(New) Challenges in Actuarial Science

Mario V. Wüthrich
RiskLab, ETH Zurich
Swiss Finance Institute Professor

July 11, 2016
R in Insurance Conference, London

Data science, data analytics and big data

"Whilst the fundamentals of analyzing data have not changed, our approach to collating and understanding data, creating accessible and useful information, developing skill sets and ultimately transforming huge and ever-growing repositories of data into actionable insights for our employers, shareholders and our communities more generally has entered a new paradigm."

Source: ASTIN Big Data/Data Analytics Working Party - Phase 1 Paper - April 2015

Data science, data analytics and big data

"Data captured and stored by industry doubles every year and 90%+ of all data in existence has been created during the last 2 years. Most of this is unstructured data such as emails, tweets and videos - every minute we send 204m emails, generate 1.8m Facebook likes, send 278k Tweets, up-load 200k photos to Facebook and 100 hours of video to YouTube."

Source: ASTIN Big Data/Data Analytics Working Party - Phase 1 Paper - April 2015

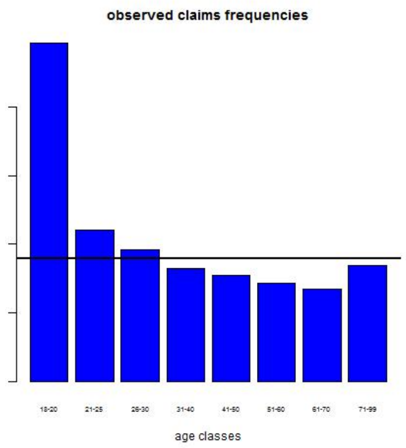
Actuarial functions

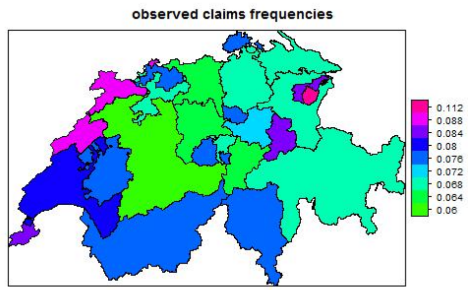
The actuary plays a central role in data analysis and predictive modeling:

- ▷ insurance pricing and product development;
- ▷ reserving and accounting;
- ▷ risk management and Nat Cat modeling;
- ▷ marketing;
- ▷ social and political process.

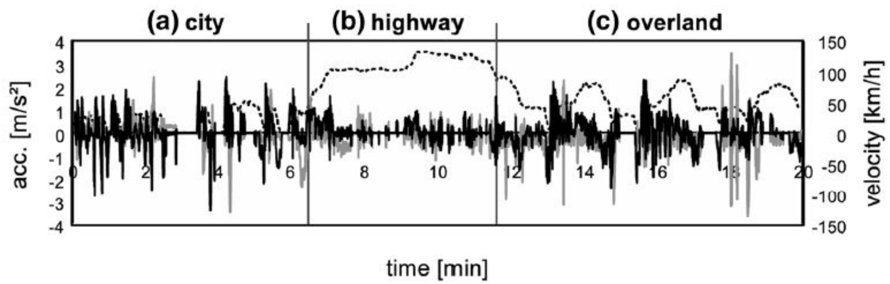
Insurance pricing and product development

* Classical actuarial pricing: GLM with a *small number of selected tariff criteria*:





* Data analytics and telematic data with *continuous data collection*:



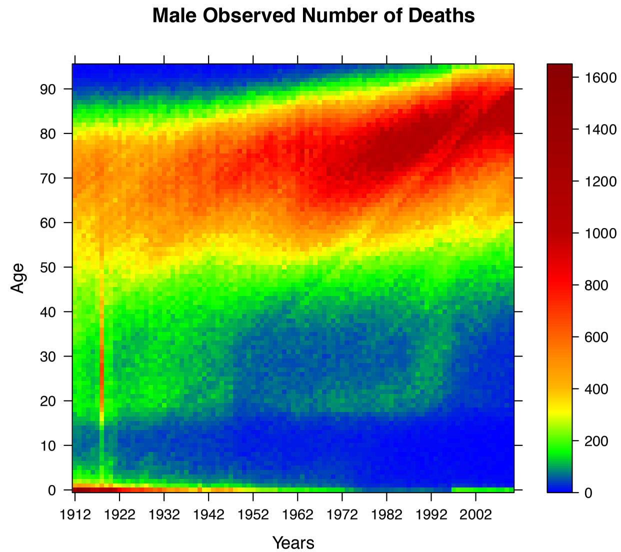
Source: Weidner et al. (EAJ 2016)

