Bayesian model averaging for mortality forecasting using leave-future-out validation

Joint Section Virtual Colloquium - IAA

Karim Barigou, Pierre-Olivier Goffard, Stéphane Loisel, Yahia Salhi

October 12, 2021

ISFA

Motivation

Mortality forecasting is subject to

- model uncertainty
- parameter uncertainty
- overdispersion

Currie (2016):

> *The RH model, with its large number of parameters, is easily the best fitting model but provides an extreme illustration of the fact that model fit on its own is not a sound basis for forecasting.*

Model selection is often based on
goodness-of-fit criterion

Motivation

Mortality forecasting is subject to

- model uncertainty → Model averaging
- parameter uncertainty → Bayesian
- overdispersion → Negative-Binomial

Currie (2016):

> *The RH model, with its large number of parameters, is easily the best fitting model but provides an extreme illustration of the fact that model fit on its own is not a sound basis for forecasting.*

Model selection is often based on
goodness-of-fit criterion → Out-of-sample validation.

This presentation

- Proposes two approaches of averaging multiple mortality models using **stacking** and **Pseudo-BMA**.
  1. A full Bayesian approach which takes model uncertainty, parameter uncertainty and overdispersion into account.
  2. We adapt the leave-one-out framework of Yao et al. (2018) to leave-future-out for mortality forecasting.
  3. Weights are assigned by minimising a leave-future-out validation criterion rather than a goodness-of-fit criterion (such as AIC or BIC).
  4. Based on a simulation study and real-data applications, stacking and Pseudo-BMA outperform the standard Bayesian model averaging based on marginal likelihood.
- Discusses the R package **StanMoMo**[1] for Bayesian mortality modelling, forecasting and averaging.

[1] `https://kabarigou.github.io/StanMoMo/`

Negative-Binomial model

Deaths at age x and time t follow the dynamics

$$D_{x,t} \mid \mu_{x,t} \overset{\text{ind}}{\sim} \text{Poisson}(\mu_{x,t} e_{x,t})$$

$$\log \mu_{x,t} = \alpha_x + \sum_{i=1}^{p} \beta_x^{(i)} \kappa_t^{(i)} + \beta_x^{(0)} \gamma_{t-x} + \log \nu_{x,t}$$

$$\nu_{x,t} \mid \phi \overset{\text{ind}}{\sim} \text{Gamma}(\phi, \phi),$$

Mortality model	Predictor $\log \mu_{x,t}$
Lee-Carter (LC)	$\alpha_x + \beta_x \kappa_t^{(1)}$
Renshaw-Haberman (RH)	$\alpha_x + \beta_x \kappa_t^{(1)} + \gamma_{t-x}$
Age-Period-Cohort (APC)	$\alpha_x + \kappa_t^{(1)} + \gamma_{t-x}$
Cairns-Blake-Dowd (CBD)	$\kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)}$
M6	$\kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)} + \gamma_{t-x}$

Negative-Binomial model

Deaths at age x and time t follow the overdispersed dynamics

$$
\begin{aligned}
D_{x,t} \mid \mu_{x,t} &\overset{\text{ind}}{\sim} \text{Poisson}(\mu_{x,t} e_{x,t}) \\
\log \mu_{x,t} &= \alpha_x + \sum_{i=1}^{p} \beta_x^{(i)} \kappa_t^{(i)} + \beta_x^{(0)} \gamma_{t-x} + \log \nu_{x,t} \\
\nu_{x,t} \mid \phi &\overset{\text{ind}}{\sim} \text{Gamma}(\phi, \phi),
\end{aligned}
$$

Mortality model	Predictor $\log \mu_{x,t}$
Lee-Carter (LC)	$\alpha_x + \beta_x \kappa_t^{(1)}$
Renshaw-Haberman (RH)	$\alpha_x + \beta_x \kappa_t^{(1)} + \gamma_{t-x}$
Age-Period-Cohort (APC)	$\alpha_x + \kappa_t^{(1)} + \gamma_{t-x}$
Cairns-Blake-Dowd (CBD)	$\kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)}$
M6	$\kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)} + \gamma_{t-x}$

Bayesian inference

Bayes' theorem: The posterior distribution is

$$p(\theta|y) = \frac{l(y|\theta)p(\theta)}{l(y)},$$

where

- $p(\theta)$ is the prior.
- $l(y|\theta)$ is the Negative-Binomial likelihood.
- $l(y)$ is the marginal likelihood.

