

A fit-for-purpose model selection framework for stochastic mortality modelling

Philipp Schulz – joint work with Arne Freimann and Jochen Ruß

- 21st International Longevity Risk and Capital Market Solutions Conference
- Rome, Italy
- September 2026



Motivation

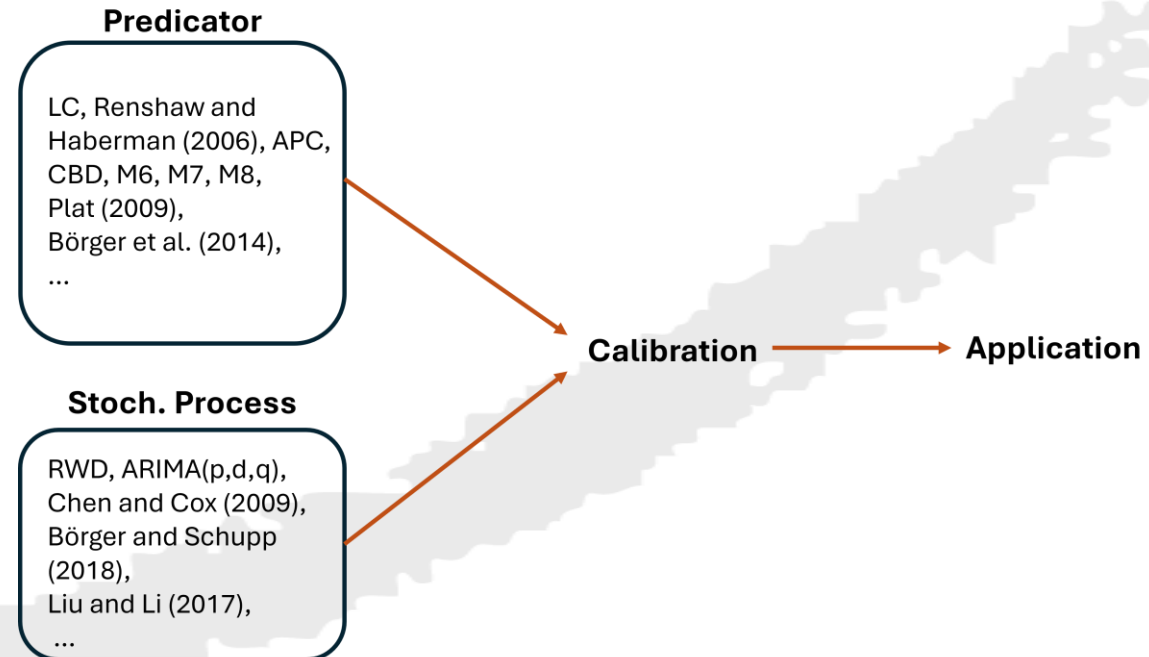
A wide variety of mortality models exists

- Numerous stochastic mortality models (like Cairns-Blake-Dowd (CBD) or Lee-Carter (LC) model)

- are used to address a wide range of applications, but

- selecting a suitable model is not straightforward since models can differ in

- predictor structure
 - stochastic process used for the extrapolation
 - identifiability constraints
 - calibration technique
 - [...]



- Model risk further intensifies this problem since

- statistical fit alone is an insufficient criterion (cf. Cairns et al., 2011),

- and model choice can affect pricing and capital requirements significantly, see e.g. Richards et al. (2020).

Motivation

Limitations of traditional approaches for model selection

- Data-driven model selection has known limitations:
 - Good historical fit does not ensure plausible forecasts, cf. Cairns et al. (2011).
 - Results depend on the specific data and statistical criterion, cf. Dowd et al. (2010a) and Dowd et al. (2010b).
 - Backtesting does not guarantee robustness to future structural breaks (e.g. medical breakthroughs).
- Recently, model averaging has come into focus, see e.g. Barigou et al. (2023) and Bimonte et al. (2025).
 - Idea: reduce the risk of selecting a “wrong” model by combining forecasts from multiple models
 - However, model averaging is limited by the candidate set:
 - no information about the suitability of the models in the candidate set for the intended application
 - may give the illusion that model risk has been reduced
- Cairns et al. (2009, 2011) discuss relevant properties of what makes a stochastic mortality model “good”. However, they treat these properties independently of the specific application.



