A Bayesian framework for addressing bias in delayed cybersecurity breach data

Fabio Viviano

Joint work with Marco Pirra & Sofia Sarubbo

Department of Economics, Statistics and Finance

University of Calabria - ITALY

E-mail: *fabio.viviano@unical.it*

UNIVERSITÀ DELLA CALABRIA

Workshop PRIN 2022

Building resilience to emerging risks in financial and insurance markets

Diamante (CS), June 12 – 13, 2025

Introduction

Understanding and quantifying **cyber risk** has become a **critical priority** for researchers, insurers, and policymakers.

$\Downarrow$

Data breaches

$\Downarrow$

Highly sensitive information

$\Downarrow$

Financial & reputational threats

Empirical studies have employed a variety of statistical methods to capture breach **frequency** and **severity**.

- Sun *et al.* (2020): hurdle models to account for zero-inflated count data;
- Edwards *et al.* (2016) and Wheatley *et al.* (2019): log-normal or heavy-tailed distributions for breach sizes;
- Li & Mamon (2023): Markov-modulated processes $\longrightarrow$ Health-related data breaches;
- McLeod & Dolezel (2018) and Hu *et al.* (2022): state-level frequency-severity models.

Motivations

The reporting delay in data breach incidents poses a significant challenge for **Incurred But Not Reported** (IBNR) studies

$\Downarrow$

Pricing & Reserving

Our aim

Model the **timing** and **reporting** of data breaches.

The idea

Develop a **Hierarchical Bayesian modeling framework** that adjusts for **reporting delays** and decomposes breach counts into interpretable **temporal**, **seasonal**, and **delay**-related components (similarly to Bastos *et al.*, 2019 for epidemiology).

Mathematical Framework

We consider three **Hierarchical Bayesian models** for **delay-adjusted** reporting of cyber breach counts.

Let $n_{t,d}$ be a random variable representing the number of **cases** that **occurred** at time $t = 1, 2, \ldots, T$ but **not reported until** $d = 0, 1, 2, \ldots, D$ time units later.

- T is the last time step for which data is available;
- D is the maximum acceptable delay.

Time \ Delay	1	2	3	4	...	D	n	
1	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$	$n_{1,4}$		$n_{1,D}$	n_1	Observed
2	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$	$n_{2,4}$		$n_{2,D}$	n_2	
...								
$T-D$	$n_{T-D,1}$	$n_{T-D,2}$	$n_{T-D,3}$	$n_{T-D,4}$		$n_{T-D,D}$	n_{T-D}	
$T-D+1$	$n_{T-D+1,1}$	$n_{T-D+1,2}$	$n_{T-D+1,3}$	$n_{T-D+1,4}$		$n_{T-D+1,D}$	n_{T-D+1}	Nowcasting
...								
$T-1$	$n_{T-1,1}$	$n_{T-1,2}$	$n_{T-1,3}$	$n_{T-1,4}$		$n_{T-1,D}$	n_{T-1}	

Mathematical Framework

Model A - Hierarchical Negative Binomial

We assume that $n_{t,d}$ is a **Negative Binomial** random variable, i.e.

$$n_{t,d} \sim \mathsf{NegBin}\left(\lambda_{t,d}, \theta\right), \quad \theta > 0,$$

where θ is the overdispersion parameter, and the mean $\lambda_{t,d}$ has a **log-linear predictor**

$$\log \lambda_{t,d} = \alpha_t + \beta_d + \gamma_{t,d} + \eta_{w(t)}, \tag{1}$$

where α_t, β_d, and $\eta_{w(t)}$ capture respectively **time**, **delay** and **seasonal** effects, while $\gamma_{t,d}$ is a **time-delay** interaction component.

