

Modeling the number of hidden events subject to observation delay

Jonas Crevecoeur^{1,3,*}, Katrien Antonio^{1,2,3,4} and Roel Verbelen^{1,3,4}

¹Faculty of Economics and Business, KU Leuven, Belgium.

²Faculty of Economics and Business, University of Amsterdam, The Netherlands.

³LRisk, Leuven Research Center on Insurance and Financial Risk Analysis, KU Leuven, Belgium.

⁴LStat, Leuven Statistics Research Center, KU Leuven, Belgium.

*Corresponding author. E-mail: jonas.crevecoeur@kuleuven.be

March 27, 2019

Abstract

Copyright © 2019 European Journal of Operational Research. This paper considers the problem of predicting the number of events that have occurred in the past, but which are not yet observed due to a delay. Such delayed events are relevant in predicting the future cost of warranties, pricing maintenance contracts, determining the number of unreported claims in insurance and in modeling the outbreak of diseases. Disregarding these unobserved events results in a systematic underestimation of the event occurrence process. Our approach puts emphasis on modeling the time between the occurrence and observation of the event, the so-called observation delay. We propose a granular model for the heterogeneity in this observation delay based on the occurrence day of the event and on calendar day effects in the observation process, such as weekday and holiday effects. We illustrate this approach on a European general liability insurance data set where the occurrence of an accident is reported to the insurer with delay.

Keywords: Risk management; Occurrence of events; Observation delay; Calendar day effects; Data analytics.

1 Introduction

In many domains within operational research analysts are interested in building a stochastic model for the occurrence of events. However, the events of interest are often observed or reported

with some delay. Analysts should account for these unobserved events since ignoring them will bias the decisions based on the stochastic model under consideration. Figure 1 visualizes this setting. We specify a well defined observation window (on the x -axis) in which we observe the creation of new objects (e.g. products or contracts). Over the course of their lifetimes some objects may experience the event of interest (object 1 and 2 in Figure 1) before a given evaluation date, and others will not (object 3 and 4 in Figure 1). Upon occurrence the event is initially hidden from the decision maker. The time that elapses between the onset of the object’s lifetime and the occurrence of the event is called the event delay. Only after a so-called observation or reporting delay the decision maker becomes aware of the existence of the event. This paper outlines a data driven strategy to predict the number of events that occurred in the past (before the evaluation date), but which are hidden at the time of evaluation and will only be observed or reported in the future. Subject 2 in Figure 1 is an example of such an event.

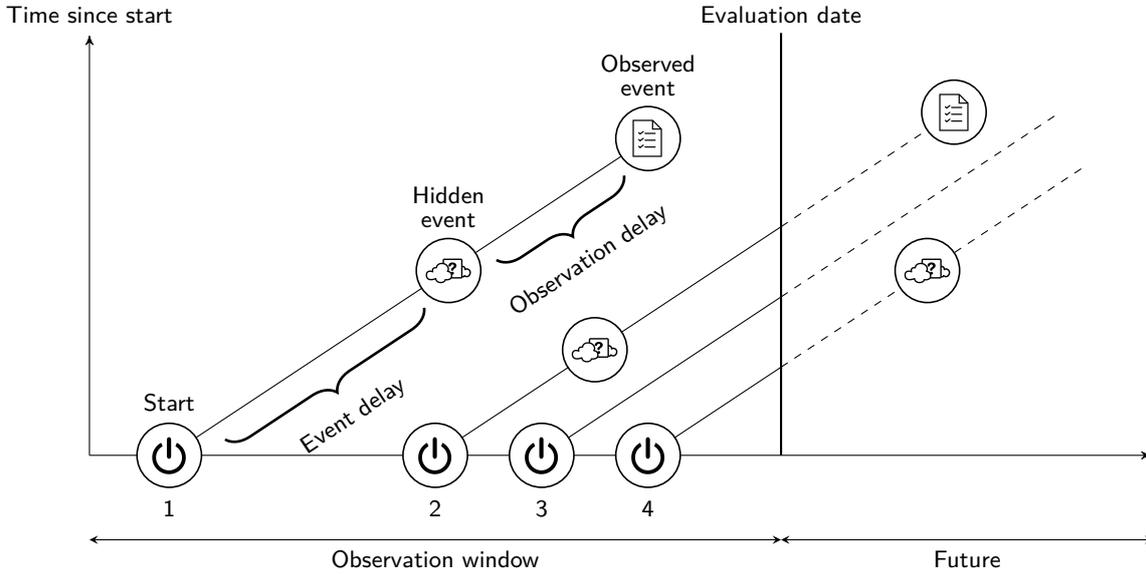


Figure 1: Occurrence and observation of events

The modeling of the time to occurrence of an event (‘the event delay’), the number of (hidden) events that occurred during a specific time window and the delay between occurrence and observation (‘the observation delay’) have been active research areas in the literature on operational research, actuarial science and epidemiology. Typical examples of applications where this predictive problem matters are: a portfolio of maintenance, warranty or insurance contracts, but also an outbreak of a specific disease fits within this framework. We highlight some relevant contributions and explain how this paper extends the existing literature.

A warranty contract requires the manufacturer to compensate the buyer for all failures occurring within the warranty period. Manufacturers hold capital for future compensations related to goods produced in the past. The amount of capital required depends on the number of defective products that have been sold. Accurate estimation of this number is complicated due to the

incompleteness of the data. The diagonal time line in Figure 1 begins when a defective product is produced. However, the warranty period only starts when the product is sold to a customer. Manufacturers are typically not aware of these sales and we consider them as a hidden events. Once the defect emerges and the customer calls his warranty contract, the manufacturer is informed of the sale ('the observed event'). Akbarov and Wu (2012) and Ye and Ng (2014) simultaneously model the time to sale and the delay between sale and failure of the product using parametric methods. Since both processes interact in the likelihood, estimation is difficult. Akbarov and Wu (2012) resolve to numerical maximization, whereas Ye and Ng (2014) use a Stochastic Expectation Maximization strategy. While these authors model the time to sale with a simple, parametric distribution without covariates, our framework accounts for the seasonal effects, promotions holidays and weather effects typically present in sales data.

Epidemiologists face a similar statistical problem when modeling the evolution of diseases (Harris, 1990; Salmon et al., 2015). In this setting, subjects are followed over time and a recent disease infection may remain unobserved due to either delay in disease diagnosis by a medical doctor or incubation time. Modeling these delays allows to take the yet unobserved infections ('the hidden events') into account and thus enables a faster and more accurate identification of disease outbreaks and epidemics (Noufaily et al., 2016).

Maintenance contracts are typically sold together with large industrial appliances. Under these contracts the manufacturer or a third party guarantees the continued use of the equipment. A machine failure ('the observed event') is often the result of previous defects ('the hidden event') which remained unobserved. These defects can be detected by on site inspections and timely repairs will prevent expensive failures or breakdowns of the machine. However, the profitability of these inspections depends largely on the number of hidden defects. Observation delay was first modelled in the context of maintenance contracts by Christer (1973), where it is called delay-time. Since then several papers have focussed on the delay-time concept. Baler and Wang (1993) model delay-time from observed failure data using maximum likelihood estimation. In this approach both the time to defect as well as the time to observation of the machine failure are tackled with parametric distributions. This literature typically assumes a constant intensity for the occurrences of defects and ignores heterogeneity in the delay-time distribution. Wang (1997) and Apeland and Scarf (2003) rely on expert opinions to formulate a fully subjective delay-time model. Wang (2010) and Berrade et al. (2018) focus on economic decision making when the delay-time distribution is known. In line with the current era of big data analytics (see Mortenson et al. (2015)), our approach goes beyond these assumptions and proposes a data driven strategy to capture heterogeneity in both the occurrence of defects as well as in the delay between a defect and its observation.

The case-study presented in this paper illustrates our data driven approach with an insurance data set where contracts are sold to policyholders. Some policyholders will be involved in an accident or other type of insured event, while others will not. In insurance parlance the delay

between the occurrence ('the hidden event') of an accident and the reporting or filing of the claim to the insurance company ('the observed event') is called the reporting delay. These delays are strongly portfolio dependent and can be substantial when the insured does not immediately notice the damage. In the remainder of the paper we only consider accidents that will eventually be reported. Accidents that are never reported do not get reimbursed and are not relevant for the balance sheet of the insurer. Once the claim is reported and accepted by the insurer, the insurer reimburses the loss with a single payment or a series of payments. Insurance companies book a reserve to be able to settle the claims that are Incurred But Not yet Reported (IBNR) and refer to this capital as the IBNR reserve. Estimating the number of claims from past exposures that will be reported beyond the evaluation date (the so-called IBNR claim counts) is crucial in setting this reserve. Motivated by computational constraints from the past, many estimation methods in insurance structure the data from Figure 1 in a two dimensional table that aggregates the number of accidents by their year of occurrence and year of reporting. We refer the reader to [Taylor \(2000\)](#); [Wüthrich and Merz \(2008\)](#); [Wüthrich and Merz \(2015\)](#) for more details on reserving with aggregate methods. Relatively few papers address the problem of specifying a model at granular level for the phenomenon sketched in Figure 1. [Badescu et al. \(2016\)](#) and [Avanzi et al. \(2016\)](#) focus on modeling the accident arrival process at a weekly level using Cox processes. These models allow to capture over-dispersion and serial dependence, which is often encountered in such occurrence data. The assumption of independence between the occurrence date and the reporting delay is a disadvantage of the models presented in [Badescu et al. \(2016\)](#) and [Avanzi et al. \(2016\)](#). [Verrall and Wüthrich \(2016\)](#) were the first to present a model for IBNR counts at a daily level, including the heterogeneity in reporting delays based on the occurrence date of the claim and the strong weekday pattern leading to less claims being reported during the weekend. This weekday pattern relates to calendar day effects in the reporting process which are difficult to model using classical techniques designed for aggregated data (see [Kuang et al. \(2008\)](#)). [Verrall and Wüthrich \(2016\)](#) provide a method to incorporate this weekday pattern for reporting delays of less than one week. [Verbelen et al. \(2017\)](#) extend this weekday pattern to reporting delays beyond the first week by separately estimating weekly and intra week reporting probabilities. Moreover, [Verbelen et al. \(2017\)](#) present the Expectation Maximization algorithm as a framework for jointly estimating the occurrence and reporting process.

Our paper models the occurrence of hidden events non-parametrically. This allows to capture fluctuations in occurrence counts (for example due to seasonality or weather conditions) without explicitly modeling these events. Moreover, extending the work of [Verrall and Wüthrich \(2016\)](#) and [Verbelen et al. \(2017\)](#) we model the observation delay in the presence of multiple covariates, including calendar day effects. Examples of such calendar day effects are: a reduction in observed events during the weekend, the effect of national holidays and seasonality in observation delay. Our strategy introduces the concept of observation exposure as an intuitive and flexible framework for incorporating (multiple) calendar day effects through regression. This approach elegantly transforms the observation delay distribution by scaling the probability of

observing an event on a certain date based on covariates. As such, the transformed observation delay distribution becomes independent of these covariates and is then modelled with a simple, parametric distribution. This makes our approach suitable to a wide range of problems.

This paper is organized as follows. Section 2 describes a statistical framework for modeling the number of hidden events subject to an observation delay. In Section 3 we illustrate this approach in a case-study involving an insurance data set. We also investigate the performance of our model in four simulated scenarios. The online appendix provides detailed expressions for implementing the model and links our approach to the non-parametric Kaplan-Meier estimator (Kaplan and Meier, 1958).

2 A granular model for the occurrence of events subject to delay

Denote by N_t the number of events occurring on date t , where $t = 1$ is the date of the first event. These events remain hidden until their observation at date s after a delay $s - t$. Let $N_{t,s}$ be the number of events that occurred on date t and are observed on date s . Since all events will be observed at some point in the future, we find

$$N_t = \sum_{s \geq t} N_{t,s}.$$

Consider an evaluation date τ at which we have to predict the number of hidden events. At τ we split the events from a past occurrence date t into observed ($s \leq \tau$) and hidden events which are not yet observed ($s > \tau$), respectively denoted by

$$N_t^{\text{Obs}}(\tau) = \sum_{s=t}^{\tau} N_{t,s} \quad \text{and} \quad N_t^{\text{Hidden}}(\tau) = \sum_{s=\tau+1}^{\infty} N_{t,s} \quad \text{for } t \leq \tau.$$

We obtain the total number of hidden events by aggregating the unobserved events from all past occurrence dates, i.e.

$$N^{\text{Hidden}}(\tau) = \sum_{t=1}^{\tau} N_t^{\text{Hidden}}(\tau) = \sum_{t=1}^{\tau} \sum_{s=\tau+1}^{\infty} N_{t,s}.$$

This total count is the number that we want to predict. Following Jewell (1990) and Norberg (1993), we formulate two distributional assumptions from which the number of hidden events can be predicted:

- (A1) The event occurrence process $(N_t)_{t \geq 1}$ follows an inhomogeneous Poisson distribution with intensity $(\lambda_t)_{t \geq 1}$.
- (A2) The observation delay is independent and identically distributed for events occurring on the same date.

