Parameter risk in time-series mortality forecasts

Kleinow, Torsten
T.Kleinow@hw.ac.uk
Department of Actuarial Mathematics and Statistics
and the Maxwell Institute for Mathematics Sciences
Heriot-Watt University, Edinburgh

Richards, Stephen J.
stephen@longevitas.co.uk
Longevitas Ltd, Edinburgh

February 15, 2016

Abstract

The projection of mortality rates is an essential part of valuing liabilities in life-insurance portfolios and pension schemes. An important tool for risk-management and solvency purposes is a stochastic projection model for mortality. We show that ARIMA models can be better representations of mortality time-series than simple random-walk models. We also consider the sometimes-overlooked issue of parameter risk in time-series models — formulae are given for decomposing overall risk into undiversifiable trend risk (parameter uncertainty) and diversifiable volatility.

Using the bootstrap approach from Pascual et al. (2004) we find that, while certain kinds of parameter risk are negligible, others are too material to ignore. In our specific mortality examples, a modification to the procedure from Pascual et al. (2004) reduced bias when bootstrapping the variance of the volatility, σ_ϵ^2. The conclusions have relevance to projection models used by insurers in the European Union under Solvency II.

Keywords: mortality projections, longevity trend risk, parameter risk, model risk, ARMA, ARIMA, Solvency II.

1 Introduction

The Solvency II regime in the European Union (EU) requires a probabilistic model for the various risks that an insurer carries on its balance sheet. In particular, Solvency II requires that an insurer holds enough reserves to cover 99.5% of scenarios which might occur over a one-year time horizon. The former ICA regime in the UK and the Swiss Solvency Test (SST) in Switzerland are defined similarly, and with the same 99.5% requirement.

An important, non-diversifiable risk for many insurers is longevity trend risk, i.e. the uncertainty over the future direction of mortality rates for annuitants. This same risk is present in many corporate defined-benefit pension schemes, although such entities are not regulated in the same manner as insurers. Longevity trend risk is the broad direction taken by mortality rates over many years, as opposed to the year-on-year fluctuations which we will refer to as the volatility. We note in passing that longevity trend risk unfolds over many years, so it does not naturally fit into the one-year time horizon required by regulators in banking and insurance. Some authors have proposed frameworks which address this: see Börger (2010), Plat (2011) and Richards et al. (2014) for examples.

As advocated by Börger (2010), the most useful tool for investigating uncertainty over longevity trend risk is a stochastic projection model. There is a wide choice of such models in the literature, and a quantitative review of some of the most commonly used ones in actuarial

work is given by Cairns et al. (2009). However, the choice of model — or models — often involves significant judgement by the analyst: a change in model can lead to material changes in the best-estimate reserves, while even within a model family there can be major differences (Richards and Currie, 2009). This phenomenon of model risk is particularly important, and it requires that analysts use more than one stochastic projection model when assessing longevity trend risk.

A key driver of capital requirements for annuity business is the uncertainty over the trend of future mortality rates, which can be measured as the variance of the mortality forecast values. In this paper we consider the decomposition of this uncertainty into two parts: (i) the uncertainty over the broad trend and (ii) the temporary volatility. In particular, we seek to investigate their respective contributions to overall uncertainty, and thus to the capital requirements for longevity trend risk. We will assume that the trend risk is synonymous with the uncertainty over model parameters (more on this later), and we will consider under which circumstances (if any) it is acceptable to ignore certain kinds of parameter uncertainty. We will illustrate with reference to the model by Lee and Carter (1992), but the basic conclusions will apply to any model which uses time-series methods to project a mortality index. We will also quantify the respective contributions to capital requirements using a value-at-risk calculation suitable for insurers operating under Solvency II.

Since parameter uncertainty is at the centre of our approach we will pay particular attention to the finite-sample distribution of the relevant estimators. We do this by applying the parametric bootstrap method of Pascual et al. (2004) for ARMA processes. We found that, in our specific mortality examples, a modification to the procedure from Pascual et al. (2004) reduced the bias when applied to obtain bootstrap samples from certain parameters, in particular the variance of the volatility.

The structure of this paper is as follows: Section 2 describes the data and model used to produce the example time series we want to forecast. Section 3 outlines the structure of a random walk for forecasting and considers the separation of forecast uncertainty into components due to parameter uncertainty and volatility. Section 4 outlines the structure of an ARMA or ARIMA process for time-series forecasting, while Section 5 considers the improvement in fit and forecast over the random-walk model. Section 6 considers an alternative ARIMA model which does not fit the data as well, but which has better forecasting properties, while Section 7 considers the implications for insurer capital requirements. Section 8 considers a practical point for actuaries in commercial work when calibrating the CMI spreadsheet for mortality projection. Section 9 discusses the results and Section 10 concludes the paper. To preserve the narrative flow of the paper, the major mathematical proofs are presented separately in the appendices.

2 Data and example mortality index

The data used for this paper are the number of deaths $D_{x,y}$ aged x last birthday during each calendar year y, split by gender. Corresponding mid-year population estimates are also given. The data therefore lend themselves to modelling the force of mortality, $\mu_{x+\frac{1}{2},y+\frac{1}{2}}$, without further adjustment. However, for brevity we will drop the $\frac{1}{2}$ and just refer to $\mu_{x,y}$.

We use data provided by the Office of National Statistics (ONS) for the population of England & Wales. For illustrative purposes we will just use the data for males. As we are primarily interested in annuity and pension liabilities, we will restrict our attention to ages 50–104 over the period 1971–2013. All death counts were based on deaths registered in England & Wales in a particular calendar year and the population estimates for 2002–2011 are those after the revision to take account of the 2011 census results. More detailed discussion of this data set, particularly regarding the current and past limitations of the estimated exposures, can be found

in Cairns et al. (2015).

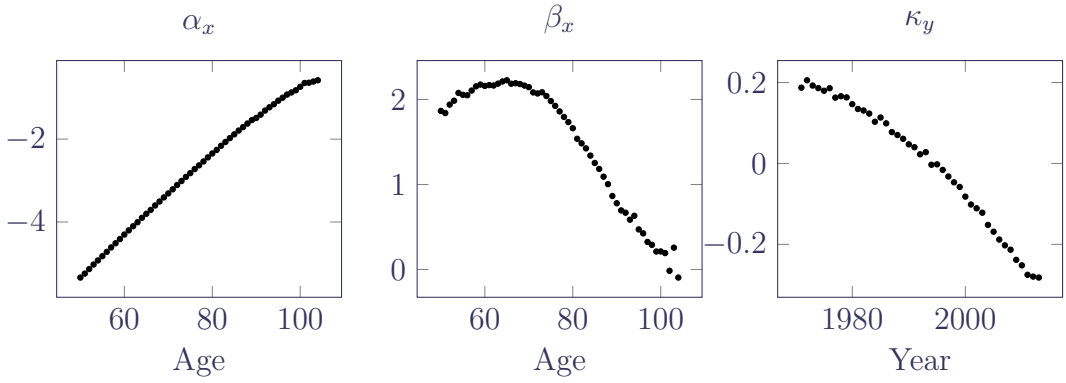
To generate an example mortality index for forecasting, we fitted the model from Lee and Carter (1992) to the data assuming a Poisson distribution for the number of deaths, i.e.

$$\begin{aligned} D_{x,y} &\sim \text{Poisson}\left(\mu_{x,y} E^c_{x,y}\right) \\ \log \mu_{x,y} &= \alpha_x + \beta_x \kappa_y \end{aligned} \tag{1}$$

where $E^c_{x,y}$ denotes the central exposure to risk at age x last birthday in calendar year y. The parameter α_x is broadly the average mortality level at age x, κ_y is the period mortality effect and β_x is the age-related modulation of κ_y. Since we will be working with only a subset of historical data, we will henceforth index κ from 1 to t, where t is the number of years of data and κ_t is the most recent observed value of κ from which point we want to project. In this paper we will use the following notation: κ_{t+h} will denote a future, yet-to-be-observed value of κ at time $t+h$; $\kappa_t(h)$ will denote a projected value at time $t+h$ where the projection is h years ahead from time t and where the underlying process parameters are known; $\hat{\kappa}_t(h)$ is the equivalent of $\kappa_t(h)$, but where the underlying process parameters are estimated. The difference between $\hat{\kappa}_t(h)$ and $\kappa_t(h)$ is therefore the parameter risk, while the difference between $\kappa_t(h)$ and κ_{t+h} is the accumulated random error over h years. These differences are used in Appendices A and C to decompose the overall risk into parameter and volatility components.

Following Brouhns et al. (2002) we estimate the parameters using the method of maximum likelihood, rather than the singular-value decomposition of Lee and Carter (1992). The Lee-Carter model is non-linear, so we fitted the model using the `gnm()` function in R (R Core Team, 2012). `gnm()` uses random constraints, so we then applied the constraints of Richards and Currie (2009), i.e. $\sum_{i=1}^{t} \kappa_i = 0$ and $\sum_{i=1}^{t} \kappa_i^2 = 1$, after fitting. Other constraint systems are possible: Lee and Carter (1992) used $\sum_x \beta_x = 1$, while Girosi and King (2008) used $\sum_x \beta_x^2 = 1$; different constraint systems produce different parameter values, but the same fitted values for $\log \mu_{x,y}$. The resulting parameter estimates are shown in Figure 1, with the estimated values for κ in Table 1. It is these κ values which must be projected, and this paper is about (i) the options available for forecasting κ, (ii) the uncertainty over these projections, and (iii) how the uncertainty may be decomposed into various sources of risk.

Figure 1: Parameter estimates for Lee-Carter model fitted to mortality data for males in England & Wales aged 50–104 over the period 1971–2013.



The κ values will be treated throughout this paper as if they are known quantities, but it is worth noting that this is a simplification; in fact, the κ values are themselves estimates, and there is thus uncertainty over their true underlying value, especially if the κ values are

Table 1: Estimates of κ from Figure 1. $t = 43$, $\sum_{i=1}^{t} \kappa_i = 0$ and $\sum_{i=1}^{t} \kappa_i^2 = 1$.

Year	κ	Year	κ	Year	κ	Year	κ
1971	0.187305	1982	0.130861	1993	0.027859	2004	-0.152008
1972	0.205557	1983	0.123310	1994	-0.003255	2005	-0.168814
1973	0.192505	1984	0.103061	1995	-0.002242	2006	-0.188083
1974	0.186180	1985	0.113493	1996	-0.015946	2007	-0.202479
1975	0.179728	1986	0.099141	1997	-0.032266	2008	-0.213225
1976	0.185944	1987	0.077453	1998	-0.046641	2009	-0.238276
1977	0.162683	1988	0.070201	1999	-0.058118	2010	-0.251942
1978	0.166125	1989	0.060964	2000	-0.082013	2011	-0.275248
1979	0.163246	1990	0.047191	2001	-0.101537	2012	-0.279736
1980	0.146532	1991	0.040114	2002	-0.110974	2013	-0.282070
1981	0.134640	1992	0.022597	2003	-0.121817		

estimated from the mortality experience of a small population. The true κ can be regarded as a hidden process, since we cannot observe κ directly and can only infer likely values given the random variation from realised deaths in a finite population. As a result, the estimated variance of the volatility will in fact be an over-estimate, as the estimated κ values are subject to two sources of variation. There is a parallel here to the concept of a Kalman filter, which models an observable process (the estimated κ) which is itself a realisation of a hidden underlying linear process (the true κ). The Kalman filter therefore allows for two types of noise: measurement error and volatility. In this paper the parameters of κ will be estimated using R's `arima()` function, which uses a Kalman filter to estimate the parameter values for an ARIMA model, but assuming that there is no measurement error.

Note that the κ values need not come from a Lee-Carter model — they may equally come from any other stochastic mortality model with a time index, for example the Age-Period model. Note also that in practical work we would smooth the α_x and β_x terms to reduce the effective number of parameters being estimated: α_x and β_x show a high degree of regularity in Figure 1, which means far fewer than fifty parameters are needed to summarise the patterns. Such a reduction in the effective number of parameters also tends to improve the properties of the forecast values of $\mu_{x,y}$ — see Delwarde et al. (2007) and Currie (2013).

The results in this paper apply to a one-dimensional time index, but the main results regarding the uncertainty about the drift of κ can be generalised to n-dimensional projection models such as the CBD family (Cairns et al., 2009).

3 Modelling κ as a random walk

We suppose that the κ mortality index in Section 2 follows a random walk with drift, i.e.

$$\kappa_{t+1} = \kappa_t + \mu_0 + \epsilon_{t+1} \tag{2}$$

where μ_0 is the (unknown) drift term and ϵ_{t+1} is an error term with zero mean and constant finite variance, σ_ϵ^2. This random-walk structure is perhaps the commonest approach to projecting κ, although it is not necessarily the most realistic. In this paper we only consider a constant value for μ_0, although van Berkum et al. (2014) consider the alternative where μ_0 remains constant for a period of time and then changes to an alternative value.

