Claims reserving in non-life insurance: old and new adventures

One World Actuarial Research Seminar

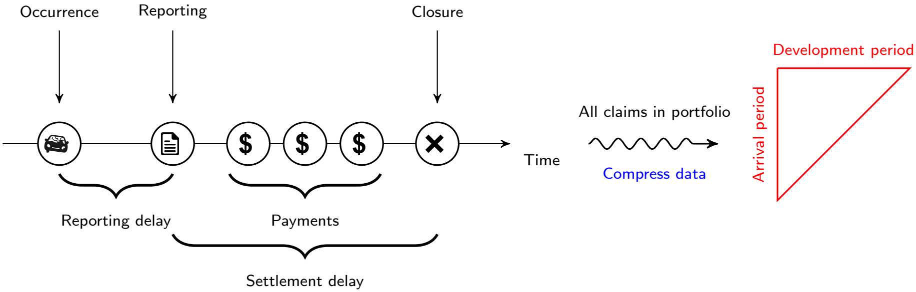
Katrien Antonio and Jonas Crevecoeur
LRisk - KU Leuven and ASE - University of Amsterdam

June 17, 2020

My personal website: `https://katrienantonio.github.io`.

Talk is based on joint work with Jonas Crevecoeur, Roel Verbelen and Gerda Claeskens.

Claim dynamics



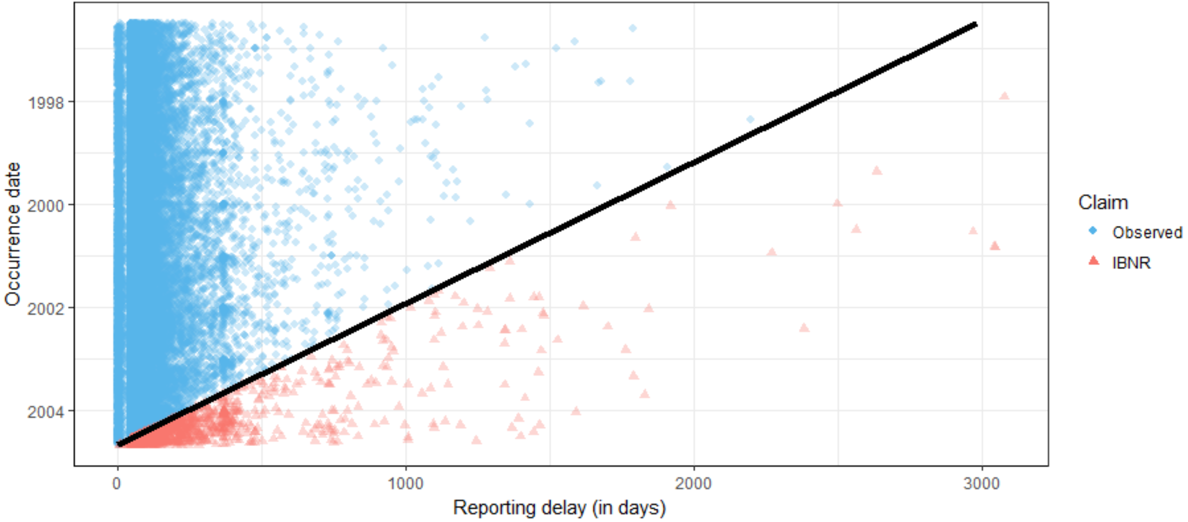
We typically aggregate the data from the time line into a run-off triangle.

Covariates largely ignored!

This webinar's mission statement

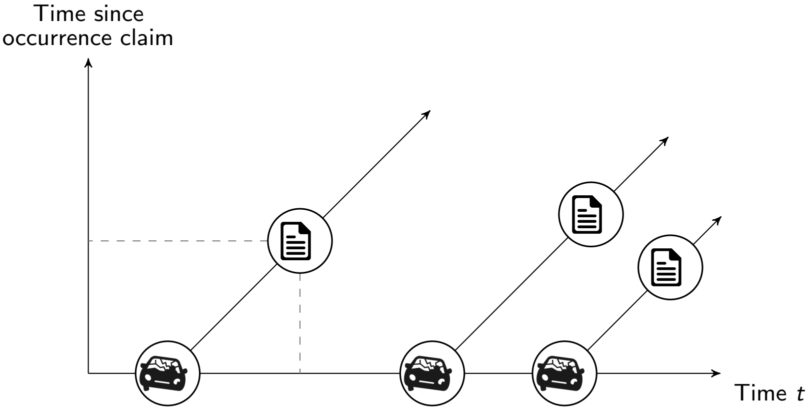
1. Launch a discussion on individual, granular data for loss reserving, and their features.
2. Sketch published research on the modeling of IBNR claim counts.
3. Sketch ongoing research on the development of RBNS claims.
4. Provide (data driven) guidance on the choice between aggregate and individual reserving for a given portfolio.
5. Reflect upon stream of academic research on data analytics for reserving.

Research on modelling IBNR claim counts



Research focus

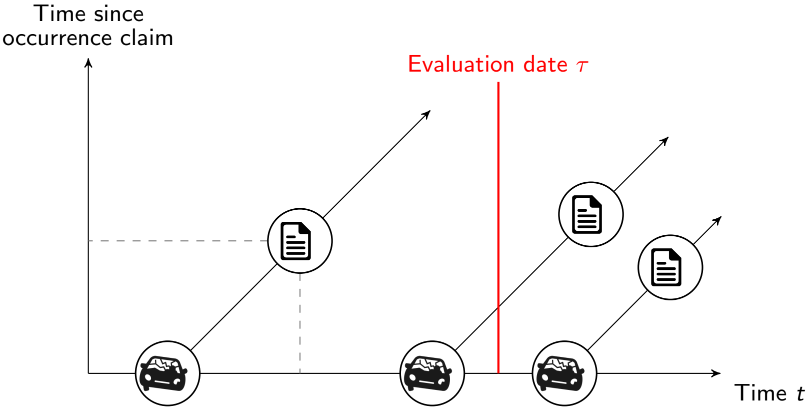
IBNR claims in a Lexis diagram setting



The insurance company is not aware (yet) of claims related to past exposures that are not (yet) reported!

