British Actuarial Journal (2025), Vol. **30**, e11, pp. 1–15
doi:10.1017/S1357321724000084

Institute and Faculty of Actuaries

SESSIONAL MEETING DISCUSSION

Robust mortality forecasting in the presence of outliers

[Institute and Faculty of Actuaries, Sessional Webinar, Monday 27 November 2023]

Moderator (Mr P. O. J. Kelliher, F.I.A.): Thank you for coming along tonight to this sessional meeting on "Robust Mortality Forecasting in the Presence of Outliers" by Dr Stephen Richards. Dr Richards is a Fellow of the Institute and Faculty of Actuaries (IFoA), and he also has a doctorate in Mortality Analysis and Projection from Heriot Watt University. He is the managing director of Longevitas, which is a specialist provider of actuarial tools and consultancy services for longevity risk and annuities. He set up Longevitas in 2006 and has had considerable success with his software now being used in the UK, USA, Canada and Switzerland. Prior to founding Longevitas, he headed Prudential's longevity analysis team, and before that he headed product pricing at Standard Life. Stephen is an honorary research fellow at Heriot Watt University and regularly publishes research addressing practical longevity issues, including numerous sessional papers. He also contributes to a regular blog on the Longevitas website, which is worth checking out for anyone with an interest in longevity. Therefore, it gives me great pleasure to invite Stephen (Richards) to present his latest paper.

Stochastic mortality models are now widely used in modern actuarial work. They play a particularly important role in the value-at-risk approach to setting capital requirements for longevity trend risk; see Richards *et al.* (2014). However, the COVID-19 pandemic has left some major outliers in the mortality data, and, without proper handling, those outliers severely distort mortality forecasts and thus distort any value-at-risk capital requirements that are based on these forecasts. The methodologies presented in tonight's paper draw heavily on research from other fields. This is because outliers have long been a problem in other disciplines, particularly in the study of economic time series. As actuaries, we can therefore benefit hugely from existing research. The existence of rigorous statistical approaches to outliers developed outside the profession means that there is no need for actuaries to ignore outliers or develop ad hoc methods for dealing with them. Tonight's paper illustrates some tried and tested methodologies from other fields that we as actuaries can apply to our work using some freely available R-libraries.

We will start with the motivation, which is obviously COVID-19. In Figure 1(b) in the sessional paper, we can see the male deaths in England and Wales from 1971 to 2020 inclusive. We can see the obvious outlier in the number of male deaths in 2020. There is a similar outlier for females in Figure 1(d) in the sessional paper, although it is not quite as pronounced as for males. There are obviously other important features in this data, such as the change of direction of death counts around 2011. However, tonight's paper is about dealing with the outliers in 2020.

The problem with outliers is that they have a number of unwanted consequences. The first is that they distort forecasts. The second is that they bias the starting points for a forecast if the most recent observations include outliers. Outliers also inflate variance and thus inflated value-at-risk capital. We will look at how COVID-19 outliers distort central projections. In Figure A1, we have a plot of the fitted mortality rates at age 75 for males in England and Wales, up to and including

2019. We have fitted a Lee–Carter model, produced a forecast, and the forecast is a reasonable extrapolation of the prior trend. The problem comes from including the 2020 data as shown in Figure A2. The outlier from 2020 so distorts the forecasting model that it takes a completely nonsensical direction. The effect of COVID-19, or outliers in general, can lead to distorted forecasts. Even if they do not distort the direction, they do distort the starting points, as we will see next.

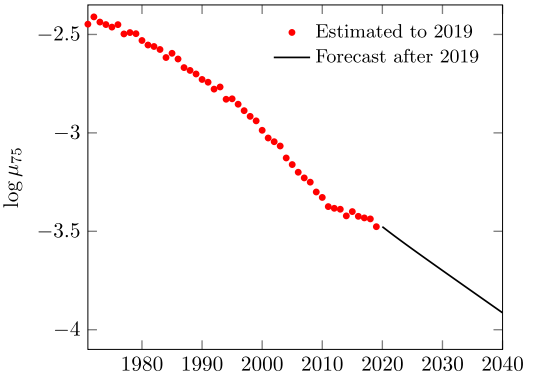

Figure A1. ARIMA forecast of *k* time index in Lee–Carter model. Source: Own calculations using data for males in England & Wales, ages 50–105, 1971–2019.

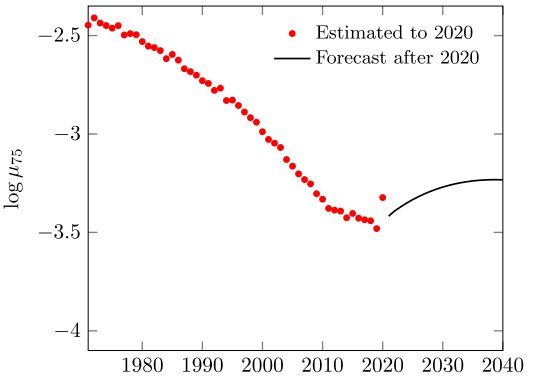

Figure A2. ARIMA forecast of *k* time index in Lee–Carter model. Source: Own calculations using data for males in England & Wales, ages 50–105, 1971–2020.

In Figure A3 we have data for females in England and Wales. We used the mortality data to 2019 and fitted a Cairns–Blake–Dowd M5 model. This gave a fairly sensible extrapolation of the forecast mortality rate. If we look at what happens when we include the 2020 data, as shown in Figure A4, the direction of the forecast is still essentially sensible, but the starting point has been biased. So, even when the direction of the forecast is not distorted, the forecast ican still be biased.

