understanding of the claim occurrence as well as reporting delay process and leads to a refined view on the future reporting times (at daily level or aggregated e.g. by future reporting weeks or months).

We indicate some possible directions for future research. First of all, we would like to

stress that the provided estimation framework involving an EM algorithm is applicable in a wide context (e.g. epidemiology and reliability engineering) whenever cases are only reported after a delay. This provides a more desirable alternative over the ad hoc methods or two-step approaches used earlier in the literature. The estimation procedure described in Section 3 is readily applicable to other contexts after specifying a suitable parametric model for the reporting delay probabilities.

It would then be interesting to explore different distributional assumptions for the daily total event counts and the reporting delay structure. The reporting delay can be easily altered within the given framework to, for instance, a zero-inflated or hurdle distribution or a more heavy-tailed distribution. Relaxing the Poisson assumption for the daily total event counts is also feasible but might complicate the E-step in which we now relied on the thinning property of Poisson distributions. The EM framework is however compatible with latent underlying processes affecting the occurrence of events such as hidden Markov models or shot noise process (see e.g. Badescu et al., 2019; Avanzi et al., 2016). Another promising approach would be to investigate how time series models for counts (see Jung and Tremayne, 2011, for an overview) could be introduced in this setting.

Acknowledgements

The authors are grateful to the editors, associate editor and the referee for the valuable comments and suggestions.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csáki (Eds.), *Second International Symposium on Information Theory*, pp. 267–281. Akadémiai Kiadó, Budapest.

Antonio, K. and R. Plat (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal 7*, 649–669.

Avanzi, B., B. Wong, and X. Yang (2016). A micro-level claim count model with overdispersion and reporting delays. *Insurance: Mathematics and Economics 71*, 1–14.

Bacchetti, P., M. R. Segal, and N. P. Jewell (1993). Backcalculation of HIV infection rates. *Statistical Science 8*(2), 82–101.

Badescu, A. L., X. S. Lin, and D. Tang (2016). A marked Cox model for the number of IBNR claims: Theory. *Insurance: Mathematics and Economics 69*, 29–37.

Badescu, A. L., X. S. Lin, and D. Tang (2019). A marked Cox model for the number of IBNR claims: Estimation and application. *ASTIN Bulletin 49*, 709–739.

Barreto-Souza, W. and A. B. Simas (2017). Improving estimation for beta regression models via EM-algorithm and related diagnostic tools. *Journal of Statistical Computation and Simulation 87*(14), 2847–2867.

Bastos, L. S., T. Economou, M. F. Gomes, D. A. Villela, F. C. Coelho, O. G. Cruz, O. Stoner, T. Bailey, and C. T. Codeço (2019). A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in Medicine 38*(22), 4363–4377.

Bates, D. and M. Maechler (2015). *MatrixModels: Modelling with Sparse And Dense Matrices*. R package version 0.4-1.

Becker, N. G. and J.-S. Cui (1997). Estimating a delay distribution from incomplete data, with application to reporting lags for AIDS cases. *Statistics in Medicine 16*(20), 2339–2347.

Bellocco, R. and I. C. Marschner (2000). Joint analysis of HIV and AIDS surveillance data in back-calculation. *Statistics in Medicine 19*(3), 297–311.

Brookmeyer, R. and M. H. Gail (1988). A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic. *Journal of the American Statistical Association 83*(402), 301–308.

Cavanaugh, J. E. and R. H. Shumway (1998). An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference 67*, 45–65.

Claeskens, G. and F. Consentino (2008). Variable selection with incomplete covariate data. *Biometrics 64*, 1062–1069.

Claeskens, G. and N. L. Hjort (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological) 34*(2), 187–220.

Crevecoeur, J., K. Antonio, and R. Verbelen (2019). Modeling the number of hidden events subject to observation delay. *European Journal of Operational Research 277*(3), 930 – 944.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Dirick, L., G. Claeskens, and B. Baesens (2015). An Akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research 241*, 449–457.

Donker, T., M. van Boven, W. van Ballegooijen, T. van't Klooster, C. Wielders, and J. Wallinga (2011). Nowcasting pandemic influenza A/H1N1 2009 hospitalizations in the Netherlands. *European Journal of Epidemiology 26*, 195–201.

England, P. D. and R. J. Verrall (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal 8*(3), 443–518.

Farrington, C. P., N. J. Andrews, A. D. Beale, and M. A. Catchpole (1996). A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 159*(3), 547–563.

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

Greene, S. K., S. F. McGough, G. M. Culp, L. E. Graf, M. Lipsitch, N. A. Menzies, and R. Kahn (2021, Jan). Nowcasting for real-time COVID-19 tracking in New York City: An evaluation using reportable disease data from early in the pandemic. *JMIR Public Health Surveill 7*(1), e25538.

Haastrup, S. and E. Arjas (1996). Claims reserving in continuous time: a nonparametric Bayesian approach. *ASTIN Bulletin 26*, 139–164.

Hachemeister, C. A. and J. N. Stanard (1975). IBNR claims count estimation with static lag functions. In *Spring Meeting of the Casualty Actuarial Society*.

Harris, J. E. (1990). Reporting delays and the incidence of AIDS. *Journal of the American Statistical Association 85*(412), 915–924.

Höhle, M. and M. an der Heiden (2014). Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics 70*(4), 993–1002.

