# Key points for Projections Toolkit Seminar


## Iain Currie, Heriot-Watt University & Longevitas


### The Lee-Carter Family

S1. CMI assured lives data, ages 40-89, years 1950-2005.

S2. Perspective plot of $\log \hat{\mu}_{ij} = \log(d_{ij}/e_{ij})$.

S3. Log(mortality) averaged over year. This is the famous Gompertz line (approximately).

S4. Log(mortality) averaged over age. This is what all the fuss is about. Average mortality is 'heading south' relentlessly.

S5. The Lee-Carter model (1992) is the father of all stochastic forecasting models. There are a few things to note about it:

(a) It is designed to forecast the whole table. It does this by reducing the 2-dimensional forecasting problem to a 1-dimensional problem. The parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$ are estimated. The age parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are then held fixed and $\boldsymbol{\kappa}$ is forecast. The great strength of the Lee-Carter is its stability; its great weakness is that it imposes a very rigid structure on the mortality table which may not always be suitable. In contrast, the 2-d P-spline methods offer a more flexible structure but can suffer from instability, particularly if the year signal is weak.

(b) The Lee-Carter model is over-parameterized and two constraints are required to fix the estimates. The forecasts are invariant with respect to the particular constraints used.

S6. The estimate of $\boldsymbol{\alpha}$ looks like average mortality by age in S3 and the estimate of $\boldsymbol{\kappa}$ has the same shape as the average mortality by year in S4. Notice that estimates of all three parameters follow an identifiable smooth pattern (which we'll take advantage of shortly).

S7. An ARIMA model has been used to forecast $\boldsymbol{\kappa}$ although a simple drift model could also be used (see *Forecasting with Time Series* later). The forecast comes with CIs.

S8. Notice that the shape of the forecasts is the same for all ages (the same as the shape of $\boldsymbol{\kappa}$ in S7).

S9. On the evidence of S6 (and S10) we can smooth some/all of the parameters. S9 reminds us how this works. The coefficients, unconstrained (solid triangles) and constrained (solid green squares) act on the B-splines in the lower panel to raise or lower each B-spline; these are then added together to give the smooth curves - wiggly when unconstrained and smooth when constrained.

S11. Mix and match smoothing. Delwarde, Denuit and Eilers (DDE) (2007) smoothed $\boldsymbol{\beta}$. They had a good reason for doing this - without smoothing the Lee-Carter can give some embarrassing crossovers, as S12 shows. Smoothing $\boldsymbol{\beta}$ will fix this most of the time and we would recommend that DDE always be used in preference to the original 1992 version. Not only are crossovers avoided (for the most part) but the forecasts change smoothly from age to age (S12, panel 3).

S13. Stephen and I (2009) used penalty forecasting in the Lee-Carter model. This introduces a new class of models. We're very keen on widening model choice since model risk is the 'great unknown'; the only way to have some kind of handle on model risk is to have a wide selection of models available. S14 compares the results from the DDE (time series) and CR (penalty) models. Some other possibilities are given in S15 and S16.

### 2D P-spline models

S4. Notice that not only does the penalty bring about the smoothing but also enables the forecasting to take place as a linear forecast of the final two coefficients (solid green squares).

S5. Think of the two mountain range profiles in 1-d combining to give the full mountain range in 2-d, as illustrated in S6.

S7. Cohorts run diagonally in an AP data matrix so we skew the matrix to make them run vertically (S8). Then we can apply row (cohort) and column (age) penalties (S9).

S11. Notice that the two models (AP and AC) give different forecasts with different widths for their CIs. Of course, we don't know which is to be preferred. This is an example of **model risk**.

### Piggyback forecasts of mortality

**Background** Piggyback forecasting is a way of using limited company data to adjust a forecast made from a large standard data set (CMI, ONS, etc). Any forecasting method can be used on the large data set and crucially the company data can be subdivided by classes of business. Slides 1-4 set the data scene.

S5. This slide shows we have a problem if we just go with the CMI forecast. Company mortality (which has been averaged over the four classes of business) is clearly heavier than the CMI. But the company data would appear to be insufficient to support a stand-alone forecast. Piggyback forecasting is designed to address this problem and it does this by adjusting the CMI forecast in a way not dissimilar to how standard table adjustment works.

S7. Log(mortality) has been averaged over year and this averaging reveals a pattern which can be seen over class of business.

S8. This is the corresponding slide when log(mortality) is averaged over age. The missing points arise because there are some company data cells with no deaths and so log(mortality) is not defined. These points have been omitted from the plot.

S9. Examination of S7 and S8 suggests these simple relationships. The key assumption is the constant in time one since this will allow forecasting to take place.

S10. We fit the company data over the existing CMI forecast. We call this *data trimming*.

S11. The top half gives the model: it's a linear adjustment by age of the existing forecast. The bottom half gives the forecast.

S12. This slide shows that the model fits quite well. Notice that there is evidence of mortality convergence at high ages, as we would expect.

S13. Again the model fits well and the key parallel assumption is clearly seen.

S14. The piggyback forecasts. The gaps between the CMI forecast and the company forecast depend obviously on the class of business; the size of these gaps is also age dependent (not shown).

### The Time Series *v* Penalty Debate

**Background** The *Projections Toolkit* offers two broad classes of forecasting methods: time series and penalty. Lee and Carter (1992) pioneered the time series approach; the penalty method is more recent (Currie, Durban and Eilers, 2004). But which method is more appropriate? Here is the argument as we see it.

S1. We have 1-dimensional data (age 70) over time (1950-2005). What do we expect to see in 2006?