The third consequence of outliers is a specifically actuarial problem. In addition to distorting the parameter estimates for the forecasting model or biasing the starting point for the forecast, outliers also inflate the variance measure, which feeds straight through to the capital requirements. In Figure A5, we show the value-at- risk capital for longevity trend risk for annuities using a Lee–Carter model and data to 2019. As per Richards *et al.* (2014), we repeatedly simulate the next year's experience, refit the model and revalue the annuity payments. We come up with rough 1% capital requirement on a 99.5% basis from ages 50 through to 90. However, if we include the 2020

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

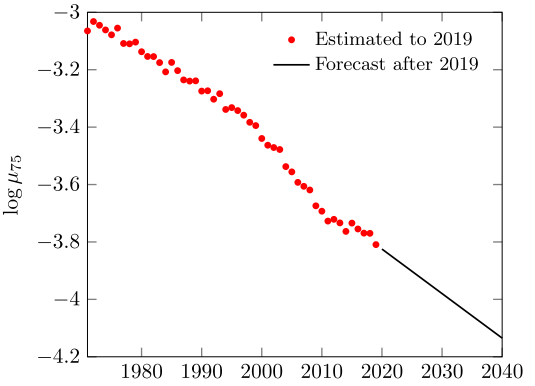

Figure A3. Bivariate random-walk forecast under M5 model. Source: Own calculations using data for females in England & Wales, ages 60–105, 1971–2019.

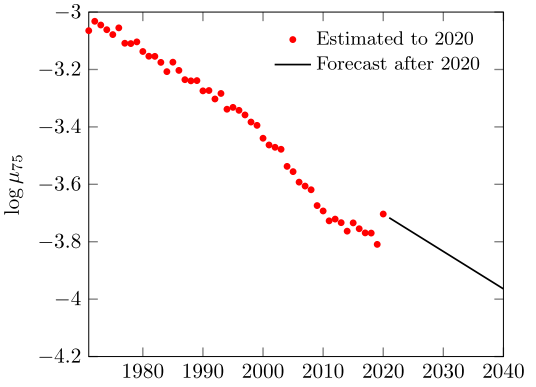

Figure A4. Bivariate random-walk forecast under M5 model. Source: Own calculations using data for females in England & Wales, ages 60–105, 1971–2020.

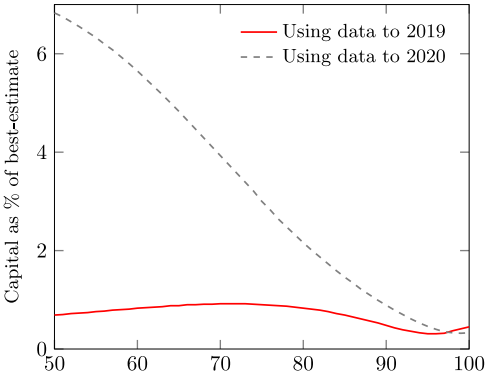

Figure A5. Value-at-risk longevity capital requirement as percentage of best-estimate annuity value. Source: Own calculations using 10,000 recalibrations of Lee–Carter model using data for males in England & Wales. Annuity cashflows discounted at 0% per annum.

data, in addition to the broken forecasting model and the biased starting point, we also have massive inflation of the variance in the process, which leads to an over-statement of the value-at-risk capital. So, we have three fundamental problems arising from outliers.

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

An outlier like that caused by the COVID pandemic in 2020 breaks almost all stochastic forecasting models in three very important ways. The question then is how we can "robustify" our forecasting models so that they are less sensitive to these outliers and can continue to produce useable results. We have a number of requirements when we robustify stochastic models. The first is that we want to remove, or at least limit, the distortion in the parameter estimates. Second, we want to be able to estimate the effect of an outlier so that we can deduct it from our mortality series and get a clean starting point for the forecast. Third, we want a robust estimate of variance so that our value-at-risk capital is not overestimated. Above all, we want an objective methodology for all of this because we need to be able to re-calibrate these models thousands of times in value-at-risk calculations.

In the paper I propose a two-stage solution. The first step is to identify the position of the outliers using some statistical tests, then to co-estimate the effect of these outliers in conjunction with the other parameters, which essentially limits the biasing effect from the outliers. Before we come to the details, we will take a quick look at some alternative methods that have been used recently. The first and most obvious solution back in 2020 was simply to not use the 2020 data and just continue using the population mortality data up to and including 2019. This was a workable temporary solution, which was widely used and made perfect sense under the circumstances.

The problem today is that ignoring 2020 and 2021 is not a workable solution from 2023 onwards. The outliers are no longer at the end of a data series that one can just trim off. At the time of writing, the outliers have moved towards the middle of the data series. Ignoring the affected data is no longer an option.

Another approach might be to weight the log-likelihood functions. The issue is that the method of maximum likelihood is quite sensitive to outliers. However, there are a number of problems with weighting a log-likelihood. First, it is entirely subjective as to which years are an outlier and should be weighted in some way. Second, the weights applied are completely arbitrary: even if two analysts agree that 2020 is an outlier, who is to decide as to whether the weight should be 50%, 25%, 0% or 80%? Third, weighting a log-likelihood does not give any estimate of the outlier effect, nor does it tel you anything about the nature of the outlier either. Is the outlier a one-off? Is it one of a series of outliers? Is it an effect that is decaying over time? Is it a step change in the process? As we will shortly see, weighting a log-likelihood also only gives limited protection against the distortion and bias of the parameters.

We will look at this by way of example. In Figure 2 in the sessional paper, the solid line shows the log-likelihood for a Normal (0, 1) random variable, that is, an observation that has been standardised by deducting the mean and dividing by the standard deviation. We present it as the loss function function, that is, the negative log-likelihood function. A standard normal random variable has a u-shaped loss function, as in Figure 2 in the sessional paper. The contribution of the variable z is quadratic. This means that an outlier has a quadratically inceasing impact on the log-likelihood. The further the outlier gets away from a sensible value, the more distortion it causes.

For this reason, a lot of early attempts to robustify time series looked at replacing the unbounded log-likelihood with some alternative function that behaved very similarly for non-outliers, but which would either reduce or eliminate the impact of outliers. An example is shown in Figure 2 in the sessional paper. The solid black line shows the Normal loss function, so values outside (−3, +3) could be regarded as extreme or outliers. The further you go away from the mean, the greater the contribution to the log-likelihood and the greater the distorting impact. Instead, ρ-functions seek to behave like a log-likelihood function close to the mean, but limit the impact of outliers. In Figure 2 in the sessional paper, the ρ-function from Equation (2) behaves very much like the normal loss function over (−2, +2) and will allow non-outliers to make the same contribution as before; However,, the ρ-function has been structured so that outliers do not have quadratically increasing effect. In the case of the ρ-function in Figure 2 in the sessional paper, we have a linear increasing contribution outside (−1, +1).