Hiabu, M., E. Mammen, M. D. Martínez-Miranda, and J. P. Nielsen (2021). Smooth backfitting of proportional hazards with multiplicative components. *Journal of the American Statistical Association*.

Jewell, W. S. (1989). Predicting IBNYR events and delays: I. Continuous time. *ASTIN Bulletin 19*, 25–55.

Jung, R. C. and A. R. Tremayne (2011). Useful models for time series of counts or simply wrong ones? *AStA Advances in Statistical Analysis 95*(1), 59–91.

Kalbfleisch, J., J. Lawless, and J. Robinson (1991). Methods for the analysis and prediction of warranty claims. *Technometrics 33*(3), 273–285.

Kalbfleisch, J. D. and J. F. Lawless (1989). Inference based on retrospective ascertainment: An analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association 84*(406), 360–372.

Kalbfleisch, J. D. and J. F. Lawless (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica 1*(1), 19–32.

Lagakos, S. W., L. M. Barraj, and V. D. Gruttola (1988). Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika 75*(3), 515–523.

Lawless, J. (1994). Adjustments for reporting delays and the prediction of occurred but not reported events. *Canadian Journal of Statistics 22*(1), 15–31.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 44*(2), 226–233.

Mack, T. (1991). A simple parametric model for rating automobile insurance or estimating IBNR claims reserves. *ASTIN Bulletin 21*(1), 93–109.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin 23*(02), 213–225.

Martínez Miranda, M. D., J. P. Nielsen, S. Sperlich, and R. Verrall (2013). Continuous chain ladder: Reformulating and generalizing a classical insurance problem. *Expert Systems with Applications 40*(14), 5588 – 5603.

McGough, S. F., M. A. Johansson, M. Lipsitch, and N. A. Menzies (2020). Nowcasting by Bayesian smoothing: A flexible, generalizable model for real-time epidemic tracking. PLOS *Computational Biology 16*(4), 1–20.

Meng, X.-L. and D. B. Rubin (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association 86*(416), 899–909.

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance. *ASTIN Bulletin 23*(1), 95–115.

Norberg, R. (1999). Prediction of outstanding liabilities II. Model variations and extensions. *ASTIN Bulletin 29*(1), 5–27.

Noufaily, A., P. Farrington, P. Garthwaite, D. G. Enki, N. Andrews, and A. Charlett (2016). Detection of infectious disease outbreaks from laboratory data with reporting delays. *Journal of the American Statistical Association 111*(514), 488–499.

Noufaily, A., Y. Ghebremichael-Weldeselassie, D. G. Enki, P. Garthwaite, N. Andrews, A. Charlett, and P. Farrington (2015). Modelling reporting delays for outbreak detection in infectious disease data. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 178*(1), 205–222.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society Series B-Statistical Methodology 61*(2), 479–482.

Pagano, M., X. M. Tu, V. D. Gruttola, and S. MaWhinney (1994). Regression analysis of censored and truncated data: Estimating reporting- delay distributions and aids incidence from surveillance data. *Biometrics 50*(4), 1203–1214.

Renshaw, A. E. and R. J. Verrall (1998). A stochastic model underlying the chain-ladder technique. *British Actuarial Journal 4*(4), 903–923.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Sellero, C. S., E. V. Fernández, W. Manteiga, X. L. Otero, X. Hervada, E. Fernández, and X. A. Taboada (1996). Reporting delay: a review with simulation study and application to Spanish data. *Statistics in Medicine 15*(3), 305–321.

Tabnak, F., H. Müller, J. Wang, J.-M. Chiou, and R. Sun (2000). A change-point model for reporting delays under change of AIDS case definition. *European Journal of Epidemiology 16*, 1135–1141.

Taylor, G. (2000). *Loss reserving: an actuarial perspective*. Kluwer Academic Publishers.

van de Kassteele, J., P. Eilers, and J. Wallinga (2019). Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained P-spline smoothing. *Epidemiology 30*, 737–745.

Verrall, R. J. and M. V. Wüthrich (2016). Understanding reporting delay in general insurance. *Risks 4*(3), 25.

Wahl, F. (2019). Explicit moments for a class of micro-models in non-life insurance. *Insurance: Mathematics and Economics 89*, 140 – 156.

Wu, S. (2013). A review on coarse warranty data and analysis. *Reliability Engineering & System Safety 114*, 1–11.

Wüthrich, M. V. and M. Merz (2008). *Stochastic claims reserving methods in insurance*, Volume 435 of *Wiley Finance*. John Wiley & Sons.

Wüthrich, M. V. and M. Merz (2015). *Stochastic Claims Reserving Manual: Advances in Dynamic Modeling*. Swiss Finance Institute Research Paper No. 15–34.

Zhu, H., S. Lee, B. Wei, and J. Zhou (2001). Case-deletion measures for models with incomplete data. *Biometrika 88*(3), 727–737.

Supplementary material for "Modeling the occurrence of events subject to a reporting delay via an EM algorithm"

Roel Verbelen[1,3,4], Katrien Antonio[1,2,3,4], Gerda Claeskens[1,3], and Jonas Crevecoeur[1,4]

[1] *Faculty of Economics and Business, KU Leuven, Belgium.*
[2] *Faculty of Economics and Business, University of Amsterdam, The Netherlands.*
[3] *LStat, Leuven Statistics Research Center, KU Leuven, Belgium.*
[4] *LRisk, Leuven Research Center on Insurance and Financial Risk Analysis, KU Leuven, Belgium.*

June 24, 2021

This Appendix collects additional figures and tables regarding the case study on insurance nowcasting, discussed in Section 4 of our paper entitled *Modeling the occurrence of events subject to a reporting delay via an EM algorithm*.

