UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2015, Vol. 8, No. 2, pp. 173–192

Proyección de la población universitaria utilizando el modelo Lee-Carter[1]

Forecasting college populations via Lee-Carter models

Norman Giraldo Gómez[a]
ndgirald@unal.edu.co

Carlos Ochoa Molina[b]
ochoacarlosandres@gmail.com

Resumen

En universidades con grandes poblaciones estudiantiles es importante tener un sistema de monitoreo de su evolución. Su correcta medición e interpretación se torna necesario para evaluar los efectos de diferentes políticas, cambios en los reglamentos estudiantiles, costos de servicios médicos, seguros, bienestar estudiantil, etc.. Para lograr estos fines, una alternativa puede consistir en establecer un modelo estadístico para la evolución de las poblaciones a través de diferentes cohortes, que es el objetivo de este trabajo. El modelo aplicado se conoce como el modelo Lee-Carter, y puede localizarse dentro del conjunto de modelos de "tablas de vida dinámicas", en el área de análisis de supervivencia actuarial. En este trabajo se aplicaron dos metodologías de ajuste para el modelo con base en datos poblacionales de una universidad pública colombiana. Se compararon los estimadores del modelo Lee-Carter por el método original de descomposición singular y por el método de maxima verosimilitud en un modelo log-bilineal Poisson, utilizando la librería `gnm` de `R`. Aunque están disponibles otras librería como `LifeMetrics` e `ilc`, el procedimiento en gnm, para modelos no lineales generalizados parece ser el más aceptado en la actualidad. En particular, se implementó una regresión no-paramétrica Loess para el parámetros κ_t para pronosticarlo a corto plazo, y se comparó con un modelo ARIMA(1,1,0) con tendencia (los modelos ARIMA se recomiendan en la literatura para la estimación de κ_t). En ambos casos se observó una evolución estocástica del parámetro κ_t durante el periodo entre los semestres 1989-02 a 2006-01, en total 34 semestres. Con los modelos ARIMA(1,1,0) y Loess se procedió a calcular 10 pronósticos para κ_t. Con el pronóstico número 10, correspondiente a la cohorte del semestre 2011-01, se calculó la evolución futura de esta

[1]Girado, N., Ochoa, C. (2015). Proyección de la población universitaria utilizando el modelo Lee-Carter. *Comunicaciones en Estadística,* **8**(2), 173-192.

[a]Profesor Asociado. Escuela de Estadística. Universidad Nacional de Colombia, sede Medellín. Colombia

[b]Ing. Adm. Estudiante Maestría en Estadística, Universidad Nacional de Colombia, sede Medellín. Colombia

cohorte durante 17 semestres, de los cuales ya se han observado 8 al final del año 2014. Al comparar los observados con los calculados los resultados son satisfactorios y los pronósticos coinciden razonablemente con los valores observados, por lo que los restantes pronósticos son confiables. La evolución de la mortalidad durante 34 semestres (17 años) permite obtener conclusiones acerca de los cambios en la población estudiantil, las cuales aparecen en la última sección del estudio.

