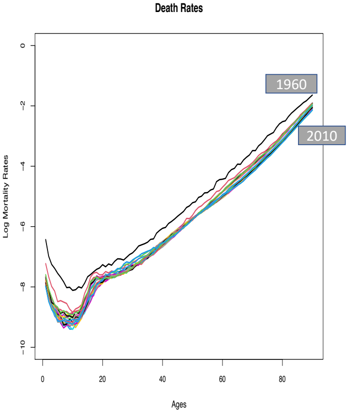
Stacked Regression Ensemble Learning for Mortality Forecasting.

Salvatory Roman Kessy
Joint Work With: M. Sherris, A. Villegas, & J. Ziveyi.

UNIVERSITY OF NEW SOUTH WALES

8 December 2020

Model Selection Dilemma



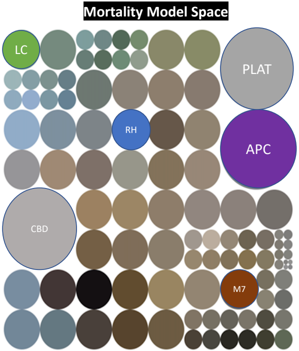


Figure 1: Model Selection Dilemma.

- What mortality model is likely to perform best?

Different Mortality Models

- Multiple mortality models capture different features of death rates such as **trends, linearity, non-linearity, curvature, and cohort effects**.

Model	Predictor (η_{xt})	Parameters
LC	$\alpha_x + \beta_x^{(1)}\kappa_t^{(1)}$	$2n_a + n_y$
RH	$\alpha_x + \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(0)}\gamma_c$	$3n_a + n_y + n_b$
APC	$\alpha_x + \kappa_t^{(1)} + \gamma_c$	$n_a + n_y + n_b$
CBD	$\kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)}$	$2n_y$
M7	$\kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)} + \left((x - \bar{x})^2 - \hat{\sigma}_x^2\right)\kappa_t^{(3)} + \gamma_c$	$3n_y + n_b$
Plat	$\alpha_x + \kappa_t^{(1)} + (\bar{x} - x)\kappa_t^{(2)} + (\bar{x} - x)^+\kappa_t^{(3)} + \gamma_c$	$n_a + 3n_y + n_b$

Table 1: Generalized Age-Period-Cohort (GAPC) mortality models. Here, year of birth is $c = t - x$, n_a is a number of age and n_y is a number of years. The functions $\beta_x^{(i)}, \alpha_x, \kappa_t^{(i)}$, and γ_c are age, period and cohort effects respectively. $\bar{x}$ is the mean age over the range of ages being used in the analysis, $\hat{\sigma}_x^2$ is the mean value of $(x - \bar{x})^2$.

- Better methods are needed.

Model Combination

- Simple Model Averaging (Shang 2012), Bayesian Model Averaging (Shang 2012) and (Kontis et al. 2017), Model Confidence Set (Shang and Haberman 2018).

- Model combination formulation:

$$\ln\left(\widehat{\mu}(x, t+h)\right)_{\text{comb}} = \sum_{m=1}^{M} w_m \ln\left(\widehat{\mu}_m(x, t+h)\right).$$

Stacking Ensemble Techniques

- Ensemble methods use **different models** to obtain **better predictive performance** than could be obtained from any of the **constituent models** (Wolpert 1992).
- If a set of models does not contain the **true prediction function**, ensembles can give a good approximation of that function (Polley and Laan 2010).
- The stacking ensemble has been successfully applied and improved the predictive accuracy on a wide range of problems:
  1. Forecasting global energy consumption (Khairalla et al. 2018).
  2. Credit risk assessment (Doumpos and Zopounidis 2007).
  3. Financial time series data sets (Ma and Dai 2016).
- Most winning teams in data science competitions have been using the stacked regression ensemble (Sill et al. 2009; Puurula, Read, and Bifet 2014; Makridakis, Spiliotis, and Assimakopoulos 2019).