A Case study

We analyze a data set with the occurrence and reporting dates of claims in a portfolio of general liability insurance policies for private individuals from a European insurance company. The data set used in this case study has also been studied in e.g. Antonio and Plat (2014) with a stochastic model for the development of claims after reporting and Crevecoeur et al. (2019) who also nowcast the IBNR claims, but exclusively put focus on the dynamics in the reporting process.

A.1 Description of the insurance data set

Exposure is available in this dataset by month from January 2000 onwards, expressed as *earned exposure*, i.e. the fraction of policies actually exposed to risk during the period. This means that a policy providing insurance coverage only during 10 days in January will contribute 10/365th to the exposure of that month. Earned exposure is not available on a daily level so instead we transform the monthly exposure to daily exposure by dividing by the number of days in each month. This accounts for the varying month lengths.

To enable out-of-time prediction, we restrict our analysis to claims that have occurred between January 1, 2000 and August 31, 2004. We then study the nowcasting problem using evaluation date, say $\tau^\star$, at the end of this time window, on August 31, 2004 and want to estimate the total incurred but not reported (IBNR) claim count, as well as the reporting dates of these IBNR claims. Based on the full data set until August 2009, 176 671 claims have occurred during the time window between January 1, 2000 and August 31, 2004. Due to a reporting delay, only 174 624 of these have been reported by the evaluation date, as depicted in blue in the daily run-off triangle in Figure 6. The remaining 2047 are IBNR claims, i.e. claims which have occurred between January 2000 and August 2004 but have only been reported after the evaluation date and before the end of the observation period. These are graphically illustrated in red in Figure 6.

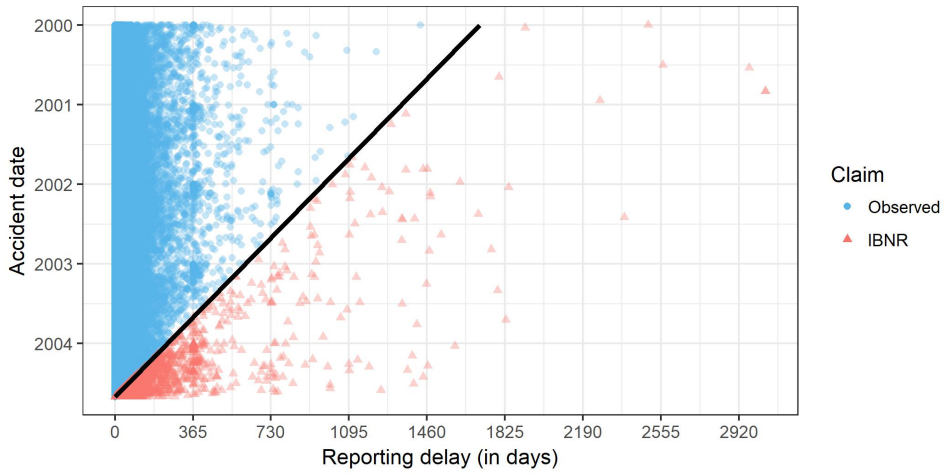

Figure 6: Daily run-off triangle of claims with occurrence dates between January 1, 2000 and August 31, 2004. The black line indicates the evaluation date, August 31, 2004. Only the claims in the upper triangle depicted as blue dots are observed at the evaluation date. The remaining claims in the lower triangle depicted as red triangles are the IBNR claims based on the full data set until August 2009 and have to be predicted.

A.2 Parametric models for the occurrence and reporting processes

Assisted by the data we specify a structure for the reporting probabilities $p_{td}(\boldsymbol{\theta}, \boldsymbol{x}_{td})$ introduced in assumption (A2'). The barplot in Figure 7a shows the empirical reporting probabilities in the first 28 days after occurrence for claims that occurred on a Monday. Reporting probabilities decrease over time and are low on Saturdays and Sundays. The claims pictured in Figure 7a all occurred on a Monday. Therefore the first weekend corresponds to delays of 5 (Saturday) and 6 (Sunday) days. Following this intra-week pattern, we structure reporting delay as the product of week probabilities and day (or intra-week) probabilities:

$$p_{td}(\boldsymbol{\theta}, \boldsymbol{x}_{td}) = p^{\mathrm{W}}_{tw}(\boldsymbol{\theta}, \boldsymbol{x}_t) \cdot p^{\mathrm{intra}}_{td}.$$

Here p^{W}_{tw} denotes the probability of reporting an event from occurrence day t in the w-th week after occurrence. The intra-week reporting probabilities are denoted with p^{intra}_{td}. The latter take values between 0 and 1 and sum to 1 over the days within the reporting week. Figure 7b indicates that the negative binomial distribution is a good fit for the empirical week probabilities.

A.3 Parameter estimates

Complementary to the results shown in the paper, we give a detailed overview and discussion of the parameter estimates obtained for various model specifications.

A.3.1 Occurrence and reporting processes: reporting week and intra-week probabilities

We focus on the occurrence model (13) in combination with a reporting specified by the reporting week probabilities in (14) and the intra-week reporting probabilities in (15). We report the estimated day probabilities from matrix P in Table 2. Only a small fraction of claims is being reported on Saturdays and nearly none on Sundays. In fact, in the entire observed part of the data, only 3 claims have been reported on Sunday.