Personalized health & mortality data

* Collection of *personalized* health and mortality data.

* Better medical knowledge about risk drivers, e.g., old-age research.

* Main questions about causality requires *interdisciplinary research*.



Reserving and accounting

* Aggregate claims triangles:

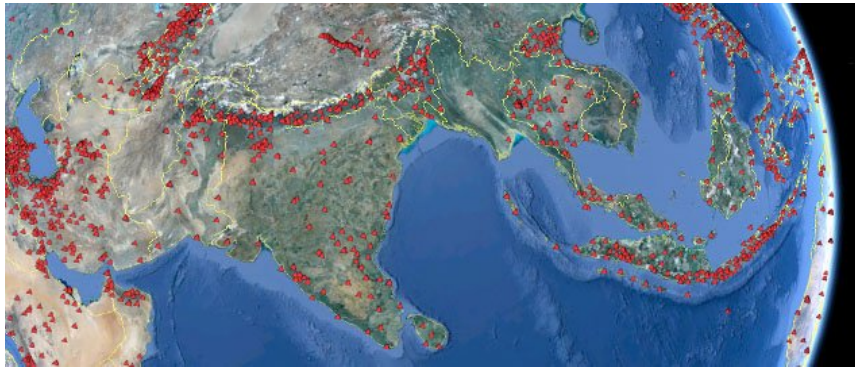
i / j	0	1	2	3	4	5	6
2007	1'052	1'321	1'700	1'971	2'298	2'645	3'003
2008	808	1'029	1'229	1'590	1'842	2'150	
2009	1'016	1'251	1'698	2'105	2'385		
2010	948	1'108	1'315	1'487			
2011	917	1'082	1'484				
2012	1'001	1'376			$C_{i,j}$ predicted		
2013	841						

* Individual claims histories:

	A	B	C	D	E	F	G	H	I	J
1	no	record date	LoB	claims code	accident	reporting	closing	status	paid	case est.
2	32	200108	3	14	20010628	20010805	20010815	0	328	600
3	33	200109	2	11	20010831	20010924	NA	1	0	4000
4	33	200111	2	11	20010831	20010924	NA	1	4625	4000
5	33	200206	2	11	20010831	20010924	20020615	0	4625	4000
6	33	200212	2	11	20010831	20010924	20021215	0	5189	4000
7	34	200306	3	21	20030515	20030625	NA	1	0	50000
8	34	200307	3	21	20030515	20030625	NA	1	43000	50000
9	34	200312	3	21	20030515	20030625	NA	1	61281	62000
10	34	200404	3	21	20030515	20030625	NA	1	50481	62000
11	34	200405	3	21	20030515	20030625	NA	1	49671	62000
12	34	200407	3	21	20030515	20030625	20040717	0	50166	50000
13	35	200411	3	14	20040821	20041124	NA	1	0	12000
14	35	200507	3	14	20040821	20041124	20050730	0	12244	12000
15	36	200202	4	10	20011025	20020220	NA	1	0	3000

Risk management and Nat Cat modeling

* Natural hazard modeling:



* Physical models, collection of sensor data, statistical models, etc.

* Efficiency: run-time is crucial (earthquake: 1 minute (?) of reaction time).

* Civil engineering: building code, catastrophe simulation, etc.

Actuarial functions

* Insurance pricing and product development
* Reserving and accounting
* Risk management and Nat Cat modeling
* Marketing

All these actuarial fields go through massive changes:

▷ These changes are data driven.