This Presentation is About . . .

- Propose a new approach of estimating the optimal weights for combining multiple mortality models using **stacked regression ensemble framework** (Wolpert 1992).

  1. Concurrently solve the problem of **model selection and estimation of the model combination** to improve model predictions (Sridhar, Seagrave, and Bartlett 1996).

  2. Tackle the **model list miss-specification limitation** associated with the BMA approach (Yao et al. 2017).

  3. Assigns weights to the individual mortality models by **minimising the cross-validation criterion.**

- Develops the mortality model combination that is **dependent on the forecasting horizon** (SriDaran et al. 2020; Rabbi and Mazzuco 2018).

Stacked Regression Ensemble

- Stacked regression ensemble **combines point predictions** from multiple mortality base learners using the weights that **optimise a cross-validation criterion** (Wolpert 1992).
- Bagging is a special case of the stacked regression ensemble.

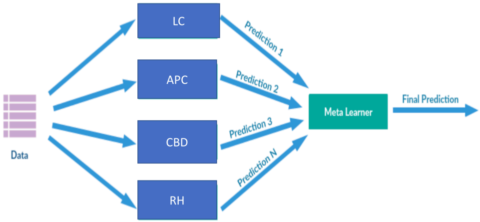


Figure 2: An example scheme of stacking ensemble learning.

Stacked Regression Ensemble

- Suppose that the h-year-ahead mortality rate forecasts from M mortality models $L_1, \dots, L_M$ are $\widehat{\mu}_1(x, t_{n_y} + h), \dots, \widehat{\mu}_M(x, t_{n_y} + h)$ for age $x \in [x_1, x_{n_a}]$ at time $t_{n_y} + h$.
- Combining weights are viewed as the linear regression coefficients:

$$\underbrace{\ln \mu(x, t_{n_y} + h)}_{\text{Dependent variables}} = \sum_{m=1}^{M} \underbrace{w_m(h)}_{\text{coefficients}} \underbrace{\ln \widehat{\mu}_m(x, t_{n_y} + h)}_{\text{covariates}},$$

- Any supervised machine learning algorithm can be used to estimate the weights by optimising the squared loss function (Wolpert 1992).
- The optimization is constrained such that these weights sum to unity.

Blcok Cross-validation

- Block cross-validation of mortality data by period (Bergmeir, Costantini, and Benítez 2014; SriDaran et al. 2020).

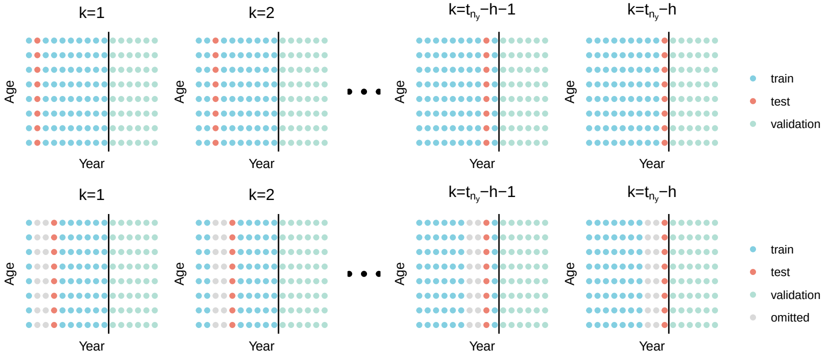


Figure 3: Iterations of cross validation for horizon one-year-ahead ($h = 1$) (top row) and three-years-ahead ($h = 3$) (bottom row).

Metadata

- Train each mortality base learners $L_1, \ldots, L_M$ on the training data set (blue) from Figure 3.
- For each base learner $L_1, \ldots, L_M$, predict the mortality rates $\widehat{\mu}(x, t+h)$ using the test set (red) from Figure 3.
- Generate level-one/metadata.