S2. In this slide if you were ask to predict a value for 2006 would you start at the trend value or at last year's value? It's a simple question but quite fundamental.

S3. This gives the time series answer (using the simple random walk with drift model for illustration).

S4. This slide reminds you how penalty forecasting works: linear extrapolation of the final two coefficients (incidentally, this explains why you should not have your knots too close together since this will make your forecast over-sensitive to the recent past).

S5. And this is how the penalty forecast works.

**Forecasting with Time Series: ARIMA and Drift Models**

S1. Lee-Carter used the simple random walk with drift when they forecast future life expectancy in the USA. This worked for the data they were using (US mortality data up to 1989). The model is very simple: on average, $\hat{\kappa}_{n+1} = \hat{\kappa}_n + \hat{a}$. Thus next year is the current year + a constant term (known as the drift). The actual observation is $\hat{\kappa}_{n+1} = \hat{\kappa}_n + \hat{a} + z_{n+1}$ where $z_{n+1} \sim \mathcal{N}(0, \sigma^2)$ is a random disturbance; $\sigma^2$ measures the volatility. Estimation is very simple and the result is shown in S2. Applied to CMI data the result seems less than satisfactory. We need something more general.

S3. ARIMA or ARIMA$(p, d, q)$ models are a very flexible family of time series models. The acronym stands for AutoRegressive Integrated Moving Average process. The general formula is

$$\text{ARIMA}(p, d, q) : \phi(B)(1 - B)^d \kappa_t = \theta(B) z_t$$

where $B$ is the *backward-shift operator*. We examine each term in turn.

(a) $d$ is the order of differencing. For example, if $d = 2$ we have

$$(1 - B)^2 k_t = (1 - 2B + B^2) k_t = k_t - 2k_{t-1} + k_{t-2}$$

(b) $\phi(B)$ is a polynomial in $B$ of order $p$. For example, if $p = 1$ and $\phi(B) = 1 - 0.2B$ (as in S3), we have (assuming for simplicity that $d = 0$),

$$\phi(B)(1 - B)^d k_t = (1 - 0.2B) k_t = k_t - 0.2k_{t-1}$$

(c) $\theta(B)$ is a polynomial in $B$ of order $q$. For example, if $q = 1$ and $\theta(B) = 1 - 0.3B$ (as in S3),

$$\theta(B) z_t = (1 - 0.3B) z_t = z_t - 0.3z_{t-1}$$

Slide S3 gives an example of how this all gets put together and shows how next year's observation depends on the previous years' observations, the order of the differencing and the previous years' innovations.

S4. The *Projections Toolkit* gives automatic ARIMA model selection via minimizing AIC (note: slide S4 shows $-$AIC so the best fitting model maximizes AIC). The *Projections Toolkit* also allows you to over-ride the best fitting model with the help of the panel of fits for $\boldsymbol{\kappa}$.

S5. A comparison of the drift model and the best fitting ARIMA model. The central ARIMA forecast looks much more realistic. The ARIMA model has many more parameters than the simple drift model; the ARIMA CI is much wider than that of the drift model (which looks optimistically narrow to my eyes).

**Overdispersion and Outlier Handling**

S2. shows the result for age 70 of fitting the 2d AP P-spline model to the ONS data with ages 40-90 and years 1961-2007. The fit to the observed mortality looks fine but there are some worrying features (the *Projections Toolkit* will alert you to these).

S3. The two main worries are (a) very small smoothing parameters and (b) some very large standardized residuals. The large residuals are evidence of *overdispersion*. This arises when there is more variation in the data than the underlying Poisson assumption allows. This is common in population data when there is *heterogeneity* among those at risk (death rates vary enormously over social class, for example). With insurance data there is the additional problem of duplicate policies (although modern de-duplication methods can help here).

S4. shows what happens when a penalty forecast is made when there is essentially no signal in the year direction (very small $\lambda_y$).

S5. The Poisson model is the backbone of mortality work (Brouhns, et al., 2002) but it assumes a homogeneous risk set. The quasi-Poisson model allows for additional variation. Overdispersion is measured by the *overdispersion parameter*, $\Psi^2$. This generalizes the Poisson model since if $\Psi^2 = 1$ the quasi-Poisson model reduces to the Poisson model.

S6. With ONS data we have $\Psi^2 = 4.47$: strong evidence of extensive overdispersion (the CMI data is more homogeneous than the ONS data so we would expect a smaller value of $\Psi^2$ and indeed we find $\Psi^2 = 1.82$). Notice we have fixed our two problems: we have much larger smoothing parameters and much smaller residuals. S7 is much more satisfactory.

S8. shows the forecast for the CMI data. With a much smaller value of $\Psi^2$ than for the ONS data the original Poisson model gives satisfactory results. The central forecasts are almost identical but there is a bonus with overdispersion: the CIs are narrower (a stiffer fit gives a more secure forecast and hence narrower CIs).

S9. Large residuals can arise in various ways: overdispersion, poorly fitting model and rogue (outlier) data points. Fitting an overdispersion parameter should fix the first problem but the other two can remain. The Lee-Carter model can fit population data poorly by the usual “hypothesis testing” criteria but this does **not** mean that the forecasts are not useful. What of rogue data points? These can arise because of some unusual event: Spanish flu, a very cold winter. The effect of such anomalous points is to reduce the size of the smoothing parameters with the effect that forecasts are more volatile and CIs wider. Removal of such points will stabilize the forecasts. The *Projections Toolkit* allows automatic deletion of data points with large residuals. This facility should be used sparingly since a large number of very large residuals might indicate problems with model fit.