Research focus

IBNR claims in a Lexis diagram setting



The insurance company is not aware (yet) of claims related to past exposures that are not (yet) reported!

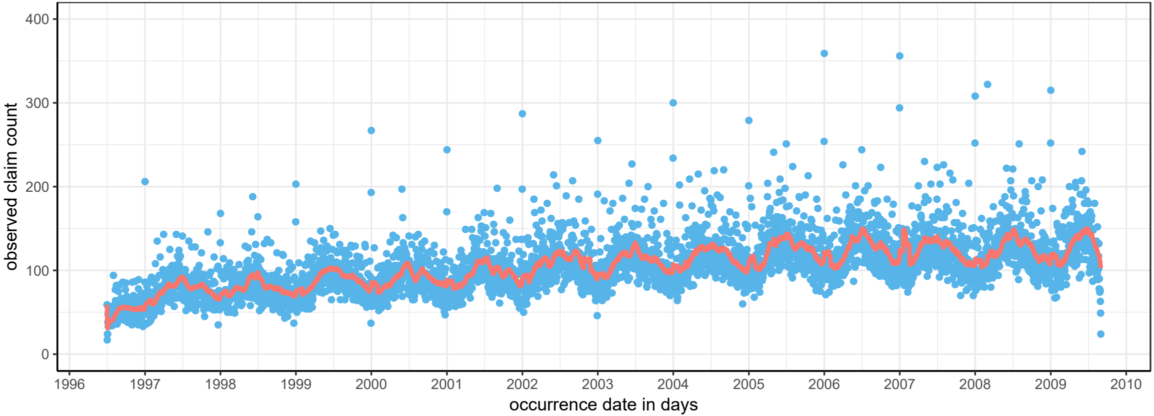
Research questions

IBNR claim counts

- Research questions with focus on IBNR claim counts:
  - How many claims occurred but are not yet reported, because reporting delay is subject to right truncation?
  - When will these IBNR claims be reported?
- Pioneering work by Ragnar Norberg (1993, 1999), (basic, first) implementation in Antonio & Plat (2014), new work in Verrall & Wüthrich (2016), all in continuous time!
- Recent contributions (a.o.) by Avanzi, Wong & Yang (2017) with a Shot Noise Cox Process, Badescu, Lin & Tang (2016) with a Hidden Markov Model.

Case study with liability claims data set

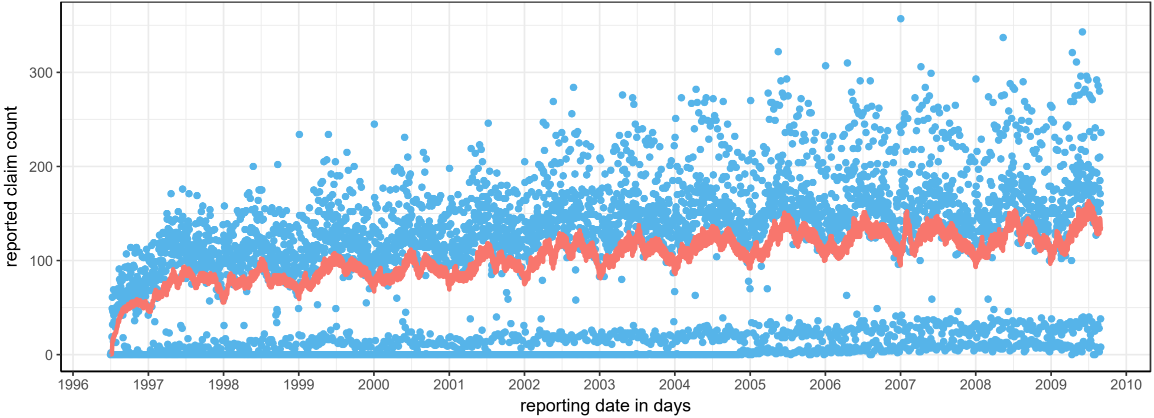
Claim occurrence process



Observation window: July 1, 1996 to August 31, 2009, MM/DD/YY format, i.e. day is natural time unit.

Case study with liability claims data set

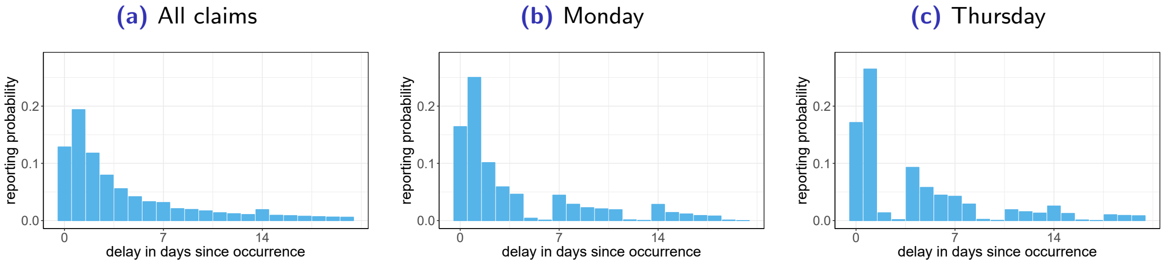
Reporting process



Case study with liability claims data set

Reporting delay

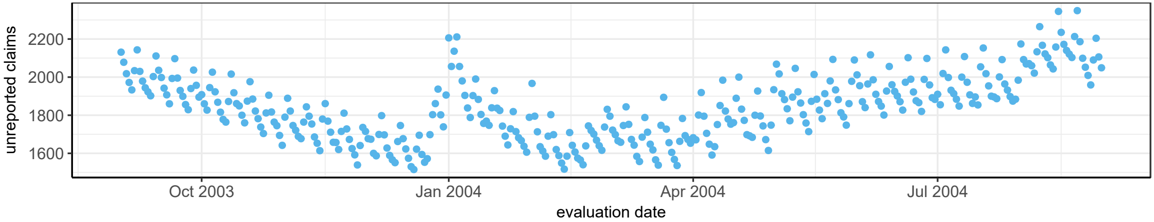
Declining pattern in reporting delay + intra-week pattern, depending on the occurrence day of the week.