▷ Is the actuarial profession ready for these changes?

▷ How can the R community support the actuarial profession in these changes?

Legal and marketing issues

- A lot of (unstructured) data is available in the internet, social media, etc.
- Personal data is collected by several stakeholders (data has a value).
- Privacy is an issue (and legislation lags behind).
- Collection of telematic data:
  - ⋆ voluntariness;
  - ⋆ often only younger drivers (for reputational and marketing reasons);
  - ⋆ price reduction of 20% motivates 70% of young drivers to install drive recorder!

  ▷ Volume is too small to do statistical analysis (see below).

On-going changes in general insurance

Many general insurance market leaders re-structure their actuarial organization.

▷ Classical actuarial pricing departments are split into 3 sub-units:

(1) pricing using classical methods (like GLM with classical tariff criteria);

(2) data science and data analytics (not necessarily actuaries and statisticians);

(3) financial controlling and business development.

▷ Competence of simplifying complex data for decision makers is split into different sub-units/modules (and different disciplines).

▷ Actuaries and statisticians lose influence if they do not compete in the new fields!

Data science and data analytics: modeling tools

- linear regression (and correlation) analysis, principal component analysis (PCA)
- generalized linear models (GLM), generalized additive models (GAM)
- classification and regression trees (CART)
- bootstrap aggregating (bagging), random forests
- boosting, support vector machine (SVM)
- Fourier analysis, hidden Markov models, particle filters
- Bayesian networks, machine learning, etc.

These (statistical) techniques have in common that they all try to extract the relevant features and try to measure their influence on the response.

Data science and data analytics: features

Determine the structural function f, such that we can write response Y as

$$Y = f(x_1, \ldots, x_p) + \varepsilon,$$

for features x_i and a (centered) measurement error ε.

Main questions:

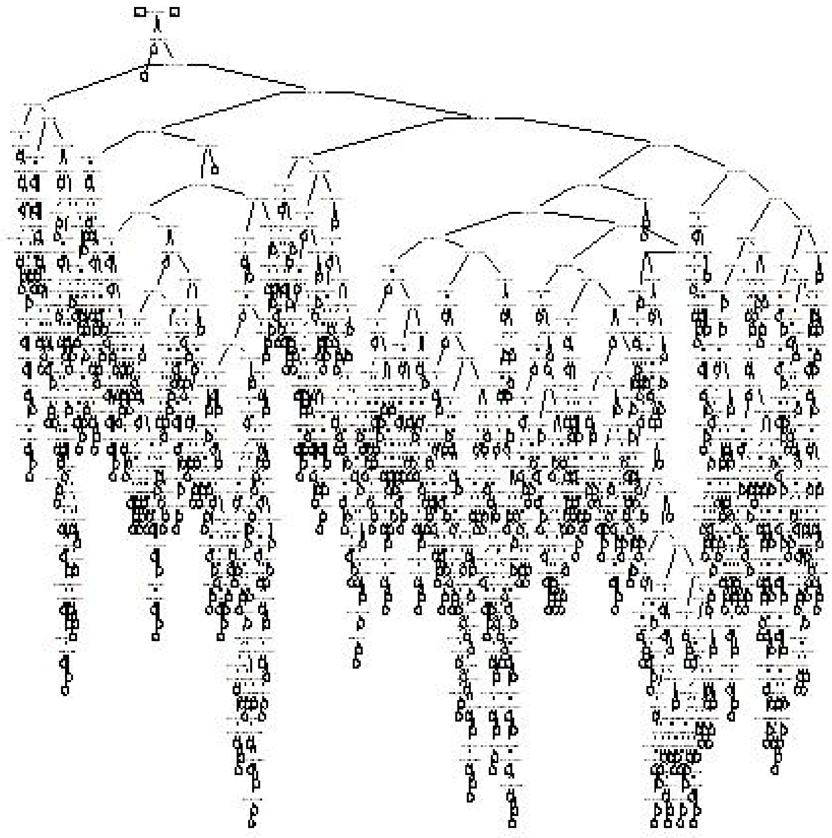
- What are the relevant features x_i to be included (with predictive power)?
- What does the structural function f look like?
- How can features x_i be constructed from (continuous) data?
- What if the response Y is rather a dynamic process (potentially non-stationary)?