	LC	RH	APC	CBD	M7	PLAT	Actual
1	-4.91	-4.90	-4.87	-4.75	-4.94	-4.91	-4.93
2	-4.87	-4.91	-4.85	-4.73	-4.95	-4.90	-4.93
3	-4.86	-4.92	-4.85	-4.73	-4.94	-4.90	-4.89
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1197	-1.48	-1.50	-1.49	-1.38	-1.50	-1.48	-1.49
1198	-1.50	-1.46	-1.46	-1.38	-1.46	-1.44	-1.44
1199	-1.50	-1.46	-1.47	-1.36	-1.42	-1.42	-1.52

$\Rightarrow \boxed{Y = Z\mathbf{w}}$

- Train a meta-learner on metadata to estimate the optimal weights of combining M mortality base models.

Meta-learners

- Non-negative Least Square Regression (Breiman 2004; Naimi and Balzer 2018):

$$\widehat{\mathbf{w}}^* = \underset{\mathbf{w}}{\text{argmin}} \sum_{t=t_1}^{t_n} \sum_{x=x_1}^{x_n} \left(\ln(\mu_{x,t}) - \sum_{m=1}^{M} w_m \ln \widehat{\mu}_m(x,t) \right)^2, \ \widehat{w}_m^* \geq 0.$$

- Ridge Regression (Leblanc et al. 2016):

$$\widehat{\mathbf{w}}^* = \underset{\mathbf{w}}{\text{argmin}} \sum_{t=t_1}^{t_{n_y}} \sum_{x=x_1}^{x_{n_a}} \left(\ln \mu(x,t) - \sum_{m=1}^{M} w_m \ln \left(\widehat{\mu}_m(x,t)\right)^{\text{cv}} \right)^2 + \lambda \sum_{m=1}^{M} w_m^2.$$

- Lasso Regression (Gunes, Wolfinger, and Tan 2017):

$$\widehat{\mathbf{w}}^* = \underset{\mathbf{w}}{\text{argmin}} \sum_{t=t_1}^{t_{n_y}} \sum_{x=x_1}^{x_{n_a}} \left(\ln \mu(x,t) - \sum_{m=1}^{M} w_m \ln \left(\widehat{\mu}_m(x,t)\right)^{\text{cv}} \right)^2 + \lambda \sum_{m=1}^{M} |w_m|.$$

Competing Model Averaging Techniques

- Bayesian Model Averaging (Hoeting et al. 1999).

$$\mathbb{P}(\Psi|\mathcal{D}) = \sum_{m=1}^{M} \mathbb{P}(\Psi|L_m, \mathcal{D})\mathbb{P}(L_m|\mathcal{D}) = \sum_{m=1}^{M} w_m \mathbb{P}(\Psi|L_m, \mathcal{D}).$$

- BMA weights using projection bias Kontis et al. (2017):

$$w_m^{\text{bias}}(h) \approx \frac{e^{-0.5|\text{Projection Bias}_m|}}{\sum_{m=1}^{M} e^{-0.5|\text{Projection Bias}_m|}}, \quad \forall m = 1, 2, \ldots, M.$$

- BMA weights using cross-validation mean square errors CVMSE:

$$w_m^{\text{CVMSE}}(h) \approx \frac{e^{-0.5\text{CVMSE}_m(h)}}{\sum_{m=1}^{M} e^{-0.5\text{CVMSE}_m(h)}}, \quad \forall m = 1, 2, \ldots, M.$$

- Model Confidence Set (Shang and Haberman 2018).

Model Selection Risk

- A Case Study: England and Wales, Males and Females.
- Human Mortality Database: 1960 to 2015 and ages 50 – 89.