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

Other ρ-functions have been proposed which cap the contribution to the log-likelihood. There is also a class of ρ-functions called re-descenders, where the contribution turns back down to zero. An extreme outlier would then make zero contribution to the log-likelihood.

Those are some alternative approaches to outliers from past and present. Now we will look at three major classes of stochastic model and how we could go about robustifying the forecasts in a more methodical and statistical way. We start with univariate mortality models. A univariate model has multiple parameter vectors, but only one of them represents a time index to forecast for projections. One example is the Lee–Carter model, shown in Equation (3) of the paper. Another example is the age-period-cohort (APC) model in Equation (4). In both cases, there are three different parameter vectors to estimate. In the Lee–Carter case, alpha and beta are estimated and then held constant for forecasting, but it is the kappa term, once it is estimated, to which we want to fit an ARIMA model and then forecast to get our projected mortality rates.

At this point it is helpful to have a more technical definition of an outlier. I have alighted on the following: "An outlier is an observation that is further away from the one-step-ahead forecast than is consistent with the noise variance in the process." To assess this, we follow a process from the first data point forward. One builds a picture as to the nature of the process and its parameters and then compares the next observation with the emerging forecast to see how extreme it is. We therefore need some kind of critical threshold to decide what counts as extreme, for example, 3.5 standard deviations away from what might be expected. In the wider literature there are typically at least four different types of outlier: innovation outliers, additive outliers, temporary changes and level shifts. Figure 3 in the sessional paper illustrates these, and we will quickly recap what these different outlier types are, and why only some of them are an issue.

In Equation (5) in the paper we define a simple moving-average process. It is not meant to represent a particular mortality process. It is merely an illustrative process to get a feel for what the different kinds of outliers are and how they look. Using the model in Equation (5), we simulate an uncontaminated moving average process in Figure 3(a) in the sessional paper and then add an innovation outlier in Figure 3(c). Because Equation (5) is a moving average process, the outlier at time 29 also leads to an outlier at time 30; there is a year-on-year correlation here. An innovation outlier happens at time 29 but, because it is an integral part of the process, it also results in downstream effects that are consistent with the nature of the process.

An innovation outlier is usually a relatively modest outlier that is still integrated into the underlying process. A mortality-related example might be a year with particularly heavy winter mortality because of an especially virulent strain of influenza in circulation. This is then followed by a summer of somewhat lighter mortality due to the effect of harvesting, as frailer individuals had died in the winter as part of the heavier mortality. As a consequence, there would be slightly lighter mortality in the half-year or so following the winter. That would be an example of an innovation outlier, which can be regarded as perhaps a slightly more extreme example of a normal event. An innovation outlier is built into the process because there are immediate downstream consequences that are consistent with the normal behaviour of the process. Our handling of these kind of outliers is simply to leave them alone.

Extreme winters with heavy mortality, and any autocorrelation effects, are a fundamental feature of seasonal mortality variation, so we would not want to remove these outliers. They are a basic feature of the process. If, however, we take the uncontaminated process in Figure 3(a) in the sessional paper and put in an additive outlier, as in Figure 3(b), this is a one-off effect that has no downstream effects. It is simply a one-off change at a particular point in time with no downstream consequences. An additive outlier is a more extreme outlier that we want to exclude if we are looking to model the underlying process. An example might be a pandemic or some other negative mortality event in a single year that does not get repeated. An event happens, there is additional mortality and that is it. There are no downstream consequences, because the outlier is not part of the process we are trying to model.

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

Alternatively, we can take the uncontaminated process in Figure 3(a) in the sessional paper and put in a "temporary change" as in Figure 3(d). This is like an additive outlier, but it takes some time for its effects to wear off. There is a step change when the first outlier occurs and then the effect decays over time before the process returns to normal. One could regard this as a short series of decaying additive outliers. The temporary change, like the additive outlier, is not integrated into the process that we are trying to estimate. It could be, for example, a war or a pandemic that lasts two or three years with some downstream consequences before things return to normal. As with the additive outliers, we would want to co-estimate the effect of these outliers to remove the bias that would otherwise result.

The last example is a level shift in Figure 3(e) in the sessional paper. It is essentially the same process, but shifted up or down from a particular point. This is a permanent change in the level of the process. For a mortality example, after German reunification in 1990, the old-age pension system was unified across East and West Germany. As a result, the mortality levels of East German pensioners, which were higher than West German pensioners, rapidly converged to the West German levels of mortality and then stayed at roughly West German levels thereafter. This is an example of a permanent shift over a relatively short period of time that then sees the mortality process evolving as before, just at a completely different level. We would not try to control for this as an outlier. Instead, we would change the nature of the model or, more likely, we would change the data period that we were using and restrict ourselves to the data period after the step change. We would not seek to remove level shift outliers, we would probably just shorten the dataset to exclude them.

How would one go about actually implementing any of this? This is surprisingly straightforward. To robustify a univariate ARIMA model for forecasting kappa, we would use the methodology from Chen and Liu (1993), who made two extremely useful contributions. The first is that they proposed a collection of four test statistics, one for each kind of outlier, that identify an outlier's location. They are objective statistical tests that tells you where the outliers are in a time series. Second, having identified the location of an outlier, Chen and Liu (1993) provided four further statistical tests that classify an outlier's type. I wholeheartedly recommend reading Chen and Liu (1993), as it is a very accessible paper.

One minor caveat is that, while one can detect an outlier anywhere in the time series using Chen and Liu's test statistics, one cannot tell the nature of the outlier if it occurs at the very end of the series. One needs to have a couple of observations after the outlier to work out what kind of outlier it is. This is a relatively minor consideration, however.