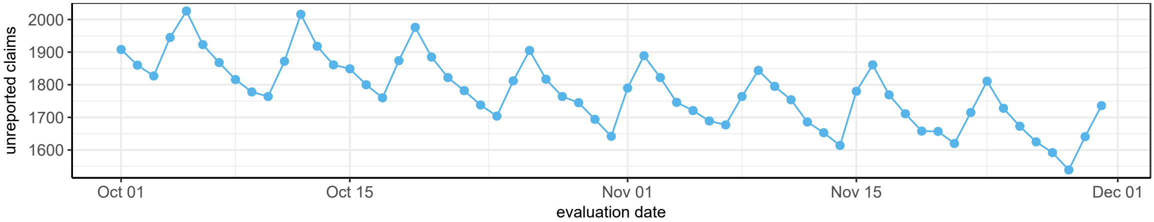


On (semi-)official holidays: drop in number of reported claims compared to daily average.

Case study with liability claims data set

Total IBNR counts





The statistical model for IBNR

Thoughts

	IBNR	RBNS
event structure	single event	multiple, recurrent events
time horizon	usually quick	longer (in years)
time granularity	in days since occurrence	in years since reporting
covariates	triangle with daily occurrences and reportings	individual claim- and policy(holder) specific
other fields	nowcasting in epidemiology; occurrence of events, observed with delay	recurrent events with marks

The statistical model for IBNR

Notations

- N_t: the (total) number of claims that occurred on day t.
- $N_{t,s}$: the number of claims from day t that are reported on day s.
- Each claim eventually gets reported, thus $N_t = \sum_{s=t}^{\infty} N_{t,s}$.
- Assumption: the $N_{t,s}$ are independent and

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

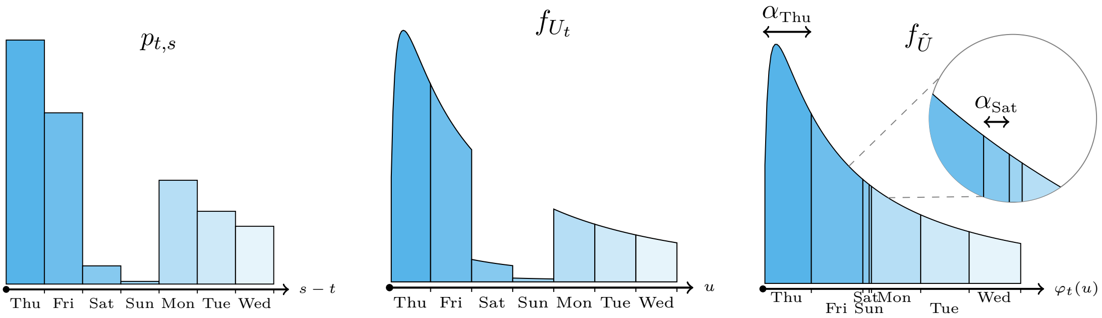
Research contributions

1. Verbelen, R., Antonio, K., Claeskens, G & Crèvecoeur, J. 2019, R&R.
    - joint estimation of occurrence process and reporting delay distribution
    - use EM to optimize the likelihood in presence of missing data
    - regression at (t, s) level.
2. Crèvecoeur, J., Antonio., K. & Verbelen, R. 2019, EJOR.
    - time-change strategy, daily reporting exposures

      focus on calendar day effects, e.g. national holidays and weekend, reporting at specific delays (e.g. 14 days, 1 year)
    - regression at (t, s) level, investigate different settings via simulation study.

Time change strategy

The idea pictured!



Time change strategy

Structuring the reporting exposures

- Use a standard distribution for $\tilde{U}$ (e.g. exponential, lognormal).
- Explain the daily reporting exposures as a function of covariates:

$$\alpha_{t,s} \quad = \quad \exp\left(\boldsymbol{x}_{t,s}^{'} \cdot \boldsymbol{\gamma}\right).$$

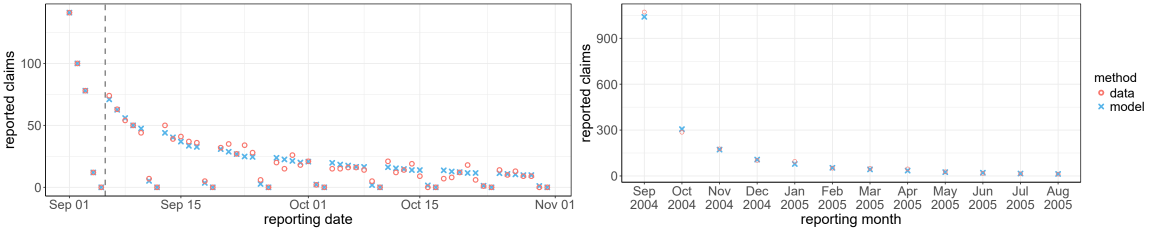
- Joint estimation of distribution $\tilde{U}$ and regression parameters to structure $\alpha_{t,s}$.

  Use maximum likelihood estimation with the likelihood of the reported claims.

Case study

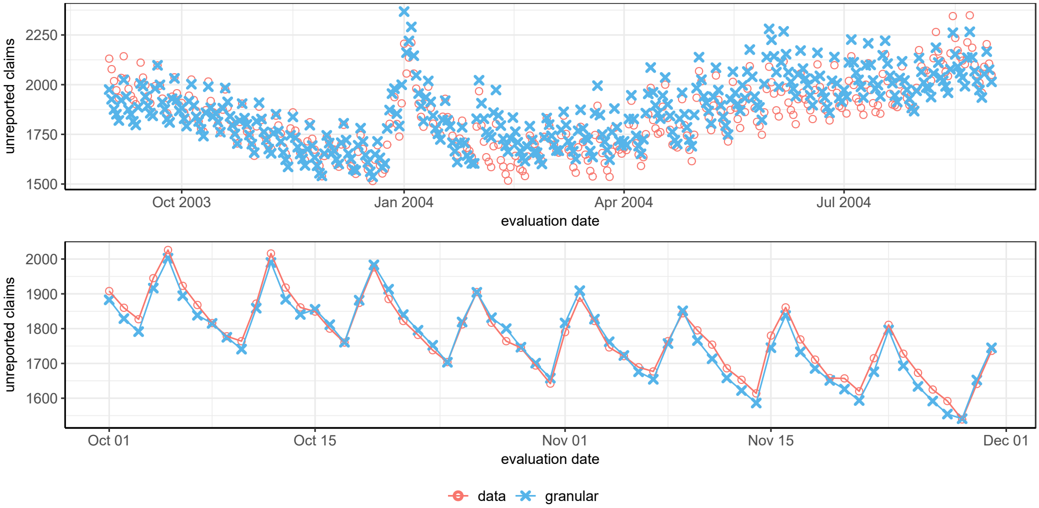
IBNR Results - first evaluation (only granular)

When are the claims that are IBNR (on August 31, 2004) reported?