Although the mathematics are straightforward, we set out the steps in detail as they make

a useful comparison for the ARIMA model in Section 4. Under the random-walk model, the realised values of κ_{t+h} (h steps ahead) are given by:

$$\kappa_{t+h} = \kappa_t + h\mu_0 + \sum_{j=1}^{h} \epsilon_{t+j} \tag{3}$$

Assuming for the moment that μ_0 is known, and given the value for κ_t, we obtain a central forecast h years ahead, $\kappa_t(h)$, by setting all the error terms ϵ_{t+j} to zero, i.e.

$$\kappa_t(h) = \kappa_t + h\mu_0 \tag{4}$$

However, since μ_0 is unknown we replace it with an estimate, $\hat{\mu}_0$, which is derived as follows:

$$\hat{\mu}_0 = \frac{1}{t-1}\sum_{i=2}^{t}(\kappa_i - \kappa_{i-1}) = \frac{\kappa_t - \kappa_1}{t-1} \tag{5}$$

where t is the number of κ terms ($t = 43$ in Table 1). In equation (5) we estimate μ_0 using an unbiased estimator $\hat{\mu}_0$ which coincides with the MLE for the mean of the Normal distribution (and for the mean of many other distributions besides).

Equations (4) and (5) gives us a forecast estimator h years ahead, $\hat{\kappa}_t(h)$:

$$\hat{\kappa}_t(h) = \kappa_t + h\hat{\mu}_0 \tag{6}$$

For the variance of $\hat{\mu}_0$ we have:

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_0) &= \mathrm{Var}\left(\frac{\kappa_t - \kappa_1}{t-1}\right) \\
&= \frac{1}{(t-1)^2}\mathrm{Var}\left(\kappa_1 + (t-1)\mu_0 + \sum_{j=2}^{t}\epsilon_j - \kappa_1\right) \\
&= \frac{\sigma_\epsilon^2}{t-1}
\end{aligned} \tag{7}$$

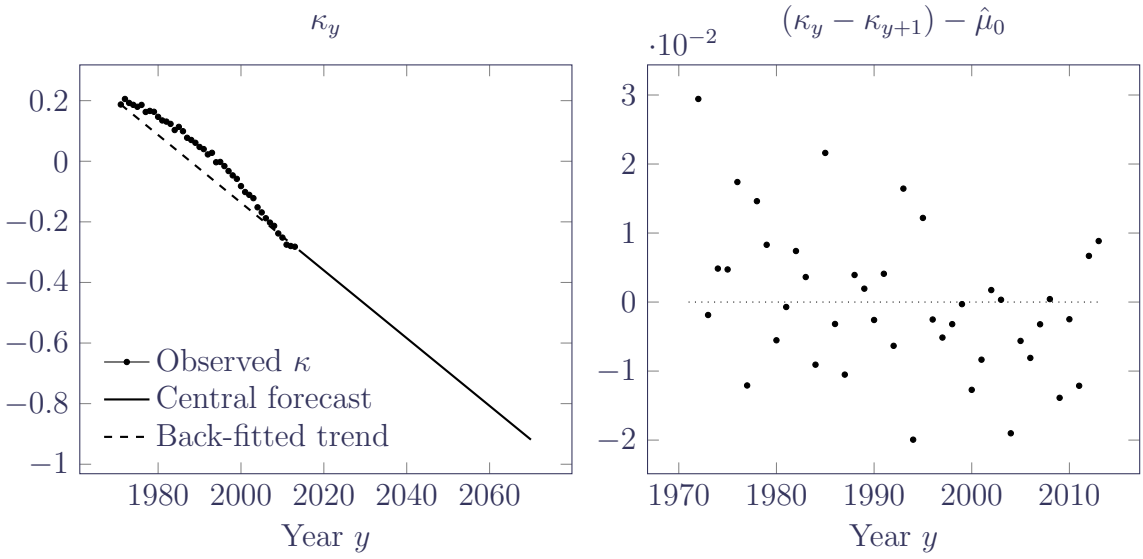
For the data in Table 1 we obtain $\hat{\mu}_0 = -0.011176$. Our estimate of σ_ϵ^2 is the appropriate sample variance, σ_ϵ^2:

$$\hat{\sigma}_\epsilon^2 = \frac{1}{t-2}\sum_{i=2}^{t}\left(\kappa_i - \kappa_{i-1} - \hat{\mu}\right)^2 \tag{8}$$

which gives $\hat{\sigma}_\epsilon^2 = 0.00011$ ($\hat{\sigma}_\epsilon = 0.010512$). Technically σ_ϵ^2 is also a parameter being estimated, and there is uncertainty over this estimate. However, for the purposes of this paper we are concerned with parameter risk which affects the trend. The estimated standard error of $\hat{\mu}_0$ is therefore 0.001622.

The resulting central forecast, $\hat{\kappa}_t(h)$, defined in equation (6) is depicted in Figure 2, where we can see that the random-walk process is a poor representation of the past behaviour of κ. As shown in equation (5), the random walk with drift just draws a line between the first and last κ terms, ignoring the non-linear pattern in between. This is not a unique finding — van Berkum et al. (2014) noted a similarly poor fit to modern Dutch data using a Lee-Carter model. We therefore need to find a better model; van Berkum et al. (2014) used a random walk with drift

Figure 2: Left panel: κ values from Figure 1 with random-walk forecast from equation (6) with $\hat{\mu}_0 = -0.011176$. The forecast line is also extended backwards, showing that the random-walk model is a poor representation of the behaviour of the observed κ process. Right panel: residuals, i.e. differenced κ minus $\hat{\mu}_0 = -0.011176$, showing that early values are more likely to have positive residuals and more recent values are more likely to have negative ones.



where μ_0 was piecewise constant in time, whereas we will consider using an ARIMA model in Section 4.

For risk management we are interested in (i) the uncertainty over $\hat{\kappa}_t(h)$, and (ii) the relative role of parameter uncertainty and volatility. We therefore look at the mean squared prediction error, i.e. $\mathrm{E}\left[\left(\hat{\kappa}_t(h) - \kappa_{t+h}\right)^2\right]$. Using the result in Appendix A we find that this breaks down into the sum of the parameter uncertainty ($h^2\sigma_\epsilon^2/(t-1)$) and the volatility component ($h\sigma_\epsilon^2$):

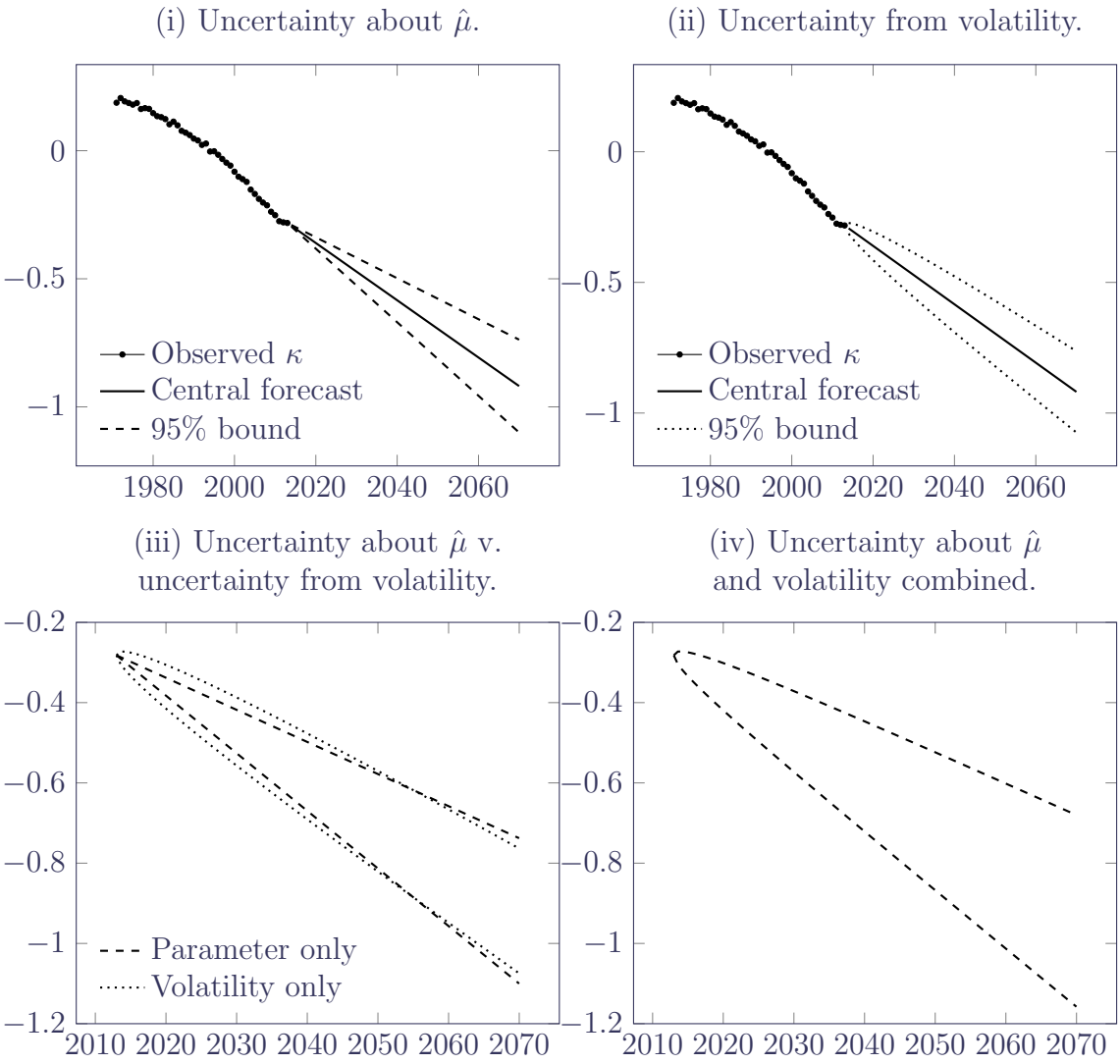
$$\mathrm{E}\left[\left(\hat{\kappa}_t(h) - \kappa_{t+h}\right)^2\right] = \underbrace{\frac{h}{t-1}h\sigma_\epsilon^2}_{\substack{\text{parameter}\\ \text{uncertainty}}} + \underbrace{h\sigma_\epsilon^2}_{\text{volatility}} \tag{9}$$

where the parameter uncertainty is the variance of $h\hat{\mu}_0$, i.e. $h^2\mathrm{Var}(\hat{\mu}_0)$ from equation (7). The two components on the right hand side of equation (9) are related as follows:

$$\text{Parameter uncertainty (variance of } h\hat{\mu}_0) = \frac{h}{t-1}\text{Volatility}$$

Thus, if the projection horizon h is small, but a long history of data (t) is used, the volatility component is much larger than the parameter uncertainty. Conversely, if the projection horizon is long and the data history is short, parameter uncertainty is the larger component of overall uncertainty. As we will see in Section 7, this is reflected in insurer capital requirements for longevity trend risk under a value-at-risk calculation. We note from Figure 3 that the two components of uncertainty have different shapes, and that uncertainty about κ_{t+h} due to volatility dominates in the shorter term. However, eventually the parameter uncertainty about $\hat{\mu}$ dominates. The crossover point is when $h/(t-1) = 1$. Sampson (1991) shows that a similar result holds for the asymptotic prediction error of ARIMA models, which we will explore in Section 4.

Figure 3: κ values with forecast from random walk with drift with (i) 95% bounds for parameter uncertainty and (ii) stochastic volatility. Panel (iii) shows the relationship between the two sources of uncertainty over the course of the projection. Note the wider range in panel (iv), reflecting the wider interval arising from including both sources of uncertainty.



4 Modelling κ as an ARMA or ARIMA process

We saw in Section 3 that the forecast uncertainty for a random walk with drift could be split into two components: parameter uncertainty and volatility. However, we also saw in Figure 2 that a random walk with drift was a poor representation of the historical pattern of κ. As we will see, κ is better represented as an ARIMA process, and the forecast uncertainty of such a process can be analogously decomposed into parameter uncertainty and volatility.

We begin by defining an ARMA(p,q) process, X_t^0, with zero mean:

$$X_t^0 = ar_1 X_{t-1}^0 + \ldots + ar_p X_{t-p}^0 + ma_1 \varepsilon_{t-1} + \ldots + ma_q \varepsilon_{t-q} + \varepsilon_t \tag{10}$$

where ε_t is a series of independent, identically distributed error terms with zero mean and constant finite variance, σ_ε^2. The p autoregressive parameters and q moving-average parameters

are to be estimated, and there will be uncertainty over these estimates which will contribute to the overall forecast uncertainty. Moreover, there is also uncertainty about p and q which define the model within the class of ARMA(p,q) processes. We will discuss the impact of choosing specific values for p and q in Section 9 when we discuss model uncertainty. However, we restrict ourselves to the family of stationary ARMA(p, q) processes for any orders p and q that we consider.