As an example we can consider the Lee–Carter model for the mortality of males in England and Wales. As we have seen, if we just use an ordinary ARIMA model for kappa, we get a nonsensical forecast. If, however, we first use the methodology of Chen and Liu (1993), it identifies 2020 as an outlier. By co-estimating the ARIMA model with an outlier effect, we get a different forecasting model that also allows us to estimate what 2020 would have looked like without the pandemic. The forecast starting point is therefore corrected, and we have far more sensible forecasts, as shown in Figure 4(b) in the sessional paper. The methodology is described in more detail in the paper; implementation details in R are given in Appendix A. We can use the methodology from Chen and Liu (1993) to robustify almost any univariate stochastic forecasting model.

The situation is similar for females, albeit 2020 only seems to bias the starting point for the forecast, as shown in the dashed line of Figure 4(a) in the sessional paper. The Chen and Liu method again identifies 2020 as the outlier, and it allows us to calculate a clean starting point for our forecast, as shown in the solid line of Figure 4(a) in the sessional paper.

Chen and Liu's procedure allows us to identify the location of outliers using an objective, quantitative criteria. If we want, it also allows us to classify the nature of the outlier. Very importantly, it allows us to estimate the outlier effect and because of this we can calculate a clean version of the series and, therefore, a clean starting point for our forecasts. As a result, the forecasting model parameters will be unbiased and undistorted, as we would hope.

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

So, that is the situation for univariate forecasting models. However, there is a very important class of forecasting models with two or more time indices. One example is the model from Cairns et al. (2006), which is essentially a bivariate random walk. The two parameter vectors, κ_0 and κ_1, are estimated and, for forecasting, we project them jointly as a bivariate random walk with drift. This gives us our forecast of mortality rates at all ages. Table 2 in the paper lists some other members of the Cairns–Blake–Dowd family, which is particularly important for actuarial work.

Such multivariate models lead to some interesting considerations. In section 5.3, I fitted an M9 model, which has five parameter vectors, of which the three kappa vectors are projected forward in time. Figure 6 in the sessional paper shows these three kappa vectors, where it is not entirely obvious that there is an outlier. We can see that the 2020 value for κ_0 is somewhat higher than the others, but it is nowhere nearly as clear that this is an outlier compared to the univariate case. Visual inspection is not necessarily a reliable way of identifying outliers. The key feature of Cairns–Blake–Dowd forecasts is that the first differences of the kappa vectors should be roughly constant. These differences are plotted in a pseudo-3D view in Figure 7 in the sessional paper. One of the issues with visual inspection is that whether or not an outlier is obvious is rather dependent on the viewing angle. Depending on how you rotate Figure 7, 2020 may or may not look like an outlier. Sometimes, one might identify a different year as an outlier. This is a reminder of why we need an objective methodology that allows us to calculate a scalar test statistic.

There are many ways in which we could reduce a multivariate problem to an objective univariate test. One is the Mahalanobis distance. Indeed, there are multivariate robustification methods from Hadi (1992, 1994) and Hadi *et al.* (2009) that use robustified versions of the Mahalanobis distance. However, the method I have used in the paper is from Galeano *et al.* (2006). I will briefly look at these alternatives in turn.

The Cairns–Blake–Dowd family projects the kappa values forward using a p-dimensional random walk with drift. This assumes that the first difference of the kappa values has a multivariate normal distribution. This means that we could take our first differences, calculate the mean and the covariance matrix, and use the Mahalanobis distance to standardise the values as in Equation (14). We subtract the mean from each distance and then, by using the inverse of the covariance matrix, we reduce each p-dimensional difference to a single scalar value D_j. Under the null hypothesis, D_j^2 will then have a chi-squared distribution with 3 degrees of freedom, which will allow us to test whether or not a particular observation is an outlier or not. If we do this for the M9 model for females, we see in Figure 8 in the sessional paper that 2020 has a larger Mahalanobis distance than the other years, but is it extreme enough to count as an outlier? The square root of the upper 1% point of the χ_3^2 distribution is 3.6, so we can see that 2020 is indeed an outlier. However, there are two other years in Figure 8 in the sessional paper that are quite close to being regarded as outliers. The critical threshold needs to be set carefully so asnot to identify too many false positives.

However, there are some problems with the Mahalanobis distance. Hadi (1992) pointed out that the Mahalanobis distance itself is not robust because it is affected by masking and swamping. Masking is where an outlier is hidden because the outlier itself has inflated or distorted the covariance matrix, which in turn is used to identify outliers. If an outlier is sufficiently extreme, it distorts the covariance matrix enough to hide itself. Similarly, non-outliers can sometimes have seemingly large Mahalanobis distances because the outlier has distorted the mean of the process. Then, once you have deducted the mean to standardise, non-outliers end up looking extreme. This may be the case for those two years in Figure 8 in the sessional paper that are quite close to the critical threshold. So, the Mahalanobis distance itself is not entirely robust. It is for that reason that Hadi (1992, 1994) presented a method of identifying outliers in multivariate data with a robust estimate of the mean and a robust estimate of the covariance matrix.

However, tonight's paper uses a quite different methodology from Galeano *et al.* (2006), who use an approach called projection pursuit. I am not going to go into the details of projection

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

pursuit; there are some references in the paper for those who want to delve into the subject more deeply. Essentially, if we think back to the loss function in Figure 2 in the sessional paper, where an outlier had a very distorting effect because the loss function was quadratic. Projection pursuit exploits the fact that, if an outlier distorts the variance (a second-order moment), then it really distorts the kurtosis (a fourth-order moment). Projection pursuit is then a search for observations that have a very distorting effect on the kurtosis. Figure 9 in the sessional paper shows an application – we have our M9 model with data up to 2020, which looks like an outlier. If we do not robustify our forecasting approach, we get the forecasts shown as dashed lines that are in broadly the right direction, but starting at the wrong point, that is, with elevated 2020 mortality continued forever. If, however, we use Galeano, Peña and Tsay's methodology, it identifies the outlier and provides us with an estimate of the outlier effect. We can then deduct this outlier effect to calculate a clean starting point for the robustified forecast, shown as the solid lines in Figure 9.