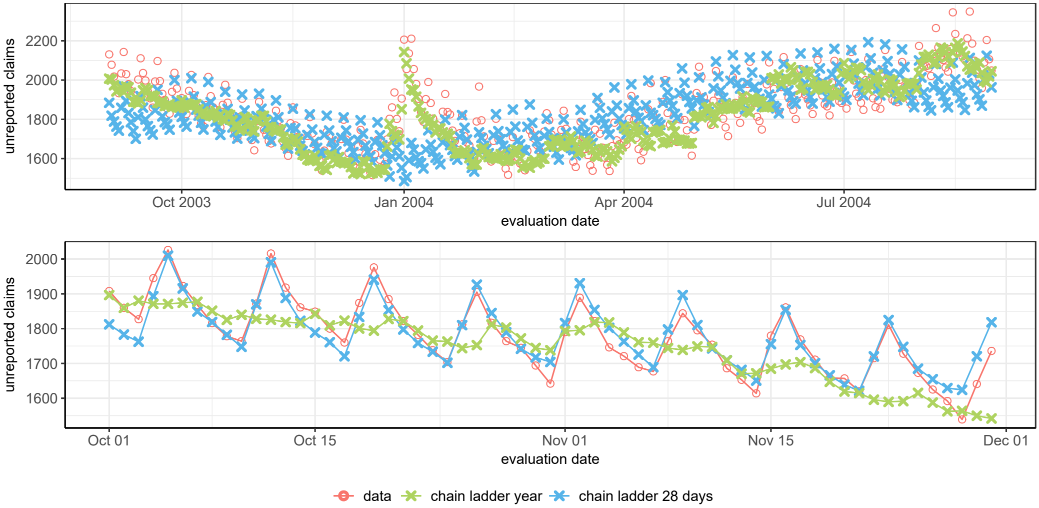
Case study

IBNR Results - second evaluation



Case study

IBNR Results - third evaluation, see Hiabu (2017, SAJ) and Martínez-Miranda, Nielsen et al. (2013)



Wrap-up

IBNR reserving

- ✓ Accurate predictions under weekday and holiday effects.
- ✓ Faster detections of changes in the occurrence/reporting process (cfr. simulation study).
- ✓ Triangle at daily level, incorporate covariates.
- ✗ Longer computation time.
- ✗ More choices (i.e. distributional assumptions, selecting variables) - should be done with care!

Run-off triangles in epidemiology

D		N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_8	N_9	N_{10}
	3	$n_{1,3}$	$n_{2,3}$	$n_{3,3}$	$n_{4,3}$	$n_{5,3}$	$n_{6,3}$	$n_{7,3}$	$n_{8,3}$	$n_{9,3}$	$n_{10,3}$
	2	$n_{1,2}$	$n_{2,2}$	$n_{3,2}$	$n_{4,2}$	$n_{5,2}$	$n_{6,2}$	$n_{7,2}$	$n_{8,2}$	$n_{9,2}$	$n_{10,2}$
	1	$n_{1,1}$	$n_{2,1}$	$n_{3,1}$	$n_{4,1}$	$n_{5,1}$	$n_{6,1}$	$n_{7,1}$	$n_{8,1}$	$n_{9,1}$	$n_{10,1}$
	0	$n_{1,0}$	$n_{2,0}$	$n_{3,0}$	$n4_{,0}$	$n_{5,0}$	$n_{6,0}$	$n_{7,0}$	$n_{8,0}$	$n_{9,0}$	$n_{10,0}$
d	t	1	2	3	4	5	6	7	8	9	10

← from van de Kassteele, Eilers & Wallinga in Epidemiology (2019)

from Bastos et al. in Statistics in Medicine (2019) →

Time	0	1	2	…	D-2	D-1	D	N	
1	$n_{1,0}$	$n_{1,1}$	$n_{1,2}$		$n_{1,D-2}$	$n_{1,D-1}$	$n_{1,D}$	N_1	Observations
2	$n_{2,0}$	$n_{2,1}$	$n_{2,2}$	…	$n_{2,D-2}$	$n_{2,D-1}$	$n_{2,D}$	N_2	
3	$n_{3,0}$	$n_{3,1}$	$n_{3,2}$		$n_{3,D-2}$	$n_{3,D-1}$	$n_{3,D}$	N_3	
⋮		⋮				⋮			
T-D	$n_{T-D,0}$	$n_{T-D,1}$	$n_{T-D,2}$		$n_{T-D,D-2}$	$n_{T-D,D-1}$	$n_{T-D,D}$	N_{T-D}	
T-D+1	$n_{T-D+1,0}$	$n_{T-D+1,1}$	$n_{T-D+1,2}$		$n_{T-D+1,D-2}$	$n_{T-D+1,D-1}$	$n_{T-D+1,D}$	N_{T-D+1}	Nowcasting
T-D+2	$n_{T-D+2,0}$	$n_{T-D+2,1}$	$n_{T-D+2,2}$		$n_{T-D+2,D-2}$	$n_{T-D+2,D-1}$	$n_{T-D+2,D}$	N_{T-D+2}	
T-2	$n_{T-2,0}$	$n_{T-2,1}$	$n_{T-2,2}$		$n_{T-2,D-2}$	$n_{T-2,D-1}$	$n_{T-2,D}$	N_{T-2}	
T-1	$n_{T-1,0}$	$n_{T-1,1}$	$n_{T-1,2}$		$n_{T-1,D-2}$	$n_{T-1,D-1}$	$n_{T-1,D}$	N_{T-1}	
T	$n_{T,0}$	$n_{T,1}$	$n_{T,2}$	…	$n_{T,D-2}$	$n_{T,D-1}$	$n_{T,D}$	N_T	
T+1	$n_{T+1,0}$	$n_{T+1,1}$	$n_{T+1,2}$		$n_{T+1,D-2}$	$n_{T+1,D-1}$	$n_{T+1,D}$	N_{T+1}	Forecasting
T+2	$n_{T+2,0}$	$n_{T+2,1}$	$n_{T+2,2}$		$n_{T+2,D-2}$	$n_{T+2,D-1}$	$n_{T+2,D}$	N_{T+2}	
⋮		⋮		…		⋮			
T+K	$n_{T+K,0}$	$n_{T+K,1}$	$n_{T+K,2}$		$n_{T+K,D-2}$	$n_{T+K,D-1}$	$n_{T+K,D}$	N_{T+K}	