UNIVERSITÀ DELLA CALABRIA

The **random effects** in Equation (1) are modelled as first-order random walks:

$$\alpha_t \sim \mathcal{N}\left(\alpha_{t-1}, \sigma_\alpha^2\right) \quad \text{where} \quad \sigma_\alpha \sim \mathcal{HN}\left(0.1^2\right)$$

$$\beta_d \sim \mathcal{N}\left(\beta_{d-1}, \sigma_\beta^2\right) \quad \text{where} \quad \sigma_\beta \sim \mathcal{HN}\left(1\right)$$

$$\gamma_{t,d} \sim \mathcal{N}\left(\gamma_{t-1,d}\, , \, \sigma_\gamma^2\right) \quad \text{where} \quad \sigma_\gamma \sim \mathcal{HN}\left(0.1^2\right)$$

The **seasonal component** $\eta_{w(t)}$ is modelled as a **Conditional Auto-Regressive** (CAR) model for monthly seasonality:

$$\eta_{w(t)} \sim \text{CAR}_{\text{RW2}}(W = 12, \sigma_\eta^2) \quad \text{where} \quad \sigma_\eta \sim \mathcal{HN}\left(1\right).$$

The **overdispersion parameter**

$$\theta \sim \text{Gamma}(\alpha_\theta, \beta_\theta) \quad \text{where} \quad \alpha_\theta, \beta_\theta \sim \text{Exp}(1).$$

Mathematical Framework

Model B: HNB with multiplicative interaction

We assume that $n_{t,d}$ is a **Negative Binomial** random variable, i.e.

$$n_{t,d} \sim \text{NegBin}\left(\lambda_{t,d}, \theta\right), \quad \theta > 0,$$

where θ is the overdispersion parameter and $\lambda_{t,d}$, in contrast to **Model A**, is defined as

$$\log \lambda_{t,d} = \alpha_t + \beta_d + \gamma_{t,d} + \eta_{w[t]} + \boldsymbol{\alpha_t \cdot \beta_d}.$$

The additional component $\alpha_t \cdot \beta_d$ captures **time-delay interactions.**

Remark

For each component we use the same structure adopted for Model A.

Model C: Hierarchical Zero-Inflated Negative Binomial

UNIVERSITÀ DELLA CALABRIA

We assume that $n_{t,d}$ is a **Zero-Inflated Negative Binomial** random variable, i.e.

$$n_{t,d} \sim \text{ZINB}\left(\lambda_{t,d}, \theta, x_{\text{zinb}}\right)$$

where θ is the dispersion parameter, $1 - x_{\text{zinb}}$ is the zero-inflation probability, and

$$\log \lambda_{t,d} = \alpha_t + \beta_d + \gamma_{t,d} + \eta_{w(t)}.$$

ZINB likelihood:

$$\mathbb{P}\left(n_{t,d} = n\right) = \begin{cases} (1 - x_{\text{zinb}}) + x_{\text{zinb}} \left(\frac{\theta}{\theta + \lambda_{t,d}}\right)^{\theta}, & n = 0 \\ x_{\text{zinb}} \cdot \text{NB}(n; \lambda_{t,d}, \theta), & n > 0. \end{cases}$$

Remark

Same structure as Model A with the addition:

$$x_{\text{zinb}} \sim \text{Beta}(1, 1).$$

Application

Data and calibration

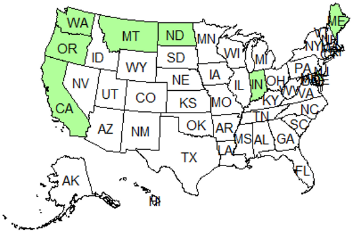
In this work, we exploit **breach data** released by the **U.S. state attorneys**.

Motivations

- Collected under legally mandated and state-specific notification laws $\longrightarrow \uparrow$ **legal consistency**;
- Reports submitted directly by the affected organizations $\longrightarrow \downarrow$ **selection biases**;
- Reporting process is typically **granular** and **timely** (even daily updates on breach occurrence, disclosure dates and number of individuals affected).

These aspects enhance the reliability of longitudinal analyses and support the detection of temporal patterns in breach activity.