Denote by $p_{t,s}$ the probability of observing an event from occurrence date t on date s . We use the notation $p_t^{\text{Obs}}(\tau)$ for the probability that an event from date t is observed by the evaluation date τ . This probability is

$$p_t^{\text{Obs}}(\tau) = \sum_{s=t}^{\tau} p_{t,s}.$$

By assumption (A1) and (A2) the conditions for the Poisson thinning property (Kingman, 1993) are satisfied. The thinning property implies that all $N_{t,s}$ are independent and

$$N_{t,s} \sim \text{Poisson}(\lambda_t \cdot p_{t,s}). \quad (1)$$

This allows us to construct the likelihood for the observed data at time τ . Let $\boldsymbol{\chi}$ denote the available data, consisting of all events that are observed on the evaluation date τ

$$\boldsymbol{\chi} = \{N_{t,s} \mid t \leq s \leq \tau\}.$$

The loglikelihood of the observed data is

$$\ell(\boldsymbol{\lambda}, \boldsymbol{p}; \boldsymbol{\chi}) = \sum_{t=1}^{\tau} \sum_{s=t}^{\tau} \left[N_{t,s} \cdot \log(\lambda_t) + N_{t,s} \cdot \log(p_{t,s}) - \lambda_t \cdot p_{t,s} - \log(N_{t,s}!) \right] \quad (2)$$

where $\boldsymbol{\lambda}$ is a vector with components λ_t for observed occurrence dates t and $\boldsymbol{p} = \{p_{t,s} \mid t \leq s \leq \tau\}$. This paper puts focus on the observation process without imposing any structure on λ_t . A straightforward computation shows that the loglikelihood in (2) is maximal for

$$\lambda_t = \frac{\sum_{s=t}^{\tau} N_{t,s}}{\sum_{s=t}^{\tau} p_{t,s}} = \frac{N_t^{\text{Obs}}(\tau)}{p_t^{\text{Obs}}(\tau)}. \quad (3)$$

Replacing λ_t by this expression the loglikelihood in (2) becomes

$$\ell(\boldsymbol{p}; \boldsymbol{\chi}) = \sum_{t=1}^{\tau} \sum_{s=t}^{\tau} N_{t,s} \cdot \log(p_{t,s}) - \sum_{t=1}^{\tau} N_t^{\text{Obs}}(\tau) \cdot \log(p_t^{\text{Obs}}(\tau)) + \text{constants}. \quad (4)$$

Up to constants this is the loglikelihood for a right truncated observation delay random variable. The truncation point is $\tau - t$, which is the maximal observed delay for an event that occurred on date t .

2.1 A time change strategy to model observation delay

We are interested in structuring the observation probabilities $p_{t,s}$ based on covariates corresponding to the occurrence date t and the reporting date s of the event. The probabilistic

nature of the data enforces the constraints

$$p_{t,s} \geq 0, \quad \forall t, s \quad \text{and} \quad \sum_{s \geq t} p_{t,s} = 1, \quad \forall t. \quad (5)$$

The proposed time change strategy transforms the reporting probabilities such that they can be linked with covariates while preserving these constraints. This transformation is depicted in Figure 2, where we consider an event that occurred on a Thursday and for which observation is less likely during the weekend.

First, we view the discrete observation delay as a realization of a continuous random variable U_t under interval censoring. This is graphically illustrated in Figure 2a (discrete setting) and 2b (continuous setting). Second, we define a time change operator φ_t which assigns a positive length $\alpha_{t,s}$, called the observation exposure, to each combination of an occurrence date t and an observation date s . This time change operator is similar to the concept of operational time, which is a common technique in continuous financial mathematics, see Swishchuk (2016). We perceive dates as having variable lengths, whereas prior to this time change an equal length of one time unit was attached to each date. The probability of observing an event on a certain date is scaled by the duration of this date, which motivates calling this length the observation exposure. We define the time-changed delay $\varphi_t(d)$ for an event with occurrence date t and an observation delay of d days as

$$\varphi_t(0) = 0 \quad \text{and} \quad \varphi_t(d) = \sum_{i=1}^d \alpha_{t,t+i-1}, \quad d \in \mathbb{N} \setminus \{0\}. \quad (6)$$

This is the sum of all observation exposures $\alpha_{t,s}$ assigned to dates in between the occurrence date t and date $t+d-1$. By applying φ_t on the observation delay random variable U_t we obtain a time-changed random variable $\tilde{U} := \varphi_t(U_t)$ which is independent of the occurrence date t of the event. The discrete observation probabilities are easily extracted from this distribution using the relation

$$\begin{aligned} p_{t,s} &= P(U_t \in [s-t, s-t+1)) \\ &= F_{\tilde{U}} \left(\sum_{i=1}^{s-t+1} \alpha_{t,t+i-1} \right) - F_{\tilde{U}} \left(\sum_{i=1}^{s-t} \alpha_{t,t+i-1} \right). \end{aligned} \quad (7)$$

Under the time change transformation the constraints (5) become

$$\alpha_{t,s} \geq 0, \quad \forall t, s \quad \text{and} \quad \sum_{s \geq t} \alpha_{t,s} = \infty, \quad \forall t.$$

We specify a regression model for the daily observation exposure as a function of covariates. We set

$$\log(\alpha_{t,s}) = \mathbf{x}'_{t,s} \cdot \boldsymbol{\gamma},$$

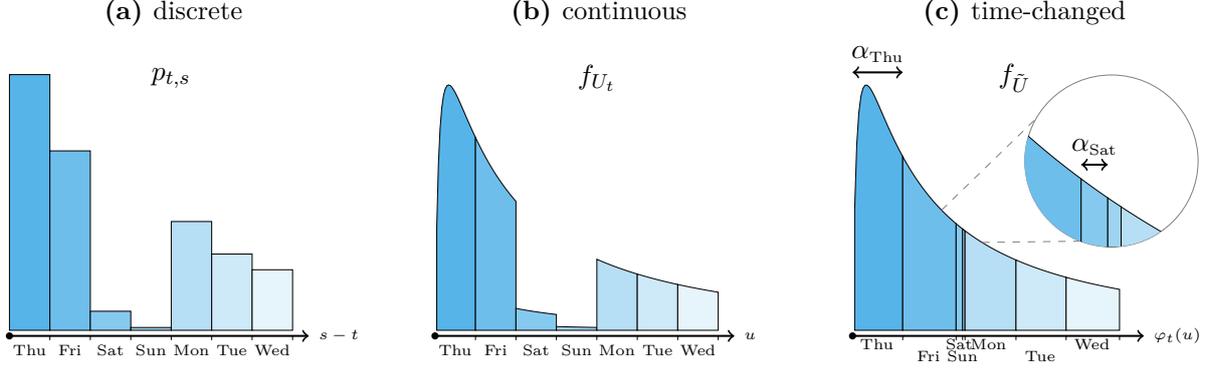


Figure 2: Observation delay distribution for an event that occurred on a Thursday. We illustrate (a) the discrete observation delay probabilities $p_{t,s}$, (b) the density of the continuous observation delay distribution U_t and (c) the density of the time-changed observation delay distribution \tilde{U} .

for a vector $\mathbf{x}_{t,s}$ of covariates related to observing on date s an event that occurred on date t and the corresponding parameter vector $\boldsymbol{\gamma}$. In contrast with classical regression methods, the reporting probabilities $p_{t,s}$ not only depend on the characteristics of the observation date, but instead take the full history between the event occurrence and observation date into account through the time change strategy.

Figure 2c illustrates this time change. Since less claims get reported during the weekend, we model observation exposure as a function of the reporting day of the week. The time change then assigns lower observation exposures to Saturday and Sunday, hereby transforming the continuous distribution from Figure 2b into a time-changed distribution that can be modeled using standard loss distributions.

2.2 Calibration

Our approach divides the observation delay model into two components. The time change transformation φ_t defined in (6) captures the heterogeneity in the observation process. This transformation is expressed by the daily observation exposures, which require the calibration of the regression parameters $\boldsymbol{\gamma}$. The time transformed observation delay \tilde{U} is modeled with a simple parametric probability distribution, where the data will assist us in choosing the best candidate. We optimize the loglikelihood in (4) with respect to $\boldsymbol{\gamma}$, i.e. we maximize

$$\begin{aligned} \ell(\boldsymbol{\gamma}; \boldsymbol{\chi}) = & \sum_{t=1}^{\tau} \sum_{s=t}^{\tau} N_{t,s} \cdot \log \left[F_{\tilde{U}} \left(\sum_{v=t}^s \alpha_{t,v} \right) - F_{\tilde{U}} \left(\sum_{v=t}^{s-1} \alpha_{t,v} \right) \right] \\ & - \sum_{t=1}^{\tau} N_t^{\text{R}}(\tau) \cdot \log \left[F_{\tilde{U}} \left(\sum_{v=t}^{\tau} \alpha_{t,v} \right) \right], \end{aligned}$$

with $\alpha_{t,v} = \exp(\mathbf{x}'_{t,v} \cdot \boldsymbol{\gamma})$. Online appendix A describes an optimization strategy for this loglikelihood that is applicable to any sufficiently smooth distribution $F_{\tilde{U}}(\cdot)$. The described strategy is generic and does not immediately take properties from the chosen distribution into account. Significant reductions in computation time can be obtained when \tilde{U} follows a standard exponential distribution. The loglikelihood then becomes

$$\begin{aligned} \ell(\boldsymbol{\gamma}; \boldsymbol{\chi}) = & - \sum_{t=1}^{\tau} \sum_{s=t}^{\tau} N_{t,s} \cdot \left(\sum_{v=t}^{s-1} \alpha_{t,v} - \log(1 - \exp(-\alpha_{t,s})) \right) \\ & - \sum_{t=1}^{\tau} N_t^{\text{R}}(\tau) \cdot \log \left(1 - \exp \left(- \sum_{v=t}^{\tau} \alpha_{t,v} \right) \right). \end{aligned} \quad (8)$$

The first line in (8) is a sum in which each term depends on a single observation exposure, $\alpha_{t,s}$. Since this facilitates computing first and second order derivatives with respect to the reporting exposure, this results in a lower computation time.

2.3 Predicting the number of hidden events

At the evaluation date τ we predict the number of events from past occurrence dates t that will be observed on future dates s . Hence our focus is on

$$N_{t,s}, \quad \text{for } t \leq \tau \text{ and } s > \tau.$$

We aggregate these future daily observation counts to find the total number of hidden events

$$N^{\text{Hidden}}(\tau) = \sum_{t=1}^{\tau} N_t^{\text{Hidden}}(\tau) = \sum_{t=1}^{\tau} \sum_{s=\tau+1}^{\infty} N_{t,s}.$$

Following the Poisson assumption in (1) each random variable $N_{t,s}$ is independently Poisson distributed with mean

$$E(N_{t,s}) = \lambda_t \cdot p_{t,s}.$$

The observation delay model developed in Section 2.1 provides estimates for the observation probabilities $p_{t,s}$, see (7)

$$\hat{p}_{t,s} = P(\tilde{U} \in [\varphi_t(s-t), \varphi_t(s-t+1)) \mid \hat{\boldsymbol{\gamma}}).$$

In (3) we proposed a pragmatic, non-parametric estimator for the claim occurrence intensity on date t , namely

$$\hat{\lambda}_t = \frac{N_t^{\text{Obs}}(\tau)}{\hat{p}_t^{\text{Obs}}(\tau)}. \quad (9)$$

This estimator depends only on the observed events and the estimated observation delay distribution. This is an advantage when the event generating process is volatile. For dates with

unexpectedly many events the number of observations will be higher and thus we correctly predict more event occurrences. On the downside, (9) is less reliable for recent dates when the denominator is close to zero or when the number of daily events is low. When the data set is small, the non-parametric estimator can be replaced by a parametric estimator following the strategy outlined in Bonetti et al. (2016) and Verbelen et al. (2017). In a parametric framework the estimator for the occurrence intensity may include the daily risk exposure, expressed as the number of policies in effect on a day. Including risk exposure increases the robustness of parametric models to evolutions in the portfolio size and may potentially improve the predictive performance of the model.

3 Case-study: reporting delay dynamics in insurance

3.1 Data characteristics

We illustrate our approach with the analysis of a liability insurance data set from the Netherlands. The same data is studied in Pigeon et al. (2013), Pigeon et al. (2014) and Godecharle and Antonio (2015) with focus on calculating reserves in discrete time, Antonio and Plat (2014) model reserves in continuous time and Verbelen et al. (2017) who propose a model for the number of hidden claim counts at a daily level. The data registers 506 235 claims related to insured events that occurred and were reported between July, 1996 and August, 2009. From these claims, we remove 75 observations with a reporting date prior to the accident date and 559 claims that are the result of transitions in the reporting system. We focus on the occurrence date of accidents and the corresponding reporting delay in days, i.e. the time (in days) between occurrence of the accident and reporting or filing of the claim to the insurer. To avoid losing valuable insights by aggregation, we study the data at a daily level. This is the most granular timescale at which the data is available.

Occurred accidents Figure 3 shows the daily number of accidents that occurred between July, 1996 and August, 2009 and initiated a claim reported to the insurance company before August 31, 2009. Since only claims reported before August 31, 2009 are observed, we see a decrease in observed event counts for the most recent dates which have a substantial number of unreported claims. Two outliers are not shown in this plot, namely 456 accidents on October 27, 2002 and 818 accidents on January 18, 2007. Both outliers correspond to a storm in the Netherlands causing many insured events.¹ The red line in this figure shows the moving average of the number of occurrences, calculated over the latest 30 days. This trend reveals a seasonal pattern in the occurrence process with more events occurring during the summer months. The

¹Details (in Dutch) about the storms by the royal national meteorological institute of the Netherlands (KNMI): <https://knmi.nl/over-het-knmi/nieuws/storm-van-27-oktober-2002-was-zwaarste-in-twaalf-jaar> and <https://knmi.nl/over-het-knmi/nieuws/de-zware-storm-kyrill-van-18-januari-2007>

trend slightly increases over time due to an increase in portfolio size. Several of the outlying observations in Figure 3 correspond to occurrences on the first of January as indicated by the vertical gray bars at the beginning of each year.