Classical approaches usually aim to select the universally “best” model given the data **independently of the concrete application.**

In contrast, we aim to determine which models are **best suited for the specific application.**

Motivation

An illustrative example (Part 1)

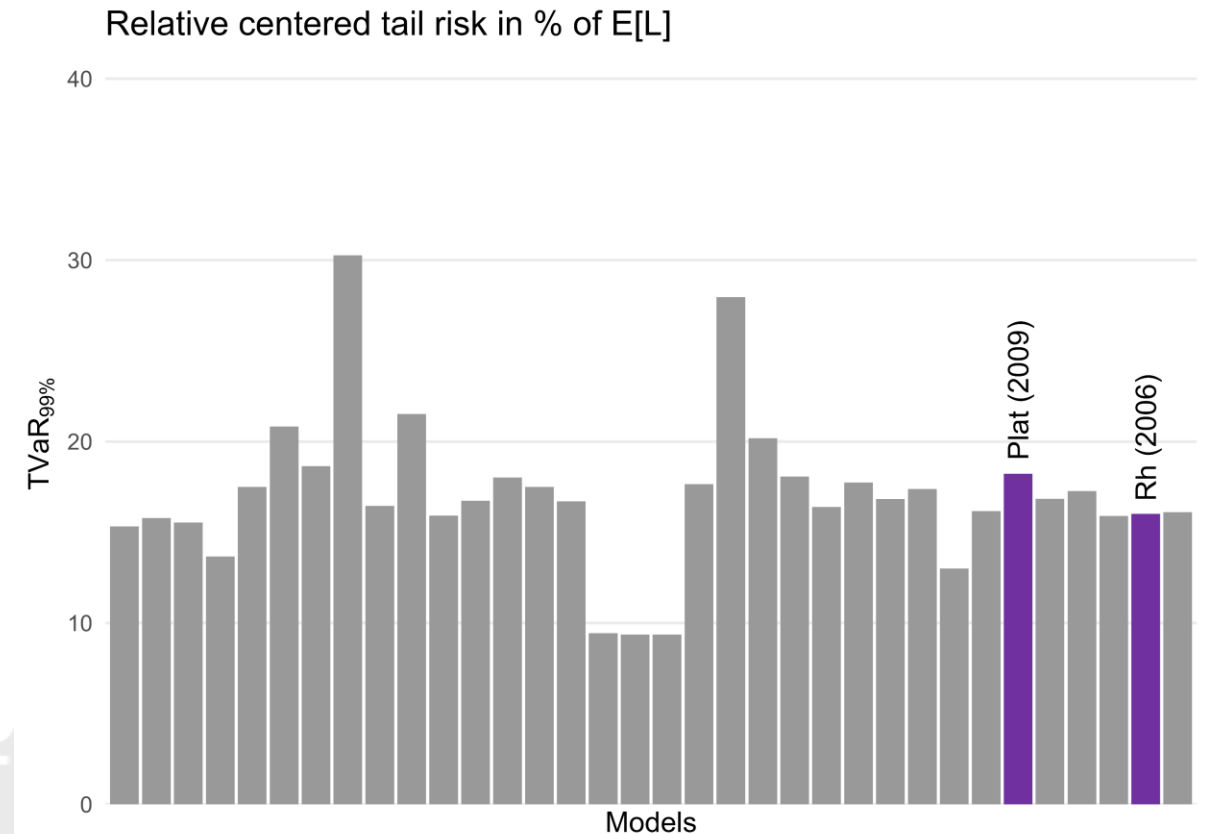
Example

- A relevant question might be: How big is the tail risk of the liabilities in a runoff portfolio of deferred annuities in the worst 1% of the outcomes?
- We analyse the distribution of centered present value of future liabilities, i.e.

$$L - E[L]$$

measured in terms of $TVaR(99\%)$.