The posterior distribution is obtained via Markov Chain Monte Carlo (MCMC) sampling:

$$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)} \sim p(\theta|y) \propto l(y|\theta)p(\theta).$$

We use **Stan**, an efficient Hamiltonian Monte Carlo (HMC) sampler available through the **rstan** package (Carpenter et al. 2017).

Bayesian model averaging

We have K models $\mathcal{M} = (M_1, \ldots, M_K)$:
$\rightarrow$ How to find the optimal combination weights w_k ?

$$D_{x,t} \sim \sum_{k=1}^{K} w_k \text{Poisson}(\mu_{x,t}^k \cdot e_{x,t})$$

- Standard approach: BMA by marginal likelihood:

$$w_k = p\left(M_k \mid y\right) = \frac{p\left(y \mid M_k\right) p\left(M_k\right)}{\sum_{k=1}^{K} p\left(y \mid M_k\right) p\left(M_k\right)},$$

where

- $p\left(M_k\right)$ is the prior model probabilities. We set $p\left(M_k\right) = 1/K$.
- $p\left(y \mid M_k\right)$ is the marginal likelihood, which can be estimated via bridge sampling.

Drawbacks: in-sample criterion, strongly prior-sensitive and flawed if the "true" model is not a model candidate.

Bayesian model averaging by leave-future-out

We consider two approaches based on leave-future-out validation:

- Data is split into two parts:
  - Training set: First N years.
  - Validation set : Last M years.
- For stacking, we maximize a log scoring rule on the validation set:

$$\max_{w} \sum_{x=x_1}^{x_n} \sum_{j=t_{N+1}}^{t_{N+M}} \log \sum_{k=1}^{K} w_k p\left(d_{x,j} \mid y_{1:N}, M_k\right)$$

where $p\left(d_{x,j} \mid y_{1:N}, M_k\right)$ is the predictive density given the training set $y_{1:N}$ and model M_k.

Bayesian model averaging by leave-future-out

We consider two approaches based on leave-future-out validation:

- Data is split into two parts:
  - Training set: First N years.
  - Validation set : Last M years.
- For Pseudo-BMA, we consider an AIC-type weighting scheme:

$$w_k = \frac{\exp\left(\text{elpd}^k\right)}{\sum_{k=1}^{K} \exp\left(\text{elpd}^k\right)}.$$

where

$$\text{elpd}^k = \sum_{x=x_1}^{x_n} \sum_{j=t_{N+1}}^{t_{N+M}} \log p\left(d_{x,j} \mid y_{1:N}, M_k\right)$$

is the expected log predictive density (Vehtari et al. 2017).

Simulation study

- **Question:** Which model is selected if the data is generated from
  - A specific model, e.g. the APC model ?
  - A mixture of models, e.g. the CBD/RH models ?
- Training: 20, 30, 40, 50 years. Validation: 10 years (2000-2009). Prediction: 10 years (2010-2019). **Data:** Belgian male mortality.
- Assessment:
  - Number of times each model has the highest weight.
  - Mean Absolute Error on the prediction set:

$$\text{MAE} = \frac{1}{41} \sum_{x=50}^{90} \frac{1}{10} \sum_{t=2010}^{2019} |d_{x,t} - e_{x,t} \widehat{\mu}_{xt}|$$