***Palabras clave*:** modelo Lee-Carter, poblaciones universitarias, proyección de poblaciones, tablas de vida.

Abstract

In Colleges with large student populations, it is important to have a system for monitoring their evolution along the time. Proper measurement and interpretation are necessary to help evaluate the effects of different policies, changes in student regulations, medical costs, insurance, student welfare, etc. To achieve these goals an alternative may be to establish a statistical model for the population evolution across different cohorts, which is the subject of this article. The model used is known as the Lee-Carter model, which can be seen as an example of "dynamic life tables models", in the area of actuarial survival analysis. In this work we applied two estimationg methodologies for adjustment of this model to real life data, from a Colombian University. The estimators were the original method of singular decomposition and the method of maximum likelihood in a log-bilinear Poisson model, using the library of `gnm` in `R`. Although there are other available libraries like LifeMetrics and ilc, the procedure in the `gnm` library, for generalized non-linear models, seems to be the most accepted at present. Additionally, we implemented a non-parametric regression of Loess type for one of the parameters in the Lee-Carter model: the κ_t, which enabled us to calculate forecasts for predicting the central rate of mortality; we compared this forecast with the ones obtained with a model ARIMA(1,1,0) with trend, given that the ARIMA models are recommended in the literature for the forecasting of κ_t. Models with the ARIMA(1,1,0) and loess is proceeded to calculate 10 forecasts for κ_t. With the forecast number 10, corresponding to the cohort semester 2011-01, we estimated the future evolution of this cohort during 17 semesters, of which had already been observed the first 8 at the end of the year 2014. When comparing the observed populations with the calculated ones results are satisfactory, and the forecasts match reasonably well with the observed values, so that the remaining forecasts are reliable. The evolution of mortality during 34 semesters (17 years) lets us obtain conclusions about the changes in the student population, which appear in the last section of the study.

***Keywords*:** college population, forecasting populations, Lee-Carter model, life tables.

1. Introducción

En universidades con grandes poblaciones estudiantiles es importante tener un sistema de monitoreo de su evolución. Su correcta medición e interpretación se torna necesario para evaluar los efectos de diferentes políticas, cambios en los reglamentos estudiantiles, costos de servicios médicos, seguros, bienestar estudiantil, etc. Para lograr estos fines una alternativa puede consistir en establecer un modelo estadístico para la evolución de las poblaciones a través de diferentes cohortes, que es el objetivo de este trabajo.

El modelo aplicado se conoce como el modelo Lee-Carter, y puede localizarse dentro del conjunto de modelos de "tablas de vida dinámicas", en el área de análisis de supervivencia actuarial. Si bien este modelo se utiliza para modelar los cambios en los patrones de la mortalidad humana a través de diferentes generaciones, encontramos que puede adaptarse al problema de modelar la evolución de las cohortes de población universitaria, con algunas modificaciones importantes. La variable respuesta analizada es la fuerza de mortalidad μ_x para un individuo de edad x, que en el contexto de este trabajo se interpreta como la fuerza de deserción-egreso, en el semestre x, es decir, la propensión a retirarse del plan de estudios estando cursando el semestre x. A mayor valor de μ_x, mayor probabilidad de retirarse. Se modifica la función para incluír un efecto de cohorte colocando $\mu_{x,\tau}$ donde $\tau = \tau_1, \ldots, \tau_{n_c}$ y $x = 1, \ldots, n_a$, donde n_c es el número de cohortes consideradas y n_a es el número máximo de semestres cursados para todas las cohortes. En este estudio, $n_c = 36$ que corresponde a 18 años y comprende el periodo entre 1989/02 y 2006/01, y $n_a = 35$. Los valores x_i se toman en semestres, indicando cuántos semestres ha completado un estudiante genérico al momento de retirarse de la Universidad. El evento de retiro puede deberse a la deserción o al completamiento de todas las asignaturas del plan de estudios. Las cohortes τ_j se refieren al semestre de ingreso a la universidad.

El modelo se plantea colocando $\log(\mu_{x,\tau}) = \alpha_x + \beta_x \kappa_\tau$, donde $\alpha_x, \beta_x, \kappa_\tau$ son los parámetros por estimar. En el modelo la fuerza de mortalidad en τ_1 es similar a la fuerza de mortalidad en las demás cohortes; la diferencia consiste en el parámetro κ_τ, que mide el efecto de la cohorte, modulado por β_x. La forma de $\exp(\alpha_x)$ define el patrón de mortalidad, pero el efecto de la cohorte lo cambia. Un aspecto importante del modelo es que κ_τ se asume como un proceso estocástico(ver Brouhns et al. (2002); 377), modelado como un ARIMA, usualmente una marcha aleatoria con tendencia (la librería `ild` asume este modelo), para propósito de calcular un pronóstico, $\widehat{\kappa}_{\tau_{n_c}+j}, j = 1, 2, \ldots, h$. En este trabajo no se aplicó el modelo de marcha aleatoria sino un modelo ARIMA(1,1,0), y además se utilizó un modelo Loess de regresión local, ver Cleveland et al. (1992), el cual proporcionó resultados diferentes y posiblemente más adecuados. Además, el suavizamiento Loess de $\widehat{\kappa}_\tau$ permitió una mejor interpretación de la variación de la mortalidad a través de las cohortes.