Having defined the ARMA(p, q) process with a zero mean in equation (10), we can now use it to define an ARMA(p, q) process with a non-zero mean, μ:

$$X = X^0 + \mu \tag{11}$$

As in Section 3, we can calculate an estimate of the mean, $\hat{\mu}$:

$$\hat{\mu} = \bar{X} = \frac{1}{t}\sum_{i=1}^{t} X_i \tag{12}$$

Since X is a stationary ARMA process, we have $\mathrm{E}[\hat{\mu}] = \mu$. Furthermore, using the derivation in Appendix B we have the following result for the variance of $\hat{\mu}$:

$$\mathrm{Var}\,(\hat{\mu}) = \frac{\mathrm{Var}(X^0)}{t} + \frac{2}{t}\sum_{k=1}^{t-1}\gamma(k)\left[1 - \frac{k}{t}\right] \tag{13}$$

where $\gamma(k) = \mathrm{Cov}(X_t, X_{t+k})$ is the auto-covariance function of X. Note that $\gamma(0) = \mathrm{Var}(X_t^0)$ by definition. Also note that (13) reduces to (7) if X is a sequence of independent random variables, since $\gamma(k) = 0$ for all $k \neq 0$ in this case.

Asymptotically, we obtain for $t \to \infty$:

$$t\mathrm{Var}\,(\hat{\mu}) \to \sum_{k=-\infty}^{\infty}\gamma(k) \tag{14}$$

if $\sum_{k=-\infty}^{\infty}|\gamma(k)| < \infty$, see Appendix B and also Theorem 7.1.1 in Brockwell and Davis (1987).

Comparing equations (7) and (13) we see that the variance of $\hat{\mu}$ can be either higher or lower than the variance of the estimated mean $\hat{\mu}_0$ of a sequence of independent random variables. Any dependencies between observations X_t and X_{t+k} can therefore either increase or decrease the uncertainty about $\hat{\mu}$, depending on whether $\sum_{k=1}^{t-1}\gamma(k)\left[1 - \frac{k}{t}\right]$ is positive or negative. For example, an AR(1) model with a negative autoregressive parameter would have lower parameter uncertainty (about the mean) than a model where X_0 is a sequence of independent variables.

In empirical studies the autocovariance function $\gamma(k)$ in equation (13) is not known. Therefore, the variance of $\hat{\mu}$ needs to be estimated. We compared the accuracy of three different approaches to obtain an estimate for $\mathrm{Var}\,(\hat{\mu})$. In our first approach we replaced $\gamma(k)$ by the sample autocovariances obtained from observed values of X. We found in simulation studies that the estimated values of $\gamma(k)$ for large values of k were unreliable (the estimator $\hat{\gamma}(k)$ has a large variance) and therefore, the estimated value of $\mathrm{Var}\,(\hat{\mu})$ was also unreliable. An alternative approach is to fit an ARMA model to X and then use the theoretical autocovariance function $\gamma(k)$ based on the estimated parameters of the fitted ARMA model. We found that this approach provided reasonable estimates of $\mathrm{Var}\,(\hat{\mu})$. However, while this second approach will give us the variance for $\hat{\mu}$, it will not give us information on possible extreme values. In insurance work we are primarily interested in extreme quantiles like the 99.5% point, so in our empirical study in Sections 5 and 6 we adopted a third approach: we used a parametric bootstrap procedure to obtain the full distribution, and in particular, the variance of $\hat{\mu}$ directly from bootstrap

realisations of $\hat{\mu}$ without the need to find the autocovariance function $\gamma(k)$ of X first. This third approach was used for consistency in the treatment of uncertainty over the ARMA parameters and μ, but also in generating a full distribution for the exploration of extremes. We will see the importance of this in Section 5.

We also note that the estimator $\hat{\mu} = \bar{X}$ is not the maximum-likelihood estimator for μ unless the ARMA process is a sequence of independent random variables. In general, the MLE for $\hat{\mu}$ will depend on the (estimated) ARMA parameters. To illustrate this point we show the derivation of the MLE for μ for an AR(p) process in Appendix D (see also Harvey (1981) for the case for an AR(1) process). We find that the MLE is very close to the mean $\bar{X}$ for small values of p. Furthermore, both estimators are unbiased. However, since the MLE for μ depends on the estimated ARMA parameters and the order of the ARMA model, different estimates for μ will be obtained when the order of the model is changed. We therefore argue that the mean $\hat{\mu} = \bar{X}$ is a more robust alternative to the MLE. Being able to estimate μ independently of other parameters also has the advantage that the process X^0 can be fitted to appropriate models using maximum likelihood, least squares or other methods without re-estimating μ. For all these reasons we prefer $\hat{\mu} = \bar{X}$ compared to the MLE for μ. A short discussion of the mean compared to the best linear unbiased estimator (BLUE) can be found in Brockwell and Davis (1987, p213). They argue that since the asymptotic variances in equation (14) are equal for both estimators, they use the simple estimator $\hat{\mu}$.

Although using the sample mean $\bar{X}$ as an estimator for $\hat{\mu}$ means that the estimated value of $\hat{\mu}$ will be independent of the fitted ARMA model, the estimated standard error of $\hat{\mu}$ will depend on the chosen ARMA model and the fitted parameters since we do not use the empirical autocovariances in equation (13) for the reasons explained earlier.

We can thus model the period effect κ in our Lee-Carter model as an ARMA(p, q) process by setting $\kappa = X$. Alternatively, and more realistically, we can model κ as an ARIMA(p, 1, q) process:

$$\kappa_t = \kappa_{t-1} + X_t = \kappa_{t-1} + X_t^0 + \mu \qquad \forall\, t > 1 \tag{15}$$

where κ_1 is a given constant or a random variable independent of the process X. The realised value of κ, h steps ahead, is as follows:

$$\begin{aligned}
\kappa_{t+h} &= \kappa_{t+h-1} + X_{t+h} \\
&= \kappa_{t+h-1} + X_{t+h}^0 + \mu \\
&= \kappa_{t+h-2} + X_{t+h-1}^0 + X_{t+h}^0 + 2\mu \\
&\;\;\vdots \\
&= \kappa_t + \sum_{i=1}^{h} X_{t+i}^0 + h\mu
\end{aligned} \tag{16}$$

with X_{t+h}^0 given by equation (10). It is possible to model κ using an ARIMA(p, d, q) process with $d > 1$, but we will restrict our attention to $d = 1$ here. Note that with $d = 1$ we are modelling improvements in κ, i.e. relative changes, whereas with $d = 2$ we would be modelling changes in the rate of improvement, i.e. acceleration or deceleration in improvements.

We can now define h-step ahead projections for κ as in the previous section, that is:

$$\begin{aligned}
\kappa_t(h) &= \kappa_t + \sum_{i=1}^{h} X_{t+i}^0 + h\mu \\
\hat{\kappa}_t(h) &= \kappa_t + \sum_{i=1}^{h} \hat{X}_{t+i}^0 + h\hat{\mu}
\end{aligned}$$

where we set all error terms $\varepsilon_{t+h} = 0$ for all $h > 0$ in equation (10) to obtain projected values for X^0, and we define $\hat{X}^0$ as in equation (10) but with the coefficients a_i and b_i, and the residuals ε up to time t being replaced by their estimated values.

The ARIMA model is a generalisation of the random-walk model — in each case $\hat{\mu}_0$ (for a random walk with drift) and $\hat{\mu}$ (for an ARIMA model) play analogous roles. However, it is interesting to note how quickly $\hat{\mu}$ can come to dominate the projected $\hat{\kappa}_t(h)$ for an ARIMA model, i.e. where the ε values are set to zero for the central forecast. For example, in an ARIMA(0,1,q) model the influence of the last observed ε terms ceases after $q + 1$ years, as per equation (10). Similarly, the closer the autoregressive parameters are to zero, the quicker $\hat{\mu}$ dominates the forecast as we discuss in detail in Section 9. Since the impact of these moving-average and autoregressive parameters can be quickly dominated by $\hat{\mu}$, it suggests that the uncertainty over them might also be dominated by the uncertainty over $\hat{\mu}$.

5 Fitting an ARIMA model for κ

To fit an ARIMA model we need to make a choice over the values of p, d and q. In the remainder of this paper we will set $d = 1$ and consider values of p and q in $\{0, 1, 2, 3\}$. We apply maximum-likelihood estimators as implemented in the statistical software R, see R Core Team (2012), to obtain estimated values of the coefficients a_i and b_i in equation (10). We then use the information criterion from Akaike (1987) with the small-sample correction from Hurvich and Tsai (1989) to select our final model. Table 2 shows the AICc values when the various ARIMA(p, 1, q) models are fitted to the data in Table 1, showing that the best-fitting model is ARIMA(1,1,2). The same model is selected if we target the BIC or the AIC without the small-sample correction. The estimated parameter values for the ARIMA(1,1,2) model are shown in Table 3.

Table 2: AICc values for various ARIMA(p, 1, q) models.

	q			
p	0	1	2	3
0	-260.16	-259.54	-260.81	-262.78
1	-260.22	-257.88	-269.83	-267.14
2	-258.10	-261.00	-267.14	-264.58
3	-258.95	-262.60	-264.17	-261.29

Table 3: Parameter estimates for ARIMA(1,1,2) model fitted to κ values in Table 1. Source: `arima()` function in R Core Team (2012) for ar_1, ma_1, ma_2 and σ_ϵ^2, equations (12) and (13) for $\hat{\mu}$.

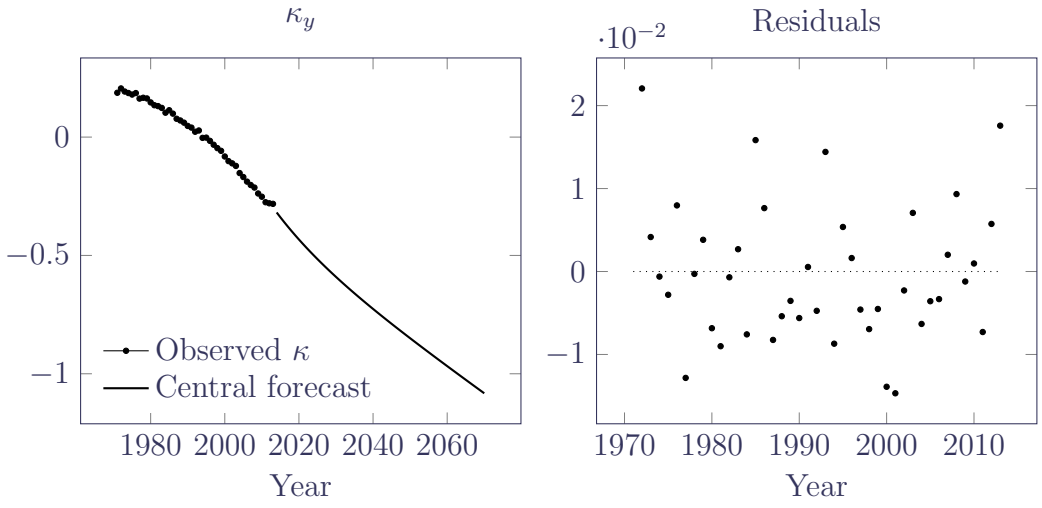
Parameter	Estimate	Standard error
ar_1	0.935	0.060
ma_1	-1.577	0.173
ma_2	0.815	0.149
σ_ϵ^2	0.000068	n/a
$\hat{\mu}$	-0.011	0.002

As mentioned earlier the estimate for μ is obtained using (12) rather than the maximum likelihood estimator implemented in R's `arima` function. The reported standard error for $\hat{\mu}$ is

therefore not based on the information matrix of the MLE, but on a bootstrap sample which we explain below. Also, while we are comfortable enough with the estimated values in Table 3, we are less comfortable with the standard errors: for example, adding more than one standard error to ar_1 would make the process non-stationary. As an alternative to these standard errors, we use bootstrapping to investigate the finite-sample distribution of ar_1, ma_1, ma_2 and σ_ϵ^2 (see Figure 5).

The forecast for the model in Table 3 is shown in Figure 4, together with the residuals. Figure 4 can be compared and contrasted directly with Figure 2. The first point of note is that the residuals look better for the ARIMA(1,1,2) model: not only are the ARIMA residuals smaller with a narrower range, but the ARIMA residuals are better distributed over time. The ARIMA model is clearly a better fit. The second point of note is that the central forecast seems to be a more natural extrapolation of the historical κ values. In particular, there is a visible curvature to the central forecast, which arises from the autoregressive component of the ARIMA model.

Figure 4: Left panel: κ values from Figure 1 with ARIMA(1,1,2) forecast from Table 3. Right panel: residuals from the ARIMA fit.



Having established that an ARIMA model is both a better fit to the κ process and seems to be a more natural extrapolation than a random walk with drift, we now turn to the subject of parameter uncertainty. Appendix C shows the decomposition of forecast uncertainty in an ARIMA model into (i) parameter uncertainty (the variance of $\hat{\mu}$, the variance of the autoregressive terms and the variance of the moving-average terms) and (ii) volatility. We are interested in the variance contributions for the various parameters, but, as noted in Appendix C, there are no closed-form expressions for the uncertainty over the parameters.

Instead, we use the bootstrap methodology described by Pascual et al. (2004). For a given ARIMA(p, d, q) model and fitted parameters, the approach is to take a short sample of the time series and simulate further values assuming the fitted parameters are known with certainty (we are only using $d = 1$ in this paper). The length of the simulated extension is such that the length of the new sequence — the fixed values plus the newly simulated values — is the same length as the original sample. An ARIMA model with the same order is fitted to the new sequence, i.e. new parameter values are estimated. The process is repeated 1,000 times, say, and the resulting distribution of each parameter value is the bootstrap estimate.