Simulation study

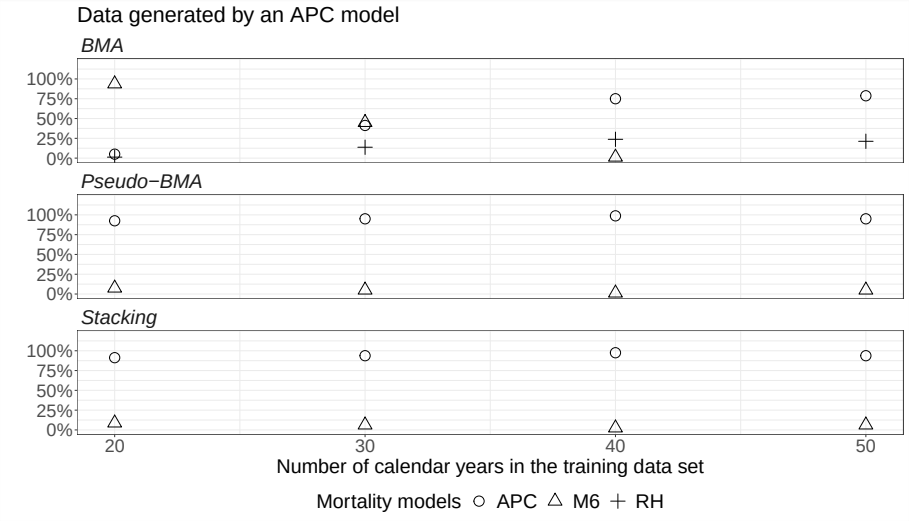


Figure 1: How many times each mortality model is selected out of 80 synthetic data sets. → BMA tends to select M6 with 20 and 30 years of training data

Simulation study

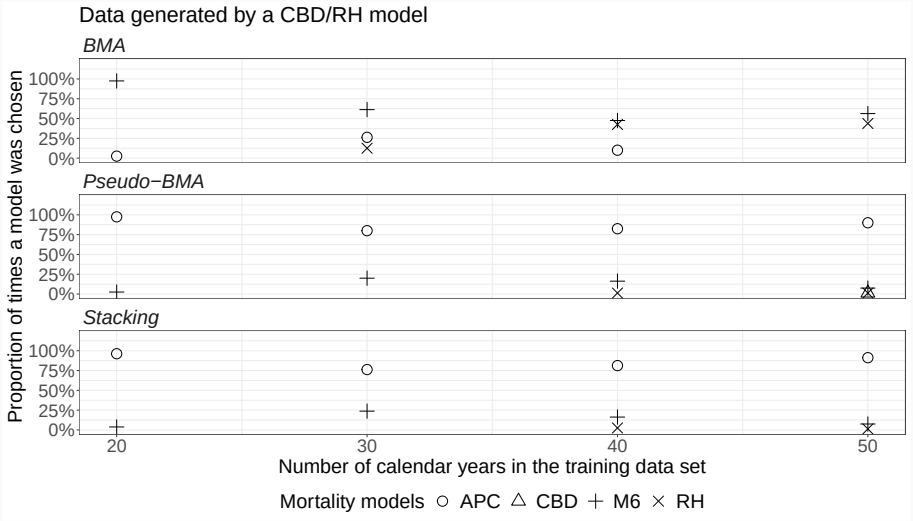


Figure 2: How many times each mortality model is selected out of 80 synthetic data sets. → Stacking and Pseudo-BMA select APC while BMA favors M6

Simulation study

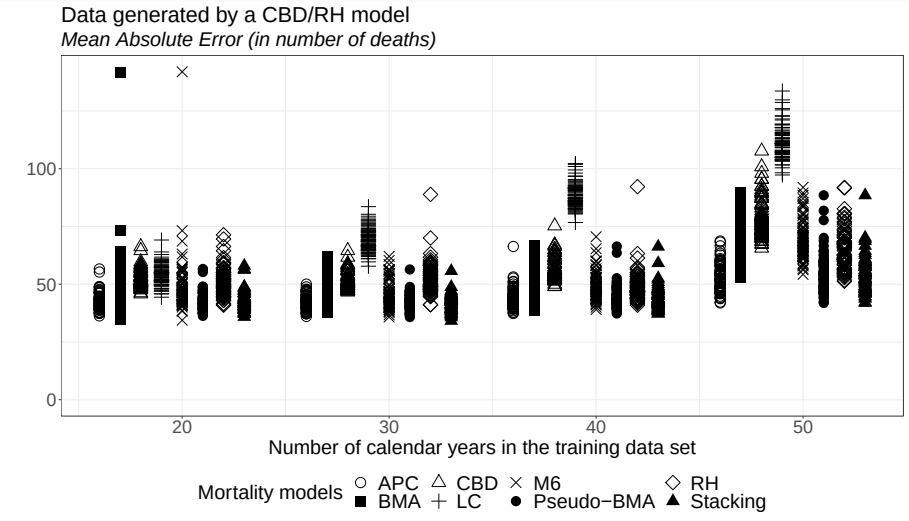


Figure 3: Mean Absolute Error over 80 data sets. → APC appears as the leading model.