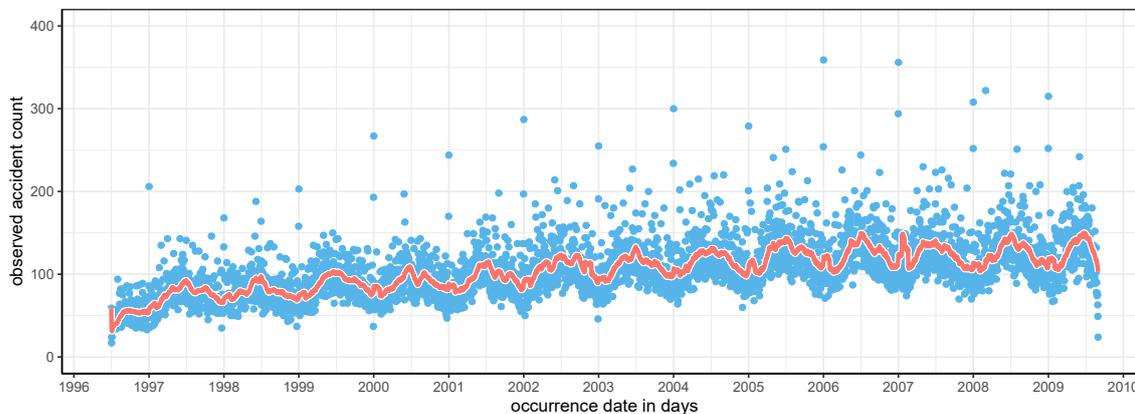


Figure 3: Daily number of accidents that occurred between July, 1996 and August, 2009 and were reported before August, 2009. The solid line shows the moving average of occurred accidents, calculated over the latest 30 dates. Two outliers are not shown on the graph: October 27, 2002 (456 accidents) and January 18, 2007 (818 accidents).

Reported claims Figure 4 shows the daily number of claims reported between July 1996 and August 2009. Again the red line shows the moving average of the number of reported claims, calculated over the latest 30 days. The seasonality in event counts observed in Figure 3 leads to a similar seasonal pattern in reported claim counts, though with a slight lag due to the delay in reporting a claim. Figure 4 reveals two regimes of reporting. On most dates many claims get reported, but there is a substantial number of dates on which few or almost no claims are reported. These dates with few reports correspond to the weekend (Saturday, Sunday) and national holidays.² This separation in two regimes is not the case for the occurrence process, since accidents continue to occur during the weekend and on holidays. We further illustrate these calendar day effects, where reporting is substantially reduced on specific dates, in Figure 5. The left hand side lists the average number of reported claims between July, 1996 and August, 2009 on ten national holidays during which all businesses are closed. These averages are compared with the overall daily average of reported claim counts over the observation period. This shows that reporting is strongly reduced on national holidays. We include two non-official holidays, New Year’s Eve and Good Friday. These dates show a slight reduction in reporting because many people take a day off from work. The reporting behavior on weekdays is shown in Figure 5b. During the weekend and especially on Sunday the number of reports is reduced. These calendar day effects motivate a model for IBNR claim counts at a daily level, capable of incorporating the weekday and holiday effect observed in our empirical analysis.

²List of national holidays in the Netherlands: <http://www.officeholidays.com/countries/netherlands/>

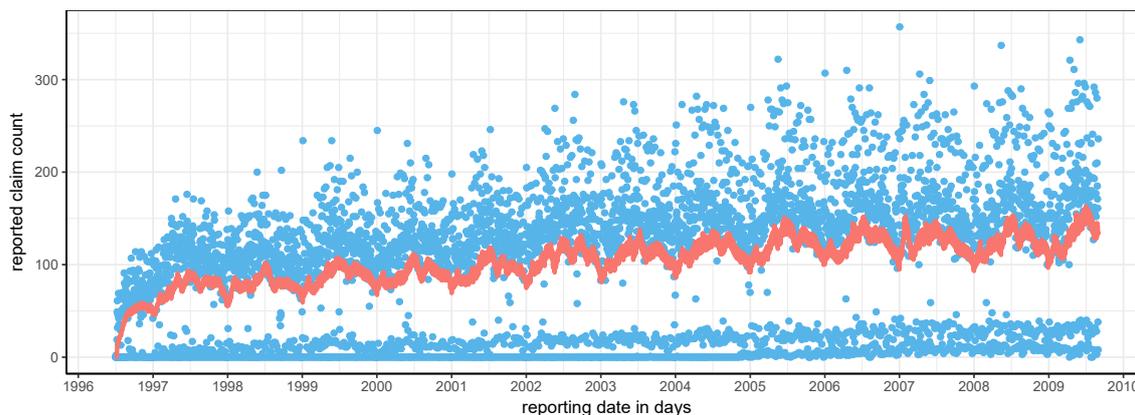


Figure 4: Daily number of claims that were reported on each date between July, 1996 and August, 2009. The solid line shows the moving average of reported claims, calculated over the latest 30 dates.

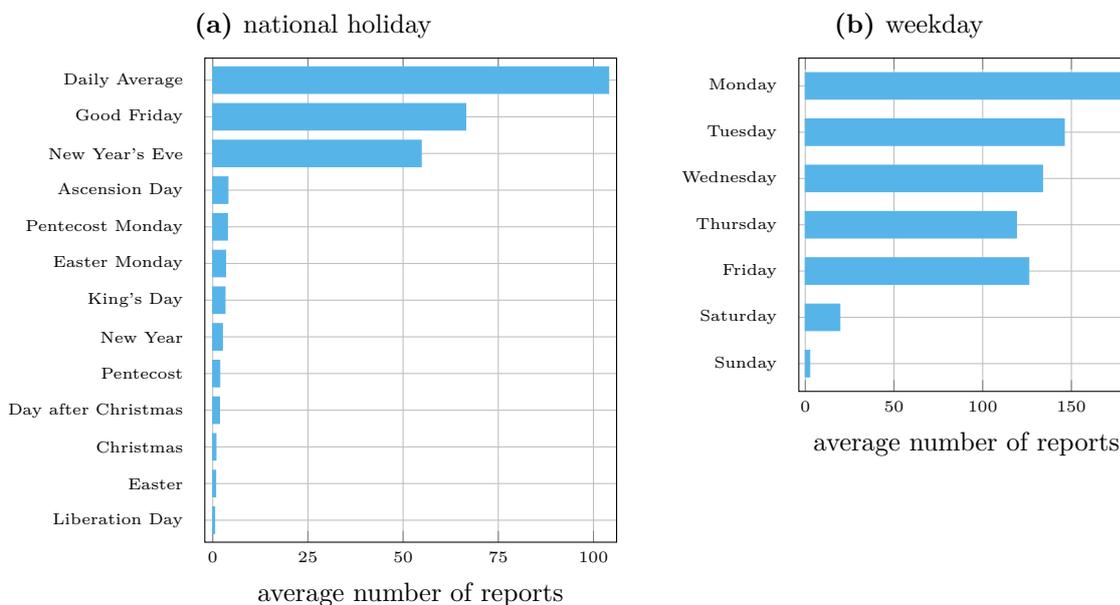


Figure 5: Average number of reported claims on (a) national holidays and (b) weekdays, calculated over all claims that occurred and were reported between July, 1996 and August, 2009.

Reporting delay Figure 6 illustrates the empirical reporting delay distribution in days over the first three weeks after the occurrence of the insured event. The empirical probability of reporting peaks the day after the claim occurred and strongly decreases afterwards. The increase in reporting after exactly fourteen days is most likely a consequence of data quality issues, where insureds who no longer recall the exact occurrence date report that the accident happened two weeks ago. The same effect to a lesser degree is visible after exactly one week. Figure 6b and Figure 6c show the empirical reporting delay distribution constructed using only accidents that occurred on Monday and Thursday, respectively. This reveals the effect of the occurrence's day of the week on the reporting delay distribution. An accident that happened on a Monday

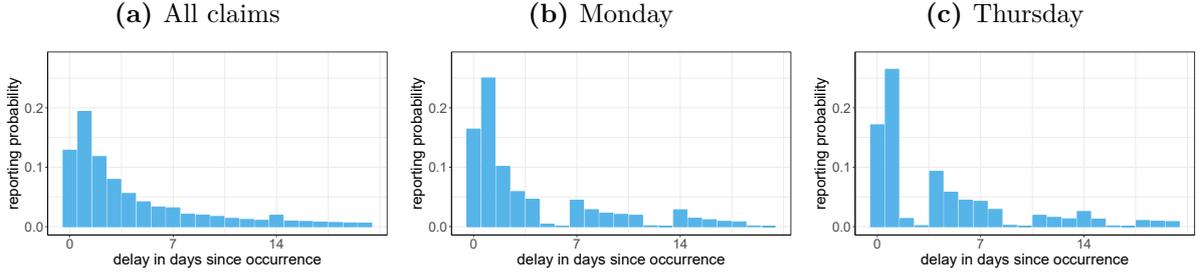


Figure 6: Empirical reporting delay distribution in days over the first three weeks after the occurrence of the claim using (a) all claims, (b) claims that occurred on a Monday and (c) claims that occurred on a Thursday.

has a decreased probability of reporting after six or seven days, since these delays correspond to Saturday and Sunday, respectively. Accidents that occurred on a Thursday show the same pattern of reporting delay, but the weekend then corresponds to a different delay. The effect of the weekend is no longer visible in the empirical distribution using all claims (Figure 6a), since the weekend then no longer corresponds to a specific reporting delay.

The number of hidden events The evaluation date refers to the date on which the insurer computes the reserve. In practice this date is often the last day of a quarter or the financial year. Figure 7 uses a rolling evaluation date to illustrate the daily number of IBNR claims. For each evaluation date we show the number of claims corresponding to insured events that occurred before this date but were reported afterwards (and before August 31, 2009, the last day of our observation period). The top panel of Figure 7 shows the daily number of IBNR claims on each evaluation date between September 1, 2003 and August 31, 2004. The number of unreported claims varies throughout the year with more unreported claims in the summer, when more accidents occur. IBNR counts peak around the start of the new year since many accidents occur on the first of January and reporting is slow due to a clustering of holidays. The bottom panel of Figure 7 zooms in on the unreported claims between October 1, 2003 and November 30, 2003. Large fluctuations in unreported claims appear when we evaluate IBNR on a daily basis. These movements follow a seven day pattern where five days of decrease in IBNR are followed by two days of strong upward movement. These upward moves correspond to the weekend when many new insured events occur, but almost no events get reported.

3.2 Model specification

We opt for computational efficiency and model the time-changed reporting delay \tilde{U} with an exponential distribution. The reporting exposures include six effects and are structured as

$$\begin{aligned} \alpha_{t,s} &= \alpha_t^{\text{occ. dom}} \cdot \alpha_t^{\text{occ. month}} \cdot \alpha_s^{\text{rep. holiday}} \cdot \alpha_s^{\text{rep. month}} \cdot \alpha_{s,s-t}^{\text{rep. dow, first week}} \cdot \alpha_{s-t}^{\text{delay}} \quad (10) \\ &= \exp \left((\mathbf{x}_t^{\text{occ. dom}})' \cdot \gamma^{\text{occ. dom}} + (\mathbf{x}_t^{\text{occ. month}})' \cdot \gamma^{\text{occ. month}} \right) \end{aligned}$$

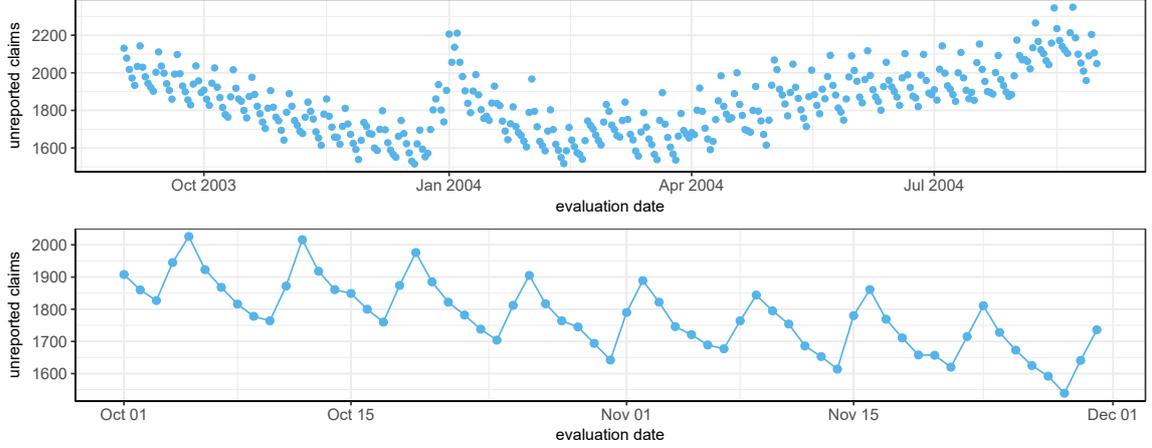


Figure 7: Number of unreported claims at each evaluation date between September 2003 and August 2004. These are the number of claims that occurred before this date, but were reported afterwards (but before the end of the observation period, i.e. August 31, 2009). The bottom panel zooms in on evaluation dates in October and November, 2003.

$$\begin{aligned}
 &+ (\mathbf{x}_s^{\text{rep. holiday}})' \cdot \gamma^{\text{rep. holiday}} + (\mathbf{x}_s^{\text{rep. month}})' \cdot \gamma^{\text{rep. month}} \\
 &+ (\mathbf{x}_{s,s-t}^{\text{rep. dow, first week}})' \cdot \gamma^{\text{rep. dow, first week}} + (\mathbf{x}_{s-t}^{\text{delay}})' \cdot \gamma^{\text{delay}}.
 \end{aligned}$$

We model the impact of the occurrence date on the reporting delay by incorporating effects for the day of the month $\alpha_t^{\text{occ. dom}}$ and the month $\alpha_t^{\text{occ. month}}$ on which the accident occurs. The holiday effect in Figure 5a is modeled by $\alpha_s^{\text{rep. holiday}}$, which distinguishes between national and unofficial holidays. Seasonal variations in reporting are captured by $\alpha_s^{\text{rep. month}}$, which scales reporting exposure based on the month in which the claim is reported. An interaction effect $\alpha_{s,s-t}^{\text{rep. dow, first week}}$ estimates the reporting exposure for combinations of a reporting delay in the first week ($s - t = 0, 1, \dots, 6$) and the day of the week on which the claim is reported. Separate weekday parameters are estimated for delays of more than one week, $s - t \geq 7$. As such, we capture the weekday effect from Figure 5a with additional flexibility in the first week after the claim occurs. Finally, $\alpha_{s-t}^{\text{delay}}$ partitions the time elapsed since the accident occurred in 23 bins according to the strategy specified in online Appendix C. These bins adapt the tail of the distribution as well as increase the probability of reporting after 14, 30 and 365 days.