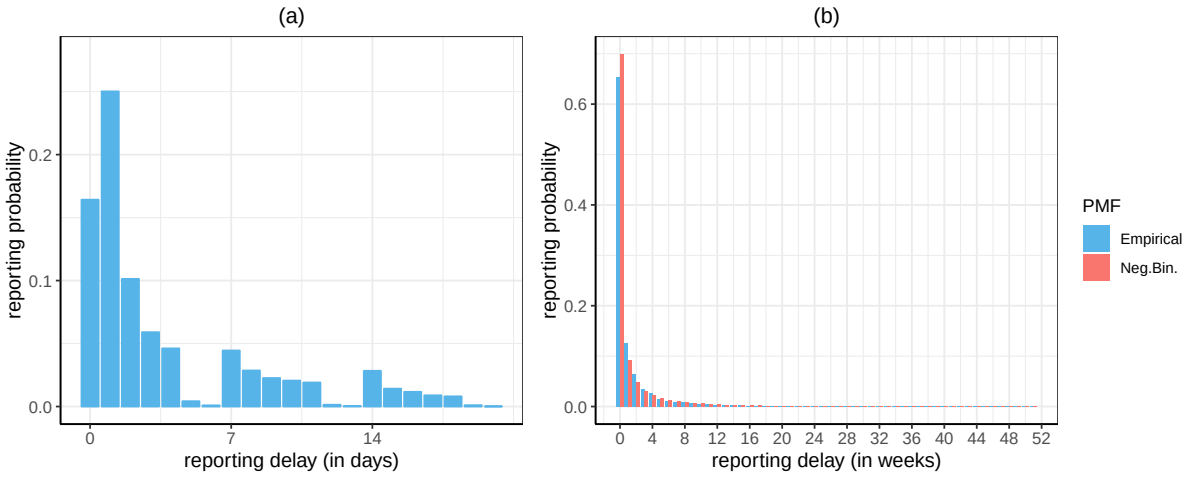

Figure 7: (a) Bar plot of the empirical reporting delay distribution in the first 28 days for claims that occurred on a Monday. (b) Bar plot of the empirical reporting delay distribution in weeks and its negative binomial fit for the first year based on claims that occurred between January 2000 and August 2004 and have been reported before August 2009.

The maximum likelihood estimates of the parameter vector $\boldsymbol{\alpha}$ from the claim occurrence model are shown in Figure 2 in the paper. Next to these, the maximum likelihood estimates of the parameter vector $\boldsymbol{\theta}$ used in the negative binomial regression model of the reporting delay in weeks are displayed in Figure 8. All estimates are shown along with 95% confidence intervals based on the inverse of the expected information matrix. Simultaneous confidence intervals are constructed in these graphs using the Bonferroni correction to adjust for multiple comparisons. For completeness, we also report that in the negative binomial model the intercept is estimated as 1.867 (with 95% confidence interval $[1.717, 2.018]$) and the dispersion parameter ϕ as 0.177.

A.4 Prediction of unreported claim counts

We illustrate how the occurrence model (13) in combination with the reporting structure defined in (14) and (15) is used to forecast the number number of unreported claims for past occurrence dates. In Figure 9a we plot point estimates and 95% simultaneous prediction intervals for N_t^{nr} with t corresponding to occurrences dates in between July 1, 2004, and August 31, 2004, i.e. the

Table 2: Maximum likelihood estimates of the day probabilities P within the reporting week. Separate reporting day probabilities are estimated for each day of the week (dow) of the occurrence date, as shown in the rows.

	wday						
dow	wday1	wday2	wday3	wday4	wday5	Saturday	Sunday
Monday	0.271	0.331	0.171	0.119	0.100	0.008	0.000
Tuesday	0.282	0.342	0.158	0.118	0.090	0.011	0.000
Wednesday	0.286	0.316	0.180	0.112	0.095	0.011	0.000
Thursday	0.278	0.337	0.156	0.114	0.097	0.019	0.000
Friday	0.303	0.264	0.160	0.120	0.096	0.057	0.000
Saturday	0.389	0.211	0.148	0.109	0.097	0.046	0.000
Sunday	0.407	0.222	0.157	0.109	0.096	0.009	0.000

