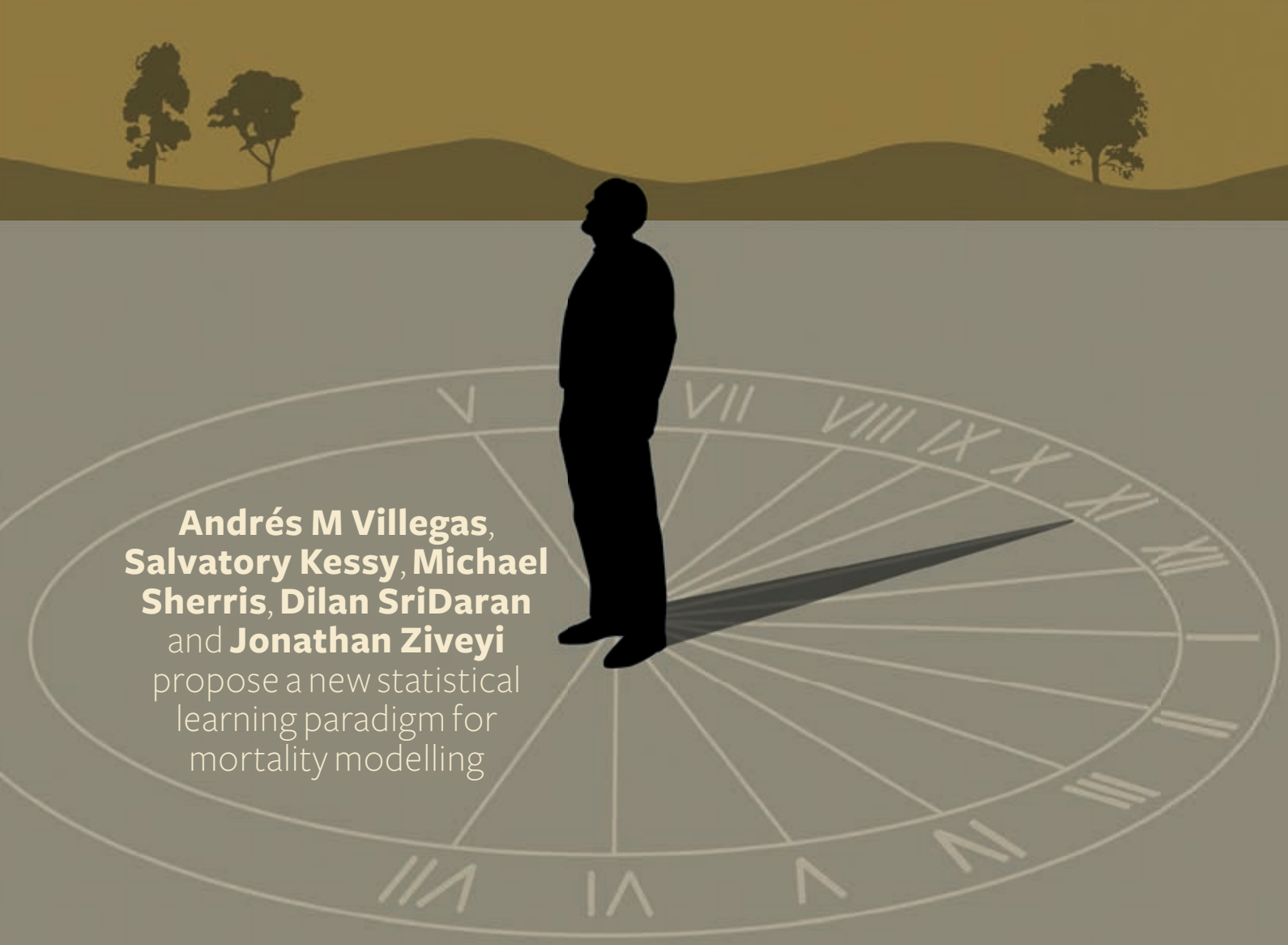


Andrés M Villegas, **Salvatory Kessy**, **Michael Sherris**, **Dilan SriDaran** and **Jonathan Ziveyi**