Following Pascual et al. (2004) our bootstrap procedure therefore consists of the following

steps, assuming that we have observed the process κ for $T+1$ years, i.e. we have observations $\kappa_0, \ldots \kappa_T$:

1. define $X_t = \kappa_t - \kappa_{t-1}$, estimate μ with $\hat{\mu} = \bar{X}$ and define $X_t^0 = X_t - \mu$ for $t = 1, \ldots, T$
2. fit an ARMA(p,q) model to obtain the vector $\theta_0 = (\hat{ar}_1, \ldots, \hat{ar}_p, \hat{ma}_1, \ldots, \hat{ma}_q, \hat{\sigma}_\varepsilon^2)$ of estimated parameters where $\hat{ar}_i$ and $\hat{ma}_i$ are the estimates for the parameters ar_i and ma_i in equation (10) and $\hat{\sigma}_\varepsilon^2$ is the estimated variance of ε_t; the estimated parameters for our data are given in Table 3
3. simulate N trajectories of an ARMA(p,q) process according to (10) with parameter vector θ_0; each trajectory $X(k) = \{X_1(k), \ldots, X_T(k)\}$ for $k = 1, \ldots, N$ is of length T starting from initial values $x_1(k), \ldots, x_p(k)$; and initial residuals $\varepsilon_1(k), \ldots, \varepsilon_q(k)$; all other residuals are drawn from a normal distribution $N(0, \hat{\sigma}_\varepsilon^2)$
4. estimate the parameter vector θ for each simulated trajectory to obtain the estimated parameter vector $\theta(k)$ for trajectory k; also estimate the mean as $\mu(k) = \bar{X}(k)$ to obtain bootstrap realisations of $\hat{\mu}$ which we then use to calculate its variance as mentioned in section 4 since the autocovariance function γ in equation (13) is not observed.

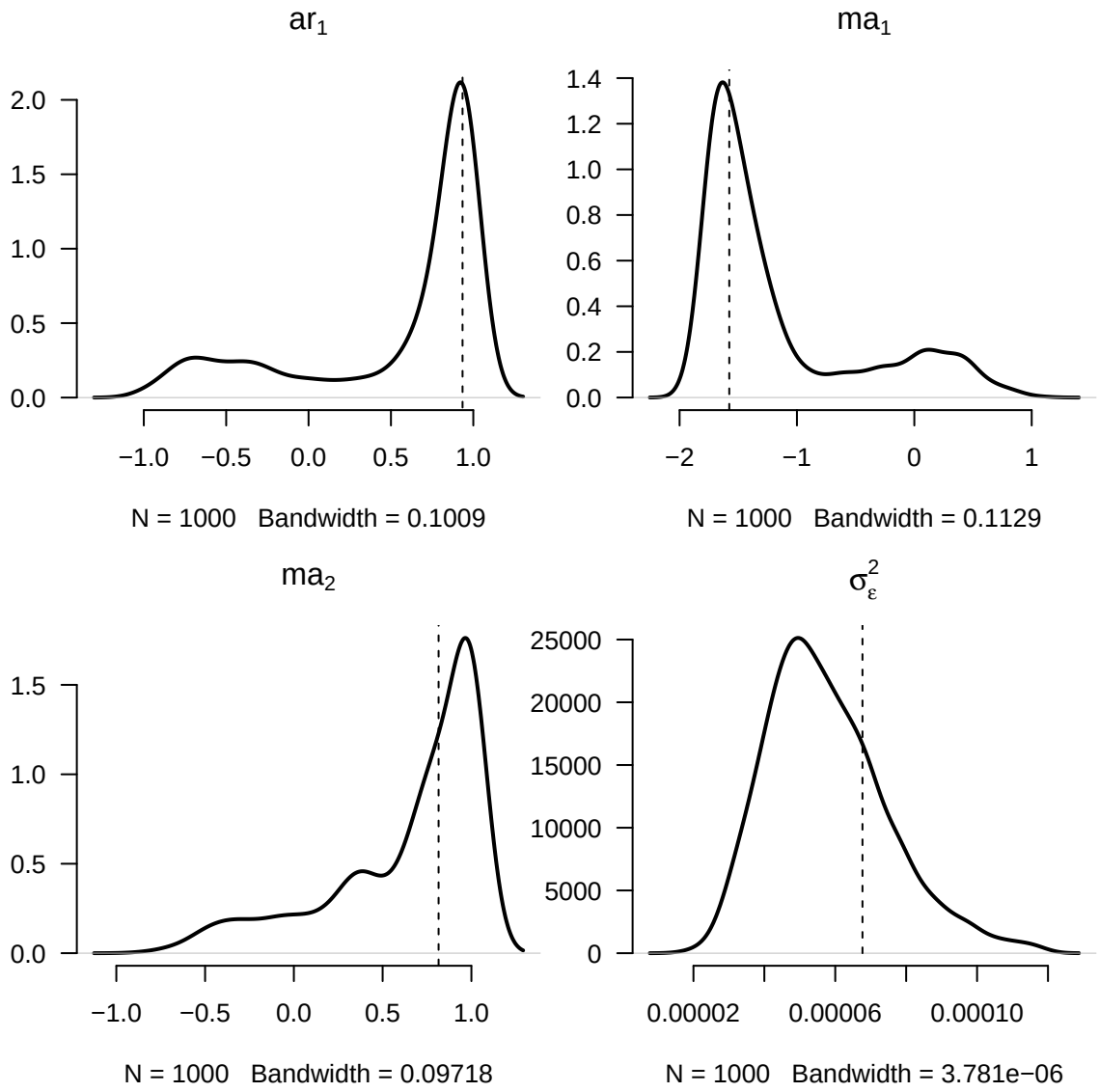
Pascual et al. (2004) used the same initial fixed sequence for each simulation in step 3, i.e. $x_1(k), \ldots, x_p(k)$ and $\varepsilon_1(k), \ldots, \varepsilon_q(k)$ in step 3 do not depend on k. In contrast, here we have randomly sampled a contiguous segment of the time series X to avoid any dependency on the choice of the fixed initial sequence. It might strike readers as unusual to not use the most recent κ values as the fixed sequence in this bootstrap procedure, but the log-likelihood in Harvey (1981, p122, equation 2.2) shows that this should have a small impact on the estimation of the autoregressive and moving-average parameters. The choice of a fixed initial sequence can have a material impact on the estimation of σ_ϵ^2, however, as shown later in the comparison of Figures 8 and 9. The resulting bootstrap estimates for the distributions of the parameters are shown in Figure 5.

In our empirical example we consider bootstrap samples with a sample size $N = 1,000$. We found that some to the trajectories simulated in step 3 of the above bootstrap procedure do not seem to correspond to a stationary ARMA(p, q) model, and therefore R's `arima` function fails to estimate the parameters. For this reason, some of the resulting bootstrap samples contain less than 1,000 simulated realisations of the estimated paramter values. Nonparametric estimates of the densities of the resulting bootstrap sample are shown in Figure 5 for the parameters of the ARIMA(1,1,2) model fitted to the observed period effect κ. For this specific example we found that two out of the 1,000 simulated trajectories for X do not correpond to a stationary ARMA(1,2) process, that is, the effective number of simulated trajectories is N=998.

We can use the results from Figure 5 to investigate the relative contribution to uncertainty from each source: (i) uncertainty over $\hat{\mu}$, (ii) uncertainty over the ARMA parameters, (iii) uncertainty from volatility and (iv) uncertainty from volatility including uncertainty over the value of σ_ϵ^2. These four cases are plotted in Figure 6. The bottom two panels show that there is no practical difference from considering the uncertainty over σ_ϵ^2, i.e. the estimate of σ_ϵ^2 in Table 3 is all we need to consider for the volatility.

Figure 6 reveals a number of important differences when compared with the relevant panels in Figure 3 for a random walk with drift. One feature is the different initial shape of the confidence bounds for volatility only: with the random-walk model, there is an immediate bulge in the first year of projection in Figure 3(ii), whereas for the ARIMA(1,1,2) model the first year shows no such bulge in Figure 6(ii). The reason lies in the contrast between structures of the models. Consider equation (9) for the random walk: with no parameter risk the first term is zero and so the variance of the random-walk forecast is linear in terms of the projection horizon, h. This is

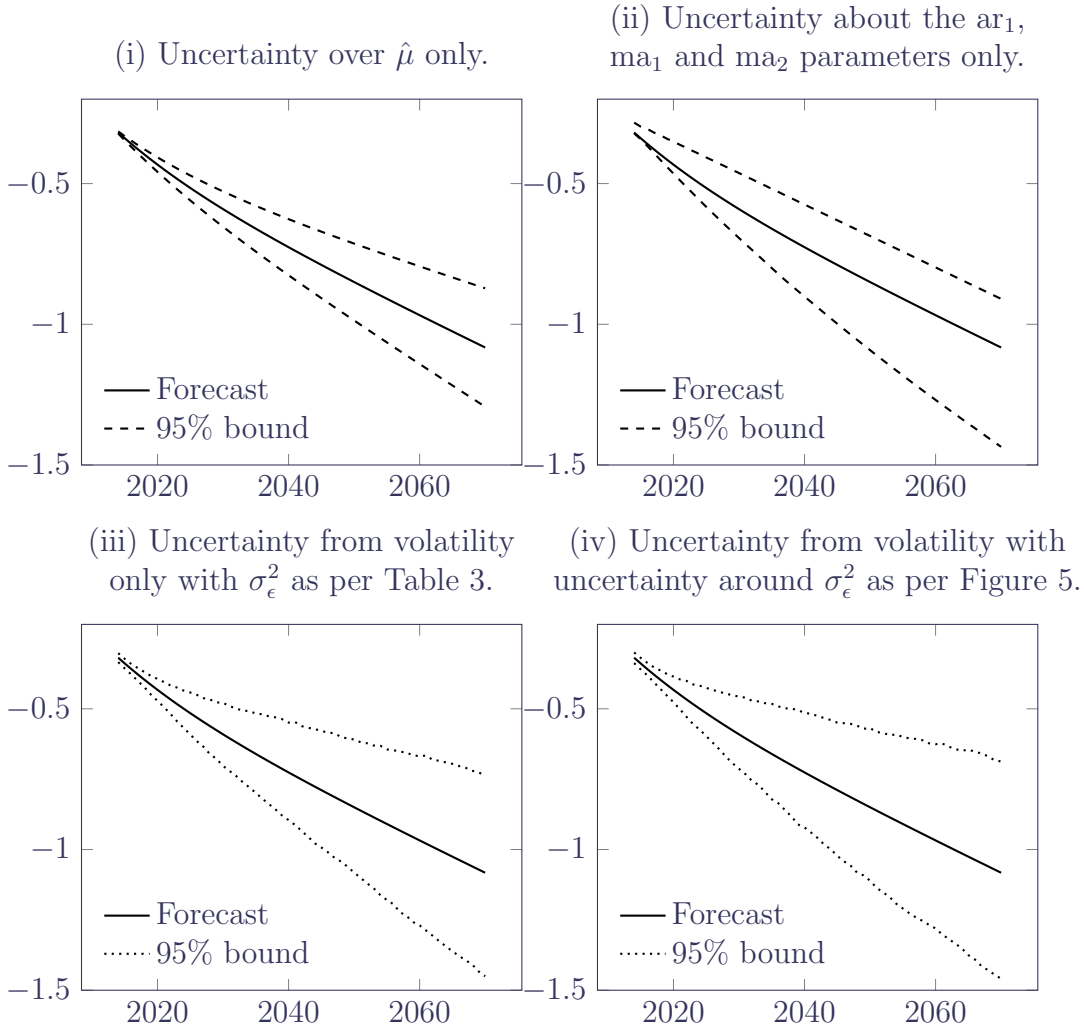
Figure 5: Estimated densities of ar_1, ma_1, ma_2 and σ_ϵ^2 parameters for ARIMA(1,1,2) model fitted to data in Table 1. The dashed vertical lines show the estimated values for each parameter from Table 3. Density estimation is done according to the procedure from Pascual et al. (2004).



because the difference $\kappa_{t+h} - \kappa_t$ involves the addition of h i.i.d. ϵ terms. This means that the standard error of the forecast increases in line with $\sqrt{h}$ and this gives a rapid initial expansion of the confidence interval, followed by slower expansion later on. The ARIMA model in equation (10) also has a projection which involves the addition of i.i.d. ϵ terms, but the structure of the model means that $\kappa_{t+h} - \kappa_t$ is not the straightforward sum of these terms. The structure of the ARIMA model introduces correlations that are not present in the drift model, both via the autoregressive and moving-average terms. These correlations mean that the variance of the forecast is not linear in terms of h, and so the standard error of the forecast does not behave like $\sqrt{h}$.