3.3 Results

3.3.1 Parameter estimates

We estimate the model parameters by maximizing the loglikelihood in (8) using 8 years of data i.e. all accidents that occurred and were reported between July 1, 1996 and September 5, 2004. The resulting training data set contains 274 187 reported claims, for which we model the reporting process using 125 parameters. Figure 8 shows the maximum likelihood estimates for

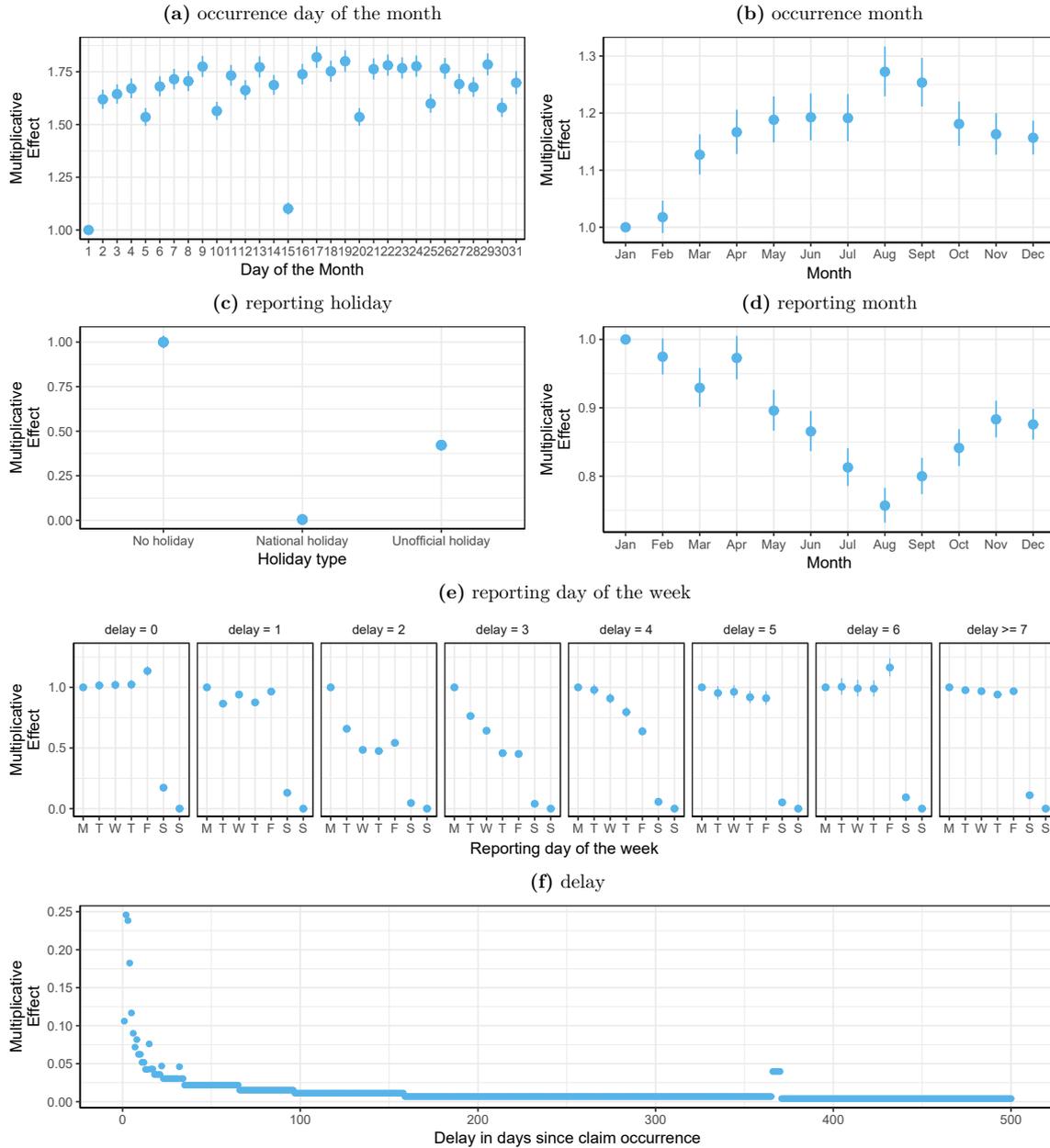


Figure 8: Maximum likelihood estimates with 95%-confidence intervals for the reporting exposure parameters $\exp(\gamma)$ in (10).

the reporting exposure parameters $\exp(\gamma)$ in (10). Together with these point estimates we plot 95%-confidence intervals derived from the Fisher information matrix for γ .

Occurrence day of month Figure 8a shows the effect of the day of the month on which the accident occurred. Reporting exposure is lower for accidents that occur on the first or fifteenth of the month, which implies that accidents from these days have a longer reporting delay. This

is most likely the result of data quality issues. Insureds who report a claim with a long reporting delay might no longer remember the exact occurrence date of the corresponding accident, which leads them to register the occurrence date at the start (first) or middle (fifteenth) of the month. This creates an increase in the average reporting delay for events that occurred on the first and fifteenth of the month. The same effect to a lesser degree is visible on the 5th, 10th, 20th, 25th and 30th of the month.

Month Two month effects are included in the reporting exposure structure. Figure 8b shows the effect for $\exp(\gamma^{\text{occ. month}})$ which considers the month in which the accident occurs. These parameters indicate that reporting is slower for accidents that occurred around the beginning of the year (January, February) and faster in the summer. Figure 8d visualizes the parameters for the reporting month, $\exp(\gamma^{\text{rep. month}})$. We observe a reduction in reporting exposure during the summer months. Slightly counterintuitive, we find that the parameters $\gamma^{\text{occ. month}}$ and $\gamma^{\text{rep. month}}$ largely offset each other for accidents that occur and get reported in the same calendar month. When combining these effects, the reduction in reporting exposure during the summer is mostly noticeable for claims that occurred before the summer months.

Holiday Figure 8c shows the effect of holidays on reporting exposure. Hardly any claim gets reported on national holidays and the reporting probability is reduced by more than 50% on unofficial holidays (Good Friday and New Year’s Eve). These estimates are of the same magnitude as the effects found in the empirical analysis in Figure 5.

Reporting day of the week We include the day of the week effect in the reporting exposure specification (10) through an interaction between the time elapsed after the accident occurred $s - t$ and the day of the week on which the claim is reported. Figure 8e shows a grouping of the estimated coefficients based on the time elapsed since the occurrence of the accident. For all delays we notice a reduction in reporting exposure during the weekend, with few reports on Saturday and almost no reports on Sunday. This interaction is important as the estimated parameters differ strongly based on the delay considered. For example, accidents that occur on Friday or Saturday are often reported on the next Monday, which corresponds to a delay of two and three days respectively. Since Monday is the reference level, the fitted parameters for other weekdays are lower at these delays. The right most panel in Figure 8e shows the effect of the reporting day of the week for delays beyond one week. For these longer delays, all working days (Mon - Fri) have a similar reporting exposure.

Delay Figure 8f shows the evolution of the reporting exposure component $\exp(\gamma^{\text{delay}})$ in (10) as a function of the time elapsed since the accident occurred. This effect scales the reporting probability at specific delays such that the time-changed reporting delay \tilde{U} better resembles

an exponential distribution. We identified 23 bins upfront based on the strategy of online appendix C. The first eight days after occurrence end up in separate bins. These short delays are important, since many claims get reported soon after their occurrence date. Moreover, Figure 8f shows that the calibrated effect changes strongly for these delays. The model also contains bins to capture the increase in reporting probability for delays of exactly 14, 21 and 31 days as well as for reporting after one year. The bin size widens when reporting delay increases. The final two bins $[158, 364]$ and $[370, \infty)$ let the model capture the tail of the distribution.

3.4 Out-of-time predictions

We predict the number of hidden events, i.e. the IBNR claim count, following the strategy outlined in Section 2.3. Because the non-parametric occurrence estimators are unreliable for recent event dates for which few events are observed, we propose a pragmatic approach to get around this drawbacks. Insurance companies use very specific evaluation dates when calculating reserves, such as the end of a quarter, semester or financial year. Typically the calculations are not performed at those exact evaluation dates, but a couple of days later (at the so-called computation date). Accordingly we predict the number of hidden events on August 31, 2004 using data until September 5, 2004. As such, the granular model predicts 2012.7 unreported claims on August 31, 2004, whereas the true number of IBNR claims (based on data until August 31, 2009) was 2049.

Future observation of hidden events Our daily model splits the total IBNR point estimate of 2012.7 claims by future reporting date. Figure 9a shows the estimated number of daily reported claims in September and October, 2004 for accidents that occurred before August 31, 2004. The dashed line in Figure 9a indicates the computation date. We do not make predictions for dates falling before the computation date as this data is observed. The model accurately predicts the low report counts during the weekend. This is the merit of adding the day of the week effect in the reporting exposure model. Also the overall reporting pattern closely matches the observed values. Figure 9b aggregates these daily report counts by month. This figure shows the estimated number of reported claims in the first twelve months following August, 2004. In these months the observed and predicted IBNR counts are very similar.

Evolution of the number of hidden events The primary focus of our granular model is estimating the total IBNR count. The top panel of Figure 10 plots the predicted number of unreported claims on each evaluation date between September, 2003 and August, 2004. Each point estimate is an out-of-time IBNR estimate obtained from the granular model calibrated on the historical data available five days after the corresponding evaluation date. We compare these estimates with the actual number of IBNR claims computed from the data until August 31, 2009. Our model recognizes the trend in IBNR counts with more unreported claims during

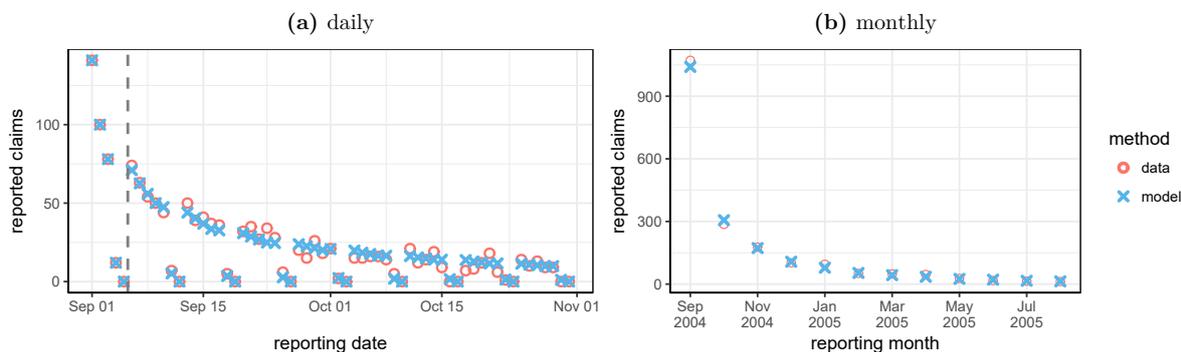


Figure 9: Out-of-time prediction of the number of reported claims for accidents that occurred before August 31, 2004. These predictions are compared with the actual number of reported claims. **(a)** Estimated at a daily level for the next two months. The dashed line indicates the last observed date (September 5, 2004). **(b)** Estimates aggregated by reporting month for the next twelve months.

the summer compared to the winter months. The model also correctly predicts an increase in IBNR claims at the start of the year (here: January 1, 2004) as a result of the holidays in this period. The middle panel of Figure 10 shows the prediction error, i.e. the difference between the predicted number of IBNR claims and the actual count. The prediction error for the granular model is centred around zero and there are no large outliers. The bottom panel of Figure 10 zooms in on the estimates for dates in October and November, 2003. This figure shows that the day of the week parameters allow the model to accurately capture the weekday pattern in IBNR counts.