propose a new statistical learning paradigm for mortality modelling

AROUND THE BLOCK

In recent decades, humanity has made significant progress in averting and delaying death. While indicative of tremendous societal progress, this longevity does present economic and social risks to healthcare, social security, private pension systems and life insurance. The main risk, however, is not the increased longevity per se, but the inherent uncertainty over the evolution of future human mortality. Dynamic mortality modelling is therefore a critical tool for actuaries, both in terms of facilitating an understanding of the past, and in providing a basis for future mortality projections.

In this article, we will discuss some of the limitations of current mortality forecasting approaches and introduce alternative data-driven tools for constructing, selecting and combining models. We have made these data-driven tools readily available through open-source code within the R packages StMoMo (**github.com/amvillegas/StMoMo**) and CoMoMo (**github.com/kessysalvatory/CoMoMo**).

The limitations of current mortality forecasting approaches

Generally, mortality rates are broken down into various functions of age, time and birth cohorts. These three features respectively aim to capture changes from growing older, systematic longevity improvements over time, and any factors that are unique to specific cohorts of individuals. Generalised age-period-cohort (GAPC) models have traditionally been favoured within actuarial spheres due to their ease of implementation and interpretability, as well as their ability to simulate future outcomes and develop probabilistic forecast intervals. GAPC models encompass a wide variety of models, including the commonly used Lee-Carter (LC) model and the Cairns-Blake-Dowd (CBD) family of models.

In such GAPC models, the force of mortality at age x in year t, $\mu_{x,t}$, can be represented as

$$\ln(\mu_{x,t}) = \alpha_x + \sum_{i=1}^{N} \beta_x^{(i)} \kappa_t^{(i)} + \gamma_{t-x},$$

where the general age shape component of mortality is captured by α_x, systematic longevity improvements across specific ages over time are reflected by $\beta_x^{(i)}$ and $\kappa_t^{(i)}$, $i = 1, \ldots, N$, and any effects embedded in an individual's year of birth are allowed for by γ_{t-x}.

In theory, an infinite number of models could be produced within this GAPC framework, ranging from simplistic models that attribute an equal mortality improvement factor to all ages, to complex models that fit several different time trends that are specific to certain ages or cohorts. Simpler models are easier to interpret but may not capture more nuanced systematic features within historical data, while overly complex models may over-fit to noise and produce volatile, spurious forecasts.

In an ideal world, we would be able to consider every possible GAPC model so that we can select the optimal model for a given dataset and desired application. However, time and computational constraints mean this is not feasible. Instead, practitioners typically rely on a small (but growing) selection of models with predefined feature specifications, some of which are shown in *Table 1*.

TABLE 1: Summary of existing GAPC models and their features. Here $\bar{x}$ is the average of the age range in the data, σ_x^2 is the variance of the age range in the data, and $(\bar{x}-x)^+ = \max\{0, \bar{x}-x\}$.