The methodologies of Chen and Liu (1993) and Galeano *et al.* (2006) cover a large proportion of the stochastic mortality models in use. Most models used by actuaries are either univariate or multivariate time series models. However, there are some other mortality forecasting models that do not use time series at all. One example is the two-dimensional penalised-spline model from Currie *et al.* (2004). This was extended by Kirkby and Currie (2010) to include the estimation of period shocks or period effects. If we fit this model to the England and Wales data for females, we get the estimated period effects in Figure 10 in the sessional paper. We have fitted a smooth surface for mortality by age and time and we also have the estimated period effects as departures from the smooth surface. Figure 10 shows a clear, large period effect for 2020, but there are also noticeable period effects in almost every other year. However, Figure 10 shows the results from an unscaled model. Kirkby and Currie (2010) further decided that the minor period effects are of relatively little interest. They then developed a scaling procedure that would smooth the minor period effects closer to zero, simply because they are of relatively little consequence. Kirkby and Currie (2010) advocated optimising the scaling factor with reference to the Bayesian information criterion (BIC), so the degree of scaling applied to the period effects was calibrated to the data set and the strength of the largest shock signal itself. Figure 12 in the sessional paper shows the result of this optimised scaling.

One of the interesting things in Figure 12 in the sessional paper is that the COVID shock in 2020 has a qualitatively different shape to Figure 10. The shock effect broadly reduces with increasing age in 2020, which contrasts with previous period effects where the additional mortality of a period effect increases quite clearly with age. The 2020 shock is therefore not just quantitatively different in terms of its size, it is also qualitatively different in terms of its shape with age.

What does this do to the forecasts? Under the 2D age-period (2DAP) penalised-spline model of Currie *et al.* (2004), if we do not carry out any kind of procedure to make the results robust, we get a nonsensical forecast shown by the dashed lines in Figure 13 in the sessional paper. If, however, we use the methodology from Kirkby and Currie (2010), we can estimate the period effects, particularly the shock effects, and because we are co-estimating a fundamental process plus the shock effect, this results in a more sensible direction forecast, as shown by the solid lines in Figure 13.

To conclude, the procedure recommended in the paper is co-estimation of outlier effects with the model parameters. This will reduce or, hopefully, eliminate the bias in the forecasting parameters. By estimating the outlier effects, we can also calculate a clean series for the forecasting starting point. Most importantly for actuaries, this will reduce the variance and stop the value-at-risk capital requirements being inflated.

For univariate forecasting models, for example the Lee–Carter and the APC models, we can use the methodology of Chen and Liu (1993). For multivariate models, including the Cairns–Blake–Dowd family and the newer Tang–Li–Tickle family of models, we can use the approach of Galeano *et al.* (2006) to robustify the forecasts. Lastly, for the two-dimensional penalised-spline model we

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

can use the approach outlined by Kirkby and Currie (2010) to estimate period effects and shocks and thus robustify our forecasts.

That concludes my presentation. Before throwing the session open to questions, I would make one final observation concerning academic research versus "practical" actuarial work. At the time that Kirkby and Currie (2010) was published, it would have struck every actuary, probably me included, as a purely academic exercise completely disconnected from everyday actuarial or commercial concerns. After all, who in 2010 needed to allow for mortality shocks like the 1919 influenza pandemic? Well, in 2023, we now have the answer – every actuary reserving for pensions and annuities! It appears that sometimes at least a decade must pass before people can truly see the value in a piece of academic research. Thank you for listening and I am now open to questions.

Moderator: Does anyone have any contributions they would like to raise?

Questioner: COVID did impact some cohorts more than others. So, you have removed a certain cohort from that. How has that impacted on the overall data set? Second, I wonder has that changed the whole structure and attributes of the remaining data set? Because that could go two ways. You have almost a healthier set of people remaining so you could get greater future improvements. Alternatively, you have people who have recovered from COVID, the survivors, and you do not know what is going to happen to them in future. We are still quite close to the end of the time series so some of that future has not come through yet. How has that been allowed for?

Dr Richards: This sort of phenomenon is perhaps clearer in higher frequency datasets. What I have used in the paper is annual data and that does not give you too much detail. If, however, you use more granular data, then you can see a lot of what you have described. In Richards (2022), I used data from a UK annuity portfolio where we had daily death data at the level of the individual. This enabled the detection of numerous mortality effects in time. It picked up seasonal effects very reliably. It picked up the COVID shocks, particularly the first shock. It also picked up the very deep trough in the summer of 2020, where mortality was unusually low following the abnormally high mortality of April and May 2020. Demographers call this "harvesting," where COVID brought forward some deaths in addition to causing many new deaths. So, the effect to which you are referring is in the data, but you can only see it if you have data that is at a high enough frequency and use something like a survival model with a flexible time effect.

If you look at population mortality following the pandemic outbreak, the UK also had a second shock in January 2021. While the mortality shocks have passed, there is still an ongoing problem with excess deaths in the UK, perhaps partly driven by lingering long COVID effects. However, at least as significant is all the missed treatment appointments during the pandemic and our much longer waiting lists. I cannot comment on the Scottish lists off the top of my head, but the English waiting lists are the highest they have been for at least fifteen years. So, there are ongoing consequences, some directly linked to the pandemic, some not. One might have expected faster improvements post-pandemic due to harvesting. However, we actually see the opposite, but for reasons not necessarily directly linked to the virus.

Prof. A. J. G. Cairns, F.I.A.: I have a question and a comment. We are now at a stage where the Office for National Statistics (ONS) is releasing or updating the exposures and the population estimates. Even before that, we had a reasonable idea about what was happening in terms of death counts, although maybe the information about death rates is not as precise. So, the question would be, as you add in these extra years, which are still in some sense outliers, how does that affect the estimates you are making and the model-fitting and detection of those outliers? The other one is a comment, which is can we use a little bit of expert judgement in terms of what types of outliers are we looking at. For example, just based on the Spanish flu 100 years earlier, we know the COVID

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

pandemic is, in some sense, a temporary shock – a "temporarily change" in your terminology. But then I also think with COVID what we can see is a small level shift in addition to that, which might be the effects of long COVID. It might be other things, such as some of the medical innovations that we might have expected in the last two or three years having been delayed, and we might never catch up with two or three years' worth of mortality improvements. So, on that basis, would you say there is a role for a little bit of expert judgement? Or do you prefer to take the pure approach?