What else is there - going beyond actuarial science?

Nowcasting the number of new symptomatic cases during infectious disease outbreaks using constrained P-spline smoothing

van de Kassteele, Eilers & Wallinga, in Epidemiology (2019).

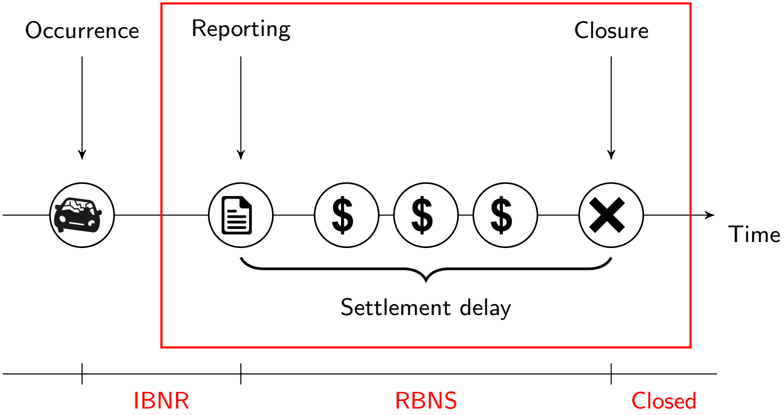
- nowcasting = assessment of the current situation based on imperfect or partial information
- illustration on a large measles outbreak in the Netherlands, May 2013 - March 2014.

A modelling approach for correcting reporting delays in disease surveillance data

Bastos et al., in Statistics in Medicine (2019).

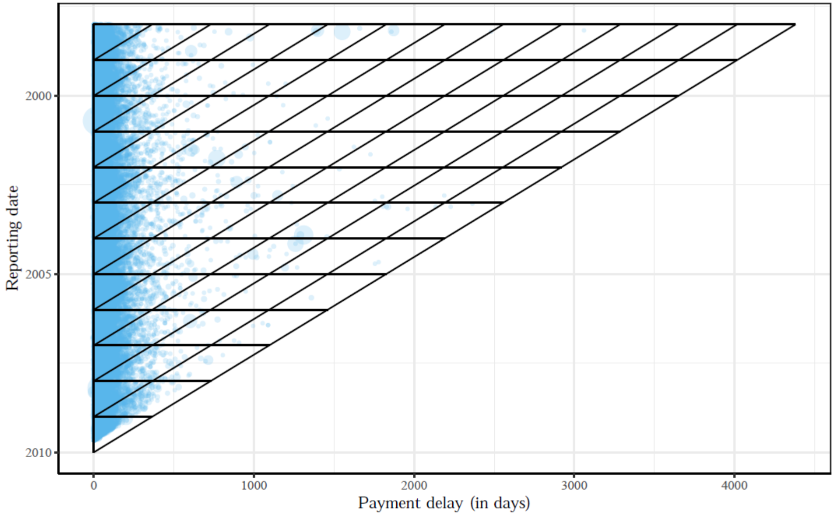
Research on the development of RBNS claims

Research focus - RBNS



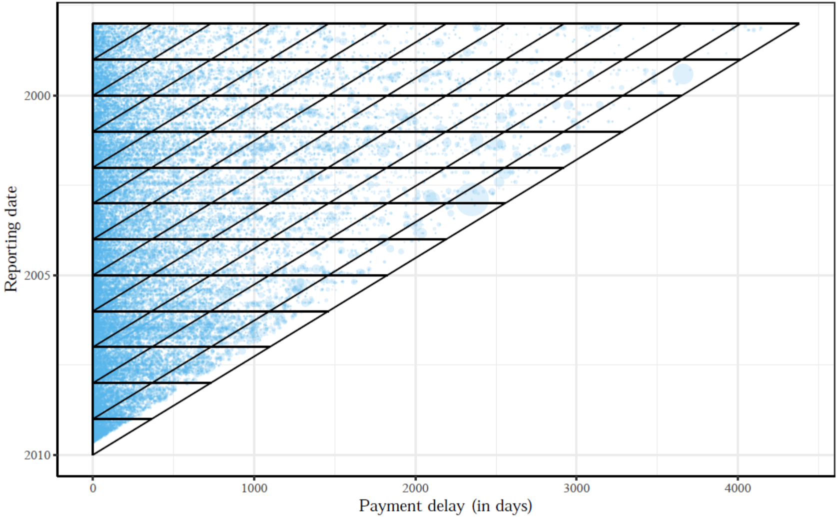
Development of an RBNS claim

Continuous time



Development of an RBNS claim

Continuous time



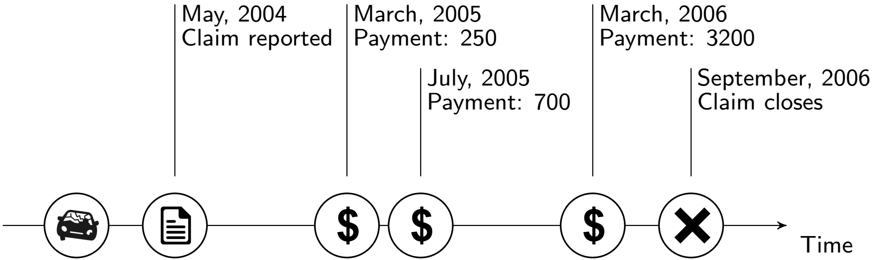
The statistical model for RBNS

Thoughts

	IBNR	RBNS
event structure	single event	multiple, recurrent events
time horizon	usually quick	longer (in years)
time granularity	in days since occurrence	in years since reporting
covariates	triangle with daily occurrences and reportings	individual claim- and policy(holder) specific
other fields	nowcasting in epidemiology, occurrence of events, observed with delay	recurrent events with marks