Benchmark with a model for aggregate data We benchmark our granular approach to Mack’s chain ladder method Mack (1993) on aggregated data, which is the industry standard in claims reserving. This method discretizes time and aggregates the observed events into a two dimensional table based on the occurrence period and the discretized reporting delay. A Poisson generalized linear model (GLM) then models the effect of the occurrence and reporting period on these aggregated records. We investigate two aggregation levels, namely aggregating based on a yearly as well as a 28 day grid. We refer to Huang et al. (2015) for a more detailed discussion on reserving with granular data versus data aggregated in two dimensional tables. Figure 11 shows the estimated IBNR counts under both chain ladder implementations evaluated on each date between September, 2003 and August, 2004. Both versions of the chain ladder detect the seasonal pattern in unreported claim counts, which is related to seasonality in the occurrence process. The end of the year holidays and corresponding increase in IBNR counts is a yearly seasonal effect in the reporting process. The chain ladder assumptions allow for seasonal effects when the period of seasonality coincides with the discretized time periods. For this reason, the yearly chain ladder method correctly predicts an increase in IBNR counts around the end of the

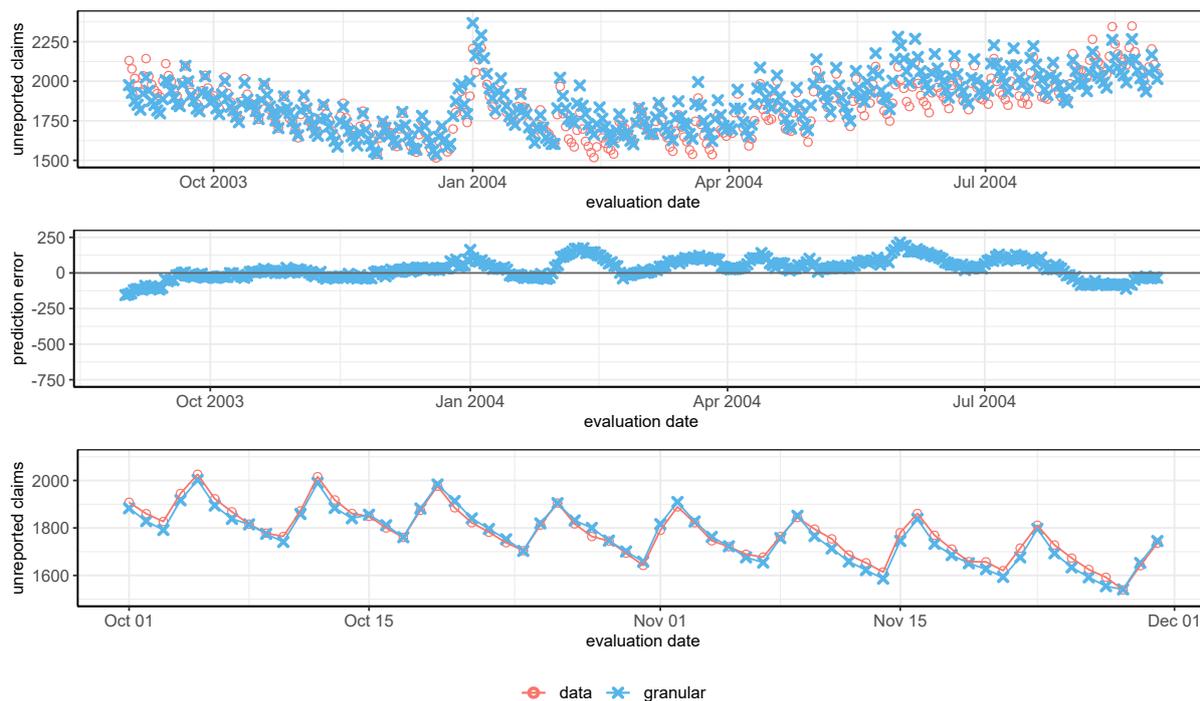


Figure 10: Out-of-time prediction of the total IBNR count by the granular reserving method for each evaluation date between September 2003 and August, 2004. These estimates are compared with the observed values using data until August, 2009. The middle panel shows the difference between the predicted and actual IBNR count. The bottom panel zooms in on the estimates in October and November, 2003.

year, whereas the 28 day chain ladder method severely underestimates IBNR counts for these dates. The bottom panel of Figure 11 zooms in on the period October to November 2003. The 28 day chain ladder method retrieves the day of the week effect, since the length of every bin is a multiple of 7 and therefore contains the same weekdays. The yearly chain ladder method has bins with either 365 or 366 days. Since both bin sizes are not divisible by 7, the yearly chain ladder method is unable to recognize the day of the week effect. This results in a systematic overestimation of IBNR counts on Fridays and an underestimation on Sunday. The middle panel of Figure 11 shows the difference between the predicted and actual IBNR count. The inability of the 28 day chain ladder to capture the holiday effect results in large underestimations around this time of the year. The yearly chain ladder overall performs better, but the prediction error is sensitive to the day of the week on which the reserve is calculated. Capturing the holiday and the day of the week effect simultaneously requires a model specified at the daily level. The chain ladder method assumes independence between the reporting delay distribution and the occurrence period of the claim. Since Figure 5 and 6 indicate that this assumption is not valid at the daily level, a daily chain ladder would not perform well. Our granular method explains both phenomena together by abandoning this independence assumption.

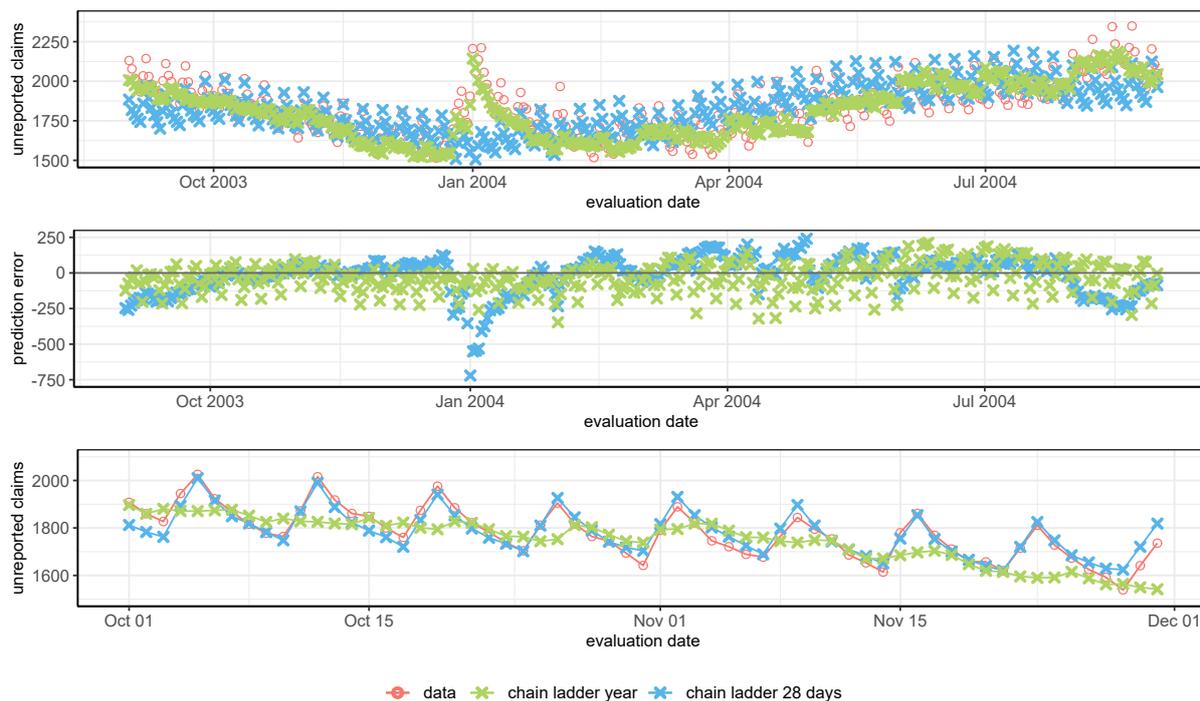


Figure 11: Out-of-time prediction of the total IBNR count by the yearly and 28 day chain ladder methods for each evaluation date between September 2003 and August, 2004. These estimates are compared with the observed values using data until August, 2009. The middle panel shows the difference between the predicted and actual IBNR count. The bottom panel zooms in on the estimates in October and November, 2003.

3.5 Scenario testing

3.5.1 Investigated scenarios

We further evaluate our approach with portfolios simulated along four different scenarios. Each scenario generates data from an insurance portfolio from January 1, 1998 onwards. Figure 12 outlines the structure of these data sets. The insurer observes the claims that are reported before the computation date (the gray area in Figure 12) and predicts the number of claims that were not yet reported on the evaluation date (the hatched area in Figure 12). We consider two evaluation dates (December 31, 2003 and August 31, 2004) to visualize the impact of holidays near the end of the year on the accuracy of IBNR claim count predictions. The four scenarios focus on characteristics of the portfolio or the claim handling process that have an impact on the total IBNR count. Figure 13 visualizes the occurrence, reporting and IBNR processes for a single simulated data set from each of the four scenarios.

Scenario 1: Baseline scenario This is the basic scenario from which the other three scenarios will slightly deviate. The occurrence of insured events follows a Poisson distribution

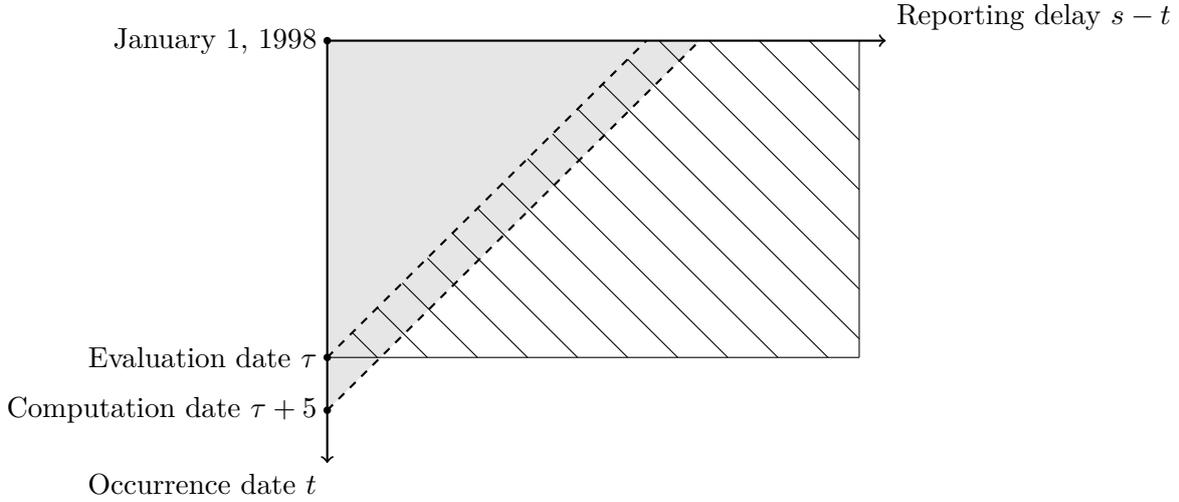


Figure 12: Structure of a simulated data set. We simulate accidents that occur between the first of January, 1998 and the computation date, together with their associated reporting delay. The gray area shows the data that is used to fit the model and to predict the hatched area, which consists of the number of unreported claims at the evaluation date τ . We obtain perfect predictions for the intersection of the gray area and the hatched area, since in this region the reported counts are observed.

with an average of 100 claims on each occurrence date. For these occurrences the reporting delay is simulated along the model specification outlined in Section 2, i.e. the distribution of the time-changed reporting delay \tilde{U} follows a lognormal distribution with density

$$f_{\tilde{U}}(u) = \frac{1}{u\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \cdot \left(\frac{\ln(u)-\mu}{\sigma}\right)^2},$$

where $\mu = 0$ and $\sigma = 1$. The daily reporting exposure depends only on the reporting date and is given by

$$\alpha_{t,s} = 0.10 \cdot (0.20)^{\mathbb{1}_{s \in \text{Sat}} + \mathbb{1}_{s \in \text{unofficial-holiday}}} \cdot (0.01)^{\mathbb{1}_{s \in \text{Sun}} + \mathbb{1}_{s \in \text{national-holiday}}},$$

where **Sat**, **Sun**, **national-holiday** and **unofficial-holiday** are the sets of all Saturdays, Sundays, national holidays and unofficial holidays respectively. As such, the reporting probability is reduced by 80% on Saturdays and unofficial holidays and by 99% on Sundays and national holidays. These effects are of the same order as those found in the exploratory data analysis, see e.g. Figure 5 in Section 3.1 and result in an average reporting delay of slightly more than three weeks. The top row of Figure 13 visualizes a simulation from this baseline scenario. The middle panel shows two regimes of reporting, where the days with few reported claims correspond to the weekend and holidays.

Scenario 2: Volatile occurrences In this scenario external causes, such as the weather, have a large effect on the number of accidents that occur on a given date. The environment

can be in two states, a good state with an average of 100 accidents per day and a bad state in which there are on average 400 accidents. The transitions between these states follow a Markov process with transition matrix

$$\begin{array}{l} \text{from/to} \\ \text{good} \\ \text{bad} \end{array} \begin{array}{cc} \text{good} & \text{bad} \\ \left(\begin{array}{cc} 0.9 & 0.1 \\ 0.6 & 0.4 \end{array} \right) \end{array}.$$

The model starts in the good state and then occasionally moves to the bad state. From this bad state there is a large probability of returning to the good state with less occurrences on average. The second row of Figure 13 (lhs) visualizes the impact of this bad state on the occurrence process. The reporting delay distribution is the one described in the baseline scenario.

Scenario 3: Low claim frequency This scenario illustrates the effect of a strong reduction in the number of occurred accidents. The occurrence process is modeled by a Poisson distribution with a daily average of two claims. The reporting model from the baseline scenario is used. This scenario is visualized in the bottom row of Figure 13. We observe that a low number of accidents leads to more volatility in the IBNR process.

Scenario 4: Online reporting In this scenario the insurer introduces an online tool for claim reporting. This online tool is launched at January 1, 2003 and increases the number of reports in the weekend and on holidays. The new reporting exposures become

$$\alpha_{t,s} = \begin{cases} 0.10 \cdot (0.20)^{\mathbb{1}_{s \in \text{Sat}} + \mathbb{1}_{s \in \text{Unofficial-holiday}}} \cdot (0.01)^{\mathbb{1}_{s \in \text{Sun}} + \mathbb{1}_{s \in \text{Holiday}}} & s < 01/01/2003 \\ 0.10 \cdot (0.50)^{\mathbb{1}_{s \in \text{Sat}} + \mathbb{1}_{s \in \text{Unofficial-holiday}}} \cdot (0.20)^{\mathbb{1}_{s \in \text{Sun}} + \mathbb{1}_{s \in \text{Holiday}}} & s \geq 01/01/2003 \end{cases}.$$

This reporting model is combined with the same occurrence process as in the baseline model, that is a Poisson process with a constant intensity of 100 claims each day. The bottom row of Figure 13 visualizes a simulation from this scenario. A vertical black line indicates the breakpoint on January 1, 2003. After the introduction of online reporting we no longer observe dates with zero reports.