Dr Richards: Everyone present will be glad to hear there is always a role for some expert judgement! I prefer to apply the expert judgement at the end of a process with clearly defined quantitative statistical tests. I do not like inserting the expert judgement early in the process. COVID is proving interesting because its effects do not neatly correspond to one of the four types of outliers. In fact, there are elements of all four types of outliers in COVID. We have the first shock in April & May 2020, which looks like an additive outlier. Then we saw the deep trough the following summer, so there was an autocorrelation aspect akin to an innovation outlier. Then, we have the post-pandemic situation. The virus still circulates, and while it does not kill nearly as many people as it used to, it has left UK society and our health system in a worse state than it was pre-pandemic. So, there are also elements of the temporary change.

These methods are useful, and they provide, I believe, a very nice objective methodology, but there is always a need for expert judgement at some point. A good example in the paper is the critical threshold for statistical tests to determine an outlier. If one picks 2.5 standard deviations as the threshold, 2020 is identified as an outlier along with nine other years, which is clearly too many. If one picks 3.5 standard deviations, only 2020 is identified as an outlier. There is a degree of expert judgement in setting the critical threshold value. Should it be 2.5, which is what a lot of economists and others use? That I would regard as too low for actuaries, but economists tend to deal with far longer time series than actuaries.

Moderator: The other question is about data.

Dr Richards: Re-statement of population estimates is a nuisance, as explored in Cairns *et al.* (2016). There was a big problem when the 2011 census led to restatements of the population estimates for the immediately preceding years. Population estimates above age 80 went down, estimated mortality rates went up, and so the implied improvements went down. What previously seemed like quite strong mortality improvements for people in their eighties ended up, due to the population estimates being restated, as weaker improvements than we first thought. I have not yet looked at the impact of the most recent updates of the population estimates.

Prof. A. D. Wilkie, F.I.A.: As you said, quite rightly, at the end of 2020, if you only knew the number of deaths for the year and not the distribution within the year, you do not know which type of outlier you have. You showed images of positive outliers, but a lot of economic series can have downwards outliers. So, you, on this basis, have got eight outliers, the four upwards and four downwards ones. You also have another possibility that has not been discussed at all, which is that the outlier just continues. You could call this an explosive or catastrophic case. These can happen in economics, with hyperinflation or a share price or an industry collapsing to zero. They can happen in plants and animal statistics with extinctions or with an invasive species taking over, as rabbits did in Australia or rhododendrons did around Loch Lomond.

My approach has been not to use a normal distribution and look for outliers, but to try and find other distributions that include the outliers because in so many of the series there are quite a lot of them. Certainly, economic ones bounce up and down a lot. Fatter-tailed distributions are appropriate, like Laplace or the hyperbolic series of distributions, which are related to Normal distributions. Normal will give a kurtosis of three, Laplace gives a Kurtosis of six, and skewness can take it up a bit.

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

With mortality though, the outliers are much more likely, I think, to be related to mortality moving upwards. I will rename slightly the old four horsemen of the apocalypse as war, famine, death, and natural disasters. All of those push mortality up a lot. It takes an awful lot of skill and expertise to bring down mortality slowly and carefully, and it is very unlikely to come down abruptly. On the other hand, share dividends or company profits can come down with a very big jump at times. So, there is an awful lot more to discuss about this, I think. Stephen (Richards) has given excellent examples of a set of ways of dealing with these issues while sticking to a normal distribution.

Dr Richards: You make some interesting points. I have given a lot of thought to the fact that these tests are symmetric and that they regard the likelihood of a positive outlier as being the same as a negative outlier. But as you say, mortality is a process that does not have a symmetric distribution. It is very quick and easy to raise mortality, but it is slow and difficult to lower it. Something needs to be done to address the fact that there is a heavier tail, but it is probably only a heavier tail on one side. There is work that could be done on a mortality-appropriate distribution that, say, has a mean of zero but a heavy tail on the positive side and a much lighter tail on the negative side. As you say, a normal distribution, or any other symmetric distribution, does not feel strictly appropriate.

Mr H. R. D. Taylor, F.I.A.: I have spent my entire career specialising in innovation and techniques for innovation. I would like to thank Stephen (Richards) for this paper which I think is not only timely but is extremely practical. It displays one of the core techniques of innovation. If there is a problem that is looking for a solution, one of the first places to look is at adjacent fields where people have come up with solutions to a similar problem. Then one can see whether those solutions can be adapted to solve the problem that you are facing. So, with that in mind, I wonder if weather forecasting might provide some insights. There is extremely complex modelling involved in forecasting weather and forecasting has improved a lot over the last few years.

I understand that for some years now, tools and techniques from the world of artificial intelligence have been applied to the complex mathematical modelling of weather. This approach is producing better results. So, one question for Stephen (Richards) is whether, somewhere down the line, this is another field where artificial intelligence might be of use given we are handling large amounts of data and large amounts of complexity in interactions between the data.

The final thought is, I remember listening to what some epidemiologists were saying about the COVID pandemic when it was happening. It seemed to me that we managed that pandemic. We are clear now and we have new techniques for vaccination. However, I am not convinced that, globally, because of the way we live now with high interconnectivity and the demands of people to have their rights and do as they please, we have seen the last pandemic in our lifetimes. Therefore, the more fundamental question is whether the effect of a pandemic on the numbers that actuaries use is really an outlier, or is it something that we are going to see more frequently? I am not suggesting we would see a pandemic every year, but the next event like this, which is health-related, may happen much sooner than any of us in this room are imagining at the moment. I think that a lot of the conditions to manage the risks of pandemics have not been addressed because we are now out of the COVID pandemic and the world is focused on other things.