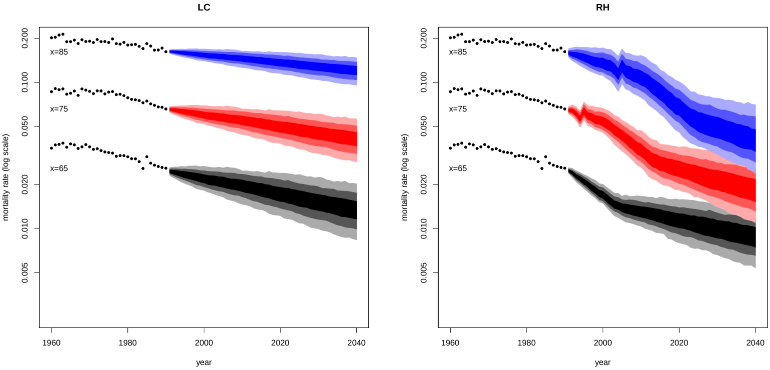


Figure 4: Fan charts for England and Wales males mortality rates at ages 65, 75, and 85. Shades in the fan represent prediction intervals at the 50%, 80% and 95% level.

Individual Mortality Models Forecasting Performance

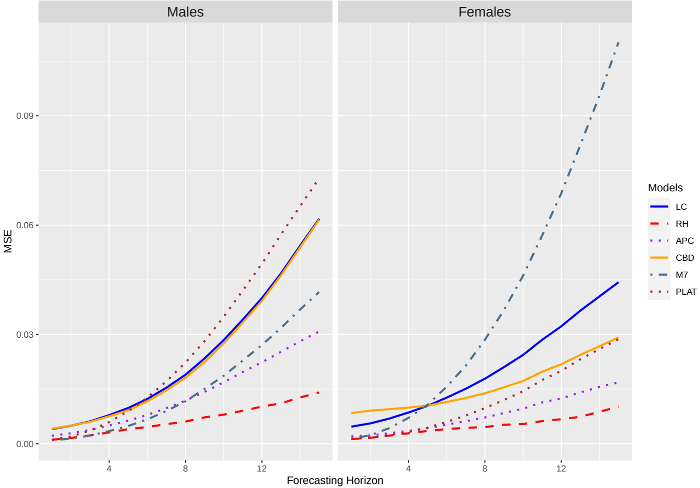


Figure 5: Mean squared errors of different mortality models for various forecasting time horizons using England and Wales males mortality data (left) and females (right).

Combination Weights for Mortality Models

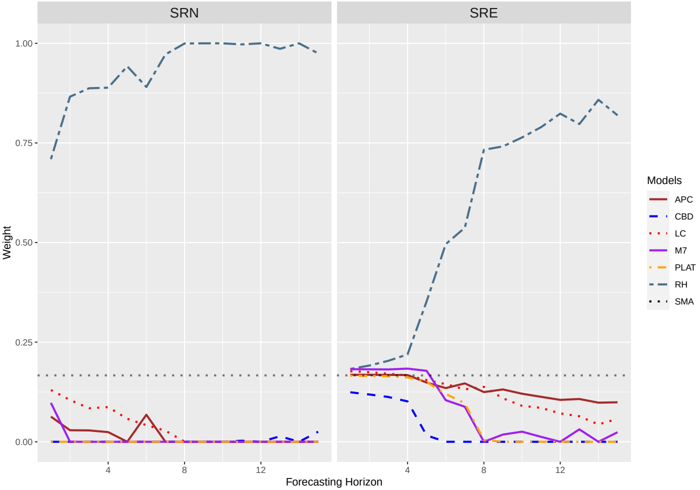


Figure 6: Horizon-specific optimal combining weights learned using different meta-learners for England and Wales males mortality data from 1960 to 1990 and ages 50 to 89.