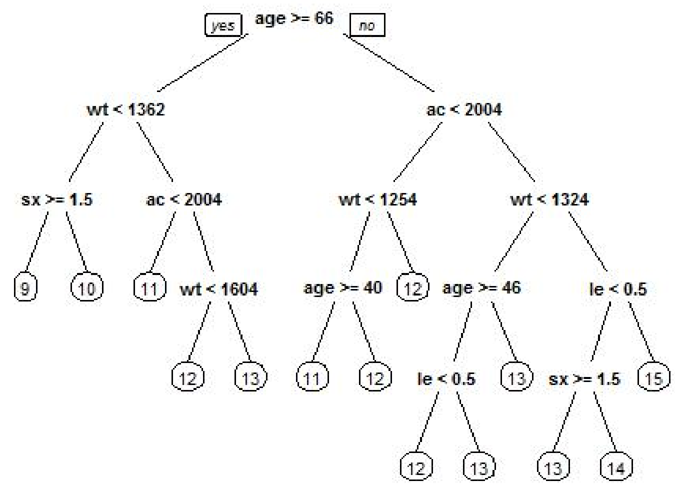
Issue in insurance

All this sounds rather simple and seems to have been done already...

... we highlight some issues in insurance pricing and reserving.

Classification and regression tree (CART)





- CART from R package `rpart` on motor insurance data.
- Crucial is the appropriate tree size: statistics and predictive power?

Claims frequencies in general insurance (1/2)

- Particular difficulty in general insurance claims frequency modeling:
  $\Longrightarrow$ yearly claims frequencies are (very) low.

- Assume that the number of claims N of a given policy

$$N \sim \text{Poisson}(\lambda v),$$

  with yearly claims frequency $\lambda = 9\%$ and yearly exposure at risk $v = 1$.

- This policy has a standard deviation (pure randomness) of

$$\text{Var}(N)^{1/2} = \sqrt{9\%} = 30\%.$$

- We need $900 = 30^2$ such independent (and identically) distributed policies to see a structural difference to a yearly claims frequency of $\lambda' = 8\%$ (for confidence bounds of 1 standard deviation)!
  $\triangleright$ Number of features x_i that explain $\lambda = \lambda(x_1, \ldots, x_p)$?

Claims frequencies in general insurance (2/2)

- Separation of systematic from random component is difficult for low frequency problems: analyze (1) bias^2, (2) estimation variance, (3) process variance

$$\mathbb{E}\left[\left(N-\widehat{\lambda}v\right)^2\right]=\left(\lambda v-\mathbb{E}[\widehat{\lambda}]v\right)^2+v^2\ \mathrm{Var}\left(\widehat{\lambda}\right)+\mathrm{Var}(N).$$

- Methods are often not robust against small changes in data:
  - ⋆ problematic in pricing process due to yearly changes in premium;
  - ⋆ problematic for accounting figures, for instance, in claims reserving;
  - ⋆ maybe fine in marketing and airfare pricing.

- Classical features may have a smoothing effect, improve predictive power and support interpretation.

▷ Use hybrid methods that combine classical features with big data features.

Practical experience about telematic (big) data

- Quality of data is often (disappointing) poor:
  - ⋆ installed devices do not work properly;
  - ⋆ frequency of data submission is not sufficient (and costly);
  - ⋆ volume of portfolio is not sufficient because insurance claims have low frequency.
- Data cleaning is costly and extensive (uses 80%+ of the time).
- Graphical tools are missing/poor.
- RAM needs to be large.

☹ Availability of data for research is an issue: data has a value.

☹ Tools and techniques of professional providers are not accessible.