Dr Richards: A lot of the techniques that I looked at were for identifying outliers in large data sets with lots of variables. Such data sets are used by people in the artificial intelligence field, for example, the procedures developed by Hadi. Hadi has published more recent work, which is directly driven by the challenges and needs of people working with data sets with very large numbers of observations, and also very large numbers of variables per observation. This research is actually very useful because people have already looked at the outlier problem and solved it. There are, however, some unique actuarial features. The type of data used by Hadi tends to have

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

observations with much higher dimensionality, whereas actuarial stochastic models tend to have only two or three dimensions. Often, papers on robust methods have analysis and tests based on hundreds of observations and 40 or more dimensions. In contrast, in the actuarial world we typically have around 50 observations and two or three dimensions for mortality forecasting. For this reason, Appendix C of the paper contains some simulation tests to show that the Galeano, Peña and Tsay approach is actually far better for differenced series than un-differenced series. This is relevant for the particular use-case that actuaries will see: low-dimensional data with a modest number of observations.

Moderator: Is there a risk that a 1-in-200 simulated years from the non-robustified models could be considered outliers using these new methods?

Dr Richards: If you took the data just until, say, 2019, you have a relatively well-behaved mortality process. If you simulate from that, you are unlikely to generate outliers unless you picked an abnormally low critical defining threshold, such as 2.25. If you simulate 10,000 processes and use a critical threshold of 2.25, you will generate quite a few outliers, but not if you picked a more sensible critical threshold like 3.5. What this question, I think, is actually driving at is, if you do not robustify the model, you fit a distorted model and then do value-at-risk simulations from that. If the variance is grossly inflated, would simulation generate such a wide spread of broken forecasts that you would either find lots of outliers or none at all? I do not know the answer to that. I think the short answer is that it is important to use robustification techniques if you have got a data set that is affected by outliers.

Mr D. J. Grenham, F.I.A.: First, at the population level mortality may not improve quickly, certainly in the developed countries. But, for example, infant mortality or child mortality in developing countries can improve quite quickly and so getting down to that granularity of data that can be useful. Second, how well do these approaches work for catastrophe events? For example, does the 1-in-200 event get worse?

Dr Richards: I think that would cut to the question of the nature of the liability for which you are reserving and calibrating capital. My paper, and these models, are all essentially done with the intention of forecasting mortality for pensioners and annuitants. There is really no risk, if you like, on the liability side from a sudden increase in mortality, there is only profit. If you are an annuity writer and there is a mortality shock, you just keep quiet and enjoy your extra profits. For term assurance business, you would adopt a completely different approach. There, you would cherish your outliers and not look to robustify against them. If you were reserving for term assurance liabilities, you would be looking at past catastrophes, such as the Spanish influenza pandemic, and you would have to craft other scenarios based around antibiotic resistance and so on. You would have a lot of speculative scenarios that you would have to generate and calibrate. If you are reserving for term assurance business, COVID provides you with a very recent example of how quickly mortality rates can shoot up.

For term-assurance business you would use your outliers to inform your reserving process in a very active way. One interesting aspect of COVID, compared to the Spanish influenza pandemic, is that the increase in mortality in 1919 was actually much higher than it was in 2020. COVID was, as a pandemic, relatively moderate in terms of its mortality impact, at least compared to the 1919 Spanish influenza pandemic.

Another curious aspect of the 1919 pandemic was the specificity about which cohorts it affected most. People aged above 60 in 1919 were largely unaffected. If you look at the profile of mortality excess in 1919, you will see next to no additional mortality for people aged 60 and above, but you will see mortality rates tripling for males in their late 20s. So, there is a very specific shark-fin shape for the Spanish influenza pandemic, but COVID excess mortality took a very different

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

shape – there was very little additional mortality for young people. On a log scale, COVID was essentially a parallel shift. Young people had low mortality rates, and a 50% increase in those low mortality rates still did not kill very many young people. However, at retirement ages, COVID had a much bigger impact, essentially the exact reverse of the Spanish influenza pandemic. You would perhaps use that outlier to inform a reserving process for term assurance, but one has to be very careful not to be lulled into a false sense of security. If you were writing term assurance business, COVID was actually quite a "good" shock because it did not generate too many additional deaths and, by extension, claims. In contrast, something like the 1919 Spanish influenza pandemic would have been far worse for a term assurance book.

Mr G. C. Wood, F.I.A.: Thank you for a very interesting and timely paper. I have been introduced to Tang–Li–Tickle models and Hermite splines, which I had never come across before I must admit, and the use of the verb "robustify." This is timely, because like a number of other colleagues in this room, I spent a few days last week in Birmingham at the annual Life Convention for life actuaries and did attend an interesting conversation on mortality forecasting where none of the main elements of this paper were discussed. It was all about applying expert judgement to the CMI model. For example, people were discussing matters like "if we have zero weights for 2020 and 2021, what shall we use for 2022, what is 2023 looking like? Do we assume a 25% weight?" There was no real science behind it at all. I dislike the use of the phrase "expert judgement." If you look at your univariate model section, you talk about critical thresholds of 3 or 3.5. If we say, "ten observations are too many, and one sounds just about right," that almost looks like reverse engineering, if you are presenting to, for example, an audit committee or a senior management committee. It is not necessarily expert judgement. I believe that if you cannot explain in the real world what it is you are assuming, then that is a failure of the technician. We should be thinking, what do we have to believe? The classic example is, of course, the COVID pandemic. Do we choose to believe it can be represented by some decay function we have seen in practice?

You talked about long COVID and delays to treatment, etc. So it would be quite easy to explain to, for example, an audit committee or a senior management committee that, say, we believe it will be two years or four years before things are back to normal, and on that basis we are doing our forecasting assuming a decay function. We should not talk about expert judgement all the time. We should not try to bamboozle audit committees with kappa functions or this estimate or that. We need to put things in ways where we can explain the underlying assumptions and why we choose to believe on that basis our forecasting is appropriate.