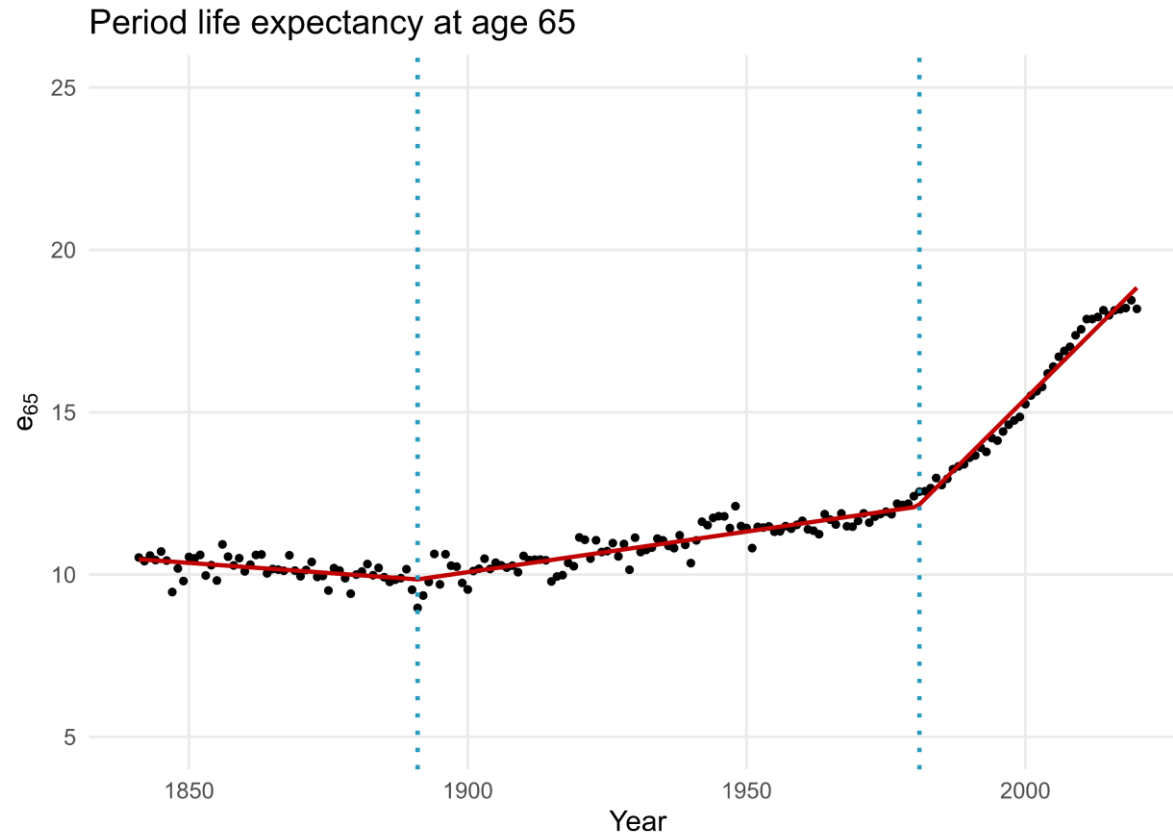
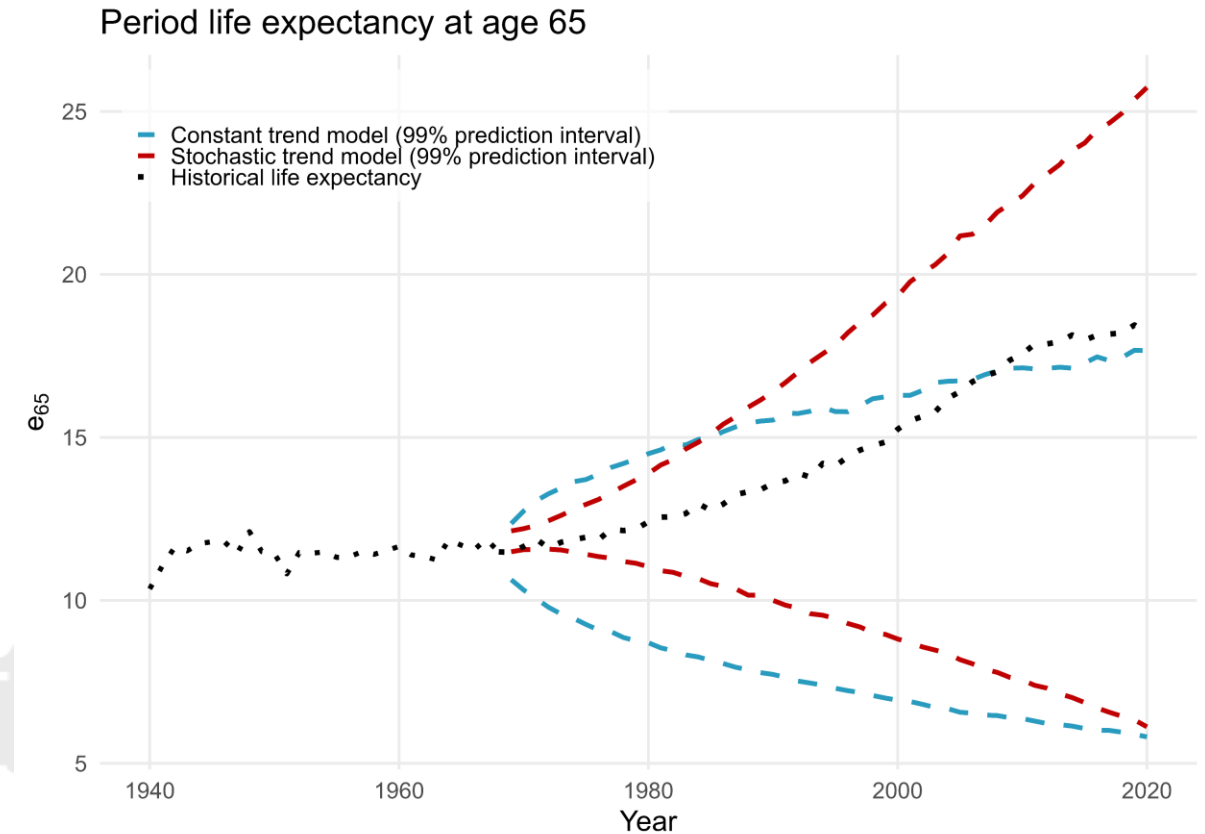
- Model choice by traditional approaches:
 - Bayesian Information Criterion (BIC) → Plat (2009)
 - Mean Squared Error (MSE) → Renshaw and Haberman (2006)



The two models selected by traditional approaches yield similar results, suggesting relatively low model risk.