Claims reserving

- Claims reserving is a dynamic process and features may also be dynamic.
- Often there is a lot of judgment and human behavior involved.
- Predictions should be robust (accounting figures).
- Non-stationarity, cyclicality and dependence is an issue in the data:
  - ⋆ emergence of new phenomena like whiplash claims;
  - ⋆ culture of claims handling units;
  - ⋆ dependence on economic factors, for instance, related to mental diseases;
  - ⋆ legal changes; local holidays, etc.

General insurance company's balance sheet

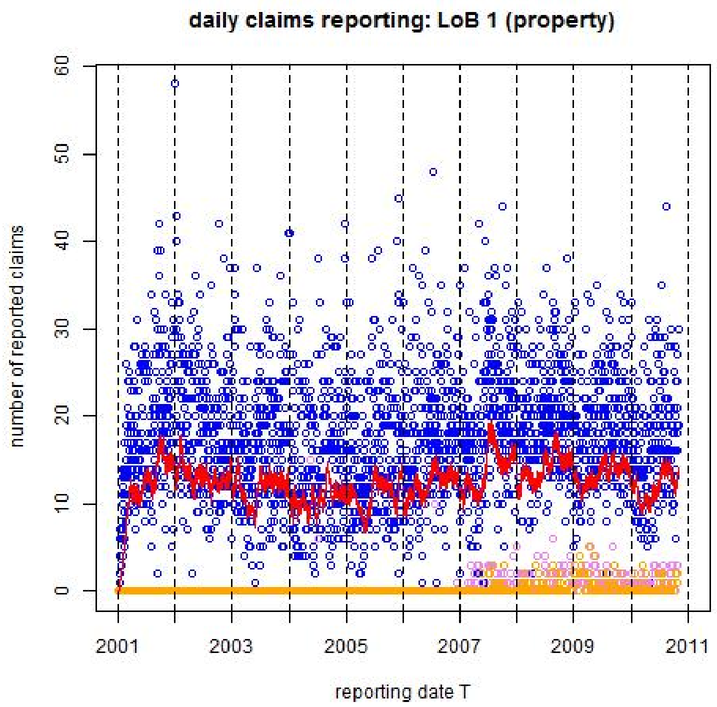
assets as of 31/12/2013	mio. CHF
debt securities	6'374
equity securities	1'280
loans & mortgages	1'882
real estate	908
participations	2'101
short term investments	693
other assets	696
total assets	13'934

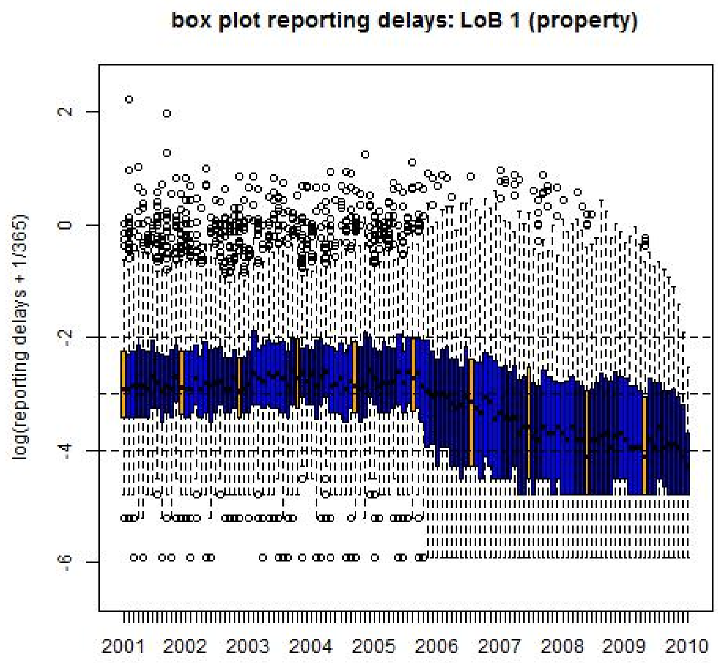
liabilities as of 31/12/2013	mio. CHF
claims reserves	**7'189**
provisions for annuities	1'178
other liabilities and provisions	2'481
share capital	169
legal reserve	951
free reserve, forwarded gains	1'966
total liabilities & equity	13'934

Claims reserves are the biggest position on the balance sheet of a general insurance company!