Final Mortality Rate Forecasts

- Use the weights generated using elastic net regression.
- Super-learner mortality model for forecasting one-year-ahead mortality rates for males:

$$\ln\left(\widehat{\mu}(x, t_{n_y+1})\right)_{\text{SRE}} = (0.18\ \widehat{\text{LC}}) + (0.18\ \widehat{\text{RH}}) + (0.17\ \widehat{\text{M7}}) + (0.18\ \widehat{\text{PLAT}}) + (0.18\ \widehat{\text{APC}}) + (0.14\ \widehat{\text{CBD}}).$$

- Super-learner mortality model for forecasting fifteen-year-ahead mortality rates for males:

$$\ln\left(\widehat{\mu}(x, t_{n_y+15})\right)_{\text{SRE}} = (0.22\ \widehat{\text{LC}}) + (0.48\ \widehat{\text{RH}}) + (0.23\ \widehat{\text{PLAT}}) + (0.04\ \widehat{\text{APC}}) + (0.04\ \widehat{\text{CBD}}).$$

- Produce the mortality forecasts from the test data using LC, RH, APC, CBD, M7, and PLAT and substitute them into the super-learner.

Performance of Stacked Regression Ensemble

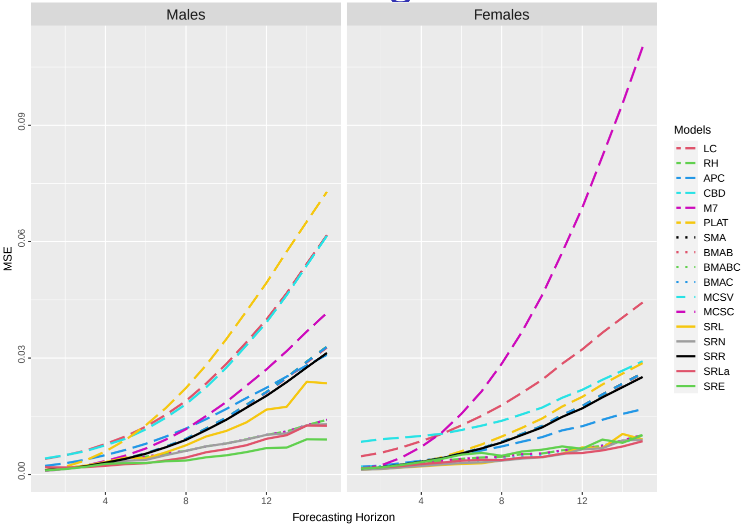


Figure 7: MSEs of the one-step-ahead to 15-step-ahead mortality rate forecasts using different mortality methods and forecast horizons for England and Wales male mortality data and females.

Base Error Reduction

- $\text{BER} = \text{MSE}_{\text{Base Learner}} - \text{MSE}_{\text{SR}}$.

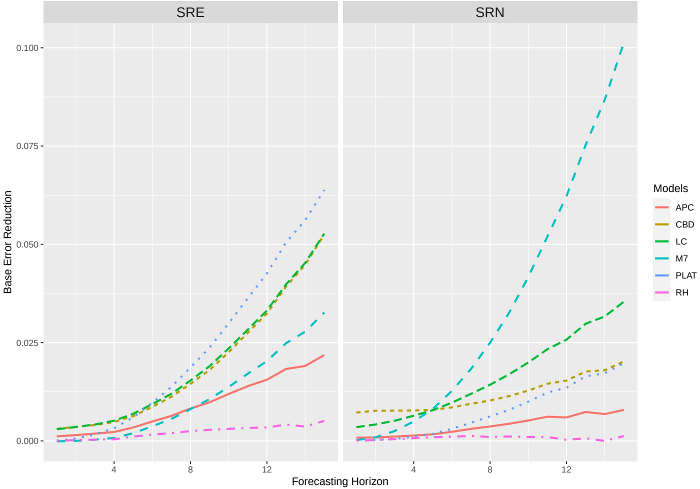


Figure 8: Base error reduction for the top-ranked model combination methods, namely SRE and SRN for males and females, respectively.

Stacked Regression Ensemble in Different Countries

Figure 9: Heat maps showing the average ranks of mortality models across different countries for males and females.