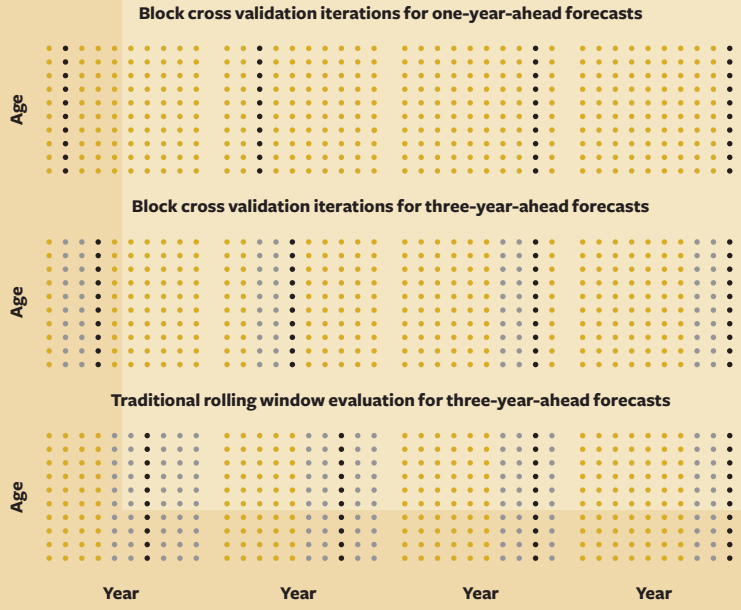
MODEL	FORMULA
Lee-Carter model (LC)	$\ln(\mu_{x,t}) = \alpha_x + \beta_x\kappa_t$
Renshaw and Haberman model (RH)	$\ln(\mu_{x,t}) = \alpha_x + \beta_x\kappa_t + \gamma_{t-x}$
Age-period-cohort model (APC)	$\ln(\mu_{x,t}) = \alpha_x + \kappa_t + \gamma_{t-x}$
Cairns-Blake-Dowd model (CBD)	$\ln(\mu_{x,t}) = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(1)}$
M7 model (M7)	$\ln(\mu_{x,t}) = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(1)} + [(x-\bar{x}) - \sigma_x^2]\kappa_t^{(1)} + \gamma_{t-x}$
Plat model (PLAT)	$\ln(\mu_{x,t}) = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(1)} + (x-\bar{x})^+\kappa_t^{(1)} + \gamma_{t-x}$

However, it is perhaps unreasonable to expect that every dataset in the world could be explained by one of just six models. Moreover, the tools available for selecting the 'optimal' model – such as the Akaike information criterion and the Bayesian information criterion – are often *ad hoc* and typically focus on in-sample goodness-of-fit. Such approaches do not consider how well a model may forecast out of sample and into the future, and therefore do not allow users to robustly determine the model's desired features and degree of complexity. In addition, model selection methods may yield contradicting information, making it difficult to select a single 'optimal' model.

Statistical learning tools for mortality modelling

To overcome these limitations, we have been exploring the use of different statistical learning tools in the context of mortality modelling and forecasting. In 'A Group Regularisation Approach for Constructing Generalised Age-Period-Cohort Mortality Projection Models' (**bit.ly/GroupReg_APC**), we propose a framework that uses group regularisation to produce bespoke GAPC models for specific datasets and applications, and in 'Mortality Forecasting Using Stacked Regression Ensembles' (**bit.ly/MortForecast_Stacked**), we propose a stacked regression approach to optimally combine existing GAPC models, rather than constructing bespoke models. Both approaches are underpinned by re-sampling techniques, which aim to determine the optimal degree of complexity for different applications by placing emphasis on how well a statistical model will generalise to unseen data.

FIGURE 1: Schematisation of block cross-validation iterations for one-year-ahead and three-year-ahead forecasts. Yellow dots represent training data, black dots test data, and grey dots omitted data.



Block cross-validation for model selection

To emphasise forecasting performance over in-sample goodness-of-fit, we propose the use of block cross-validation techniques that are specifically tailored to a user's forecasting horizon of interest, as demonstrated in *Figure 1*. Mortality data typically comes in matrix form, with rows and columns denoting unique ages and calendar years respectively. In block cross-validation, this data is iteratively divided into training data, to fit the mortality models, and test data, to evaluate their out-of-sample performances. The test data can take different widths to represent different forecasting horizons. For example, when predicting one-year-ahead out-of-sample mortality rates, test data should be defined as one-year blocks (as in the top row of *Figure 1*), and it should be defined as three-year blocks when predicting three-years-ahead rates (as in the middle row of *Figure 1*).

One interesting aspect of our cross-validation approach is that we use data both to the left and right of each test set to train the model, in contrast to the more common rolling window out-of-sample evaluation (as in the bottom row of *Figure 1*). This allows us to conduct significantly more test sets for a given dataset than if we were restricted to fitting models using only data to the left, since this would place a limit on how far to the left test sets can occur. Moreover, we minimise the portion of data that is left completely unused (the grey dots in *Figure 1*), thereby extracting as much information as possible and producing more precise estimates of the out-of-sample prediction error.