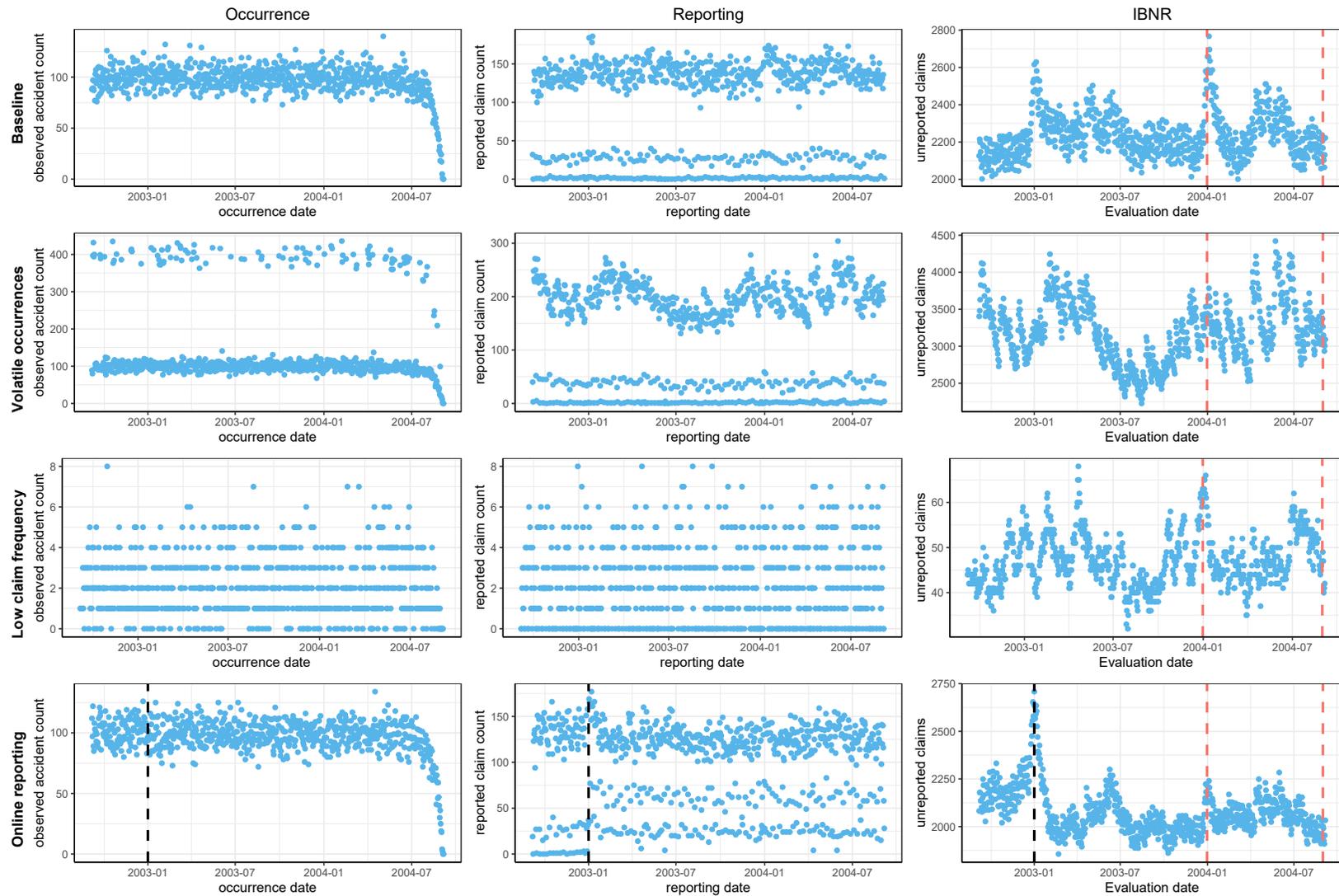


Figure 13: Each row visualizes a simulated data set from one of the four scenarios. The left column shows the daily number of accidents that were reported by August 31, 2004 (cf. Figure 3). The middle column shows the daily number of reported claims (cf. Figure 4). The right column visualizes the number of unreported accidents using a rolling evaluation date (cf. Figure 7). The red dashed lines in the IBNR plots indicate the evaluation dates of December 31, 2003 and August 31, 2004.

3.5.2 Calibrated models: granular versus aggregate

We compare the accuracy of the predictions of the hidden event counts using three models, namely the exact granular model from which we simulated the data, an approximate granular model and a model for yearly aggregated data. The historical information (gray area in Figure 12) is used to predict the number of IBNR claims (hatched area in Figure 12). Under the granular approach these predictions naturally extend to delays beyond those yet observed, whereas in the aggregate approach we limit the prediction window to the longest observed delay. We consider a gap of five days between the computation and the valuation date. The observations from these five days improve the prediction of the occurrence intensities λ_t and the reporting probabilities $p_{t,s}$, whereas there is no straightforward way to incorporate this data in the method for yearly, aggregated data. The ability to use this additional data is one of the advantages of the granular approach.

Exact granular model We use our knowledge of the shape of the distribution and reporting exposure structure behind the various scenarios and calibrate the exact same model for reporting delay on the historical data. Hence we estimate the variance parameter in the lognormal distribution for the smoothed reporting delay \tilde{U} and the parameters γ for the covariate effects in the reporting exposures $\alpha_{t,s}$. The reporting exposure $\alpha_{t,s}$ changes the scale of the time axis which is similar to the effect of the scale parameter $\exp(\mu)$ of the lognormal distribution. We avoid identifiability issues by setting μ equal to zero. The occurrence process is modeled non-parametrically as described in Section 2.

Approximate granular model This model considers the more realistic situation where the insurer wants to fit the model of Section 2, but is unaware of the exact underlying distribution. Motivated by computational benefits the insurer chooses an exponential distribution for the smoothed reporting delay \tilde{U} , and structures the reporting exposures as

$$\begin{aligned} \alpha_{t,s} &= \alpha_s^{\text{dow}} \cdot \alpha_s^{\text{holiday}} \cdot \alpha_{s-t}^{\text{delay}} \\ &= \exp((\mathbf{x}_s^{\text{dow}})' \cdot \gamma^{\text{dow}} + (\mathbf{x}_s^{\text{holiday}})' \cdot \gamma^{\text{holiday}} + (\mathbf{x}_{s-t}^{\text{delay}})' \cdot \gamma^{\text{delay}}). \end{aligned} \quad (11)$$

In this specification α_s^{dow} captures the day of the week effect, $\alpha_s^{\text{holiday}}$ identifies national and unofficial holidays and $\alpha_{s-t}^{\text{delay}}$ adapts reporting exposure based on the time elapsed since the claim occurred. For a single simulated data set we bin reporting delay in 13 bins according to the strategy outlined in online appendix C. These same bins are then reused to construct the delay covariate for all other simulations. In the fourth scenario (online reporting), we estimate different parameter values for the parameters γ^{dow} and γ^{holiday} for reporting dates before and after January 1, 2003.

Scenario	Eval. date	exact granular		approx. granular		chain ladder	
		$\mu(PE)$	$\sigma(PE)$	$\mu(PE)$	$\sigma(PE)$	$\mu(PE)$	$\sigma(PE)$
Baseline	31 Dec 2003	-0.09	3.17	4.85	2.75	2.70	2.17
	31 Aug 2004	-0.01	2.75	-0.18	2.82	1.20	2.36
Volatile occurrences	31 Dec 2003	0.11	2.64	5.01	2.93	0.16	15.52
	31 Aug 2004	-0.04	2.27	-0.20	2.51	-0.82	14.90
Low claim frequency	31 Dec 2003	-0.69	23.89	4.42	20.85	1.65	16.25
	31 Aug 2004	-2.30	20.19	-2.52	20.72	-1.33	17.96
Online reporting	31 Dec 2003	-0.13	3.12	2.93	3.07	-12.46	2.91
	31 Aug 2004	0.02	2.80	0.73	2.89	-7.00	2.68

Table 1: Evaluation of the performance of the exact granular model, the approximate granular model and the chain ladder method across four different scenarios and two evaluation dates.

A model for aggregated data: the chain ladder The chain ladder method described in Section 3.4 is the industry standard for predicting the number of unreported claims. We aggregate the simulated data by calendar year and benchmark our granular approach to the chain ladder method on this aggregated data.

3.5.3 Results and discussion

We evaluate the performance of the reserving models by predicting the total number of IBNR claims at the evaluation date, which corresponds to the hatched area in Figure 12. This prediction is compared with the actual number of unreported claims as observed in the simulated data set. We simulate 1000 data sets and calibrate the three models outlined in Section 3.5.2 on each of these. The prediction accuracy is measured by the percentage error (PE), i.e.

$$PE = 100 \cdot \frac{N^{IBNR}(\tau) - \widehat{N^{IBNR}}(\tau)}{N^{IBNR}(\tau)}.$$

Positive percentage errors reflect underestimation, whereas negative values indicate an overestimation of IBNR counts. Table 1 shows the mean and standard deviation of the percentage error for the two granular models and the chain ladder method. In Figure 14 boxplots of the percentage error visualize the model performance across the four scenarios.

Impact of evaluation date We observe in all four scenarios an increase in unreported claims on New Year’s Eve (see the last column in Figure 13). This is the result of multiple holidays at the end of the year, which prevents clients from reporting their claim. We compare the average percentage error in Table 1 on December 31, 2003 and August 31, 2004 to quantify the impact of these holidays on prediction accuracy. The exact granular model fits the distributional specification that was used in the simulation. Therefore this model can perfectly capture the

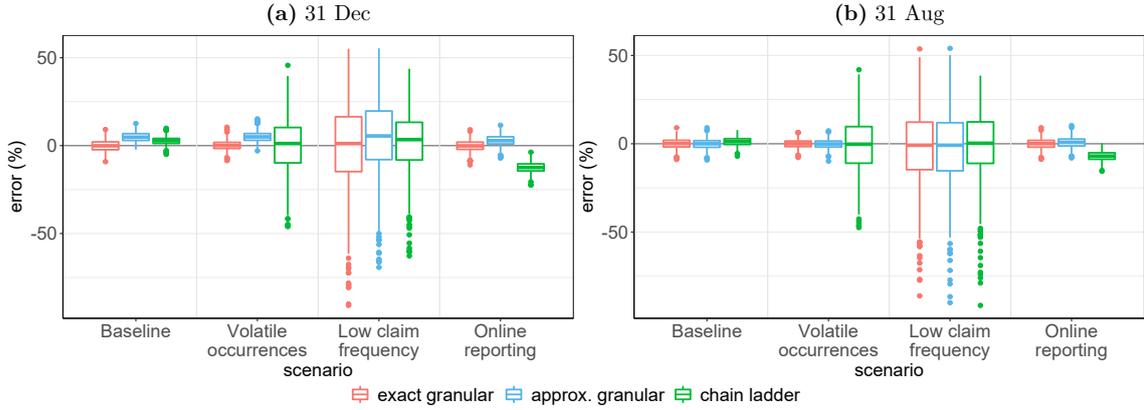


Figure 14: Boxplots of the Percentage Error (PE) of the IBNR estimate across the four scenarios and on both evaluation dates.

effect of holidays and has an average error close to zero on both dates. Seasonal effects do not violate the chain ladder assumptions when their seasonal cycle coincides with the chain ladder period. Since the end of the year holidays can be seen as a yearly seasonal event they do not affect the prediction accuracy in the yearly chain ladder method. This explains the fairly similar errors on both evaluation dates for the chain ladder method. Table 1 reveals an underestimation of IBNR counts for the approximate granular model on December 31 across all four scenarios. The data is simulated with a lognormal distribution for the smoothed reporting delay, whereas in the approximate granular model we fit an exponential distribution. Since these distributions are quite different, we include a delay effect $\alpha_{s-t}^{\text{delay}}$ in (11). This effect can increase the reporting probability at specific delays, hereby moving the time-changed data closer to an exponential distribution. However, the delay covariate can not remove all differences between these distributions and this leads to a small underestimation on December 31, 2004 in all scenarios. For all three models the choice of evaluation date does not influence the standard deviation of the percentage error.

Baseline The top row of Figure 13 visualizes a single data set from the baseline scenario. Both the occurrence and reporting process are stable. This leads to a yearly periodical pattern in IBNR counts, which is easy to predict. Since all three models perform well (see Figure 14), there is no reason to replace the chain ladder method by a granular model in this scenario.

Volatile occurrences The range of IBNR values encountered throughout a year is much wider in this scenario compared to the other three scenarios. Table 1 and Figure 14 show that the performance of the granular models is in line with their performance in the baseline scenario. The occurrence process has little effect on the prediction accuracy, since we model the occurrence process non-parametrically. The chain ladder method performs well on average, but the standard deviation has risen compared to the baseline scenario. In over half of the cases the

chain ladder produces an error of more than 10% when predicting the number of unreported claims. The chain ladder method aggregates claims by occurrence year, hereby losing the exact occurrence information. When the model was in the bad state on the evaluation date, this leads to large underestimations of total IBNR counts. This scenario identifies an unstable accident occurrence process as a reason for considering a granular model.