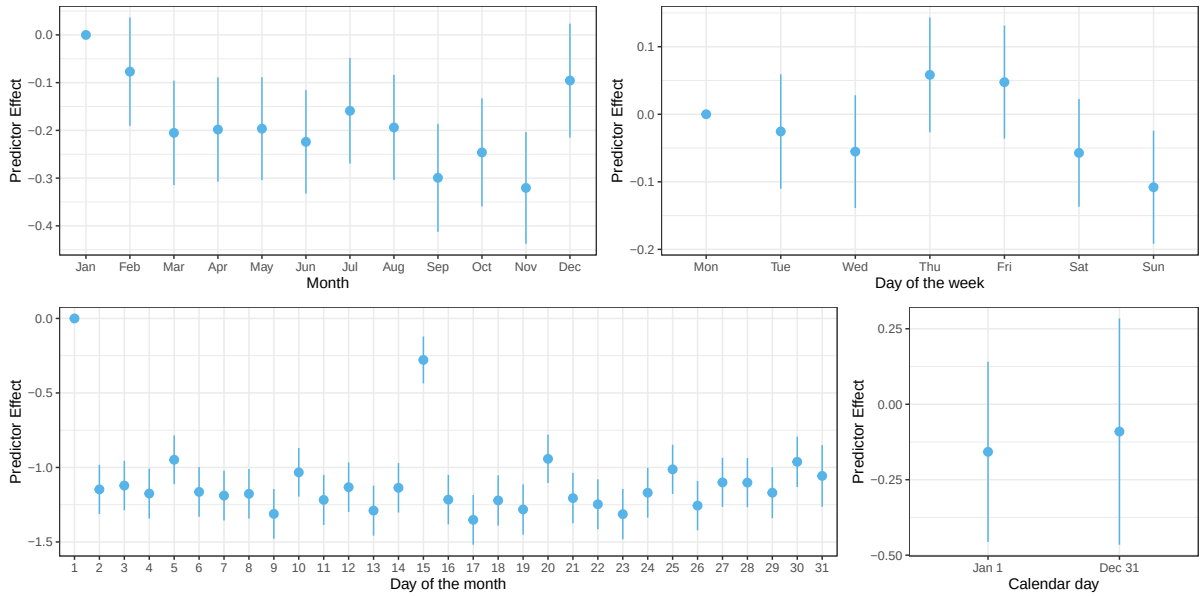

Figure 8: Maximum likelihood estimates and 95% simultaneous confidence intervals for $\boldsymbol{\theta}$ corresponding to the categorical effects of the month, the day of the week and the day of the month of the occurrence date in the negative binomial reporting delay model.

last two months from our training period. The predictions follow the same trend as the actual IBNR claim counts derived from the full data set until August 2009. We notice how IBNR claims are elevated on the first day and middle of each month, in line with our earlier findings. In Figure 9b (resp. Figure 9c) we group the occurrence dates by weeks (resp. months) prior to the evaluation date and show the IBNR claim count predictions corresponding to the past 26 weeks (resp. 12 months). We notice how, also over longer time spans, the predictions by occurrence week or month follow the pattern observed in the actual unreported counts.

A.5 Modelling the intra week probabilities using a reverse time strategy

We consider the occurrence model specified in (13) in combination with a reporting delay distribution with reporting week probabilities in (14) and intra-week probabilities obtained via the reverse time strategy in the presence of covariates, as specified in (16). Figure 10 shows the estimates for the occurrence model in (13), the estimates for the reporting week probabilities are shown in Figure 11 and the intra-week reporting probabilities obtained via the reverse time strategy are in Figure 12.

A.6 Directly modelling the reporting intensity

We now consider two nowcasting models that directly structure the reporting intensity. These models are discussed in Section 2.2.2 of the paper, where the N_{td} are independent Poisson distributed random variables with mean λ_{td}. Figure 13 shows the parameter estimates for the specification from (18) and Figure 14 displays the parameter estimates for the regression structure in (19).

A.7 Comparing nowcasting models with a moving window evaluation

In addition to the results shown in Figure 5 in the paper (with evaluation dates between November 15 to February 15, 2004), we show the estimates of the total IBNR claim count with moving