Motivation

Digression: Trend change risk as a major risk driver in long-term mortality modelling



Motivation

An illustrative example (Part 2)

- A relevant question might be: How big is the tail risk of the liabilities in a runoff portfolio of deferred annuities in the worst 1% of the outcomes?
- We analyse the distribution of centered present value of future liabilities, i.e.

$$L - E[L]$$

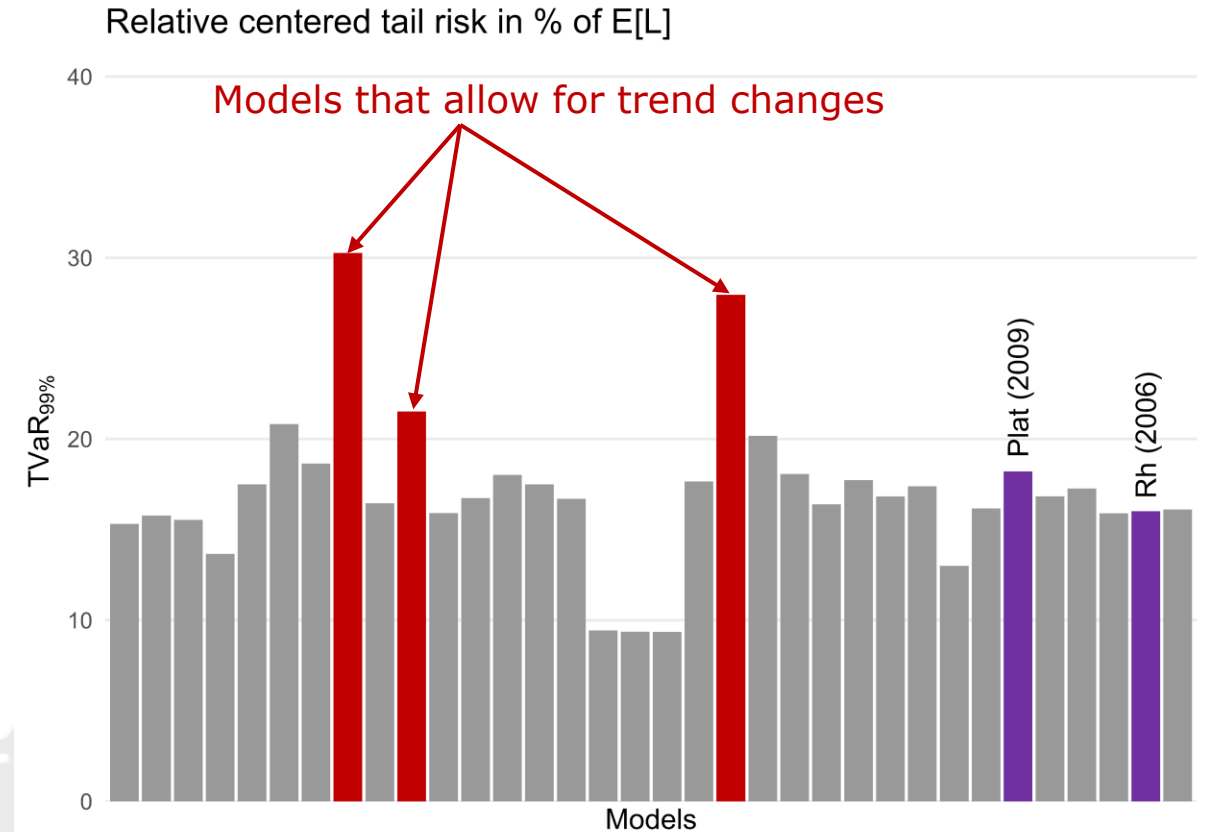
measured in terms of $TVaR(99\%)$.

- Model choice by traditional approaches:

- **BIC** → Plat (2009)

- **MSE** → Renshaw and Haberman (2006)

- Since this runoff perspective requires the projection of mortality rates in the far future, a suitable model should **allow for mortality trend changes**.