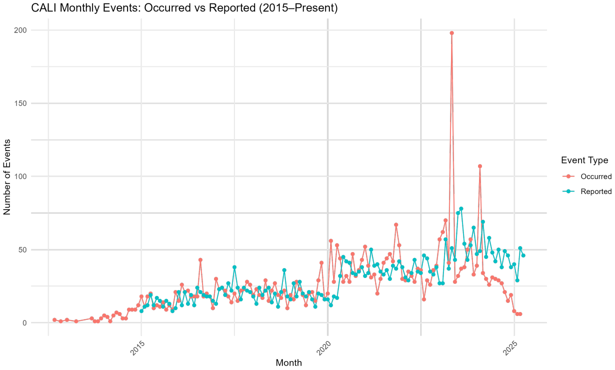
The dataset



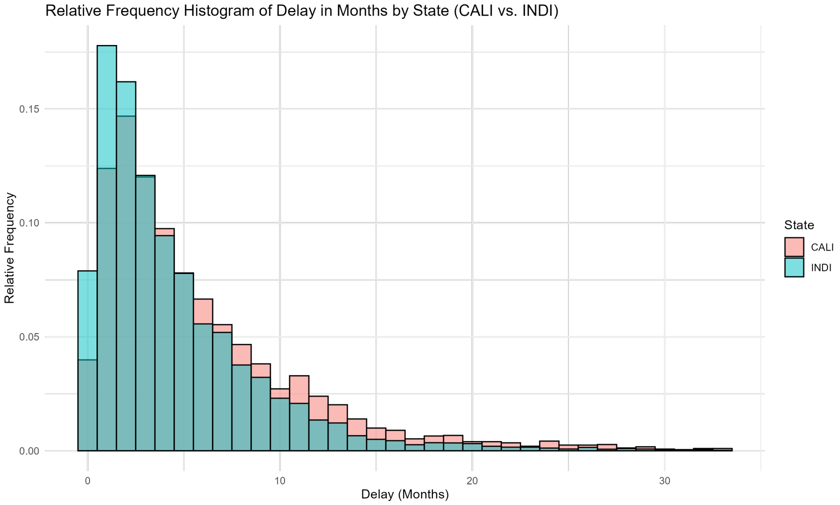
	STATE	NOTIFICATION TO AG	Beg Report	End Report	Beg Occ	End Occ	# Obs	Size
1	CALIFORNIA	January 1, 2012	20/01/2012	19/07/2024	05/07/2007	13/06/2024	4,096	**NO**
2	DELAWARE	April 14, 2018	07/12/2020	26/07/2024	22/02/2019	06/06/2024	280	YES
3	INDIANA	2006	18/12/2013	07/05/2024	01/01/2000	24/04/2024	9,778	YES
4	MAINE	2005	01/12/2012	11/09/2020	22/09/1999	17/08/2020	3,070	YES
5	MONTANA	October 1, 2015	06/05/2015	12/08/2024	01/01/1995	21/07/2024	5,721	YES
6	NORTH DAKOTA	April 13, 2015	02/01/2019	25/07/2022	01/01/2012	28/06/2022	289	YES
7	OREGON	January 1, 2016	30/10/2015	16/08/2024	01/04/2008	26/06/2024	1,148	**NO**
8	WASHINGTON	July 24, 2015	11/08/2015	22/07/2024	01/04/2008	13/06/2024	1,356	YES
						total	25,738	

We focus on a high-activity jurisdiction, i.e. **California**, which provides data richness and regulatory relevance.

We use **monthly aggregated breach reports** from 2015 through December 2024.



Delay: California vs Indiana



Posterior distribution estimations implied by **Model A**, **Model B** and **Model C** are obtained through

$$\Downarrow$$

Markov-Chain Monte Carlo sampling

(R packages `nimble` + `doparallel`)

Setup

- Chains: 3
- Burn-in sample: 1×10^6
- Total iterations: 2.5×10^6
- Thinning parameter: 10

$\Longrightarrow$

Computationally intensive

$$\Downarrow$$

Integrated Nested Laplace Approximation

$$\Downarrow$$

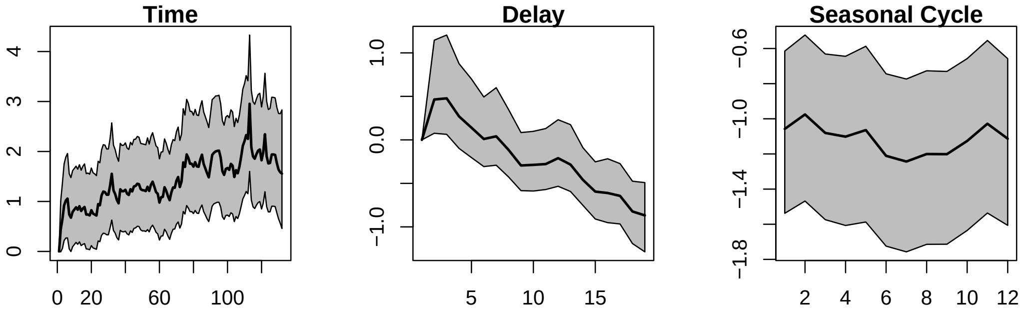
Fast and accurate inference even in high-dimensional settings