Performance of BMA for different countries

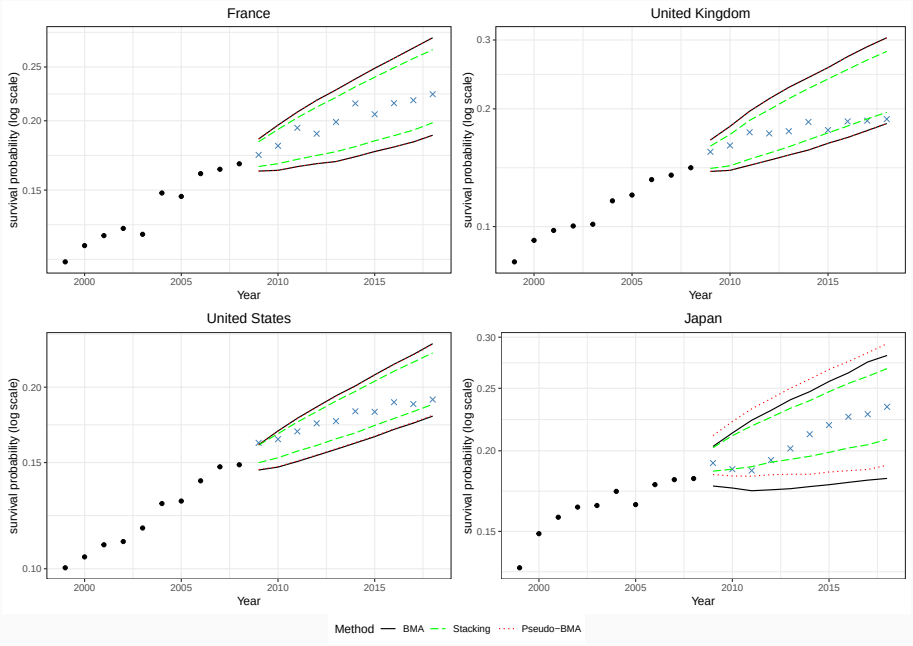
- Countries: France, UK, USA and Japan. Ages: 50-90.
- Assessment:
  - 95% prediction intervals of log death rates.
  - Survival probability at age 50 until age 90.
  - Mean Absolute Error on the holdout data.

<table>
<tr><th></th><th>1979-1998</th><th>1999-2008</th><th>2009-2018</th></tr>
<tr><td>BMA</td><td colspan="2">Fitting</td><td rowspan="3">Prediction</td></tr>
<tr><td>Stacking</td><td>Fitting</td><td>Validation</td></tr>
<tr><td>Pseudo-BMA</td><td>Fitting</td><td>Validation</td></tr>
</table>

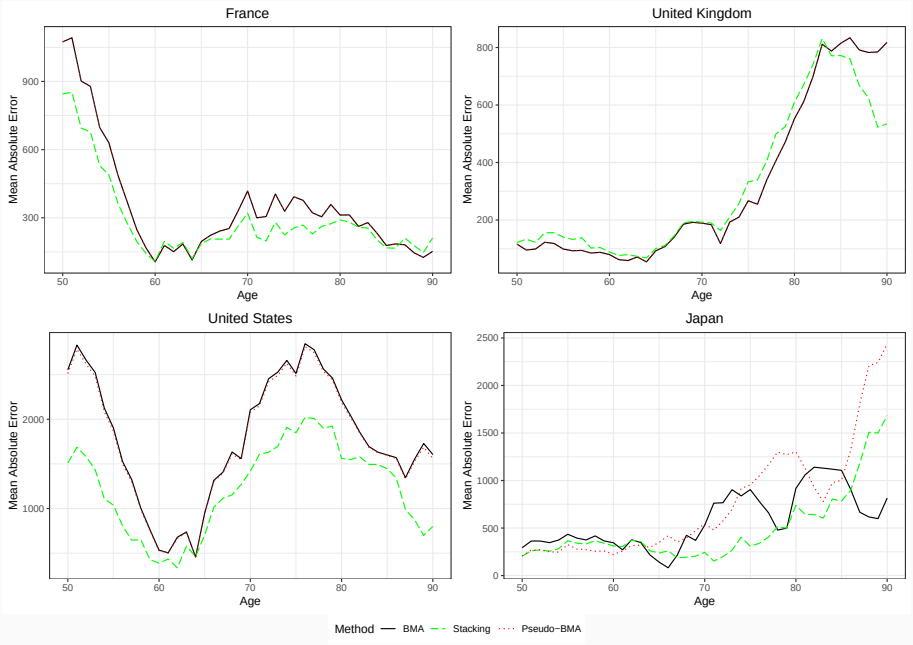
Performance of BMA for different countries: Weights