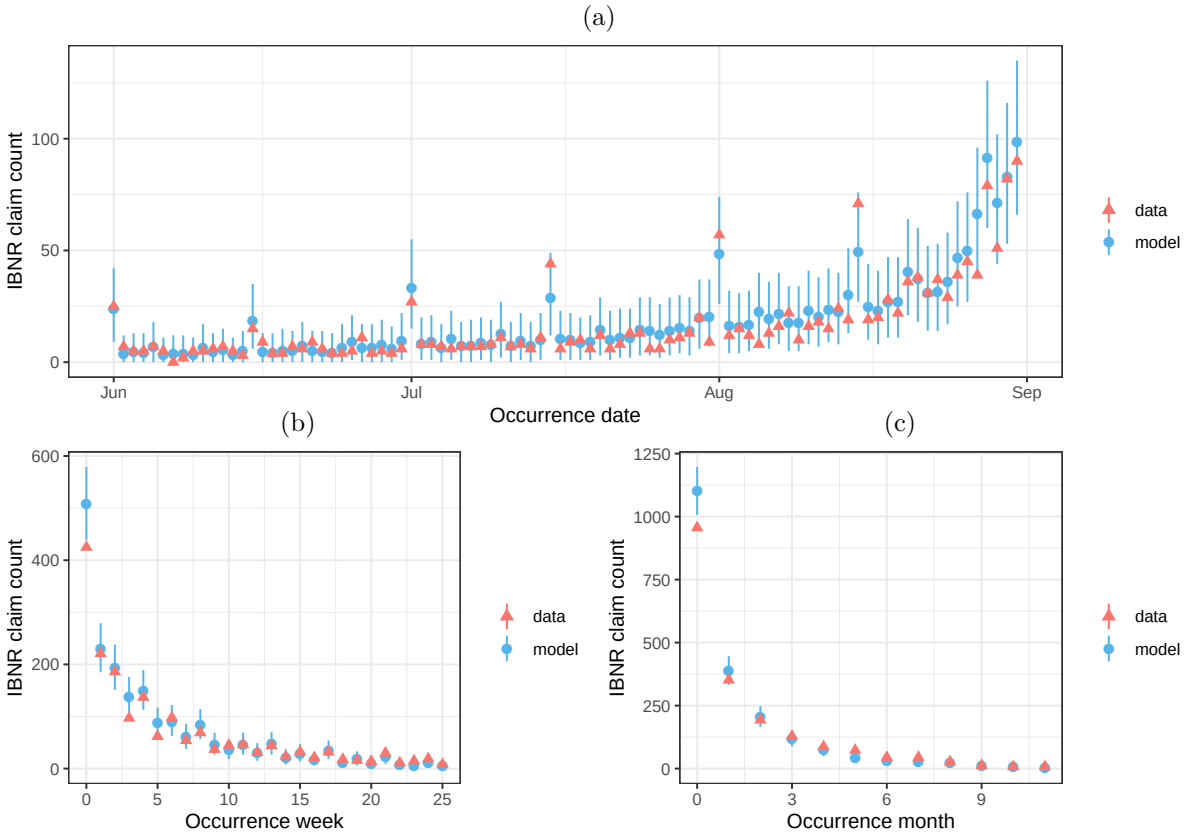

Figure 9: Predictions of the IBNR claim counts and simultaneous 95% prediction intervals by occurrence date. (a) Daily, for occurrence dates in between July 1 and August 31, 2004; (b) weekly (7 days) for the past 26 weeks; (c) monthly (30 days) for the past 12 months.

evaluation dates between August 31, 2003, and August 31, 2004. Figure 15a shows predictions obtained with the occurrence specification in (13) and the reporting model in (14) and (17). Results obtained with the yearly chain ladder method are displayed in Figure 15b. Panel 15c shows the estimates for the IBNR claim counts obtained with a nowcasting model that directly specifies the reporting intensity along (18), whereas the model leading to panel 15d uses (19).

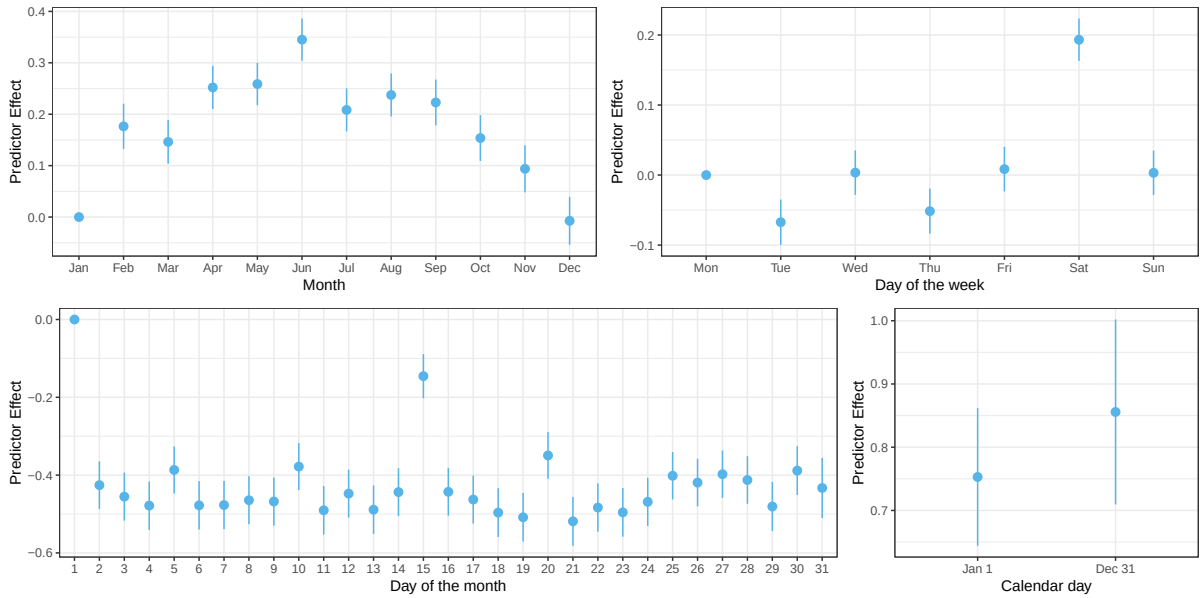

Figure 10: Maximum likelihood estimates for the parameters used in the occurrence model (13) jointly estimated with (14) and (17). 95% simultaneous confidence intervals are constructed using the Bonferroni correction and the inverse of the expected information matrix.