Stacked Regression Ensemble in Different Countries

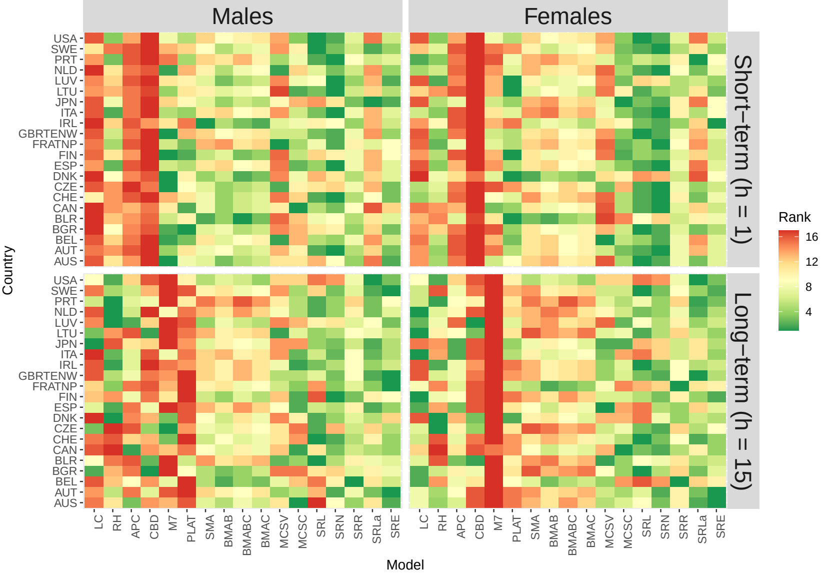


Figure 10: Heat maps showing the average ranks of mortality models across different horizons and countries for males and females.

Conclusion

- Using 44 populations from the Human Mortality Database, stacking mortality models increases predictive accuracy.
- Stacked regression (SR) achieved an average accuracy of $13-49\%$ and $20-90\%$ over the individual mortality models for males and females.
- SR also achieved better predictive accuracy than other model combination methods.
- The weights for combining the individual mortality models vary depending on the meta-learner, forecasting horizon, country, and gender.
- Estimating weights or choosing the individual mortality models via cross-validation proves to be a crucial step.
- Our results confirm the superiority of SR over the individual and other model combination methods in forecasting the mortality rates.

Future work

- Selecting a meta-learner based on the mortality data features (Talagala, Hyndman, and Athanasopoulos 2018).
- Add more mortality models to the family of the GAPC models.
- Develop a model combination that simultaneously generates the central mortality projections and their corresponding probabilistic distributions to the mortality rate forecasts
- Learning the optimal weights using the integrated cross-validated predictions.
- Develop the **CoMoMo** package for mortality model combinations.

Thank You!

Contact: s.kessy@unsw.edu.au

References

Bergmeir, Christoph, Mauro Costantini, and José M. Benítez. 2014. "On the Usefulness of Cross-Validation for Directional Forecast Evaluation." *Computational Statistics & Data Analysis* 76 (August): 132–43. https://doi.org/10.1016/j.csda.2014.02.001.

Breiman, Leo. 2004. "Stacked Regressions." *Machine Learning* 24 (1): 49–64. https://doi.org/10.1007/bf00117832.

Doumpos, Michael, and Constantin Zopounidis. 2007. "Model Combination for Credit Risk Assessment: A Stacked Generalization Approach." *Annals of Operations Research* 151 (1): 289–306. https://doi.org/10.1007/s10479-006-0120-x.

Gunes, Funda, Russ Wolfinger, and Pei-Yi Tan. 2017. "Stacked Ensemble Models for Improved Prediction Accuracy." *Sas*, 1–19.

Hoeting, Jennifer A, David Madigan, Adrian E Raftery, and Chris T Volinsky. 1999. "Bayesian Model Averaging: A Tutorial," 36.

Khairalla, Mergani A, Xu Ning, Nashat T AL-Jallad, and Musaab O El-Faroug. 2018. "Short-Term Forecasting for Energy Consumption Through Stacking Heterogeneous Ensemble Learning Model." *Energies* 11 (6). https://doi.org/10.3390/en11061605.