Dr Richards: I have two responses. First, the reason none of this was discussed at the Life Convention was that the organisers declined my paper when I offered it to them! Second, regarding expert judgement, what I am advocating is transparency so that people are always clear on what is being done and why. I like objective methods because they follow mechanistic procedures that have an underlying statistical rationale. Chen and Liu (1993) is an excellent example of this. This is a very clever example of how to fit a model and use the residuals to work out further detail about what is going on. I allude in the paper to aspects of setting, for example, the critical value for deciding what is an outlier. I give an example where, if you pick 2.25, you get too many outliers. But who am I to say what counts as too many? I would agree that it is important to be able to explain to the users of the model, but I think, probably, what we are both looking for is not so much expert judgement as clarity and transparency and the removal of subjectivity and obfuscation. So, perhaps what I am most against is fudging things in the background, where it is not clear to anyone downstream what was done and why, or what the rationale was for it because then it becomes hidden. It becomes a black box that nobody understands. I do not think I disagree with the points you are making, but I like the transparency of the process. I also like the automation – for tasks like value-at-risk capital simulations, one cannot subjectively intervene in every single one of ten thousand scenarios, inspect it and decide that that one is an outlier whereas

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

another is not. We need a quantitative process that can just run 10,000 times. That particular requirement, for me, makes it really useful to have a quantitative as well as a statistical and objective approach. You can then program it and then just let it run.

Questioner: I completely agree with your last point. I think that having that transparent methodology for identifying an outlier is much needed. It is a problem which I have heard many of my colleagues in Club Vita talk about. Indeed, it is a problem that has manifested itself in different ways in different countries. It is interesting you have used all this wonderful UK data, but we see different patterns in the USA and Canada. I would encourage anybody who wants to test out any of the approaches to try to repeat them in Canada because the Canadian experience was remarkably good in 2020 and 2021, and then it went horribly wrong in 2022. It looks like they had the patience of saints for two years, and then all the social interaction got going again in 2022 and the excess mortality jumped from 2% or 3% in the previous two years to 15% in 2022, or thereabouts.

I wanted to say how welcome and timely the paper is because the growth in the pension risk transfer market and the amount of tradable longevity risk that is now coming to market is massive. We are starting also to see longevity risk being traded in markets other than the UK. It is important for the people who are underwriting that risk that we have a competitive market within it. Having a plurality and diversity of opinions is, in my view, to be encouraged, because without that difference of opinion you would not get people taking different views about where to price the risk. We should not be trying to encourage everybody to end up all settling on exactly the same value. That is a really negative outcome, I think, from my perspective.

On the subject of expert wisdom or judgement, I would encourage us to spend our time, not so much on the mathematics and the experts trying to override the mathematics, but more on trying to understand what is in the pipeline of longevity improvements, or dis-improvements, as they call them in North America. You touched on the stress in the NHS in the UK, but you could equally talk about the weight loss drugs that are being introduced at the moment, and their potentially dramatic effect. In the UK something like 30% of people are deemed to be obese at the moment. Those weight loss drugs, if you could afford them, would have a really significant impact on mortality.

Moderator (reading written question): The paper was a well-researched piece of work that will be of practical use to actuaries less familiar with the underlying statistical methods. On page six, you talk about the innovative outliers are left in because they are part of the underlying process, but others are excluded. Why would the level shift outlier be excluded?

Dr Richards: It is not so much that we would exclude the level shift. We would basically trim the data series and only use the most recent relevant data for a consistent level. You would not include data for a process that had such an obvious step jump with the classical stochastic projection models. There are some variants of models that allow broken trends, but that is a very specialist subset. I think most people would probably just trim the data series and lose the level shift that way.

Moderator: On page eight you referenced the COVID shocks occurring just over 100 years after the Spanish flu. These are two 5-sigma events in 100 years, so the model is either wrong or underestimates the true rate of variation of the innovation process. I agree with what you are saying here, but I wonder if there is an issue caused by measuring mortality in calendar years? Both the Spanish flu and COVID events were single pandemics that lasted longer than a calendar year, influencing two years in each case but could easily have impacted three years or one year in each case.

https://doi.org/10.1017/S1357321724000084 Published online by Cambridge University Press

Dr Richards: Both the Spanish influenza pandemic of 1918–19 and the COVID pandemic in 2020–21 had double spikes in the UK. The Spanish influenza pandemic had one major spike in late 1918 and then another equally dramatic shock in early 1919 (with a possible earlier spike in 1918). It is regarded as one event but with multiple shocks. COVID is very similar. It is regarded as one pandemic, but the UK mortality has two clear spikes. These events are multi-year events. The true impact of them can only be seen with higher frequency data. Most of the 2020 additional mortality is concentrated in April and May. Using annual statistics very much understates the sheer intensity of these mortality shocks. I am a strong advocate of using higher frequency data to properly investigate the intensity of some of these events.

Moderator: Thank you, Stephen (Richards). That is all the time we have for questions. I would like to thank Stephen for a very informative presentation. I read the paper and I found the presentation clarified my understanding so thank you.

References

Cairns, A.J., Blake, D., Dowd, K. & Kessler, A.R. (2016). Phantoms never die: living with unreliable population data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **179**(4), 975–1005.

Cairns, A.J.G., Blake, D. & Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance*, **73**, 687–718. https://doi.org/10.1111/j.1539-6975.2006.00195.x

Chen, C. & Liu, L.M., 1993. Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, **88** (421), 284–297.

Currie, I.D., Durban, M. & Eilers, P.H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, **4**(4), 279–298.

Galeano, P., Peña, D. & Tsay, R.S. (2006). Outlier detection in multivariate time series by projection pursuit. *Journal of the American Statistical Association*, **101**(474), 654–669.

Hadi, A.S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **54**(3), 761–771.

Hadi, A.S. (1994). A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **56**(2), 393–396.

Hadi, A.S., Imon, A.R. & Werner, M. (2009). Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics*, **1**(1), 57–70.

Kirkby, J.G. & Currie, I.D. (2010). Smooth models of mortality with period shocks. *Statistical Modelling*, **10**(2), 177–196.

Richards, S.J. (2022). Allowing for shocks in portfolio mortality models. *British Actuarial Journal*, **27**, 1–22. https://doi.org/10.1017/S1357321721000180.

Richards, S.J., Currie, I.D. & Ritchie, G.P. (2014). A value-at-risk framework for longevity trend risk. *British Actuarial Journal*, **19**(1), 116–139. https://doi.org/10.1017/S1357321712000451

Cite this article: Robust mortality forecasting in the presence of outliers. *British Actuarial Journal*. https://doi.org/10.1017/S1357321724000084