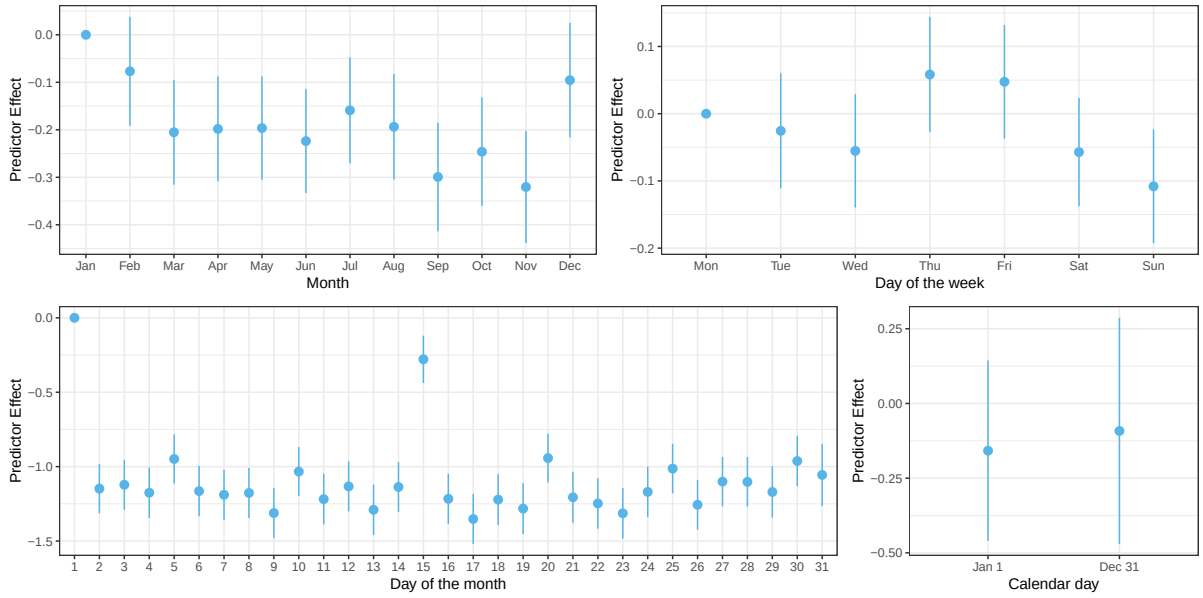

Figure 11: Maximum likelihood estimates for the parameters used in the reporting week model (14) jointly estimated with (13) and (17). 95% simultaneous confidence intervals are constructed using the Bonferroni correction and the inverse of the expected information matrix.

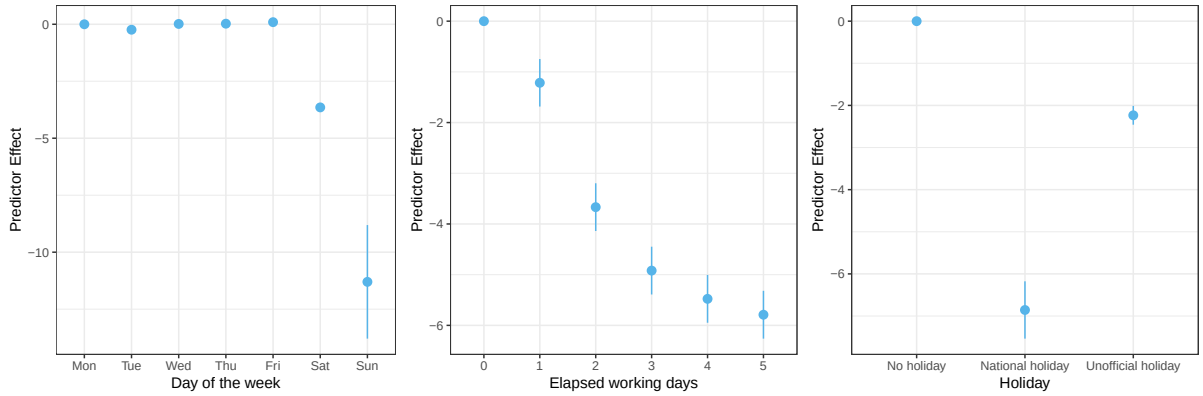

Figure 12: Maximum likelihood estimates for the parameters used in the intra-week reporting model (17) jointly estimated with (13) and (14). 95% simultaneous confidence intervals are constructed using the Bonferroni correction and the inverse of the expected information matrix.

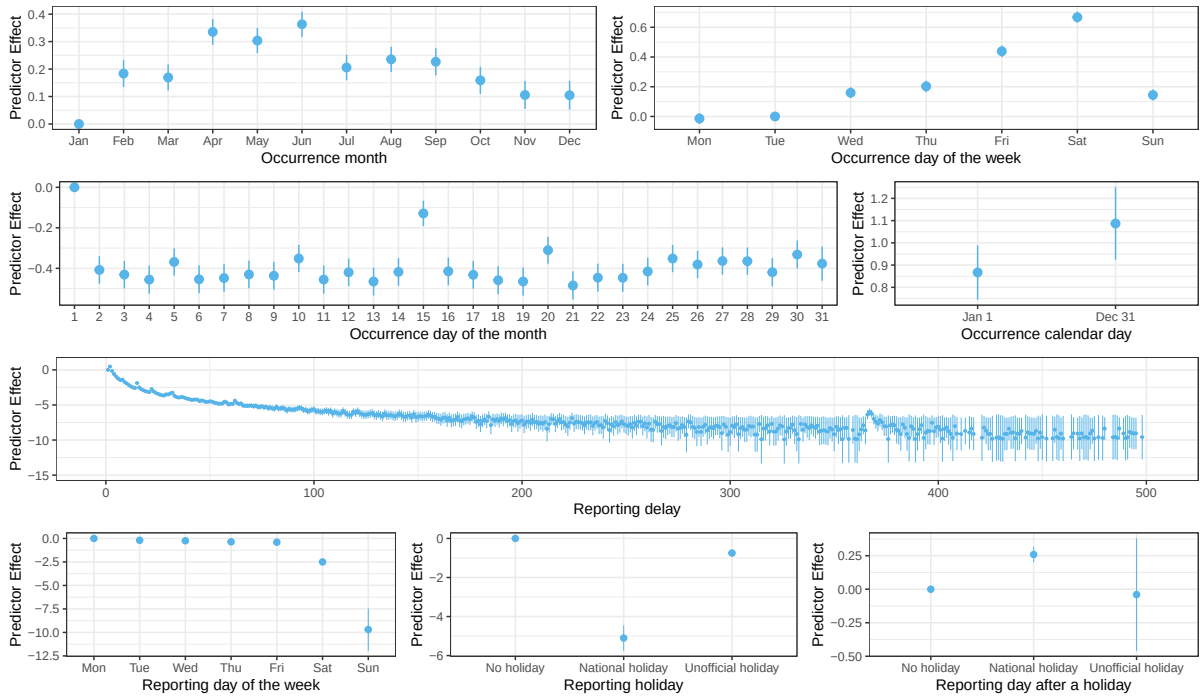

Figure 13: Maximum likelihood estimates for the parameters used in (18). 95% simultaneous confidence intervals are constructed using the Bonferroni correction and the inverse of the expected information matrix.