	France			UK		
	BMA	Stacking	Pseudo-BMA	BMA	Stacking	Pseudo-BMA
LC	0	0.093	0	0	0	0
RH	1	0.750	1	0	0.342	0
APC	0	0.157	0	0	0	0
CBD	0	0	0	0	0	0
M6	0	0	0	1	0.658	1
	USA			Japan		
	BMA	Stacking	Pseudo-BMA	BMA	Stacking	Pseudo-BMA
LC	0	0	0	0	0.292	0
RH	1	0.548	0.982	1	0.239	0
APC	0	0.452	0.018	0	0.468	1
CBD	0	0	0	0	0	0
M6	0	0	0	0	0	0

Performance of BMA for different countries: survival probability



Performance of BMA for different countries: MAE



Performance of BMA for different countries: Scoring rules

- Scoring rules are functions $S(F, y)$ that evaluates the accuracy of a forecast distribution F, given an outcome y was observed.
- We consider the logarithmic score and the continuous ranked probability score (CRPS), respectively given by

$$\text{LogS}(F, y) = -\log(f(y))$$
$$\text{CRPS}(F, y) = \int_{\mathbb{R}} (F(z) - \mathbb{1}\{y \leq z\})^2 \, \mathrm{d}z$$

  where F admits a PDF f.
- Based on our MCMC samples, we can estimate the scores via e.g. the R package **scoringRules** (Jordan et al. 2019).

Performance of BMA for different countries: Scoring rules

	France		UK	
	Log score	CRPS	Log Score	CRPS
BMA	-4.585	0.988	**-5.275**	2.109
Stacking	**-4.748**	**0.848**	-5.006	**1.965**
Pseudo-BMA	-4.585	0.988	**-5.275**	2.109
LC	-4.943	1.050	-4.701	1.840
RH	-4.585	**0.988**	-4.743	1.854
APC	**-5.402**	1.441	-5.019	2.256
CBD	-4.183	3.860	-4.401	**1.539**
M6	-3.799	3.298	**-5.275**	2.109
	USA		Japan	
	Log score	CRPS	Log Score	CRPS
BMA	-3.815	1.793	**-5.166**	**1.084**
Stacking	**-3.956**	**1.441**	-4.972	1.430
Pseudo-BMA	-3.804	1.771	-5.158	2.035
LC	-3.168	4.166	**-5.276**	1.164
RH	-3.815	1.793	-5.166	**1.084**
APC	**-5.551**	**0.938**	-5.158	2.035
CBD	-3.703	2.165	-4.972	2.577
M6	-5.151	1.404	-4.539	2.068

Covid-type effect on BMA weights

Question: What is the effect of a Covid-type perturbation ?

- For the years 2016 and 2017, we assume that there is a uniform death increase of 5% across all ages:

$$d_{x,t}^{\text{new}} = (1+\beta)\, d_{x,t},$$

with $\beta = 0.05$ for $t = 2016, 2017$.

- The increase in deaths is then compensated with a year of lower mortality. We assume a death decrease of 2% across ages:

$$d_{x,t}^{\text{new}} = (1-\beta)\, d_{x,t},$$

with $\beta = 0.02$ for $t = 2018$.

We compute the life expectancy at age 50:

$$e_{50:\overline{40|},t} = \sum_{k=1}^{40} {}_kp_{50,t}$$

Covid-type effect on BMA weights

Question: What is the effect of a Covid-type perturbation ?

- For the years 2016 and 2017, we assume that there is a uniform death increase of 5% across all ages:

$$d_{x,t}^{\text{new}} = (1+\beta)d_{x,t},$$

with $\beta = 0.05$ for $t = 2016, 2017$.

- The increase in deaths is then compensated with a year of lower mortality. We assume a death decrease of 2% across ages:

$$d_{x,t}^{\text{new}} = (1-\beta)d_{x,t},$$

with $\beta = 0.02$ for $t = 2018$.