Low claim frequency The occurrence frequency is reduced from an average of hundred daily claims to only two claims. The third row of Figure 13 visualizes a data set from this scenario. Since on average only two accidents occur per day, our predictions for the intensities λ_t in the occurrence process are less reliable. As seen in Figure 14 this leads to large prediction errors for all models. This uncertainty follows mostly from the Poisson assumption (A1) in the data generation process. The coefficient of variation $\frac{\sigma}{\mu}$ for a Poisson distribution with intensity λ is given by $\frac{1}{\sqrt{\lambda}}$. A lower intensity in the Poisson process corresponds with a larger coefficient of variation and thus more uncertainty in the data. We conclude that accurate estimation of the number of hidden events is only possible when the expected number of events is sufficiently large.

Online reporting On January 1, 2003 the insurer introduces an online tool to report claims, which creates a breakpoint in the reporting process. The granular model performs well on both evaluation dates, since we estimate different exposure parameters after the breakpoint. Both evaluation dates correspond with around one year of post breakpoint data, which is insufficient for applying the chain ladder method. Therefore, we calibrate the chain ladder method on all the available data, which leads to an overestimation of the IBNR counts. This scenario illustrates the benefits of a granular reserving model, when breakpoints can be identified in the data.

4 Conclusion

We propose a new method to model the number of events that occurred in the past, but which are not yet registered due to an observation delay. Our approach provides an elegant and flexible framework for modeling the observation delay subject to calendar day covariates by introducing the concept of observation exposure. This framework can be applied for predicting the future cost of warranties, pricing maintenance contracts and many other applications in operational research where events are observed with a delay. We illustrate our method in an extensive insurance case-study. Compared to methods designed for aggregated data our granular approach has three advantages. First of all, introducing covariates gives insight into the observation process. Second, our granular model can predict the expected number of observations for each future date. This enables the detection of changes in the reporting process in a fast way. Third, by introducing covariates the predictive performance is less sensitive to the chosen evaluation date.

The simulation study further identifies a volatile occurrence process and breakpoints in the event observation process as important arguments for choosing a data driven, granular model as developed in this paper.

5 Acknowledgments

The authors thank the anonymous referees and the associate editor for useful comments which led to significant improvements of the paper. This work was supported by the Argenta Research chair at KU Leuven; KU Leuven's research council [project COMPACT C24/15/001]; the agency for Innovation by Science and Technology (IWT) [grant number 131173]; and Research Foundation Flanders (FWO) [grant number 11G4619N].

References

- Akbarov, A. and Wu, S. (2012). Warranty claims data analysis considering sales delay. *Quality and Reliability Engineering International*, 29, 113–123. <https://doi.org/10.1002/qre.1302>.
- Antonio, K. and Plat, R. (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, 7, 649–669. <https://doi.org/10.1080/03461238.2012.755938>.
- Apeland, S. and Scarf, P. (2003). A fully subjective approach to modelling inspection maintenance. *European Journal of Operational Research*, 148, 410–425. [https://doi.org/10.1016/S0377-2217\(02\)00356-9](https://doi.org/10.1016/S0377-2217(02)00356-9).
- Avanzi, B., Wong, B., and Yang, X. (2016). A micro-level claim count model with overdispersion and reporting delays. *Insurance: Mathematics and Economics*, 71, 1–14. <https://doi.org/10.1016/j.insmatheco.2016.07.002>.
- Badescu, A. L., Lin, X. S., and Tang, D. (2016). A marked Cox model for the number of IBNR claims: Theory. *Insurance: Mathematics and Economics*, 69, 29–37. <https://doi.org/10.1016/j.insmatheco.2016.03.016>.
- Baler, R. and Wang, W. (1993). Developing and testing the delay-time model. *The Journal of the Operational Research Society*, 44, 361–374. <https://doi.org/10.2307/2584414>.
- Berrade, M., Scarf, P., and Cavalcante, C. (2018). Conditional inspection and maintenance of a system with two interacting components. *European Journal of Operational Research*, 268, 533–544. <https://doi.org/10.1016/j.ejor.2018.01.042>.
- Bonetti, M., Cirillo, P., Musile Tanzi, P., and Trincherò, E. (2016). An analysis of the number of medical malpractice claims and their amounts. *PLOS ONE*, 11(4), 1–30. <https://doi.org/10.1371/journal.pone.0153362>.
- Christer, A. (1973). Innovatory decision making. In White, D. and Brow, K., editors, *The Role and Effectiveness of Theories of Decision in Practice*, pages 369–377. Hodder and Stoughton, London.

- Godecharle, E. and Antonio, K. (2015). Reserving by conditioning on markers of individual claims: a case study using historical simulation. *North American Actuarial Journal*, 19(4), 273–288. <https://doi.org/10.1080/10920277.2015.1046607>.
- Harris, J. E. (1990). Reporting delays and the incidence of AIDS. *Journal of the American Statistical Association*, 85(412), 915–924. <https://doi.org/10.2307/2289588>.
- Huang, J., Qiu, C., Wu, X., and Zhou, X. (2015). An individual loss reserving model with independent reporting and settlement. *Insurance: Mathematics and Economics*, 64, 232 – 245. <https://doi.org/10.1016/j.insmatheco.2015.05.010>.
- Jewell, W. S. (1990). Predicting IBNYR events and delays. *ASTIN Bulletin*, 20, 93–111. <https://doi.org/10.2143/AST.19.1.2014914>.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481. <https://doi.org/10.1080/01621459.1958.10501452>.
- Kingman, J. F. C. (1993). *Poisson Processes*. Oxford Studies in Probability. Oxford University Press.
- Kuang, D., Nielsen, B., and Nielsen, J. P. (2008). Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika*, 95(4), 979–986. <https://doi.org/10.1093/biomet/asn026>.
- Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin*, 23, 213–225. <https://doi.org/10.2143/AST.23.2.2005092>.
- Mortenson, M. J., Doherty, N. F., and Robinson, S. (2015). Operational research from taylorism to terabytes: A research agenda for the analytics age. *European Journal of Operational Research*, 241(3), 583 – 595. <https://doi.org/10.1016/j.ejor.2014.08.029>.
- Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance. *ASTIN Bulletin*, 23(1), 95–115. <https://doi.org/10.2143/AST.23.1.2005103>.
- Noufaily, A., Farrington, P., Garthwaite, P., Enki, D. G., Andrews, N., and Charlett, A. (2016). Detection of infectious disease outbreaks from laboratory data with reporting delays. *Journal of the American Statistical Association*, 111(514), 488–499. <https://doi.org/10.1080/01621459.2015.1119047>.
- Pigeon, M., Antonio, K., and Denuit, M. (2013). Individual loss reserving with the multivariate skew normal distribution. *ASTIN Bulletin*, 43, 399–428. <https://doi.org/10.1017/asb.2013.20>.
- Pigeon, M., Antonio, K., and Denuit, M. (2014). Individual loss reserving using paid–incurred data. *Insurance: Mathematics and Economics*, 58, 121–131. <https://doi.org/10.1016/j.insmatheco.2014.06.012>.
- Salmon, M., Schumacher, D., Stark, K., and Höhle, M. (2015). Bayesian outbreak detection in the presence of reporting delays. *Biometrical Journal*, 57, 1051–1067. <https://doi.org/10.1002/bimj.201400159>.
- Swishchuk, A. (2016). *Change of time methods in quantitative finance*. SpringerBriefs in Mathematics. Springer.

- Taylor, G. (2000). *Loss reserving: an actuarial perspective*. Kluwer Academic Publishers.
- Verbelen, R., Antonio, K., Claeskens, G., and Crevecoeur, J. (2017). Predicting daily IBNR claim counts using a regression approach for the occurrence of claims and their reporting delay. Working paper. Available at <https://lirias.kuleuven.be/handle/123456789/580750>.
- Verrall, R. J. and Wüthrich, M. V. (2016). Understanding reporting delay in general insurance. *Risks*, 4(3). <https://doi.org/10.3390/risks4030025>.
- Wang, W. (1997). Subjective estimation of the delay time distribution in maintenance modelling. *European Journal of Operational Research*, 99, 516–529. [https://doi.org/10.1016/S0377-2217\(96\)00318-9](https://doi.org/10.1016/S0377-2217(96)00318-9).
- Wang, W. (2010). A model for maintenance service contract design, negotiation and optimization. *European Journal of Operational Research*, 201, 239–246. <https://doi.org/10.1016/j.ejor.2009.02.018>.
- Wüthrich, M. V. and Merz, M. (2008). *Stochastic claims reserving methods in insurance*, volume 435 of *Wiley Finance*. John Wiley & Sons.
- Wüthrich, M. V. and Merz, M. (2015). *Stochastic Claims Reserving Manual: Advances in Dynamic Modeling*. Available at SSRN: <https://ssrn.com/abstract=2649057>.
- Ye, Z. and Ng, H. K. T. (2014). On analysis of incomplete field failure data. *The Annals of Applied Statistics*, 8, 1713–1727. <https://doi.org/10.1214/14-AOAS752>.

Supplementary material for
 “Modeling the number of hidden events subject to observation
 delay”

Jonas Crevecoeur^{1,3,*}, Katrien Antonio^{1,2,3,4} and Roel Verbelen^{1,3,4}

¹Faculty of Economics and Business, KU Leuven, Belgium.

²Faculty of Economics and Business, University of Amsterdam, The Netherlands.

³LRisk, Leuven Research Center on Insurance and Financial Risk Analysis, KU Leuven, Belgium.

⁴LStat, Leuven Statistics Research Center, KU Leuven, Belgium.

*Corresponding author. E-mail: jonas.crevecoeur@kuleuven.be

March 27, 2019

A Maximum likelihood estimation of observation exposure parameters

We model a parameter vector γ which structures the observation exposures.

$$\begin{aligned} \ell(\gamma; \boldsymbol{\chi}) &= \sum_{t=1}^{\tau} \sum_{s=t}^{\tau} N_{t,s} \cdot \log(p_{t,s}) - \sum_{t=1}^{\tau} N_t^{\text{R}}(\tau) \cdot \log(p_t^{\text{R}}(\tau)) \\ &= \sum_{t=1}^{\tau} \sum_{s=t}^{\tau} N_{t,s} \cdot \log(F_{\tilde{U}}(\varphi_t(s-t+1)) - F_{\tilde{U}}(\varphi_t(s-t))) \\ &\quad - \sum_{t=1}^{\tau} N_t^{\text{R}}(\tau) \cdot \log(F_{\tilde{U}}(\varphi_t(\tau-t+1))), \end{aligned} \tag{12}$$

where

$$\varphi_t(d) = \sum_{v=t}^{t+d-1} \exp(\boldsymbol{x}'_{t,v} \boldsymbol{\gamma}).$$

No analytical solution exists for the optimal parameters γ and numerical optimization is required. We use the Newton-Raphson algorithm to maximize the likelihood (12). The Newton-Raphson algorithm updates the parameter estimates iteratively as follows

$$\hat{\boldsymbol{\gamma}}^{(k+1)} = \hat{\boldsymbol{\gamma}}^{(k)} - \boldsymbol{H}^{-1}(\hat{\boldsymbol{\gamma}}^{(k)}) \cdot \boldsymbol{S}(\hat{\boldsymbol{\gamma}}^{(k)}). \tag{13}$$

In this formula \mathbf{S} denotes the score vector and \mathbf{H} is the Hessian of the loglikelihood in (12), i.e. the vector of first order and the matrix of second order partial derivatives respectively. Below we derive the expression for the first and second order derivatives of the loglikelihood when $F_{\tilde{U}}$ is a known twice continuously differentiable distribution function. The components of the score vector \mathbf{S} are

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\gamma}, \boldsymbol{\xi}; \boldsymbol{\chi})}{\partial \gamma_i} &= \sum_{t=1}^{\tau} \sum_{s=t}^{\tau} \frac{N_{t,s}}{p_{t,s}} \cdot \left[f_{\tilde{U}}(\varphi_t(s-t+1)) \cdot \frac{\partial \varphi_t}{\partial \gamma_i}(s-t+1) - f_{\tilde{U}}(\varphi_t(s-t)) \cdot \frac{\partial \varphi_t}{\partial \gamma_i}(s-t) \right] \\ &\quad - \sum_{t=1}^{\tau} \frac{N_t^R(\tau)}{p_t^R(\tau)} \cdot f_{\tilde{U}}(\varphi_t(\tau-t+1)) \cdot \frac{\partial \varphi_t}{\partial \gamma_i}(\tau-t+1), \end{aligned}$$

where $f_{\tilde{U}}(\cdot)$ denotes the density function of $F_{\tilde{U}}(\cdot)$ and

$$\begin{aligned} p_{t,s} &= F_{\tilde{U}}(\varphi_t(s-t+1)) - F_{\tilde{U}}(\varphi_t(s-t)) \\ p_{t,s}^R(\tau) &= F_{\tilde{U}}(\varphi_t(\tau-t+1)). \end{aligned}$$

The derivatives of the time change operator φ_t with respect to $\boldsymbol{\gamma}$ are

$$\frac{\partial}{\partial \gamma_i} \varphi_t(s-t+1) = \sum_{v=t}^s x_{t,v,i} \cdot \alpha_{t,v}$$

where $x_{t,s,i}$ is the covariate value of the i -th parameter for reporting on date s for a claim that occurred on date t . The Hessian \mathbf{H} is given by