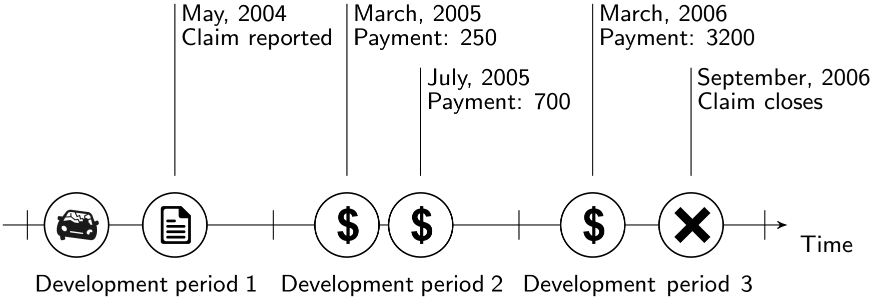
The statistical model

RBNS claim development in discrete time



The statistical model

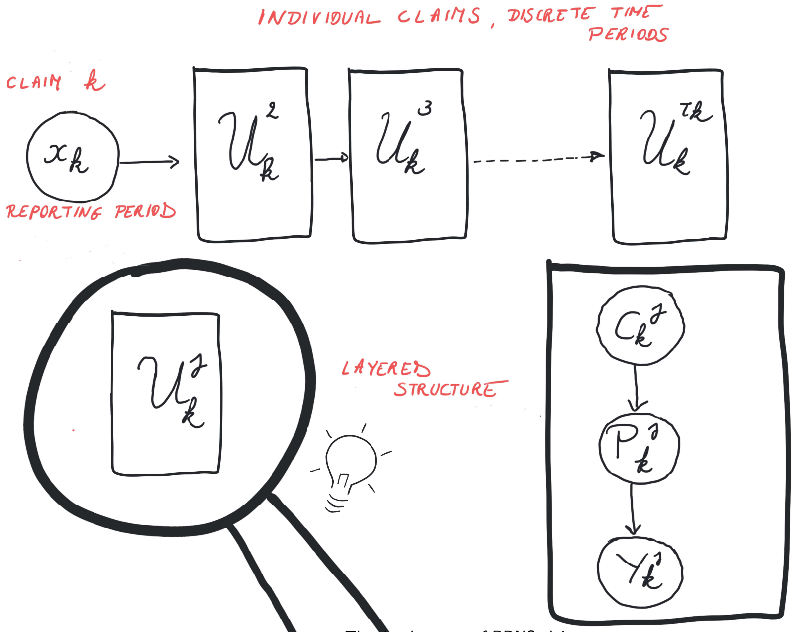
RBNS claim development in discrete time



- Claim reported
- Payment: 0

- Payment: 950

- Payment: 3200
- Claim closed


INDIVIDUAL CLAIMS, DISCRETE TIME PERIODS
CLAIM k
x_k
REPORTING PERIOD
$\mathcal{U}_k^2$
$\mathcal{U}_k^3$
$\mathcal{U}_k^{\tau_k}$
$\mathcal{U}_k^j$
LAYERED STRUCTURE
C_k^j
P_k^j
Y_k^j

A hierarchical reserving model for RBNS claims

Layers

- Index the individual claims by k and the development periods by j.
- Our approach is modular or layered:
  - $\boldsymbol{x}_k$ denotes the (observed, fixed) claim information available at the end of the first development period, i.e. the reporting period

    e.g. cause of claim, policy(holder) covariates, initial case estimate
  - $\boldsymbol{U}_k^j$ is the vector with claim k's updated information in development period j

    depends on portfolio at hand, e.g. $\boldsymbol{U}_k^j = (C_k^j, P_k^j, Y_k^j)$ with a settlement indicator C_k^j, a payment indicator P_k^j and payment size Y_k^j.
- Need more modular components $\Rightarrow$ extend $\boldsymbol{U}_k^j$!

A hierarchical reserving model for RBNS claims

Hierarchical structure

- Update vectors $\boldsymbol{U}_k^j$ for claim k are observed from development period 2 to τ_k:

$$\mathcal{R}^{\texttt{Obs}} = \{\boldsymbol{U}_k^j \mid k = 1, \ldots, n, j = 2, \ldots, \tau_k\}.$$

- We introduce a time dynamic hierarchical structure (see Frees & Valdez, 2008, JASA):

$$\begin{aligned}\mathcal{L}\left(\mathcal{R}^{\texttt{Obs}}\right) &= \prod_{k=1}^{n} f\left(\boldsymbol{U}_k^{(2)}, \ldots, \boldsymbol{U}_k^{(\tau_k)} \mid \boldsymbol{x}_k\right)\\ &= \prod_{k=1}^{n}\prod_{j=2}^{\tau_k} f\left(\boldsymbol{U}_k^{j} \mid \boldsymbol{U}_k^{(2)}, \ldots, \boldsymbol{U}_k^{(j-1)}, \boldsymbol{x}_k\right).\end{aligned}$$

Thus, future development depends on the past.