Group regularisation for model construction

To avoid limiting ourselves to a finite number of GAPC models, we consider the use of regularisation techniques to explore a much richer array of potential features. Conceptually, the idea behind our approach is to first define a large basket of candidate features that could explain different patterns and trends. For example, one simple feature could aim to capture a systematic mortality improvement over time that affects all ages equally. Another, more complex, feature could resemble a 'put' option function by fitting a time trend to only ages below a certain value, with the impact of the trend increasing as age decreases. In our example below, we propose 27 unique features; however, our package StMoMo allows users to easily add elements to or remove elements from this basket using their own judgment.

To fit the models, we exploit the fact that GAPC models can be expressed as a generalised linear model of the form

$$Y = \boldsymbol{X}\boldsymbol{\beta} = \sum_{j=0}^{N+1} \boldsymbol{X}_j \boldsymbol{\beta}_j,$$

where each $\boldsymbol{X}_j$ and $\boldsymbol{\beta}_j$, $j = 0, 1, \ldots, N+1$, represent, respectively, the design matrix and vector of coefficients associated with a unique feature in the basket. This allows us to use a group regularisation algorithm to estimate the coefficients of the model. Group regularisation is like traditional ordinary-least squares (OLS) or maximum likelihood estimation (MLE), except that it includes a penalty term (λ) that biases groups of model coefficients to zero. When $\lambda = 0$, no penalty is applied, and fitting is identical to the OLS or MLE setting. This means that all coefficients attached to the basket of features will be non-zero, yielding a highly complex model that will likely over-fit the data. At the other extreme, when $\lambda = \infty$, all coefficients will be shrunk to zero, meaning all features from the basket are 'removed' and we are left with an extremely simplistic model.

In our paper, we use the aforementioned cross-validation framework to determine the optimal value of λ and, by extension, which features from the basket to include in a model for each dataset and application.

Stacked regression ensembles for model combination

However, cross-validation offers more information than merely identifying which single model or set of features is the best. It also permits the estimation of combinations of models, which can yield more precise out-of-sample forecasts than their individual constituents. Therefore, as an alternative to picking a single best model with the lowest cross-validation error, in 'Mortality Forecasting Using Stacked Regression Ensembles' we use the cross-validation predictions to develop a stacked regression ensemble that

FIGURE 2: Regularisation path of mortality models for England and Wales males fitted to ages 50-89 for the years 1960-2018. The dashed vertical lines indicate the optimal level of regularisation at selected forecasting horizons, determined using block cross-validation.