In Figure 3(iii) we see that the parameter uncertainty over $\hat{\mu}$ in a drift model eventually dominates the uncertainty due to volatility. In contrast, a comparison of Figure 6(i) with either panel (iii) or (iv) suggests that the uncertainty due to volatility remains dominant. Finally,

Figure 6: Forecast κ values from ARIMA(1,1,2) model in Table 3 with 95% bounds for various kinds of uncertainty.



particular comment is required for Figure 6(ii); at first glance, the confidence interval looks wrong for the central projection — the confidence interval is nowhere near symmetric around the central projection, not even for the first year. However, there is no mistake — the strange-looking confidence bounds in relation to the central projection are a consequence of the highly skewed distribution of the ar_1 parameter in Figure 5. For the ARIMA(1,1,2) model, the largest single component of parameter uncertainty is the uncertainty over the ARMA parameters, particularly the uncertainty over the ar_1 parameter.

6 Alternative ARIMA models for κ

In Section 5 we used an ARIMA(1,1,2) model because this produced the lowest AICc when fitted to the data (see Table 2). However, Figure 5 shows that the estimated parameter values are not robust - the ar_1 parameter in particular is close to 1, where values of 1 or more would make the process X non-stationary. A consequence of this lack of robustness is the large impact of the uncertainty over the ar_1, ma_1 and ma_2 parameters on simulated future mortality scenarios, as shown in in Figure 6(ii). This suggests that goodness of fit to past data should perhaps

not be the sole criterion for model selection. To illustrate this, we show some results for the ARIMA(1,1,0) model, i.e. a pure integrated AR(1) model with drift term. Table 2 shows that this is a materially poorer fit to the past data than the ARIMA(1,1,2) model used in Section 5, and that the AICc is almost identical to the AICc of the random-walk model.

The estimated parameter values for the ARIMA(1,1,0) model are shown in Table 4. We note that the estimated value for σ_ε^2 is very similar to the estimated value of σ_ϵ^2 in the random-walk model, which is in line with the very similar AICc values for those two models. This parameter is significantly smaller in the ARIMA(1,1,2) model indicating that the two moving-average parameters improve the fit of the model. We also find that the estimated ar_1 parameter is very different from the estimated ar_1-parameter in the ARIMA(1,1,2) model. This leads to a very different behaviour of the central projection, as shown in Figure 7. Indeed, for the ARIMA(1,1,0) model we find that the ar_1 parameter is rather close to zero, so that the uncertainty about this parameter is not relevant for the uncertainty about the central mortality projection — see Figures 8 and 10. The central projection for the ARIMA(1,1,0) model is almost identical to the central projection obtained with a random-walk model.

Table 4: Parameter estimates for ARIMA(1,1,0) model fitted to κ values in Table 1. Source: `arima()` function in R Core Team (2012) for ar_1 and σ_ϵ^2, equation (12) for $\hat{\mu}$. The standard error of $\hat{\mu}$ is obtained from the sample standard deviation of the bootstrap sample $\{\mu(k), k = 1, \ldots, N\}$.

Parameter	Estimate	Standard error
ar_1	-0.259	0.166
σ_ϵ^2	0.000102	n/a
$\hat{\mu}$	-0.011	0.002

Figure 7: Left panel: κ values from Figure 1 with ARIMA(1,1,0) forecast from Table 4. Right panel: residuals from the ARIMA(1,1,0) fit.

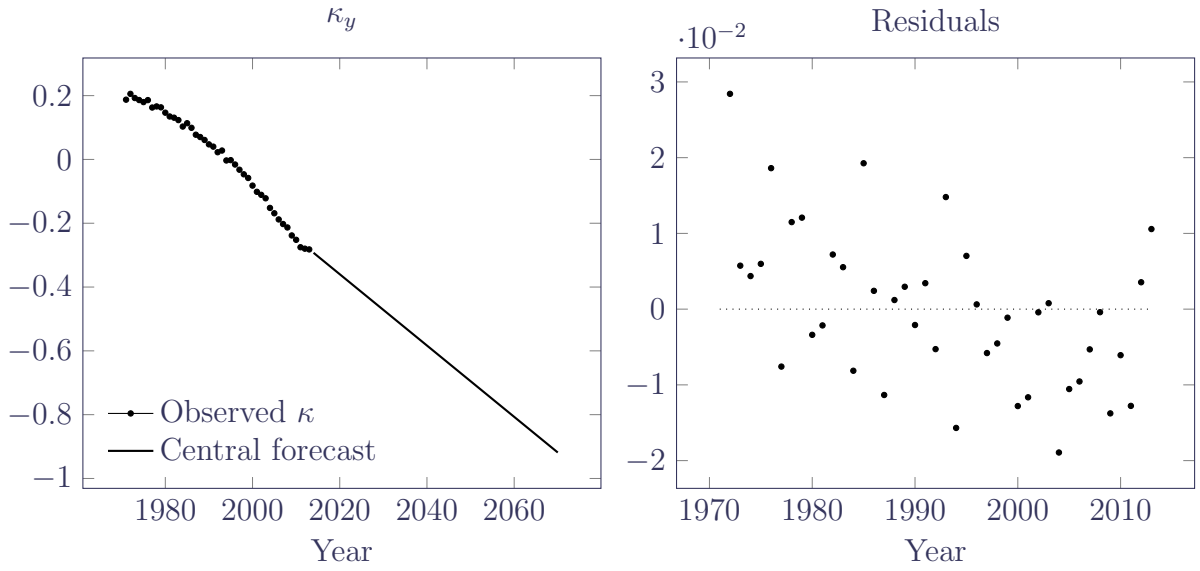


Figure 8: Estimated densities of ar_1 and σ_ϵ^2 parameters for ARIMA(1,1,0) model fitted to data in Table 1. The dashed vertical lines show the estimated values for each parameter from Table 4. Density estimation is done according to the procedure from Pascual et al. (2004).

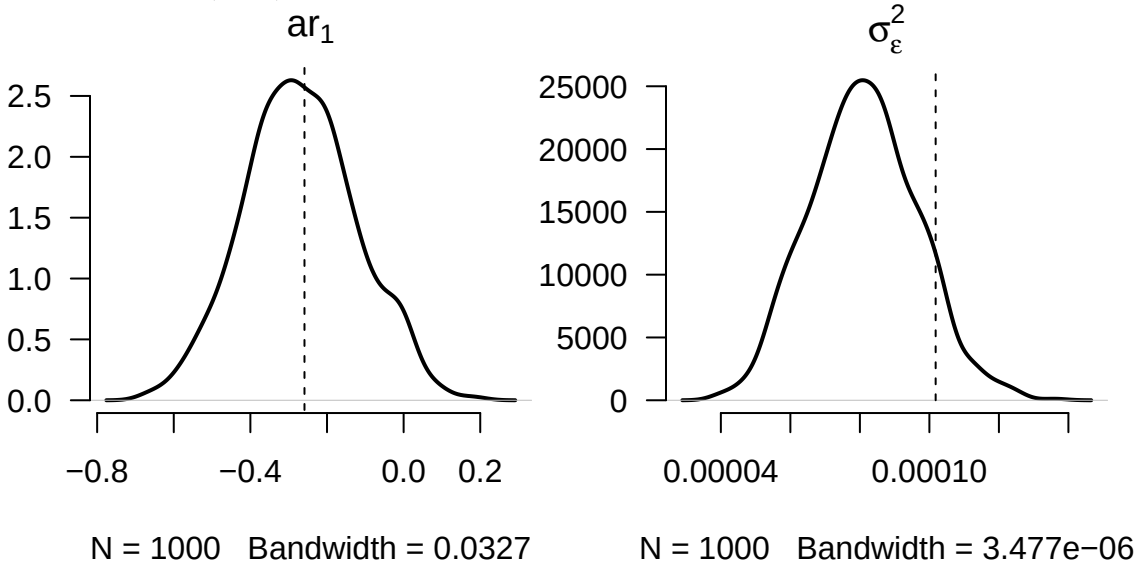


Figure 9: Estimated densities of ar_1 and σ_ϵ^2 parameters for ARIMA(1,1,0) model fitted to data in Table 1. The dashed vertical lines show the estimated values for each parameter from Table 4. In contrast to Figure 8, the bootstrapping procedure of Pascual et al. (2004) is modified to randomly select initial sub-series from the data; a comparison with Figure 8 shows that this has reduced the bias for σ_ϵ^2.

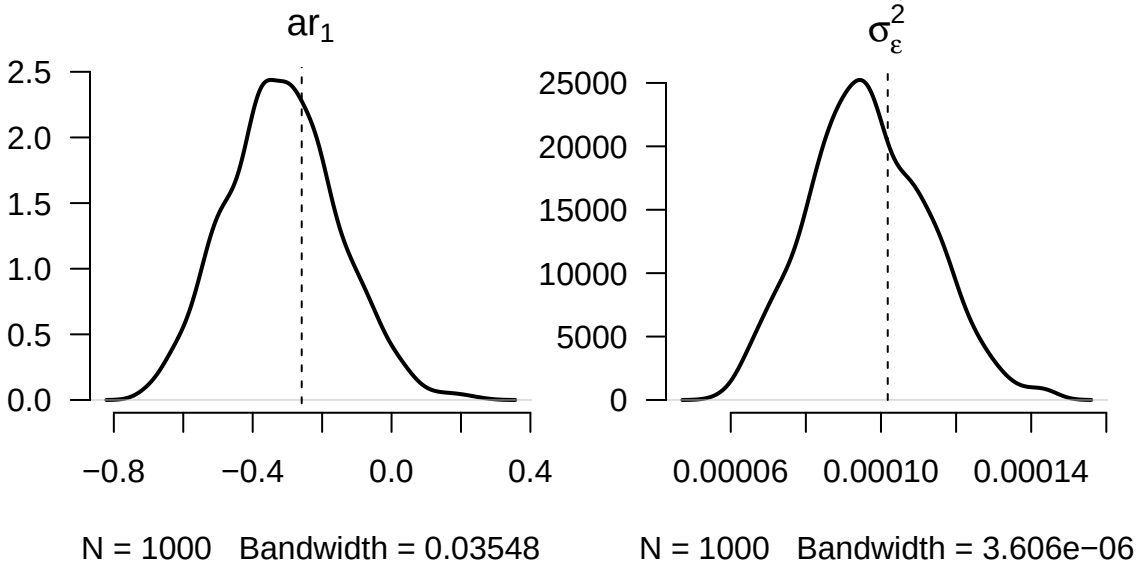
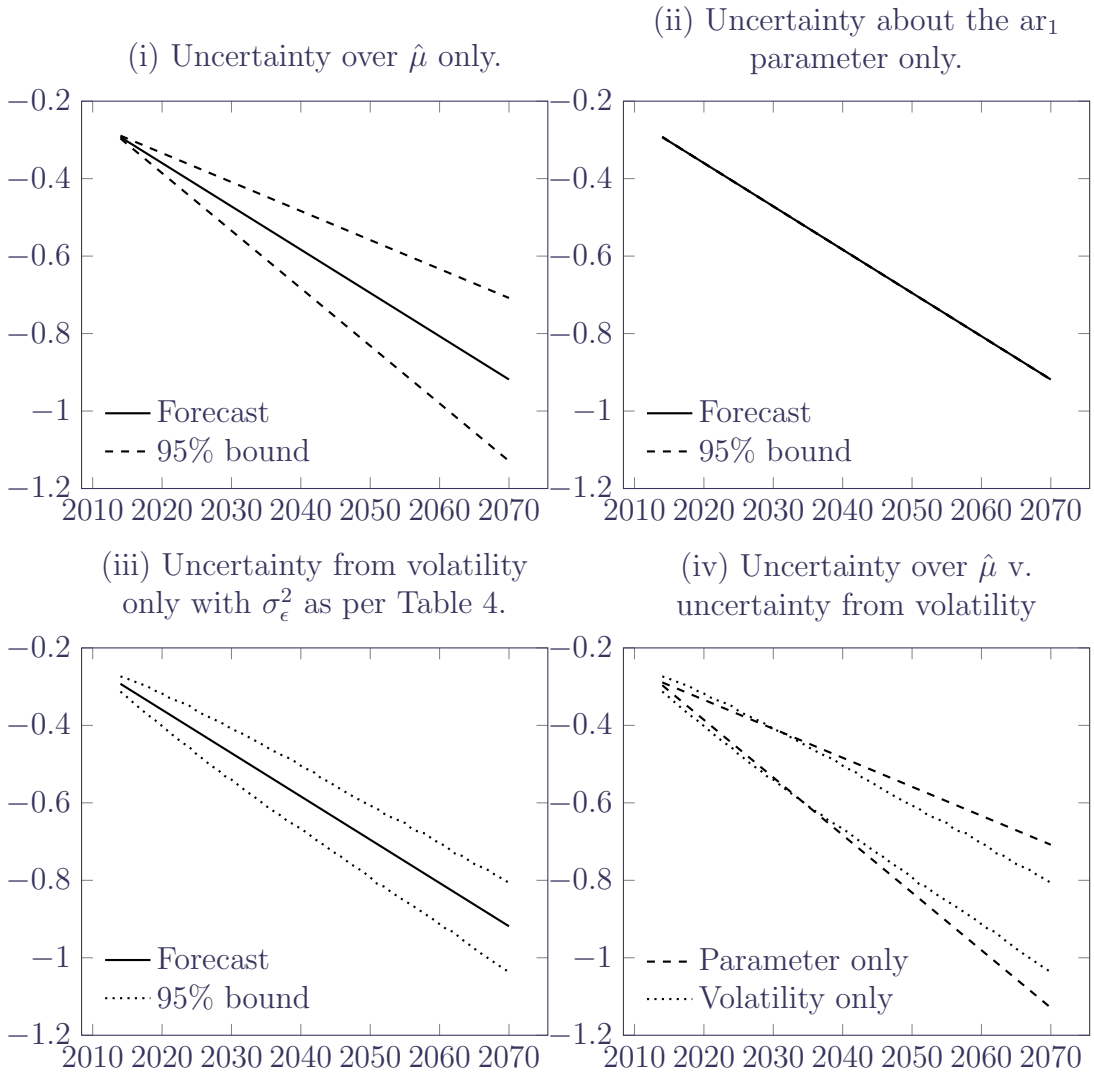


Figure 10: κ values with forecast from ARIMA(1,1,0) model with 95% bounds for various kinds of uncertainty.