We compute the life expectancy at age 50:

$$e_{50:\overline{40}|,t} = \sum_{k=1}^{40} {}_kp_{50,t}$$

Covid-type effect on BMA weights

Question: What is the effect of a Covid-type perturbation ?

- For the years 2016 and 2017, we assume that there is a uniform death increase of 5% across all ages:

$$d_{x,t}^{\text{new}} = (1+\beta)d_{x,t},$$

with $\beta = 0.05$ for $t = 2016, 2017$.

- The increase in deaths is then compensated with a year of lower mortality. We assume a death decrease of 2% across ages:

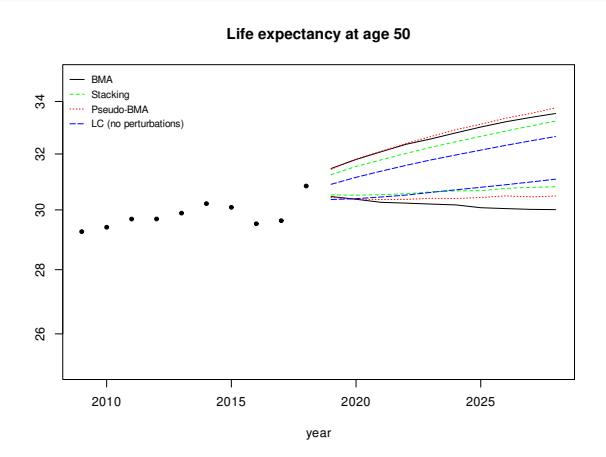
$$d_{x,t}^{\text{new}} = (1-\beta)d_{x,t},$$

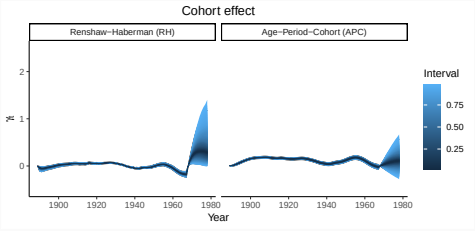
with $\beta = 0.02$ for $t = 2018$.

We compute the life expectancy at age 50:

$$e_{50:\overline{40}|,t} = \sum_{k=1}^{40} {}_kp_{50,t}$$

Covid-type effect on BMA weights





Conclusion

- The full Bayesian Negative-Binomial model averaging considers overdispersion, model and parameter uncertainty.
- Simulation study: Stacking and Pseudo-BMA outperform the standard BMA.
  1. Stacking and Pseudo-BMA select the right model when the model is well specified.
  2. Stacking and Pseudo-BMA have better predictive accuracy when the model is misspecified.
- Using data of four countries, stacking achieves overall better predictive accuracy than Pseudo-BMA and BMA.
  Hence, we recommend stacking.

StanMoMo: Bayesian Mortality Modelling with Stan

- Install the *StanMoMo* R package

```
install.packages("StanMoMo",repos=c("https://cloud.r-project.org",
"https://kabarigou.github.io/drat"),type = "binary")
```

- Estimate and forecast mortality models

```
fitLC=lc_stan(death = deathFR, exposure=exposureFR,
validation=10, forecast = 10, family = "poisson")
```

Also available: `rh_stan`, `apc_stan`, `cbd_stan`, `m6_stan`.

- Compute stacking and Pseudo-BMA weights

```
model_weights<-mortality_weights(list(fitLC,fitRH,fitAPC))
```

- Also functions to download the mortality data, for simulation study, for statistical and convergence analysis.
- **Main website**: https://kabarigou.github.io/StanMoMo/

References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017), 'Stan: A probabilistic programming language', *Journal of statistical software* **76**(1).

Currie, I. D. (2016), 'On fitting generalized linear and non-linear models of mortality', *Scandinavian Actuarial Journal* **2016**(4), 356–383.

Jordan, A., Krüger, F. & Lerch, S. (2019), 'Evaluating probabilistic forecasts with scoringRules', *Journal of Statistical Software* **90**(12), 1–37.

Vehtari, A., Gelman, A. & Gabry, J. (2017), 'Practical Bayesian model evaluation using leave-one-out cross-validation and waic', *Statistics and computing* **27**(5), 1413–1432.

Yao, Y., Vehtari, A., Simpson, D., Gelman, A. et al. (2018), 'Using stacking to average Bayesian predictive distributions (with discussion)', *Bayesian Analysis* **13**(3), 917–1007.