La evolución de los métodos de estimación del modelo Lee-Carter inicia con la metodología de la descomposición en valores singulares en Lee & Carter (1992).

Luego pasa por un momento importante con la re-formulación del modelo Lee-Carter como un modelo log-bilineal Poisson, utilizando modelos lineales generalizados, en Brouhns et al. (2002). Esta formulación permite variantes tales como utilizar distribuciones binomial negativa y binomial, en lugar de la Poisson, que no se exploraron en este trabajo. Asimismo, generó la creación de varias librerías en R para estimación con base en esta propuesta, como la librería `ilc`, creada por Butt & Haberman (2009) y la librería `LifeMetrics` en Cairns (2007). Posteriormente se incluyó el modelo Lee-Carter dentro de un conjunto muy amplio de modelos no lineales generalizado, en un trabajo realizado por Currie (2014) y Currie (2013), desarrollo acompañado de un potente, moderno y completo *software* de estimación en R, la librería `gnm`, creada por Turner & Firth (2007). Por otro lado, las generalizaciones del modelo inicial comprenden básicamente la incorporación de un efecto de cohorte en Renshaw & Haberman (2009) adicional (efecto aditivo) al término $\beta_x\kappa_t$ del modelo original. Un modelo alterno a Lee-Carter lo desarrolló Wilmoth (1993). Otra modificación consiste en suavizar los estimadores de α_x,β_x y κ_t con *splines* en Delwarde et al. (2007).

El modelo Lee-Carter tiene aplicaciones, por ejemplo, en demografía y en el modelamiento del riesgo de longevidad y sus efectos en distintos tipos de seguros de vida y en pensiones. Citamos solamente algunas por ser un tema muy desarrollado. Primero, una exposición muy completa del modelo está en Koissi & Shapiro (2008). Con relación al riesgo de longevidad, que consiste en que las previsiones para reservas en seguros de vida y pensiones se realizan con base en tablas de vida para una generación determinada, pero se aplican para las siguientes, las cuales pueden experimentar variaciones en sus patrones de mortalidad, tales como un aumento significativo en la esperanza de vida. Tales variaciones pueden ser estimadas mediante el modelo Lee-Carter. Revisiones muy completas del modelo con relación al riesgo de longevidad están en Brouhns & Denuit (2002). Aplicaciones en demografía comprenden estudios sobre la variación de la mortalidad en diferentes países. También se ha comparado el modelo Lee-Carter con el modelo aditivo doble multiplicativo propuesto por Wilmoth (1993) para pronóstico de la mortalidad en México.

La organización del artículo es la siguiente. En la sección 2 se introducen las definiciones actuariales de las funciones de la tabla de vida y su aplicación al caso de poblaciones universitarias por cohortes. En la sección 3 se introduce el modelo Lee-Carter. En la sección 4 se introducen las metodologías de estimación de descomposición en valores singulares. En la sección 5 se introduce la estimación de máxima verosimilitud para un modelo log-bilineal Poisson y su implementación en R con la función `gnm()`. En la sección 6 están los resultados del estudio. La sección 7 se presentan las conclusiones.

2. Las funciones de la tabla de vida

Las definiciones siguientes pueden encontrarse en cualquier texto de análisis actuarial de contingencias de vida, por ejemplo Bowers et al. (1986) o Huertas (2001), pero para efectos del presente trabajo se recapitulan en términos de supervivencia durantes sucesivos semestres de un estudiante genérico.

Se define X como la duración de la permanencia de un estudiante en la universidad desde su ingreso hasta su egreso definitivo; asumimos que es una variable aleatoria continua y no negativa, con función de distribución F. Si un estudiante hace una reserva de cupo y reingresa después de uno o dos semestres, el valor de X incluirá estos semestres.