A hierarchical reserving model for RBNS claims

Hierarchical structure

- As a last step, we introduce a layered hierarchical structure for $\boldsymbol{U}_k^j$:

$$\mathcal{L}\left(\mathcal{R}^{\texttt{Obs}}\right) = \prod_{k=1}^{n}\prod_{j=2}^{\tau_k}\prod_{l=1}^{s} f\left(U_{k,l}^j \mid \boldsymbol{U}_k^{(2)},\ldots,\boldsymbol{U}_k^{(j-1)}, U_{k,1}^j,\ldots,U_{k,l-1}^j, \boldsymbol{x}_k\right),$$

  with s the number of layers in the update vector.

- For example, with $\boldsymbol{U}_k^j = (C_k^j, P_k^j, Y_k^j)$ we focus on the three essential building blocks from Antonio & Plat (2014)!

- The framework incorporates static (via $\boldsymbol{x}_k$) as well as dynamic features.

A hierarchical reserving model for RBNS claims

Three layers $\boldsymbol{U}_k^j = (C_k^j, P_k^j, Y_k^j)$

- C_k^j is one if claim k settles in development period j and zero otherwise

$$C_k^j \,|_{\boldsymbol{U}_k^{(2)},\ldots,\boldsymbol{U}_k^{(j-1)},\boldsymbol{x}_k} \sim \texttt{Bernoulli}\left(p\left(\boldsymbol{U}_k^{(2)},\ldots,\boldsymbol{U}_k^{(j-1)},\boldsymbol{x}_k\right)\right).$$

- P_k^j is one if there is a payment for claim k in development period j and zero otherwise

$$P_k^j \,|_{\boldsymbol{U}_k^{(2)},\ldots,\boldsymbol{U}_k^{(j-1)},C_k^j,\boldsymbol{x}_k} \sim \texttt{Bernoulli}\left(q\left(\boldsymbol{U}_k^{(2)},\ldots,\boldsymbol{U}_k^{(j-1)},C_k^j,\boldsymbol{x}_k\right)\right).$$

- Y_k^j is the payment size, given $P_k^j = 1$. The payment size is gamma distributed with mean

$$E(Y_k^j \mid \boldsymbol{U}_k^{(2)},\ldots,\boldsymbol{U}_k^{(j-1)},C_k^j,P_k^j,\boldsymbol{x}_k) = \mu(\boldsymbol{U}_k^{(2)},\ldots,\boldsymbol{U}_k^{(j-1)},C_k^j,P_k^j,\boldsymbol{x}_k).$$

A hierarchical reserving model for RBNS claims

Model calibration

Guidelines for the model calibration:

- you can use your preferred predictive model (e.g. GLM or Gradient Boosting Machine)
- apply k-fold cross validation (to prevent overfitting) and a weighted likelihood

$$\prod_{k=1}^{n} \prod_{j=2}^{\tau_k} w_j \cdot f\left(U_{k,l}^{j} \mid \boldsymbol{U}_k^{(2)}, \ldots, \boldsymbol{U}_k^{(j-1)}, U_{k,1}^{j}, \ldots, U_{k,l-1}^{j}, \boldsymbol{x}_k\right),$$

per layer l in the model.

The weights tackle the covariate shift (or imbalance), e.g. more claims with later development years in lower vs. upper ‘triangle’.

Bridging aggregate and individual reserving

The choice between an individual and an aggregate reserving model then depends (in a data driven way) on:

- the covariates included in the predictive model, e.g. for layer l

$$E(U_{k,l}^{j}) = \tilde{\alpha}_{i_k+r_k,l} \cdot \beta_{j,l} \cdot \exp\left\{\phi\left(\boldsymbol{U}_k^2, \ldots, \boldsymbol{U}_k^{j-1}, U_{k,1}^{j}, \ldots, U_{k,l-1}^{j}, \boldsymbol{x}_k\right)\right\},$$

  with $i_k + r_k$ the reporting year of claim k.

For layer l, if $\phi(.) = 0$ (i.e. no relevant covariates) summing the individual, claim-specific updates reduces to a triangle with the multiplicative chain ladder structure (since reporting).

Connections with the literature (a selection)

Denuit & Trufin (2017, 2018), Wahl et al. (2019), RBNS in Double CL

Reserving by combining multiple runoff triangles for closure, payment and size.

Wüthrich (2018), special issue Risks edited by Taylor (2020)

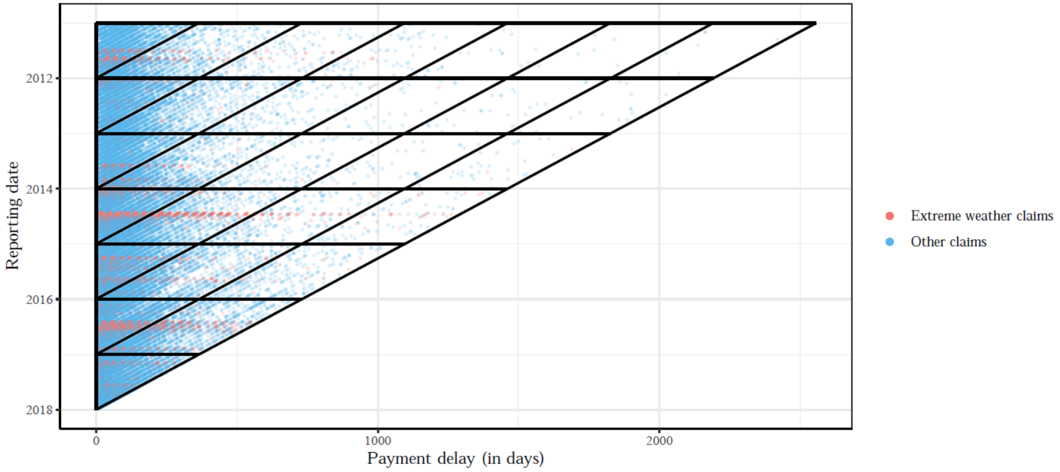
Machine learning methods (e.g. regression tree, GBM) for the payment and closure indicator.

Larsen (2007)

GLM fitted for each of the components (C, P and Y) in the development process.