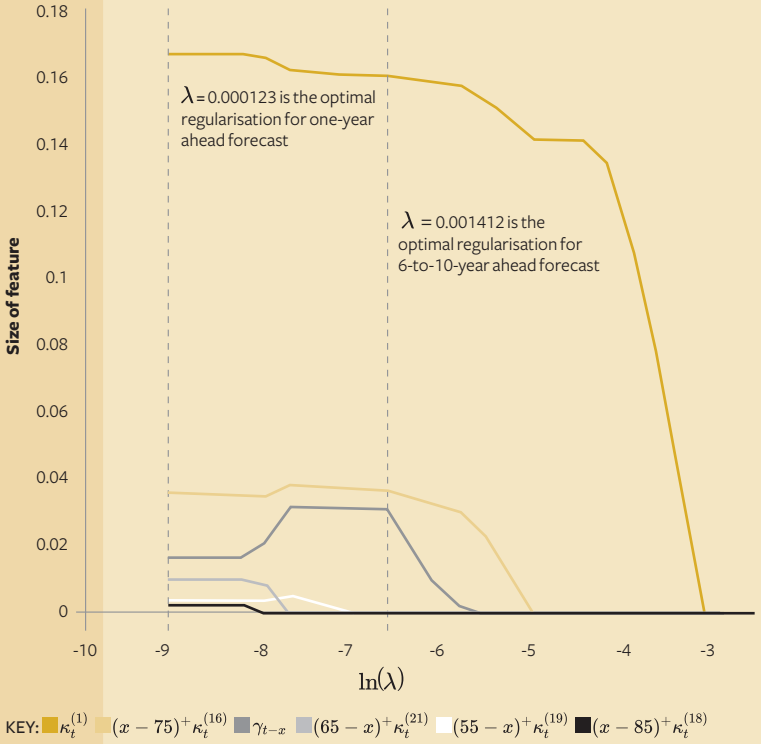
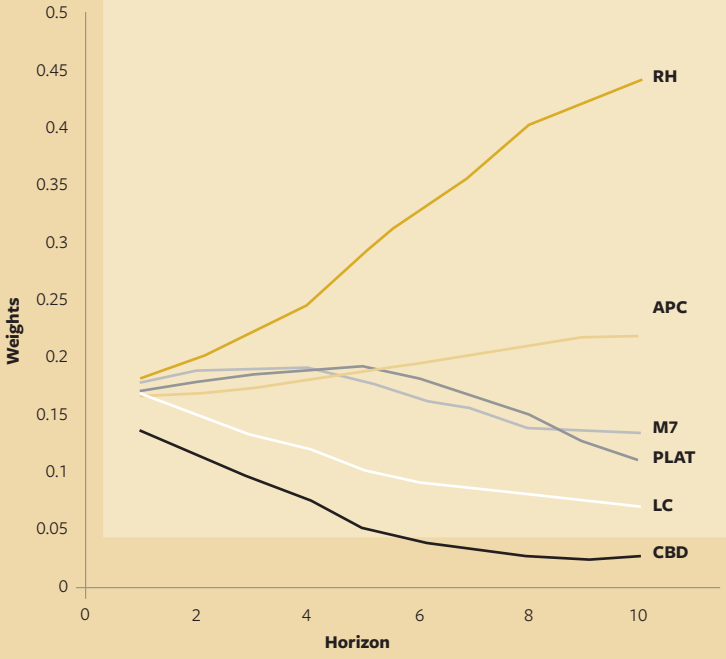


FIGURE 3: Horizon-specific optimal combining weights learned using stacked regression with elastic-net regularisation for England and Wales male mortality data from 1960-2018 and ages 50-89.



combines the predictions from multiple individual mortality models.

Stacked regression ensembles combine several diverse individual models into one powerful prediction function via a secondary meta-learning process, whereby the weights assigned to each individual model are optimised to reflect the ability of each model to generalise to new data. Owing to its empirical and theoretical benefits, stacked regression ensembles are often used by winning teams in data science competitions. Stacked regression ensemble methods have also been successfully applied, with improved predictive accuracy, to a range of problems – such as credit risk assessment, forecasting global energy consumption, financial time series forecasting, and prediction of infectious disease epidemics.

In a stacked regression ensemble, the combination weights are viewed as coefficients of a linear regression problem, in which the (log) mortality rates are treated as the dependent variable and the point predictions from individual mortality models are treated as the independent variables. Assuming that there are M models to combine, this is

$$\underbrace{\ln(\mu_{x,t+h})}_{\text{Dependent variable}} = \sum_{m=1}^{M} \underbrace{w_m(h)}_{\text{coefficients}} \underbrace{\ln(\hat{\mu}^m_{x,t+h})}_{\text{covariates}}$$

where $\ln(\mu_{x,t+h})$ are observed mortality rates and $\ln(\hat{\mu}^m_{x,t+h})$ are the mortality rate predictions from the individual models generated via cross-validation. For each forecast horizon h, the combination weights $w_m(h)$, $m = 1, \ldots, M$ can be learned using any supervised statistical learning algorithm, such as lasso regression, ridge regression or elastic-net penalisation.

Demonstration using male mortality data for England and Wales

To demonstrate these tools, we consider the Human Mortality Database's mortality data for males in England and Wales aged 50-89, for the years 1960-2018.