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\gamma}; \boldsymbol{\chi})}{\partial \gamma_i \partial \gamma_j} &= \sum_{t=1}^{\tau} \sum_{s=t}^{\tau} \frac{N_{t,s}}{p_{t,s}} \cdot \left[f'_{\tilde{U}}(\varphi_t(s-t+1)) \cdot \frac{\partial \varphi_t}{\partial \gamma_i}(s-t+1) \cdot \frac{\partial \varphi_t}{\partial \gamma_j}(s-t+1) \right. \\ &\quad - f'_{\tilde{U}}(\varphi_t(s-t)) \cdot \frac{\partial \varphi_t}{\partial \gamma_i}(s-t) \cdot \frac{\partial \varphi_t}{\partial \gamma_j}(s-t) \\ &\quad \left. + f_{\tilde{U}}(\varphi_t(s-t+1)) \cdot \frac{\partial \varphi_t}{\partial \gamma_i \partial \gamma_j}(s-t+1) - f_{\tilde{U}}(\varphi_t(s-t)) \cdot \frac{\partial \varphi_t}{\partial \gamma_i \partial \gamma_j}(s-t) \right] \\ &\quad - \sum_{t=1}^{\tau} \sum_{s=t}^{\tau} \frac{N_{t,s}}{p_{t,s}^2} \cdot \left[f_{\tilde{U}}(\varphi_t(s-t+1))^2 \cdot \frac{\partial \varphi_t}{\partial \gamma_i}(s-t+1) \cdot \frac{\partial \varphi_t}{\partial \gamma_j}(s-t+1) \right. \\ &\quad + f_{\tilde{U}}(\varphi_t(s-t))^2 \cdot \frac{\partial \varphi_t}{\partial \gamma_i}(s-t) \cdot \frac{\partial \varphi_t}{\partial \gamma_j}(s-t) \\ &\quad - f_{\tilde{U}}(\varphi_t(s-t+1)) \cdot f_{\tilde{U}}(\varphi_t(s-t)) \cdot \frac{\partial \varphi_t}{\partial \gamma_i}(s-t+1) \cdot \frac{\partial \varphi_t}{\partial \gamma_j}(s-t) \\ &\quad \left. - f_{\tilde{U}}(\varphi_t(s-t+1)) \cdot f_{\tilde{U}}(\varphi_t(s-t)) \cdot \frac{\partial \varphi_t}{\partial \gamma_i}(s-t) \cdot \frac{\partial \varphi_t}{\partial \gamma_j}(s-t+1) \right] \\ &\quad - \sum_{t=1}^{\tau} \frac{N_t^R(\tau)}{p_t^R(\tau)} \cdot \left[f'_{\tilde{U}}(\varphi_t(\tau-t+1)) \cdot \frac{\partial \varphi_t}{\partial \gamma_i}(\tau-t+1) \cdot \frac{\partial \varphi_t}{\partial \gamma_j}(\tau-t+1) \right] \end{aligned}$$

$$+ f_{\tilde{U}}(\varphi_t(\tau - t + 1)) \cdot \frac{\partial \varphi_t}{\partial \gamma_i \partial \gamma_j}(\tau - t + 1) \Bigg] \\ + \sum_{t=1}^{\tau} \frac{N_t^R(\tau)}{p_t^R(\tau)^2} \cdot f_{\tilde{U}}(\varphi_t(\tau - t + 1))^2 \cdot \frac{\partial \varphi_t}{\partial \gamma_i}(\tau - t + 1) \cdot \frac{\partial \varphi_t}{\partial \gamma_j}(\tau - t + 1),$$

where the second order derivatives of φ_t with respect to γ are

$$\frac{\partial}{\partial \gamma_i \partial \gamma_j} \varphi_t(s - t + 1) = \sum_{v=t}^s x_{t,v,i} \cdot x_{t,v,j} \cdot \alpha_{t,v}$$

The Newton-Raphson algorithm in (13) models the observation exposure parameters γ . Together with the observation parameters, the simulation study of Section 3.5 estimates the variance parameter σ in the lognormal time-changed distribution. The Newton-Raphson algorithm in (13) can easily be extended to this case, where the distribution function of $F_{\tilde{U}}$ depends on parameters.

B Simulation procedure

We outline the algorithm that was used to generate data sets from the four scenarios specified in Section 3.5.1. This algorithm combines a model for the occurrence of events with a model for the observation delay as described in Section 2. We divide the algorithm in three steps.

Step 1. Occurrence We first generate the number of occurred events. The number of daily events follows a Poisson distribution

$$N_t \sim \text{Poisson}(\lambda_t),$$

where the intensity λ_t is obtained from the occurrence process specification for the scenarios in Section 3.5.

Step 2. Observation We now simulate the observation date for each occurred event. Combining equation (6) and (7), we can write the probability that an event from date t is observed on date s as

$$p_{t,s} = P \left(\tilde{U} \in \left[\sum_{v=t}^{s-1} \alpha_{t,v}, \sum_{v=t}^s \alpha_{t,v} \right) \right).$$

We define the observation date random variable

$$S_t = \min_s \left\{ s \in \mathbb{N} \mid \sum_{v=t}^s \alpha_{t,v} > \tilde{U} \right\}. \quad (14)$$

This expression transforms the time-changed observation delay random variable into the associated observation date. Consequently S_t satisfies $P(S_t = s) = p_{t,s}$. For each event that occurred on date t we generate a realization from the distribution of \tilde{U} . We obtain the corresponding observation date by replacing the random variable \tilde{U} in (14) by this sampled value.

Step 3. Truncation With steps 1 and 2 we have simulated an observation date for each occurred event. We split this data set into observed and hidden events. We use the data set with observed events to calibrate the model and to predict the number of hidden events. The hidden events are kept only for evaluating the prediction accuracy.

C A standard distribution for the time changed observation delay

Modeling the time-changed observation delay with an exponential distribution has significant computational benefits. Therefore, this section puts focus on the use of the exponential distribution as a standard distribution for modeling the time-changed observation delay \tilde{U} . Since the exponential distribution is light-tailed it is less suited for long or heavy-tailed delays. We outline a strategy for addressing this weakness of the exponential distribution.

Our strategy bins the possible observation delays ($s-t = 0, 1, \dots$) and categorizes these bins with a delay covariate x_{s-t}^{delay} . This covariate is then included in the observation exposure specification. For each bin we estimate a parameter to capture its effect on observation exposure. These parameters can strongly reshape the distribution, hereby overcoming many of the disadvantages of the exponential distribution. We present a maximum likelihood driven binning strategy in Appendix C.1 and then Appendix C.2 derives the same bins by linking our approach to the non-parametric Kaplan-Meier estimator (Kaplan and Meier, 1958).

C.1 Binning observation delay

Our binning strategy maximizes the loglikelihood in (8) when the observation exposures depend only on the time elapsed since the event occurred, i.e.

$$\alpha_{t,s} = \exp(\gamma^{\text{delay}} \cdot x_{s-t}^{\text{delay}}) = \exp(\gamma^{s-t}),$$

where we estimate for each delay $s-t$ a separate parameter γ^{s-t} . Furthermore we neglect the last term in (8), capturing the effect of the right truncation. Under these restrictions, the

loglikelihood to optimize is

$$\ell(\boldsymbol{\gamma}; \boldsymbol{\chi}) = - \sum_{t=1}^{\tau} \sum_{v=t}^{\tau-1} \left(\sum_{s=v+1}^{\tau} N_{t,s} \right) \cdot \exp(\gamma^{v-t}) + \sum_{t=1}^{\tau} \sum_{s=t}^{\tau} N_{t,s} \cdot \log(1 - \exp(-\exp(\gamma^{s-t})))$$

We compute the derivatives of $\ell(\boldsymbol{\gamma}; \boldsymbol{\chi})$ with respect to the observation exposure parameter γ^d for positive delays $d \in \mathbb{N}$

$$\frac{\partial \ell(\boldsymbol{\gamma}; \boldsymbol{\chi})}{\partial \gamma^d} = - \exp(\gamma^d) \cdot \sum_{t=1}^{\tau-d-1} \sum_{s=t+d+1}^{\tau} N_{t,s} + \frac{\exp(\gamma^d)}{\exp(\exp(\gamma^d)) - 1} \cdot \sum_{t=1}^{\tau-d} N_{t,t+d}.$$

Both sums in this expression have a logical interpretation. The first sum ($\sum_{t=1}^{\tau-1-d} \sum_{s=d+t+1}^{\tau} N_{t,s}$) counts the number of observed events with a delay longer than d days, whereas the second sum ($\sum_{t=1}^{\tau-d} N_{t,t+d}$) counts all events with a delay of exactly d days. These derivatives are zero when

$$\exp(\gamma^d) = - \log \left(1 - \frac{|\text{delay} = d|}{|\text{delay} \geq d|} \right), \quad (15)$$

where $|\text{delay} = d|$ denotes the number of events observed with a delay of d days and $|\text{delay} > d|$ the number of events with a delay of more than d days.

We propose to bin the observation delay by grouping delays for which (15) is approximately constant. Figure 15 visualizes this approach for the liability insurance data set discussed in Section 3. This figure shows in red the estimated delay parameters using approximation (15). The top panel shows the estimates for delays up to 31 days, whereas the parameters for larger delays (up to 400 days) are shown in the bottom panel. Based on this knowledge observation delay is grouped in 23 bins, separated by vertical gray bars in Figure 15. We use more bins for short delays, since for these delays (15) differs strongly. Moreover, many accidents have a short observation delay, which makes these first delays more important. As expected, this binning strategy identifies an increase in observation probability after exactly one year. In Section 3 we structure these bins in a categorical delay covariate x_{s-t}^{delay} and estimate observation delay in a maximum likelihood framework. In Figure 15 the fitted parameters are plotted in blue. These parameters deviate from those found using approximation (15), since other covariate effects were estimated simultaneously. However, the maximum likelihood estimates are close to the approximate values which makes this approximation suitable for choosing initial values in the calibration.

C.2 A link with the Kaplan-Meier estimator

We show that under the binning strategy of Appendix C.1 the time changed model has the same flexibility as the Kaplan-Meier estimator and is as such suitable for modelling a wide range of portfolios.

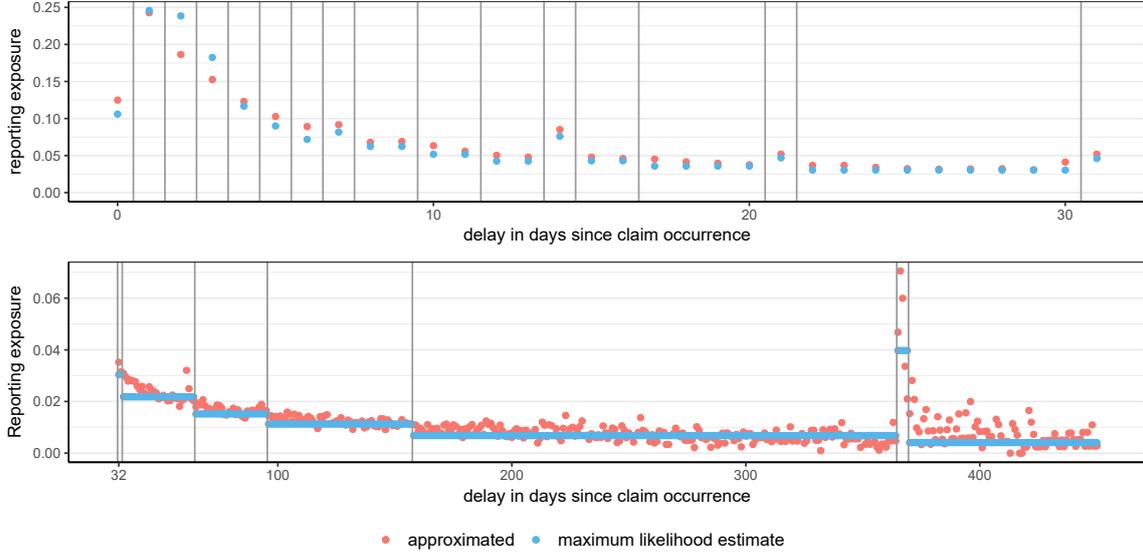


Figure 15: Observation exposure estimates for the delay effect during the first month after the accident occurrence (top) and longer delays (bottom). In red, we show estimates obtained for each delay using (15). The vertical lines indicate the chosen bins. Maximum likelihood estimates for the observation delay parameter corresponding to each bin in the regression structure proposed in Section 3.2 are plotted in blue.

The Kaplan-Meier estimator for the survival function of the observation delay random variable is

$$P(\widehat{\text{delay}} > d) = \prod_{i=0}^d \left(1 - \frac{|\text{delay} = i|}{|\text{delay} \geq i|} \right), \quad (16)$$

When we model the time-changed observation delay distribution \tilde{U} using an exponential distribution then the survival probability for an event from occurrence day t is

$$\begin{aligned} P(\text{delay} > d \mid \text{occ. day} = t) &= P(\tilde{U} \geq \varphi_t(d+1)) \\ &= 1 - F_{\tilde{U}} \left(\sum_{i=1}^{d+1} \alpha_{t,t+i-1} \right) \\ &= \prod_{i=0}^d \exp(-\alpha_{t,t+i}). \end{aligned} \quad (17)$$

Notice the similarity between this expression and the Kaplan-Meier estimator in (16). When the observation exposure only depends on the time passed since the occurrence of the event, i.e. $\alpha_{t,t+i} := \alpha_i$, then

$$P(\text{delay} > d) = \prod_{i=0}^d \exp(-\alpha_i),$$

where α_i is the observation exposure at delay i . This expression no longer depends on the

occurrence date t of the event. The Kaplan-Meier estimator is retrieved when

$$\alpha_i = -\log\left(1 - \frac{|\mathbf{delay} = i|}{|\mathbf{delay} \geq i|}\right). \quad (18)$$

Since $\alpha_i = \exp(\gamma^i)$, this is the same estimator we found in (15) through maximum likelihood estimation. This show that by estimating a separate delay parameter for each delay ($d = 0, 1, \dots$) we obtain a model with the same flexibility as the non-parametric Kaplan-Meier estimator.