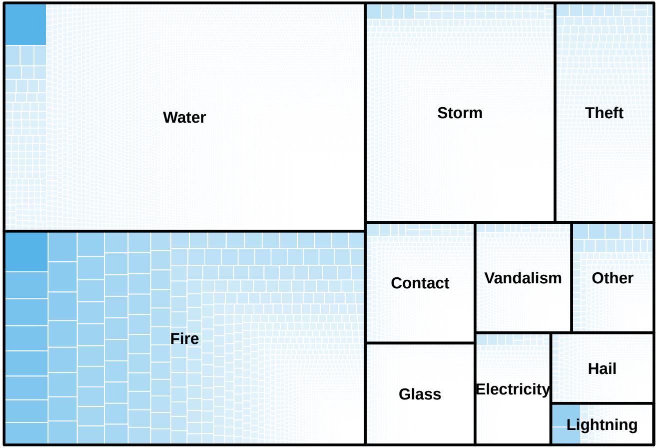
Data exploration

Home insurance claims - global events

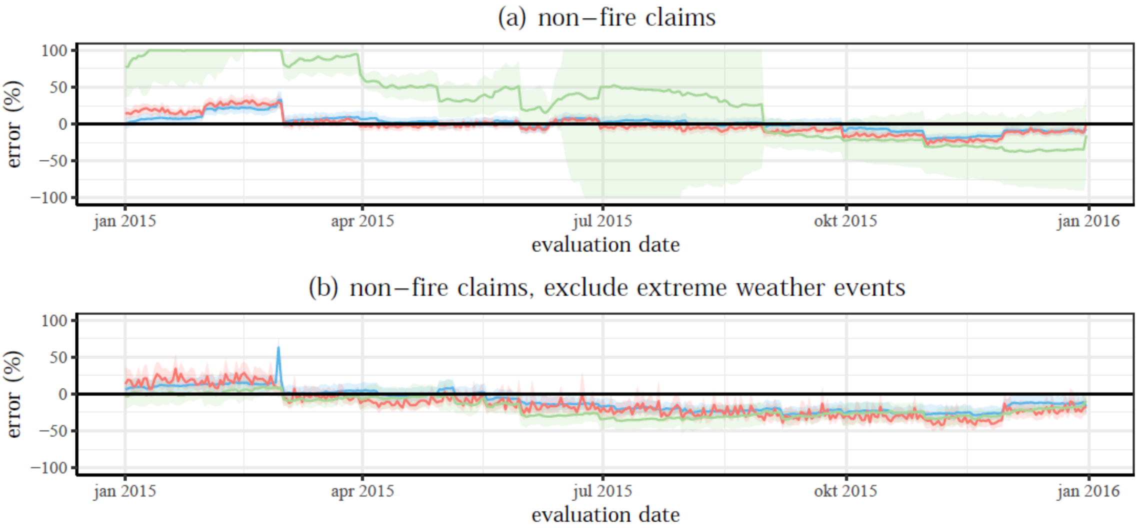


Data exploration

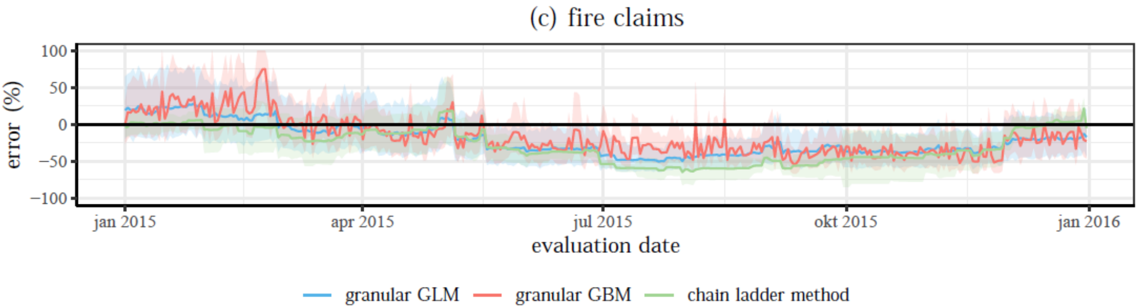
Home insurance claims - treemap



Multiple evaluation dates



Multiple evaluation dates



$$\text{Percentage Error} = 100 \cdot \frac{\texttt{predicted}-\texttt{actual}}{\texttt{actual}}.$$

Overall performance

Portfolio	hierarchical GLM		hierarchical GBM		chain ladder	
	$\mu(PE)$	$\mu(\|PE\|)$	$\mu(PE)$	$\mu(\|PE\|)$	$\mu(PE)$	$\mu(\|PE\|)$
non-fire claims	0.92	**7.32**	-1.80	10.23	33.89	51.31
non-fire claims, exclude extreme weather	-9.76	**14.90**	-14.28	20.18	-18.10	19.07
fire-claims	-20.82	**26.44**	-16.42	26.50	-28.41	29.76

Average performance is expressed as the mean percentage error and the mean absolute percentage error.

Take home insights

- ✓ Structure the (highly) scattered literature on analytics for loss reserving.
- ✓ Hybrid strategy, take data-driven position between individual and aggregate.
- ✓ *Less is more*, unify pricing and reserving methodology (e.g. GLMs, GBMs).
- ✓ Lessons to learn from the machine learning literature.
- ✓ Use multiple evaluation dates, instead of single out-of-time.
- ✓ Use multiple portfolios, *no free lunch*.

More information

For more information, please visit:

LRisk website, www.lrisk.be

https://katrienantonio.github.io

Thanks to

ageas

ARGENTA

Vlaams Supercomputer Centrum

fwo Research Foundation Flanders
Opening new horizons

References

For an overview of the literature, please see the references in:

- Verbelen, R., Antonio, K., Claeskens, G & Crèvecoeur, J. 2019.
  Modeling the occurrence of events subject to a reporting delay via an EM algorithm
  R&R, working paper online at https://arxiv.org/abs/1909.08336.
- Crèvecoeur, J., Antonio., K. & Verbelen, R. 2019.
  Modeling the number of hidden events subject to observation delay
  European Journal of Operations Research, working paper online at https://arxiv.org/abs/1801.02935.
- Crèvecoeur, J. & Antonio., K. 2020.
  A generalized reserving model bridging the gap between pricing and individual reserving
  Working paper at https://arxiv.org/abs/1910.12692; a new version will be available soon.