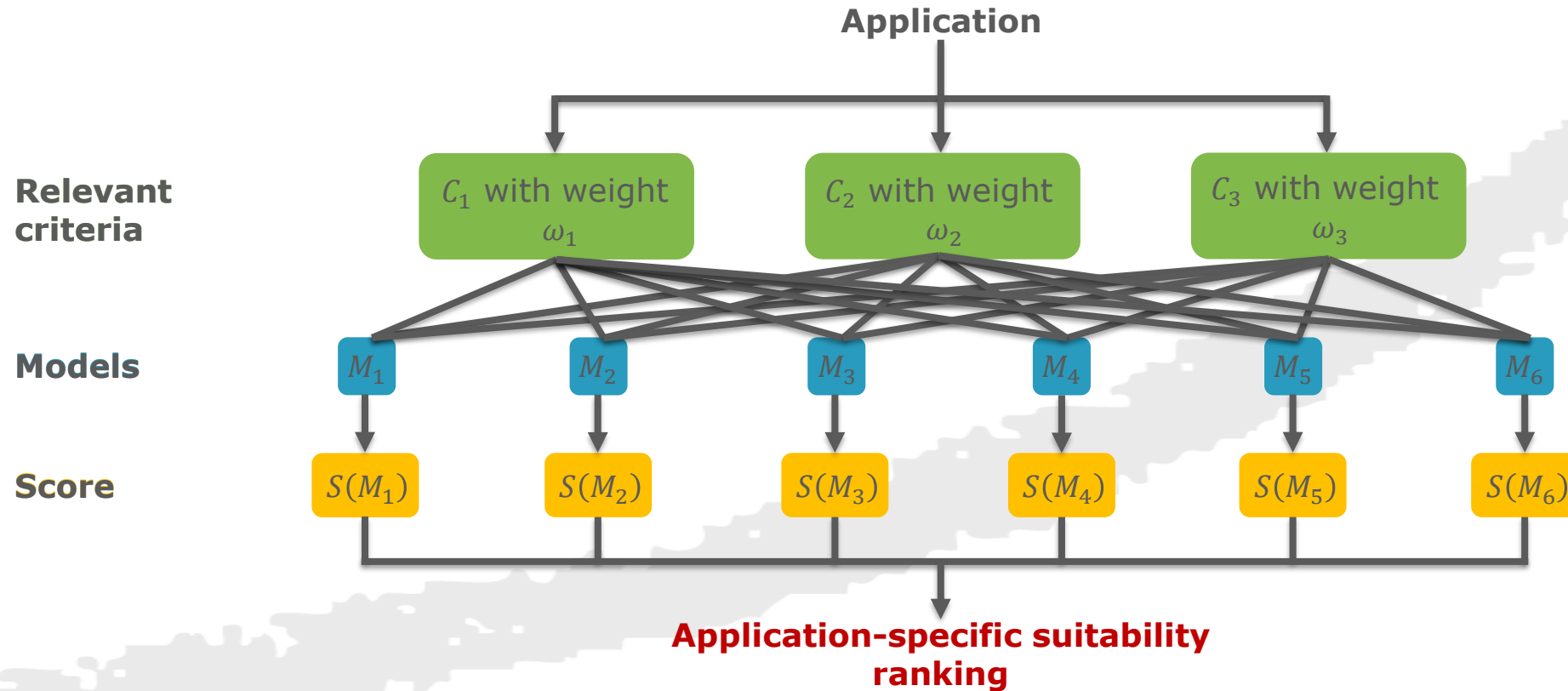
Model selection should account for application-specific risk drivers. Otherwise, selected models may have a good statistical fit but are not **“fit-for-purpose”**.

→ **Application-driven model selection**

An application-driven model selection framework

When is a model “fit-for-purpose”?

- Idea: Any application induces requirements (essential, important,..., nice-to-have) which a model should fulfil.
 - If an essential requirement is not fulfilled, the model should not be considered.
 - We rank models according to how well they fulfil these requirements.



Application 1: Tail risk assessment of a longevity book

Example

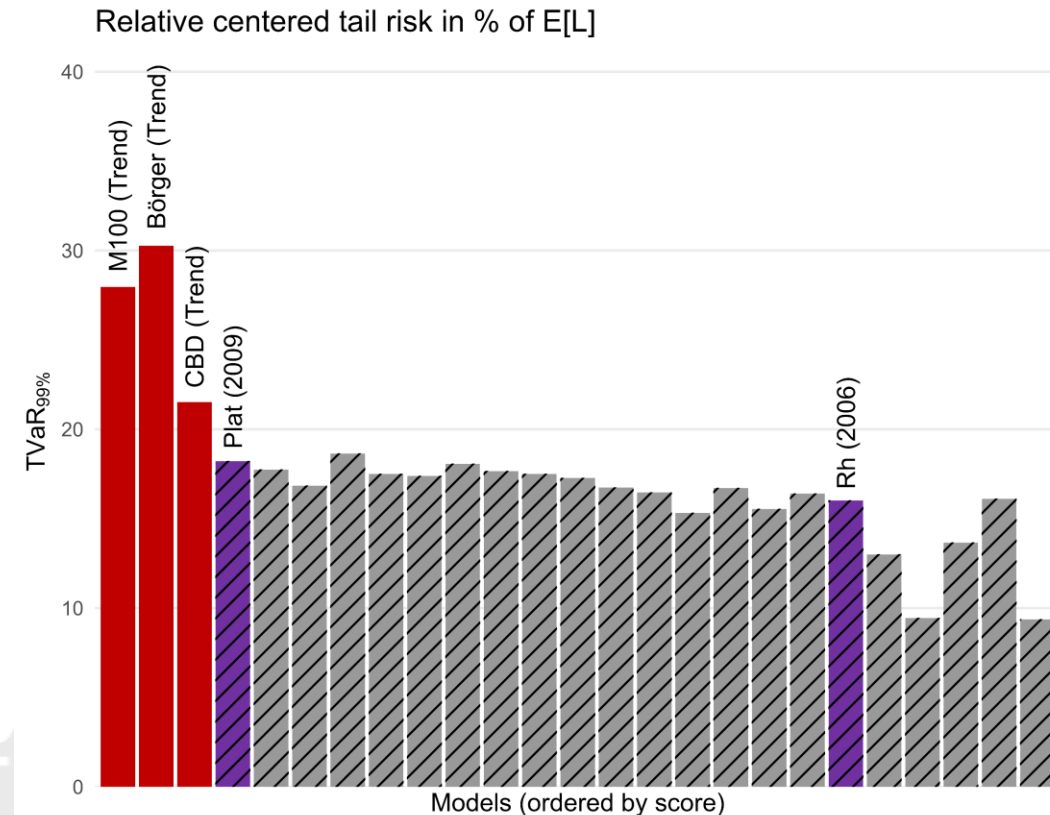
- Consider a stylised reference insurer with closed portfolio of deferred lifelong annuities:
 - The portfolio has been built over the past 30 years by selling the same type of contract to the same number of customers.
 - Retirement age 65, deferral period of 15 years , 2% interest rate
- Return to previous question: How big is the tail risk of the liabilities in a runoff portfolio of deferred annuities in the worst 1% of the outcomes?
 - Consider the $TVaR(99\%)$ of the centered present value of future liabilities, i.e. the $TVaR(99\%)$ of
$$L - E[L]$$
- Analysing the problem:
 - $TVaR(99\%)$ in the (long-term) runoff perspective is **dominated by trend uncertainty**, not short-term volatility, see e.g. Börger and Schupp (2018). We should account for
 - potential **changes of the mortality trend in the future**,
 - **uncertainty in the starting trend** and in the **trend parameters** (cf. Cairns et al., 2006 and Börger et al., 2021).
 - The distribution of the liability depends on the **joint development of future mortality rates** across all ages ≥ 50 .
 - Failure to **separate cohort and period effects** may move cohort patterns into period components.