For the group regularisation framework, we consider a basket of 27 features, which include polynomials up to order 10 and 'put' and 'call' option functions at ages 55, 60, ..., 80, 85. To search for the optimal level of regularisation, we consider λ values in the range $[e^{-9}, e^{-2.5}]$. As shown in *Figure 2*, this range of λ defines a regularisation path containing different GAPC mortality models with varying degrees of complexity.

Small values of λ correspond to low levels of regularisation and, subsequently, more complex models. For example, to the left of *Figure 2* when $\lambda = 0.000123$, we start with a model that includes six of the 27 initial features. Besides a general mortality improvement term $\kappa_t^{(1)}$ and a cohort effect γ_{t-x}, this also includes terms that capture the idiosyncrasies of mortality trends at different ages via call and put features $(x-75)^+\kappa_t^{(16)}$, $(x-85)^+\kappa_t^{(18)}$, $(55-x)^+\kappa_t^{(19)}$ and $(65-x)^+\kappa_t^{(21)}$. As the value of λ increases, the additional regularisation eliminates some of the age-period terms, resulting in more parsimonious models. For example, when $\lambda = 0.001412$, we obtain a model that only includes the $\kappa_t^{(1)}$ term capturing the general level of mortality, the $(x-75)^+\kappa_t^{(16)}$ term capturing the differential mortality improvement of individuals older than age 75, and the cohort effect γ_{t-x}.

According to our cross-validation approach, the complex $\lambda = 0.000123$ model is 'optimal' for the one-year forecasting horizon, while the simpler $\lambda = 0.001412$ is 'optimal' for six-to-10-year forecasting horizons. In general, from our empirical analyses and simulations, we have determined that the longer the forecasting horizon, the higher the value of the 'optimal' λ, the fewer the number of period terms, and the more parsimonious the selected model specification.

In *Figure 3* we show the results of applying a stacked regression ensemble to combine the six mortality models in *Table 1* for the same dataset. The stacked regression ensemble learns horizon-specific optimal weights for combining these individual mortality models at different forecasting horizons so that the mortality models with stronger predictive strength for a given horizon receive higher weights. The super-learner mortality model for forecasting one-year-ahead rates is given by combining the individual mortality models LC, RH, APC, CBD, M7 and PLAT using their corresponding weights 0.17, 0.18, 0.17, 0.14, 0.18 and 0.17 respectively, with the weights changing to 0.07, 0.44, 0.22, 0.03, 0.13 and 0.11 for the 10-year-ahead forecast. These weights reflect that the RH and APC models generalise well to new unknown future mortality data for this population and so get higher weights that increase with the forecasting horizon. In contrast, the weaker performing LC, CBD, M7 and PLAT models receive corresponding smaller weights that decrease with the forecasting horizon.

A BESPOKE APPROACH

In our two academic papers, we have carried out exhaustive analyses of the out-of-sample performance of both the group regularisation and the stacked regression approaches, using more than 40 populations from the Human Mortality Database. These empirical analyses have revealed that no predefined model can perform well for every dataset and forecasting horizon, with our bespoke regularised models and model combinations consistently outperforming existing GAPC models. Moreover, our empirical analyses reaffirmed the commonly held heuristic that simpler models outperform their complex counterparts as forecasting horizons increase.

However, we must stress that while we have obtained very strong results across many datasets, we do not claim to provide all the answers or the perfect models. We simply aim to provide alternative data-driven approaches to thinking about mortality modelling through the lens of statistical learning techniques.

ANDRÉS VILLEGAS
is senior lecturer in risk and actuarial studies at UNSW Sydney, Australia

SALVATORY KESSY
is PhD student in risk and actuarial studies at UNSW Sydney, Australia

MICHAEL SHERRIS
is professor in risk and actuarial studies at UNSW Sydney, Australia, and chief investigator at CEPAR

DILAN SRIDARAN
is senior consultant at EY

JONATHAN ZIVEYI
is associate professor in risk and actuarial studies at UNSW Sydney, Australia