7 Impact on value-at-risk-style capital requirements

The confidence intervals in Figures 3 and 6 are based on a run-off approach to projection. However, under the Solvency II regulatory regime in the European Union, insurers using internal models need to calculate reserves using a one-year, value-at-risk (VaR) approach. In this section we look at the impact of the various sources of uncertainty on capital requirements calculated using the VaR methodology of Richards et al. (2014). The VaR methodology involves using the fitted projection model to repeatedly simulate one year's additional mortality experience, then refitting the model and using the resulting updated central forecast to value an annuity liability.

For each of 1,000 VaR simulations we simulate the extra year's experience using three alternative combinations. The first option (labelled "Volatility only" in Table 5) means σ_ϵ^2 is set to its estimated value and one new ϵ error term is simulated for the moving-average process for each of the 1,000 simulations; all other parameters — ARMA parameters and the mean — are fixed at their estimated values and there is no parameter uncertainty. The second option

(labelled "Trend risk only" in Table 5) means the ARMA parameters and the mean (drift) term come from a single bootstrap realisation; σ_ϵ^2 is set to zero and there is therefore no volatility in the moving-average process. The third option (labelled "Trend risk and volatility combined" in Table 5) means the ARMA parameters and the mean term come from a single bootstrap realisation, together with σ_ϵ^2 set to its estimated value and one error term simulated for the moving-average process. Splitting the sources of uncertainty in this way allows us to examine the size of the various contributions made to the overall capital requirements.

The results for a value-at-risk assessment of longevity trend risk are given in Table 5. We have used a Lee-Carter model without smoothing fitted to the mortality data for males aged 50–104 in England & Wales over 1971–2013. All ARIMA models have been fitted with a non-zero mean. The factors in Table 5 are for temporary, continuously-paid annuities to a single life aged 70 at outset, with cashflows discounted at 2.5% per annum.

Table 5: Sensitivity of VaR capital requirements to various ARIMA models and the sources of uncertainty included. There is one bootstrap realisation of the ARMA and μ parameters per VaR calculation, but not all bootstrap realisations produce stationary models.

Model	Bootstrap simulations	Volatility	Trend uncertainty	$\bar{a}^{50\%}_{70:\overline{35}\rceil}$	$\bar{a}^{99.5\%}_{70:\overline{35}\rceil}$	VaR capital
ARIMA(0,1,0)	1000	Yes	No	12.50	12.70	1.62%
	1000	No	Yes	12.50	12.54	0.33%
	1000	Yes	Yes	12.49	12.72	1.79%
ARIMA(0,1,1)	1000	Yes	No	12.51	12.68	1.32%
	1000	No	Yes	12.51	12.56	0.34%
	1000	Yes	Yes	12.51	12.69	1.44%
ARIMA(1,1,0)	1000	Yes	No	12.51	12.68	1.36%
	1000	No	Yes	12.51	12.55	0.31%
	1000	Yes	Yes	12.51	12.69	1.43%
ARIMA(1,1,1)	1000	Yes	No	12.50	12.67	1.37%
	1000	No	Yes	12.51	12.56	0.37%
	1000	Yes	Yes	12.51	12.69	1.44%
ARIMA(1,1,2)	1000	Yes	No	12.59	12.87	2.25%
	994	No	Yes	12.53	12.61	0.63%
	994	Yes	Yes	12.53	12.87	2.70%

Table 5 shows that for many simple models the uncertainty over the ARMA parameter values and the mean makes only a modest additional contribution to the capital requirements. Value-at-risk calculations are driven by the variability of mortality experience over a one-year horizon and how the model fit responds to this. The reason why the volatility makes the largest contribution in Table 5 is the same point as made at the end of Section 3 for the random walk: the relative influence of volatility and parameter risk depends on the projection horizon and the length of the data series. In the case of VaR calculations, the projection horizon is just one year, thus maximising the influence of the volatility. This applies even for the ARIMA(1,1,2) model, where the parameter uncertainty shown in Figure 6 is very large. In contrast to the other models in Table 5, the parameter uncertainty of the ARIMA(1,1,2) model in Figure 6(ii) leads to a material additional capital requirement. This is perhaps surprising for the model which best fits the data, as demonstrated by Table 2. However, Table 5 gives another hint as to why the ARIMA(1,1,2) produces the greatest capital requirements: it is the only model

where the bootstrapped ARMA parameter values sometimes produced non-stationary models, as evidenced by the reduced number of usable parameter sets from bootstrap simulation (994 instead of 1,000). The ARIMA(1,1,2) model might fit the data best, but it is less stable and produces higher capital requirements as a result.

8 Comparison with CMI projections

The CMI is the part of the UK actuarial profession that produces reference mortality tables and occasional mortality forecasts for use by insurers and pension schemes. One such forecasting tool is described in Continuous Mortality Investigation (2009), and it has been updated more or less annually since. This tool takes the form of an Excel spreadsheet and it is in wide use throughout the UK, both for annuity portfolios and pension schemes.

The CMI spreadsheet works with relative mortality-improvement rates, rather than mortality rates. A mortality improvement is defined as $1 - q_{x,t}/q_{x,t-1}$, where $q_{x,t}$ is the probability that an individual alive aged x at the start of year t will die before the end of the year. See Willets (1999) for more detailed discussion of mortality improvements in insured portfolios and the wider population.

The CMI spreadsheet operates by user input of the assumed long-term rate of mortality improvement and the current rates of mortality improvement are then blended towards this value. There are over a thousand parameters which can be used by the user to customise the rate of blending by calendar year, age or year of birth. The CMI spreadsheet is therefore a deterministic targeting method of projection, as opposed to a stochastic model with parameters set by fitting to experience data.

The setting of the long-term mortality-improvement rate in the CMI spreadsheet is done subjectively, although some users set it with reference to average rates of improvement in the relevant population. We can relate this long-term rate to the means of the models in this paper as follows:

$$\begin{aligned}
1 - \frac{q_{x,t}}{q_{x,t-1}} &\approx 1 - \frac{\mu_{x,t}}{\mu_{x,t-1}} \\
&= 1 - \frac{e^{\alpha_x + \beta_x \kappa_t}}{e^{\alpha_x + \beta_x \kappa_{t-1}}} && \text{From equation(1)} \\
&= 1 - e^{\beta_x(\kappa_t - \kappa_{t-1})} \\
&= 1 - [1 + \beta_x(\kappa_t - \kappa_{t-1}) + \ldots] && \text{Expanding } e^x \text{as a Maclaurin series} \\
&\approx -\beta_x(\kappa_t - \kappa_{t-1}) && (17)
\end{aligned}$$

Using equation (2), the approximate mortality improvement in equation (17) becomes $-\beta_x(\mu_0 + \epsilon_t)$, which has expected value $-\beta_x\mu_0$. The mean of a random walk model, $\hat{\mu}_0$, could therefore only be appropriate for setting the long-term improvement rate in the CMI projection when it is modulated by the appropriate value of β_x. From Figure 1 we see that $\beta_x > 2, \forall x \in \{54, 55, \ldots, 75\}$. Using $\hat{\mu}_0 = -0.011176$, a best-estimate long-term rate for the CMI spreadsheet using the random-walk model would therefore be in excess of 2% for ages 50–78 inclusive, although the appropriate long-term rate declines sharply for higher ages: it is under 1% for ages over 89 and is effectively zero by age 100. A stressed value of the long-term rate for reserving purposes could be obtained by adding the appropriate number of standard errors to $\hat{\mu}_0$. However, whether it makes sense to take a parameter from a random-walk model and use it as a parameter for the CMI spreadsheet is a matter for actuarial judgement, especially in conjunction with the thousand other parameters which can be varied.

9 Discussion

We have considered three ARIMA(p,1,q) models for the period effect κ in the previous sections: ARIMA(0,1,0) (the random-walk model), ARIMA(1,1,2) and ARIMA(1,1,0). All three seem to be reasonable models to use for generating scenarios for future mortality. In this section we discuss some of the advantages and disadvantages that each model offers, focusing in particular on the impact of the selected model on mortality projections.

All three models have in common that they are integrated of order one, and therefore, the first difference of κ is modelled as a stationary process with a non-zero mean. If follows that the improvement rate of projected mortality rates will eventually approach this mean. However, the models differ significantly in the speed with which the improvement rate converges to this long term value. The autoregressive coefficient, ar_1, in equation (10) plays a central role for the behaviour of mortality projections. For a random-walk model we have $ar_1 = 0$, and for the other two models the estimated values of ar_1 are reported in Tables 3 and 4.

The importance of ar_1 can be seen when central projections are considered, i.e. where we set future error terms ε to zero. The projected values $X_t^0(h)$ based on the model in equation (10) and observations up to time t are then given by:

h	ARIMA(0,1,0)	ARIMA(1,1,0)	ARIMA(1,1,2)
1	$X_t^0(1) = ar_1 X_t^0$		$X_t^0(1) = ar_1 X_t^0 + ma_1\varepsilon_t + ma_2\varepsilon_{t-1}$
2	$X_t^0(2) = ar_1^2 X_t^0$		$X_t^0(2) = ar_1 X_t^0(1) + ma_2\varepsilon_t$
> 2	$X_t^0(h) = ar_1^{h-2} X_t^0(2)$		

Therefore, if $ar_1 = 0$ (the random-walk model) then $X_t^0(h) = 0$ for all values of h. For the other two models, long-term projections depend on the value of ar_1 and the two-years-ahead forecast $X_t^0(2)$ which in turn is determined by the last observations of κ. Consequently, if ar_1 is rather large ($ar_1 \approx 1$) the central projection of X^0 converges slowly to zero, and the increment of κ tends to μ only very slowly. This is the case for the ARIMA(1,1,2) model, and it can be seen in Figure 4 that the decrease in κ is stronger for the first few projected values and only for longer term projections κ behaves like a linear function with slope μ. On the other hand, the estimated ar_1 parameter is rather small for the ARIMA(1,1,0) model and, therefore, the impact of past κ values on central projections vanishes very quickly, see the virtually linear projection of κ in figure 7.

Turning to a quantitative comparison of the proposed models, we start with the random walk model in Section 3 and the ARIMA(1,1,0) model in Section 6. We note that the AICc values for those two models in Table 2 are almost identical. In addition, the central projections shown in Figures 2 and 7 are very similar and can be considered identical for practical purposes. Moreover, the uncertainty about ar_1 in equation (10) has no significant impact on the width of the prediction interval for κ as can be seen in Figure 10. The reason for those similarities is that the estimated ar_1 is rather close to zero. As a result, the process X^0 in equation (10) converges very quickly to zero when the error terms are set to zero, as can be seen from the above formulae for the central projections. However, comparing the difference between the width of the prediction intervals in Figures 3(iii) and 10(iii) we find that volatility is less important in the ARIMA(1,1,0) model compared to the random-walk model, and that therefore, the uncertainty about the drift becomes the leading source of uncertainty for relatively short forecast horizons.

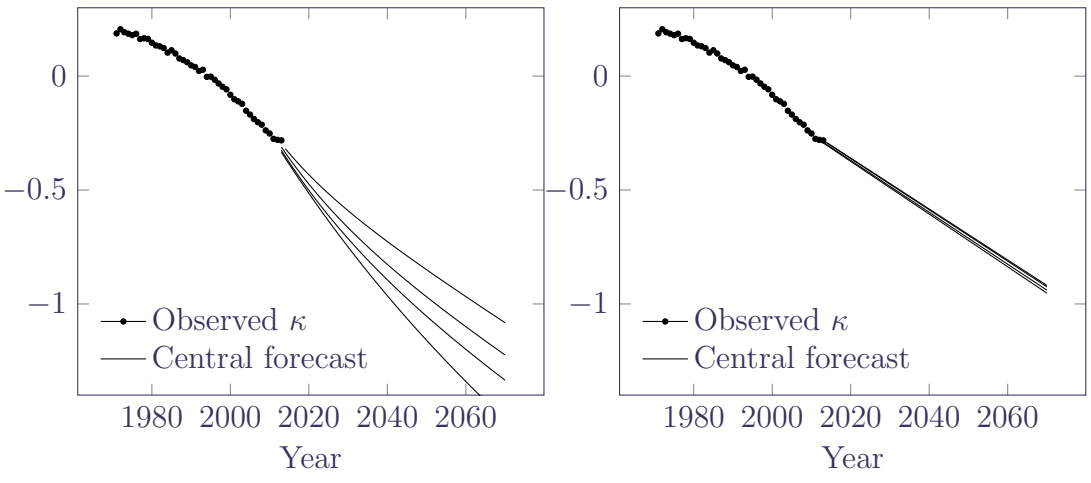
In contrast to the rather similar random-walk and ARIMA(1,1,0) models, the ARIMA(1,1,2) model behaves very differently. The estimated ar_1 is close to 1, showing that the inclusion of moving-average terms has a significant impact on the estimated value of ar_1 in equation (10). The convergence of X^0 to its long term mean 0 is therefore much slower with a large value of

ar_1. As a result, the projected value of $X_t^0(h)$ depends on the two-steps-ahead forecast $X_t^0(2)$, which is the last projected value still depending on the MA-parameters. Therefore, since the estimated ar_1 is rather large, uncertainty about this parameter, but also the last observed values of κ have a strong impact on long term projections in the ARIMA(1,1,2) model.