Application 1: Tail risk assessment of a longevity book

Criteria and model ranking

■ Relevant criteria:

- Essential:
 - allowing for future mortality trend changes
 - accounting for uncertainty in the starting trend and in trend parameters
- Important:
 - robustness of model parameters to (very) small changes in the calibration period
 - non-trivial correlation of mortality improvements across ages
 - no crossover mortality rates (biological plausibility)
 - ability to separate cohort and period effects
 - goodness-of-fit for ages ≥ 50
- Nice-to-have: transparency



The models deemed most suitable from a technical perspective yield results that are very different from those produced by “random” or “best fit”.

Application 2: Best-estimate projection over one year

Example

- Consider the previous portfolio of deferred annuities with retirement age 65
- In response to a management enquiry at t_0 , an actuary is asked to estimate the expected number of cashflow-relevant deaths in the deferred annuity portfolio during the next year.
- The metric of interest is thus

$$E[D] = \sum_{x=65}^{\omega-1} N_{annuity_x} \cdot E[q_{x+1,t_0+1}|F_{t_0}]$$

where $N_{annuity_x}$ is the size of the cohort aged x at time t_0 and ω is the maximum age.

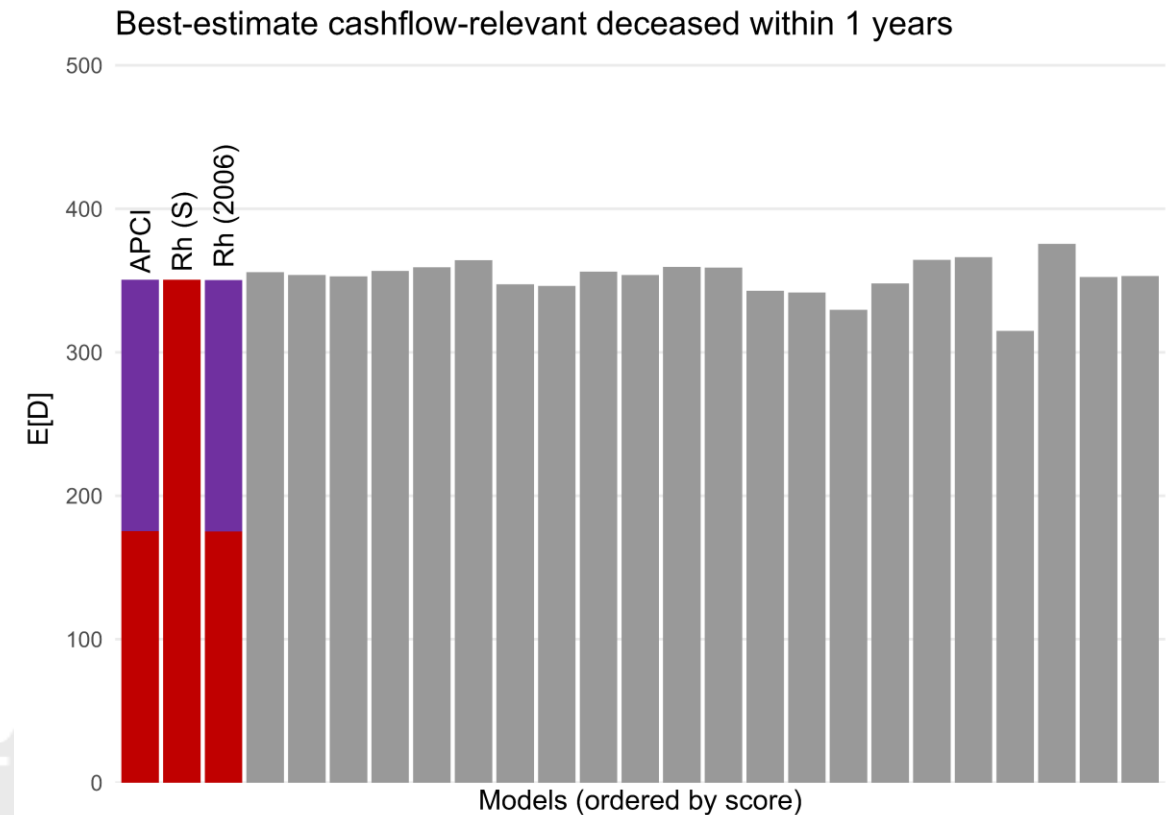
- Analysing the problem:
 - $N_{annuity_x} = 0$ for $x > 79$. Extremely old ages are thus not relevant for this application.
 - For one-year projections, the fit at **the jump-off has a greater impact than the extrapolated trend**.
 - For projections over one year, one would expect a **steady continuation of the recent mortality trend**.
 - To have confidence in $E[D]$, $E[q_{x+1,t_0+1}|F_{t_0}]$ should be **stable to small changes in the calibration period**.
 - Only the behaviour of the simulated rates in year $t_0 + 1$ is relevant for $E[D]$. For projections over one year, **cross-over mortality rates seem implausible**.