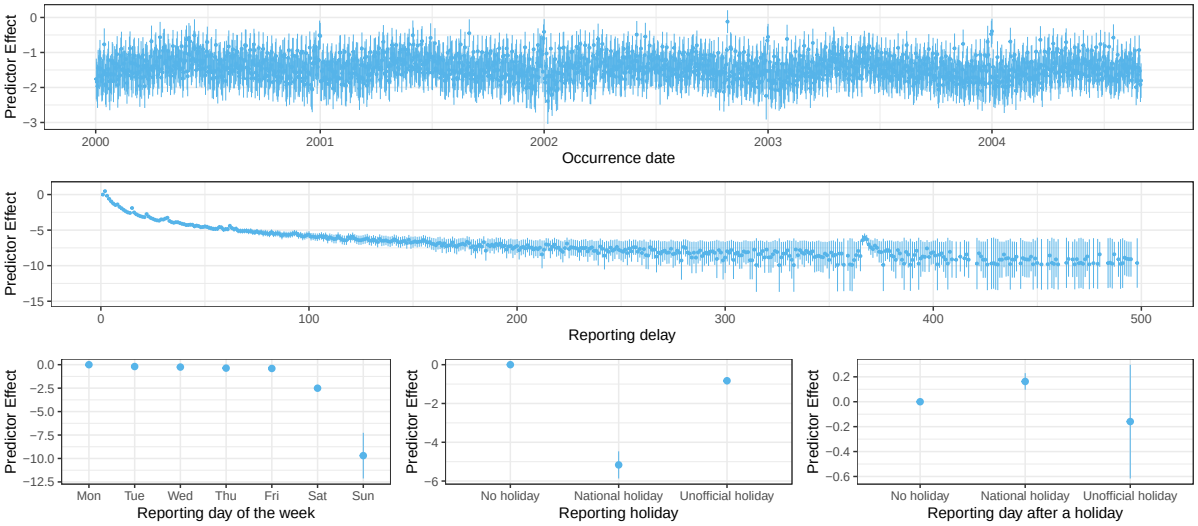

Figure 14: Maximum likelihood estimates for the parameters used in (19). 95% simultaneous confidence intervals are constructed using the Bonferroni correction and the inverse of the expected information matrix.

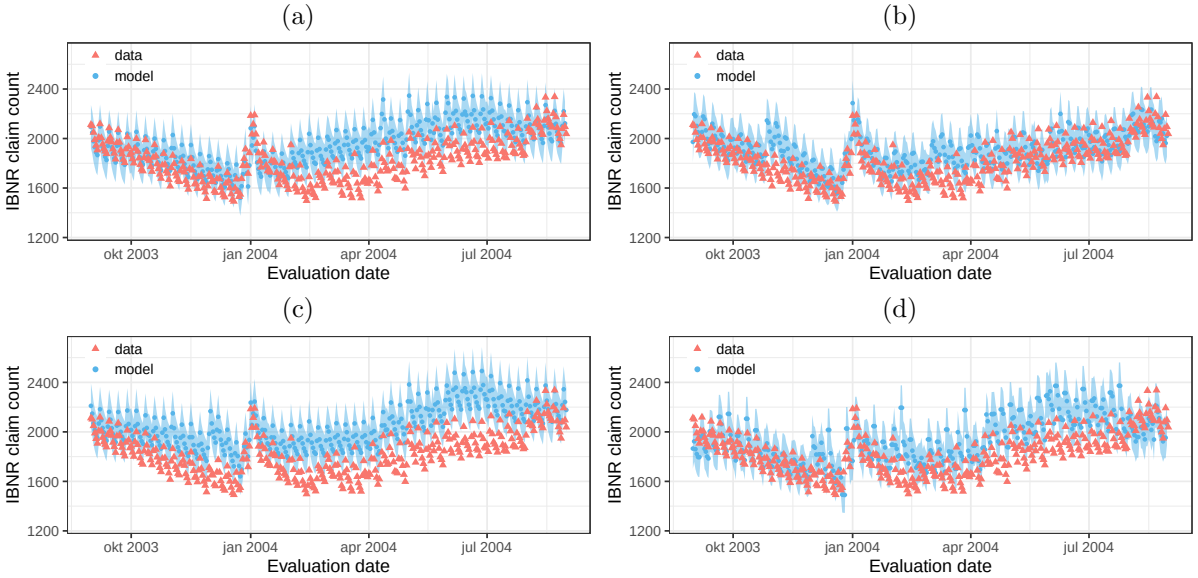

Figure 15: Predictions and 95% simultaneous prediction intervals of the total IBNR claim counts for varying evaluation dates from August 31, 2003 to August 31, 2004. (a) uses (13), (14) and (17), (b) yearly chain ladder method (20), (c) direct specification of reporting intensity in (18), (d) direct specification of reporting intensity in (19). Red triangles: data, blue circles: model.