This raises questions about the robustness of the central projection of the ARIMA(1,1,2) model. From the density plot for $\hat{ar}_1$ in Figure 5 we observe that the bootstrap realisations of the estimator $\hat{ar}_1$ are concentrated at values close to one with a rather small variance, which seems to be particularly small compared to the variance of the same estimator in the ARIMA(1,1,0) model, see Figure 8. However, the above arguments show that even small variations in the ar_1 coefficient might have a substantial impact on the central projection of mortality rates since ar_1 is close to 1. This observation is confirmed by the wide prediction intervals in Figure 6(ii) showing that projected mortality rates are very sensitive to changes in ar_1.

The sensitivity of projected rates with respect to the last observed values of κ is best illustrated by considering the different central projections obtained when a small number of observations is removed from the sample. In Figure 11 we show the central projections obtained from the complete sample $y = 1971, \ldots, 2013$ and the central projections obtained when one, two or three years are removed from the end of the sample, that is κ_y is observed for y from 1971 to 2012 (one year removed), 2011 (two years removed) or 2010 (three years removed). While the central projections obtained from subsamples do not change much if an ARIMA(1,1,0) model is considered, the changes of the central projections based on the ARIMA(1,1,2) are significant.

Figure 11: Sensitivity of central projections when up to three years are removed from the end of the sample. Left panel: ARIMA(1,1,2) model. Right panel: ARIMA(1,1,0) model.



We conclude that the inclusion of the two moving-average parameters improved the fit of the model as measured by the AICc, but the robustness of mortality rate projections suffers. The choice of an appropriate time series model for κ is therefore a problem which requires actuarial judgement and should be based on well defined objectives. If models are chosen with robustness of projections as a selection constraint then the best fitting ARIMA(1,1,2) model should not be chosen. However, if only short term projections are required then the ARIMA(1,1,2) seems to be the most appropriate of all time series models that we have considered, but one should be aware that the improvement rate of projected mortality rates will be different from the estimated μ in the short to medium term for that model.

10 Conclusions

In this paper we showed how an ARIMA model can be a more realistic representation than a random walk with drift for the index of mortality. We found that in both cases the overall risk can be decomposed into parameter uncertainty and volatility. In an ARIMA process for mortality forecasting, we found that selecting a model on the basis of fit can lead to projections where the uncertainty over the ARMA parameters is too material to ignore. Furthermore, the uncertainty over the ARMA parameters can lead to additional capital requirements for insurers under a value-at-risk assessment mandated by Solvency II. For other, non-optimally fitting ARIMA models the uncertainty from the ARMA parameters was negligible, and the impact on insurer capital requirements was small.

Acknowledgments

Torsten Kleinow acknowledges financial support from Netspar under project LMVP 2012.03.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrica 52*, 317–333.

Börger, M. (2010). Deterministic shock vs. stochastic value-at-risk: An analysis of the Solvency II standard model approach to longevity risk. *Blätter DGVFM 31*, 225–259.

Brockwell, P. J. and R. A. Davis (1987). *Time Series: Theory and Methods*. Springer Verlag.

Brouhns, N., M. Denuit, and J. K. Vermunt (2002). A Poisson log-bilinear approach to the construction of projected lifetables. *Insurance: Mathematics and Economics 31(3)*, 373–393.

Cairns, A., D. Blake, K. Dowd, and A. Kessler (2015). Phantoms never die: Living with unreliable mortality data. *Journal of the Royal Statistical Society, Series A*.

Cairns, A. J. G., D. Blake, K. Dowd, G. D. Coughlan, D. Epstein, A. Ong, and I. Balevich (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal 13(1)*, 1–35.

Continuous Mortality Investigation (2009). *User Guide for The CMI Mortality Projections Model: 'CMI 2009'*. Continuous Mortality Investigation.

Currie, I. D. (2013). Smoothing constrained generalized linear models with an application to the Lee-Carter model. *Statistical Modelling 13(1)*, 69–93.

Delwarde, A., M. Denuit, and P. H. C. Eilers (2007). Smoothing the Lee-Carter and Poisson log-bilinear models for mortality forecasting: a penalized likelihood approach. *Statistical Modelling 7*, 29–48.

Girosi, F. and G. King (2008). *Demographic Forecasting*. Princeton University Press.

Granger, C. W. J. and P. Newbold (1977). *Forecasting Economic Time Series*. Academic Press.

Harvey, A. C. (1981). *Time Series Models*. Philip Allan Publishers Limited.

Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika 76*, 297–307.

Lee, R. D. and L. Carter (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association 87*, 659–671.

Pascual, L., J. Romo, and E. Ruiz (2004). Bootstrap predictive inference for arima processes. *Journal of Time Series Analysis 25*(4), 449–465.

Plat, R. (2011). One-year value-at-risk for longevity and mortality. *Insurance: Mathematics and Economics 49(3)*, 462–470.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Richards, S. J. and I. D. Currie (2009). Longevity risk and annuity pricing with the Lee-Carter model. *British Actuarial Journal 15(II) No. 65*, 317–365 (with discussion).

Richards, S. J., I. D. Currie, and G. P. Ritchie (2014). A value-at-risk framework for longevity trend risk. *British Actuarial Journal 19 (1)*, 116–167.

Sampson, M. (1991). The effect of parameter uncertainty on forecast variances and confidence intervals for unit root and trend stationary time-series models. *Journal of Applied Econometrics 6*(1), 67–76.

Shiryayev, A. N. (1984). *Probability*. Springer Verlag.

van Berkum, F., K. Antonio, and M. Vellekoop (2014). The impact of multiple structural changes on mortality predictions. *Scandinavian Actuarial Journal 2014*, 1–23.

Willets, R. C. (1999). Mortality in the next millennium. *Staple Inn Actuarial Society, London*.

Appendices

A Decomposition of forecast uncertainty for a random walk

We decompose the mean squared prediction error of the random walk in Section 3:

$$\begin{aligned}
\mathrm{E}\left[\left(\hat{\kappa}_t(h)-\kappa_{t+h}\right)^2\right] &= \mathrm{E}\left[\left(\hat{\kappa}_t(h)-\kappa_t(h)+\kappa_t(h)-\kappa_{t+h}\right)^2\right] \\
&= \mathrm{E}\left[\left(\hat{\kappa}_t(h)-\kappa_t(h)\right)^2\right]+\mathrm{E}\left[\left(\kappa_t(h)-\kappa_{t+h}\right)^2\right] \\
&\quad +2\mathrm{E}\left[\hat{\kappa}_t(h)-\kappa_t(h)\right]\mathrm{E}\left[\kappa_t(h)-\kappa_{t+h}\right] \\
&= \mathrm{E}\left[\left(\hat{\kappa}_t(h)-\kappa_t(h)\right)^2\right]+\mathrm{E}\left[\left(\kappa_t(h)-\kappa_{t+h}\right)^2\right] \\
&\quad +2\mathrm{E}\left[\hat{\kappa}_t(h)-\kappa_t(h)\right]\mathrm{E}\left[-\sum_{j=1}^{h}\varepsilon_{t+j}\right] \\
&= \underbrace{\mathrm{E}\left[\left(\hat{\kappa}_t(h)-\kappa_t(h)\right)^2\right]}_{\text{parameter uncertainty}}+\underbrace{\mathrm{E}\left[\left(\kappa_t(h)-\kappa_{t+h}\right)^2\right]}_{\text{volatility}} \qquad (18)
\end{aligned}$$

where we use the conditional independence of $\hat{\kappa}_t(h)$ and κ_{t+h} given κ_t. The conditional independence follows from the fact that $\hat{\kappa}_t(h)$ is a function of the error terms $\{\varepsilon_1,\ldots,\varepsilon_t\}$ while κ_{t+h} depends on $\{\varepsilon_{t+1},\ldots,\varepsilon_{t+h}\}$. A similar argument can be made for ARMA and ARIMA processes in Appendix C, assuming that $\hat{\mu}$ only depends on past values of ε_t.

Equation (18) decomposes the forecast uncertainty into components for parameter uncertainty and stochastic volatility. Since the estimator for $\hat{\mu}_0$ in equation (5) is unbiased, we use equation (6) to get the following for the parameter uncertainty in equation (18):

$$\mathrm{E}\left[\hat{\kappa}_t(h)\right]=\kappa_t+h\mathrm{E}\left[\hat{\mu}_0\right]=\kappa_t(h) \qquad (19)$$

Using equation (6) again, the component of the mean squared prediction error of the random-walk forecast which is due to parameter uncertainty is therefore:

$$\begin{aligned}
\mathrm{E}\left[\left(\hat{\kappa}_t(h)-\kappa_t(h)\right)^2\right] &= \mathrm{Var}(\hat{\kappa}_t(h)) \\
&= h^2\mathrm{Var}(\hat{\mu}_0) \\
&= h^2\frac{\sigma_\epsilon^2}{t-1} \\
&= \frac{h}{t-1}h\sigma_\epsilon^2 \qquad (20)
\end{aligned}$$

where the result for $\mathrm{Var}(\hat{\mu}_0)$ is given in equation (7). The standard error of trend forecast is therefore linear in h, the projection horizon. We can similarly derive an expression for the component of the mean squared prediction error due to volatility in equation (18):

$$\mathrm{E}\left[\left(\kappa_t(h)-\kappa_{t+h}\right)^2\right]=\mathrm{E}\left[\left(\sum_{j=1}^{h}\varepsilon_{t+j}\right)^2\right]=h\sigma_\epsilon^2 \qquad (21)$$

where we note that the standard deviation of the volatility component is proportional to $\sqrt{h}$. Comparing equations (20) and (21) we see the following relationship:

$$\text{Parameter uncertainty (variance)} = \frac{h}{t-1}\text{Volatility uncertainty (variance)}$$

As noted in Section 2, this ignores model risk and the fact that κ is estimated, not directly observed.

B The variance of $\hat{\mu}$ in an ARMA(p, q) process

Derivation of the variance of $\hat{\mu}$ for a stationary ARMA(p, q) process as defined in (10) and (11) where we denote by $\gamma(i-j) = \text{Cov}\,(X_i, X_j)$ the auto-covariance function of X (or X^0):

$$\begin{aligned}
\text{Var}\,(\hat{\mu}) &= \text{Cov}\left(\frac{1}{t}\sum_{i=1}^{t} X_i, \frac{1}{t}\sum_{j=1}^{t} X_j\right) \\
&= \frac{1}{t}\sum_{i=1}^{t} \text{Cov}\left(X_i, \frac{1}{t}\sum_{j=1}^{t} X_j\right) \\
&= \frac{1}{t^2}\sum_{i=1}^{t}\sum_{j=1}^{t} \text{Cov}\,(X_i, X_j) \\
&= \frac{1}{t^2}\sum_{i=1}^{t}\sum_{j=1}^{t} \gamma(i-j) \\
&= \frac{1}{t^2}\sum_{k=-(t-1)}^{t-1} (t-|k|)\gamma(k) \\
&= \frac{1}{t}\left[\sum_{k=-(t-1)}^{t-1} \gamma(k) - 2\sum_{k=1}^{t-1}\frac{k}{t}\gamma(k)\right] \\
&= \frac{\gamma(0)}{t} + \frac{2}{t}\left[\sum_{k=1}^{t-1}\gamma(k) - \sum_{k=1}^{t-1}\frac{k}{t}\gamma(k)\right] \\
&= \frac{\text{Var}(X^0)}{t} + \frac{2}{t}\sum_{k=1}^{t-1}\gamma(k)\left[1-\frac{k}{t}\right] \qquad (22)
\end{aligned}$$

If $\sum_{k=1}^{\infty}|\gamma(k)| < \infty$ then it follows from Kronecker's lemma that $\sum_{k=1}^{t-1}\frac{k}{t}\gamma(k) \to 0$ for $t \to \infty$, see for example, Shiryayev (1984) p. 365. Multiplying equation (22) by t and then letting $t \to \infty$ we obtain:

$$\lim_{t\to\infty} t\text{Var}\,(\hat{\mu}) = \text{Var}(X^0) + 2\sum_{k=1}^{\infty}\gamma(k). \qquad (23)$$

C Decomposition of forecast uncertainty for ARMA and ARIMA processes

Since we restrict ourselves to stationary ARMA(p,q) models we can use the equivalent infinite moving-average representation of the process X^0 in (10):

$$X_t^0 = \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t-j} \tag{24}$$

with $\alpha_0 = 1$. An infinite moving-average representation can always be found for a stationary ARMA(p,q) process — see Granger and Newbold (1977, p25). Explicit formulae for the coefficients α_j can be found in Harvey (1981, p38).