Application 2: Best-estimate projection over one year

Criteria and model ranking

■ Relevant criteria:

- good in-sample fit at the right edge of the calibration window
- smooth transition between historical and projected mortality improvement rates
- mortality rates should increase with age for relevant ages
- robustness of model parameters to (very) small changes in the calibration period
- transparency



The models ranked highly by the framework all have a good in-sample fit at the jump-off and extrapolate the latest mortality trend well. Consequently, their results are very similar.

Application 3: Maximising diversification benefits

Example

- Reference insurer with two closed portfolios:
 - The previous portfolio of deferred annuities and
 - a term life portfolio
 - sold to individuals aged 35,
 - a contract duration of 30 years
 - and 2% interest rate.
- Question: How many term life contracts should the insurer have to maximise the diversification benefit with respect to the $TVaR_{95\%}$ between the two books, i.e. $Hedge := 1 - \frac{TVaR_{95\%}(L_{annuity} + L_{term})}{TVaR_{95\%}(L_{annuity}) + TVaR_{95\%}(L_{term})}$?
- Analysing the problem:
 - Hedge is a functional of the joint tail of $(L_{annuity}, L_{term})$. The numerator is driven directly by the model-implied **dependence between mortality improvements across ages**.
 - Analogous to application 1, the annuity tail is dominated by **persistent, compounding deviations in long-term improvement rates**.
 - The effect of **transitory mortality shocks** does not accumulate over long time horizons, their impact on the annuity tail is negligible.
 - Failure to **separate cohort and period effects** may move cohort patterns into period components.

Application 3: Maximising diversification benefits

Criteria and model ranking

■ Relevant criteria:

■ Essential:

- non-trivial correlation of mortality improvements across ages

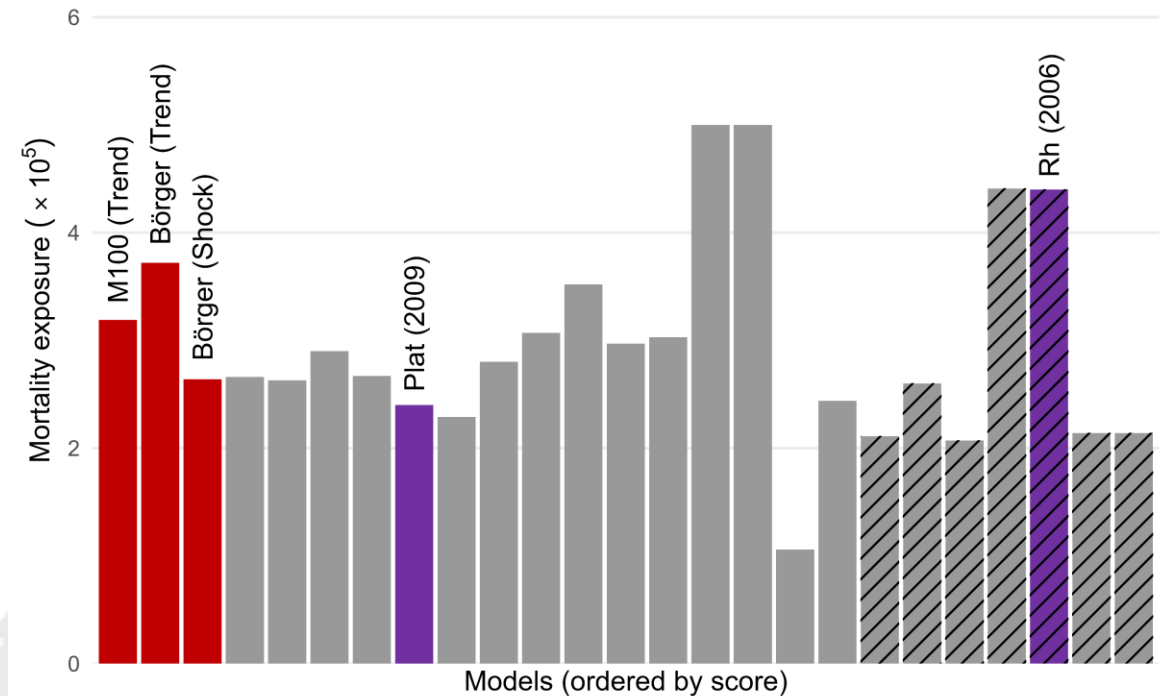
■ Important:

- allowing for changes in the underlying mortality trend
- ability to separate cohort and period effects
- goodness-of-fit for ages ≥ 35
- no crossover mortality rates (biological plausibility)
- robustness of model parameters to (very) small changes in the calibration period

■ Nice-to-have:

- ability to generate transitory mortality jumps
- transparency

Optimal mortality exposure maximising the diversification benefit



Highly ranked models allow for more complex correlation structures. Furthermore, the highest-ranked models are designed for risk analysis over long-time periods and have a good fit for all ages.

Conclusion

Summary and possible extensions

- Application-specific criteria often play a minor role in model selection. This can result in the chosen model not being “fit-for-purpose”.
- We therefore propose an application-driven model selection framework for stochastic mortality modelling.
 - Idea: Every application induces a set of criteria that a model suited for it should fulfil.
 - Ranking these criteria by relevance and scoring models produces an **application-specific model ranking**.
- We illustrate the framework on three different actuarial applications.
 - Each application has **different relevant criteria**, resulting in **different models** being considered most suitable for **different applications**.
 - The idea of a **universally “best” model is rejected** in this framework.
- Our framework reduces the risk of biased results by applying models that lack essential properties for the specific application.
 - In some sense, this can be seen as the **avoidable part of model risk**.
 - However, this also means that the **framework is only as good as the individual applying it**.
- Possible extensions:
 - applying the framework to other model classes
 - conducting a systematic robustness study where the selected models under multiple criteria weighting methods are compared

References

- Barigou, K., Goffard, P., and Salhi, Y. (2023). Bayesian model averaging for mortality forecasting using leave-future-out validation. *International Journal of Forecasting*, 39(2):674-690.
- Bimonte, G., Russolillo, M., Shang, H. L., and Yang, Y. (2025). Mortality model ensemble via Shapley value. *Decision in Economics and Finance*, 48:1131-1159.
- Börger, M., Fleischer, D., and Kuksin, N. (2014). Modeling the mortality trend under modern solvency regimes. *ASTIN Bulletin*, 44(1):1-38.
- Börger, M., and Schupp, J. (2018). Modeling trend processes in parametric mortality models. *Insurance: Mathematics and Economics*, 78:369-380.
- Börger, M., Russ, J., and Schupp, J. (2021). It takes two: Why mortality trend modelling is more than modelling one mortality trend. *Insurance: Mathematics and Economics*, 99:222-232.
- Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1-25.
- Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., and Khalaf-Allah, M. (2011). Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, 48(3):355-367.
- Chen, H., and Cox, S. H. (2009). Modeling mortality with jumps: Applications to mortality securitization. *Journal of Risk and Insurance*, 76(3):727-751.
- Dowd, K., Cairns, A. J., Blake, D., Coughlan, G. D., Epstein, D., and Khalaf-Allah, M. (2010a). Backtesting stochastic mortality models: An ex post evaluation of multiperiod-ahead density forecasts. *North American Actuarial Journal*, 14:281-298.
- Dowd, K., Cairns, A. J., Blake, D., Coughlan, G. D., Epstein, D., and Khalaf-Allah, M. (2010b). Evaluating the goodness of fit of stochastic mortality models. *Insurance: Mathematics and Economics*, 47:255-265.
- Liu, Y., and Li, J. S-H. (2017). The locally linear Cairns-Blake-Dowd model: A note on delta-nuga hedging of longevity risk. *ASTIN Bulletin*, 47(1):79-151.
- Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45:393-404.
- Renshaw, A. E., and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, 38:556-570.
- Richards, S. J., Currie, I. D., Kleinow, T., and Ritchie, G. P. (2020). Longevity trend risk over limited time horizons. *Annals of Actuarial Science*, 14:262-277.

Institute for Finance and Actuarial Services (ifa)

Contact information

Philipp Schulz
+49 (731) 20 644-268
p.schulz@ifa-ulm.de