Source: Annual Report 2013 of AXA-Winterthur Versicherungen AG

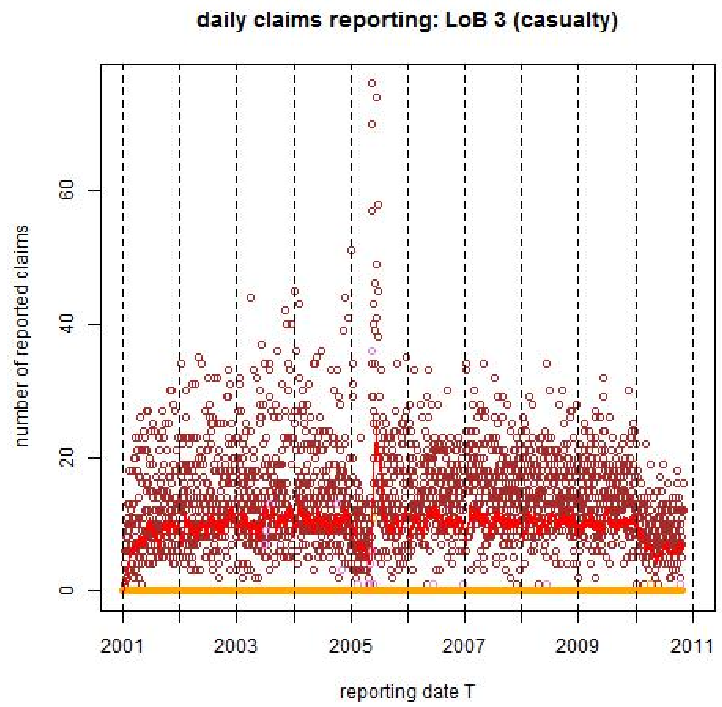
Non-stationarity in claims reporting

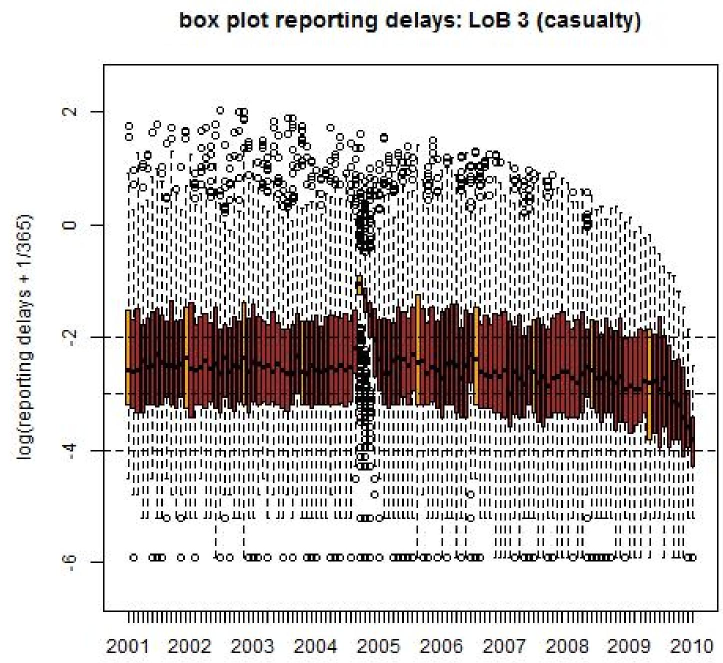




▷ Smarter things than aggregated claims reserving triangles...

Data issues in claims reporting





▷ Smarter things than aggregated claims reserving triangles...

How can the R community support actuaries?

- Develop new (and better) graphical tools.
- Provide support in education of data analytics and statistics.
- Run-time and cost efficient coding and procedures (RAM is an issue).
- Better documentation of the R packages:
  many packages do not run the described algorithms...
- Insurance adapted illustrations and examples.
- Open access data to develop and test tools.

Do not miss the connection, the change has already started!