As with the random-walk model, predicted values given κ_t and X_t^0 are obtained by setting all $\epsilon_k = 0$ in (24) for $k > t$:

$$X_t^0(h) = \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t+h-j} \mathbb{I}_{j \geq h} \tag{25}$$

$$\kappa_t(h) = \kappa_t + \sum_{i=1}^{h} X_t^0(i) + h\mu \tag{26}$$

Ignoring uncertainty about the ARMA parameters, we have the following:

$$\hat{\kappa}_t(h) = \kappa_t + \sum_{i=1}^{h} X_t^0(i) + h\hat{\mu} \tag{27}$$

As we are restricting ourselves to stationary ARMA processes, we can use the infinite moving-average representation of X^0 to express the prediction error for κ:

$$\begin{aligned}
\kappa_{t+h} - \hat{\kappa}_t(h) &= \Big(\kappa_{t+h} - \kappa_t(h)\Big) + \Big(\kappa_t(h) - \hat{\kappa}_t(h)\Big) \\
&= \sum_{i=1}^{h} \big(X_{t+i}^0 - X_t^0(i)\big) + h(\mu - \hat{\mu}) \qquad (28) \\
&= h(\mu - \hat{\mu}) + \sum_{i=1}^{h} \sum_{j=0}^{\infty} \alpha_j \Big[\epsilon_{t+i-j} - \epsilon_{t+i-j} \mathbb{I}_{j \geq i}\Big] \\
&= h(\mu - \hat{\mu}) + \sum_{i=1}^{h} \sum_{j=0}^{i-1} \alpha_j \epsilon_{t+i-j} \qquad (29)
\end{aligned}$$

The second term in equation (29) only depends on $\epsilon_{t+1}, \ldots, \epsilon_{t+h}$ while the first term only depends on $\varepsilon_1, \ldots, \varepsilon_t$. The two terms are therefore independent. We can use this to derive an expression for the mean squared prediction error as follows:

$$\begin{aligned}
\mathrm{E}\left[\Big(\kappa_{t+h} - \hat{\kappa}_t(h)\Big)^2\right] &= h^2 \mathrm{E}\left[\Big(\mu - \hat{\mu}\Big)^2\right] + \mathrm{E}\left[\Big(\sum_{i=1}^{h} \sum_{j=0}^{i-1} \alpha_j \varepsilon_{t+i-j}\Big)^2\right] \\
&= h^2 \mathrm{Var}\left(\hat{\mu}\right) + \mathrm{E}\left[\Big(\sum_{i=1}^{h} \sum_{j=0}^{i-1} \alpha_j \varepsilon_{t+i-j}\Big)^2\right] \qquad (30)
\end{aligned}$$

since $\mathrm{E}[\hat{\mu}] = \mu$. We have an expression for the first term on the right-hand side of equation (30) from the result in equation (22). For the second term in equation (30), we note that $\mathrm{E}\left[\sum_{i=1}^{h}\sum_{j=0}^{i-1}\alpha_j\varepsilon_{t+i-j}\right] = 0$ and so (setting $k = i - j$):

$$\sum_{i=1}^{h}\sum_{j=0}^{i-1}\alpha_j\varepsilon_{t+i-j} = \sum_{k=1}^{h}\left(\sum_{i=0}^{h-k}\alpha_i\right)\varepsilon_{t+k} \tag{31}$$

and so:

$$\mathrm{Var}\left(\sum_{i=1}^{h}\sum_{j=0}^{i-1}\alpha_j\varepsilon_{t+i-j}\right) = \sigma_\varepsilon^2\sum_{k=1}^{h}\left(\sum_{i=0}^{h-k}\alpha_i\right)^2$$

We can then further progress with equation (30) using the result in (22):

$$\begin{aligned}\mathrm{E}\left[\left(\kappa_{t+h} - \hat{\kappa}_t(h)\right)^2\right] &= h^2\mathrm{Var}\left(\hat{\mu}\right) + \sigma_\varepsilon^2\sum_{k=1}^{h}\left(\sum_{i=0}^{h-k}\alpha_i\right)^2 \\ &= h^2\left(\frac{\mathrm{Var}(X^0)}{t} + \frac{2}{t}\sum_{k=1}^{t-1}\gamma(k)\left[1 - \frac{k}{t}\right]\right) + \sigma_\varepsilon^2\sum_{k=1}^{h}\left(\sum_{i=0}^{h-k}\alpha_i\right)^2\end{aligned} \tag{32}$$

For large values of t, i.e. a long history, the mean squared prediction error can be approximated using equation (23):

$$\mathrm{E}\left[\left(\kappa_{t+h} - \hat{\kappa}_t(h)\right)^2\right] \approx \frac{h^2}{t}\left(\mathrm{Var}(X^0) + 2\sum_{k=1}^{\infty}\gamma(k)\right) + \sigma_\varepsilon^2\sum_{k=1}^{h}\left(\sum_{i=0}^{h-k}\alpha_i\right)^2 \tag{33}$$

Note that a random walk with drift is an ARIMA(0,1,0) process, i.e. X is a white noise process with mean μ. This allows us to use equation (32) to obtain the following:

$$\mathrm{E}\left[\left(\kappa_{t+h} - \hat{\kappa}_t(h)\right)^2\right] = \frac{h^2}{t}\sigma_\varepsilon^2 + h\sigma_\epsilon^2 = \left(\frac{h}{t} + 1\right)h\sigma_\epsilon^2 \tag{34}$$

since $\alpha_0 = 1$, $\alpha_k = 0$ and $\gamma(k) = 0$ for all $k > 0$. In each of the equations (32), (33) and (34), the first term on the right-hand side corresponds to uncertainty about the drift, μ, while the second term corresponds to uncertainty about future innovations ε (the volatility of the process κ).

It should be noted that both terms in (32) depend on the ARMA parameters α_i and the variance σ_ε^2, which we have assumed to be known with certainty. As mentioned earlier, in an empirical study those parameters will need to be replaced by appropriate estimates, which adds further uncertainty. Moreover, if the parameters α_i are considered to be uncertain we need an extra term in equation (28) for the prediction error, namely:

$$\sum_{i=1}^{h}\left(X_t^0(i)-\hat{X}_t^0(i)\right)=\sum_{i=1}^{h}\left(\sum_{j=0}^{\infty}\alpha_j\varepsilon_{t+i-j}\mathbb{I}_{j\geq i}-\sum_{j=0}^{\infty}\hat{\alpha}_j\varepsilon_{t+i-j}\mathbb{I}_{j\geq i}\right)$$
$$=\sum_{i=1}^{h}\sum_{j=0}^{\infty}\left(\alpha_j-\hat{\alpha}_j\right)\varepsilon_{t+i-j}\mathbb{I}_{j\geq i}$$
$$=\sum_{i=1}^{h}\sum_{j=i}^{\infty}\left(\alpha_j-\hat{\alpha}_j\right)\varepsilon_{t+i-j}$$

There are no finite-sample analytical expressions for the forecast density of $\hat{X}_t^0(i)$. We therefore use the bootstrap procedure proposed by Pascual et al. (2004), albeit with a modification to randomly select the initial sequence; this avoids bias in the estimation of σ_ϵ^2, as shown by contrasting Figures 8 and 9.

D MLE for drift parameter

An alternative estimator for the drift parameter μ can be found from the likelihood function of an ARMA model. We will only show this estimator for an AR(p) process with no moving-average term. To simplify the derivation we assume that the first p values of X are fixed, or that we ignore the contribution to the likelihood from those initial values — see Harvey (1981). Using the definition of an ARMA(p,0) process in equations (10) and (11), and assuming normally distributed error terms, we find the log-likelihood function of all parameters as in Harvey (1981):

$$\begin{aligned}
l(X_t,X_{t-1},\ldots,X_1;\theta) &= \log L(X_t,X_{t-1},\ldots,X_1;\theta)\\
&= C-\frac{\sigma^2}{2}\sum_{i=p+1}^{t}\left(X_i-\mu-ar_1(X_{i-1}-\mu)-\ldots-ar_p(X_{i-p}-\mu)\right)^2\\
&= C-\frac{\sigma^2}{2}\sum_{i=p+1}^{t}\left(-\sum_{k=0}^{p}ar_k(X_{i-k}-\mu)\right)^2\\
&= C-\frac{\sigma^2}{2}\sum_{i=p+1}^{t}\left(\sum_{k=0}^{p}ar_k(X_{i-k}-\mu)\right)^2 \qquad (35)
\end{aligned}$$

where $ar_0=-1$, and C is a constant that is independent of μ. Note that the constant C would depend on μ if we were to assume that the first p values are drawn from the stationary distribution of the process rather than being fixed. However, this will only have a small impact on the estimated value of μ.

Maximising the likelihood function with respect to μ is therefore equivalent to finding the least-squares estimator for μ, that is.

$$\hat{\mu}_{\text{LS}}=\text{argmin}_\mu\sum_{i=p+1}^{t}\left(\sum_{k=0}^{p}ar_k(X_{i-k}-\mu)\right)^2$$

We find for the score function for μ from equation (35):

$$
\begin{aligned}
U(\mu) &= \frac{\partial}{\partial \mu} l(X_t, X_{t-1}, \ldots, X_1; \theta) \\
&= -\frac{\partial}{\partial \mu} \frac{\sigma^2}{2} \sum_{i=p+1}^{t} \left(\sum_{k=0}^{p} a r_k (X_{i-k} - \mu) \right)^2 \\
&= -\sigma^2 \sum_{i=p+1}^{t} \left\{ \left(\sum_{k=0}^{p} a r_k (X_{i-k} - \mu) \right) \frac{\partial}{\partial \mu} \left(\sum_{k=0}^{p} a r_k (X_{i-k} - \mu) \right) \right\} \\
&= -\sigma^2 \sum_{i=p+1}^{t} \left\{ \left(\sum_{k=0}^{p} a r_k (X_{i-k} - \mu) \right) \left(-\sum_{k=0}^{p} a r_k \right) \right\} \\
&= \sigma^2 \left(\sum_{k=0}^{p} a r_k \right) \sum_{i=p+1}^{t} \left(\sum_{k=0}^{p} a r_k X_{i-k} - \mu \sum_{k=0}^{p} a r_k \right) \\
&= \sigma^2 \left(\sum_{k=0}^{p} a r_k \right) \left(\sum_{k=0}^{p} a r_k \left[\sum_{i=p+1}^{t} X_{i-k} \right] - (t-p) \mu \sum_{k=0}^{p} a r_k \right)
\end{aligned}
$$

It follows that we find $\hat{\mu}_{\text{MLE}}$ from solving:

$$
0 = \left(\sum_{k=0}^{p} \left[a r_k \sum_{i=p+1}^{t} X_{i-k} \right] - (t-p) \mu \sum_{k=0}^{p} a r_k \right)
$$

The solution to this equation is given by:

$$
\begin{aligned}
\hat{\mu}_{\text{MLE}} &= \frac{\sum_{k=0}^{p} \left[a r_k \sum_{i=p+1}^{t} X_{i-k} \right]}{(t-p) \sum_{k=0}^{p} a r_k} = \frac{\sum_{k=0}^{p} \left[-a r_k \sum_{i=p+1}^{t} X_{i-k} \right]}{(t-p) \sum_{k=0}^{p} (-a r_k)} \\
&= \frac{\sum_{i=p+1}^{t} X_i - a r_1 \sum_{i=p+1}^{t} X_{i-1} - \ldots - a r_p \sum_{i=p+1}^{t} X_{i-p}}{(t-p) \left(1 - \sum_{k=1}^{p} a r_k \right)} \\
&= \frac{\sum_{i=p+1}^{t} X_i - a r_1 \sum_{i=p}^{t-1} X_i - \ldots - a r_p \sum_{i=1}^{t-p} X_i}{(t-p) \left(1 - \sum_{k=1}^{p} a r_k \right)}
\end{aligned}
$$

which we can also rewrite as:

$$
\hat{\mu}_{\text{MLE}} = \sum_{k=0}^{p} \left[\frac{a r_k}{\sum_{k=0}^{p} a r_k} \frac{\sum_{i=p+1}^{t} X_{i-k}}{(t-p)} \right] \tag{36}
$$

We therefore find that the MLE $\hat{\mu}_{\text{MLE}}$ is a weighted average over the mean values over shifted $t-p$ observations, where the weights are determined by the coefficients of the AR(p) process. We also find that $\hat{\mu}_{\text{MLE}}$ is approximately equal to the mean $\hat{\mu}$. The approximation becomes obvious when the means $(\frac{1}{t-p}) \sum_{i=p+1}^{t} X_{i-k}$ are replaced by the overall mean $\frac{1}{t} \sum_{i=1}^{t} X_i$ in equation (36). Note that the difference between the mean $\hat{\mu}$ and the MLE $\hat{\mu}_{\text{MLE}}$ is decreasing for p decreasing and/or t increasing.