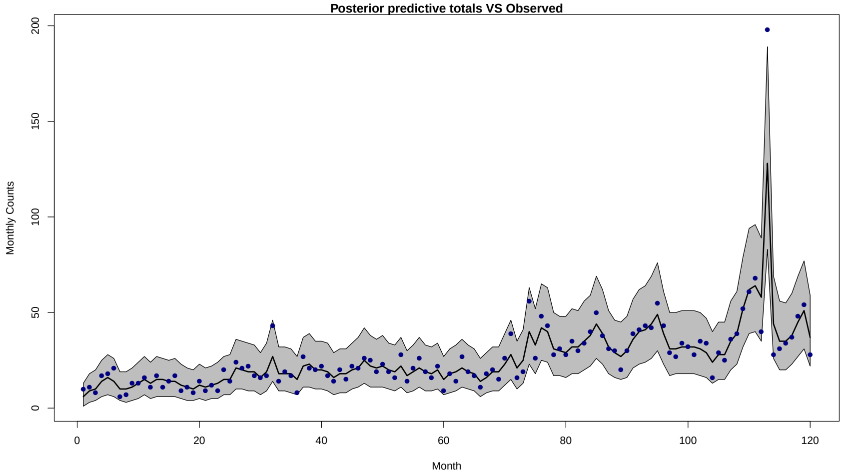
Numerical Results

The analysis

- Model selection → **Goodness of fit metrics**;
- Graphical comparison → **Posterior predictive distributions**;
- **IBNR estimates** ⟶ Comparison wrt Chain-Ladder method.

	WAIC	RMSE	MAE	Coverage	Rhat	ESS
Model A	3326.13	1.410	0.817	98.94%	1.009	2620
Model B	3270.93	1.336	0.774	99.35%	1.002	3891
Model C	3328.20	1.410	0.817	99.02%	1.003	2680


Time
0
1
2
3
4
0
20
60
100
Delay
1.0
0.0
−1.0
5
10
15
Seasonal Cycle
−0.6
−1.0
−1.4
−1.8
2
4
6
8
10
12


Posterior predictive totals VS Observed
Monthly Counts
0
50
100
150
200
0
20
40
60
80
100
120
Month

IBNR Predictions	MAE	RMSE
Chain Ladder	8.709	13.726
Model A	1.373	5.554
Model B	1.431	5.479
Model C	1.353	5.477

Conclusion

This work introduces a **Hierarchical Bayesian model** ⟶ **IBNR cyber incidents**.

Advantages

- Breach counts decomposed into **temporal**, **seasonal** and **delay**-adjusted components;
- High predictive **accuracy**;
- **Outperform** traditional methods (e.g., the Chain-Ladder approach).

Limitation

- MCMC is computationally intensive ⟶ **INLA**.

Future extension

- Modelling jointly frequency & severity ⟶ **Reserving**.

Thank you for your attention!

UNIVERSITÀ DELLA CALABRIA

- Bastos, L.S., Economou, T., Gomes, M.F., Villela, D.A., Coelho, F.C., Cruz, O.G., Stoner, O., Bailey, T., Codeço, C.T. (2019). A modelling approach for correcting reporting delays in disease surveillance data. *Statistics in medicine*, **38**(22), pp. 4363–4377.
- Edwards, B., Hofmeyr, S., Forrest, S. (2016). Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity*, **2**(1), pp. 3–14.
- Hu, K., Levi, R., Yahalom, R., Zerhouni, E.G. (2022). Supply chain characteristics as predictors of cyber risk: A machine-learning assessment. Available at https://api.semanticscholar.org.
- Li, Y., Mamon, R. (2023). Modelling health-data breaches with application to cyber insurance. *Computers & Security*, **124**.
- McLeod, A., Dolezel, D. (2018). Cyber-analytics: Modeling factors associated with health-care data breaches. *Decision Support Systems*, **108**, pp. 57–68.

- Sun, H., Xu, M., Zhao, P. (2020). Modeling malicious hacking data breach risks. *North American Actuarial Journal*, **25**, pp. 484–502.
- Wheatley, S., Hofmann, A., Sornette, D. (2019). Data breaches in the catastrophe framework & beyond. Available at https://api.semanticscholar.org/CorpusID:155100199.