Denotamos por $S(x) := 1-F(x)$ la función de supervivencia. Si $x \geq 0$ consideremos la variable aleatoria $T(x) := X - x|X > x$ con distribución $G(t) := P(T(x) \leq t)$ para $t \geq 0$. La función $G(t)$ se interpreta como la probabilidad de retiro antes de t semestres habiéndose observado una permanencia de x semestres. La notación actuarial para esta función es ${}_tq_x$. Igualmente ${}_tp_x := 1-{}_tq_x$. Cuando $x = 0$ se tiene que $G(t) = F(t)$, y cuando $t = 1$ se escribe $q_x = {}_1q_x$. Igual con p_x. La función de fuerza de mortalidad o fuerza de deserción-egreso, es la base para la definición de modelos paramétricos. Esta función se define como $\mu_x := \frac{1}{S(x)}\frac{dF(x)}{dx}$, y se cumple la siguiente identidad

$$ {}_tp_x = \exp(-\int_0^t \mu_{x+s}ds). \tag{1}$$

Definimos μ_x como la fuerza de retiro o deserción-egreso. Cada retiro se debe a una deserción o a un egreso, por lo que resulta que existen dos riesgos competitivos (decrementos múltiples): deserción y egreso, es decir, $\mu_x = \mu_x^{(d)} + \mu_x^{(e)}$. Sin embargo, los datos utilizados en este estudio no proveen información acerca de cuál riesgo actúa cuando se reporta un retiro. Por esta razón, no incluímos análisis acerca de deserción y egresos. Igualmente, tampoco se tienen datos según el género, por lo que no se puede conocer el efecto de este factor.

Asumiendo una población inicial de l_0, por ejemplo, el total de estudiantes que ingresan en un semestre determinado, se define una variable aleatoria binomial, $\mathcal{L}(x)$ tal que $\mathcal{L}(x) \sim \text{Bin}(l_0, S(x))$. El número de estudiantes de la cohorte que alcanzan a matricularse en el semestre x se define como el valor esperado de $\mathcal{L}(x)$, $l_x := E(\mathcal{L}(x)) = l_0 S(x)$.

También, el número de estudiantes que se retiran entre los semestres x y $x+1$ es una variable aleatoria binomial, $\mathcal{D}(x) \sim Bin(l_0, S(x) - S(x+1))$. Su valor esperado se denota por $d_x = l_0(S(x) - S(x+1)) = l_x - l_{x+1}$. Por tanto, $d_x/l_x = (S(x) - S(x+1))/S(x) = q_x$. De (1) se sigue esta identidad:

$$ l_{x+1} = l_x \exp(-\int_0^1 \mu_{x+s}ds). \tag{2}$$

En lo que sigue asumimos el supuesto de fuerza de mortalidad constante por tramos, que establece:

$$\mu_{x+s} = \mu_x,\ 0 \leq s < 1. \tag{3}$$

Por tanto, $l_{x+1} = l_x e^{-\mu_x}$ y $q_x = 1 - e^{-\mu_x}$. La función l_x proporciona el total de estudiantes que empiezan el semestre x; pero el total de los que empezaron este semestre y no han alcanzado a terminarlo es diferente, se indica por L_x y se puede aproximar por $L_x = l_x - d_x/2$. La función L_x se denomina total de expuestos al riesgo, o exposición, también denotada por E_x. La función $q_x = d_x/l_x$ se denomina *tasa de mortalidad*. La función $m_x = d_x/L_x$ se denomina *tasa central de mortalidad*. Se comprueba que aproximadamente, $m_x = \mu_{x+1/2}$.

Introducimos ahora el concepto de *cohorte*. En el análisis de supervivencia su definición se da en términos de decrementos por decesos en la población. Aquí utilizamos decrementos por retiros definitivos.

Definición 2.1. *Una cohorte semestral es un grupo de estudiantes que tienen las siguientes características:*

1. *El grupo consiste inicialmente de l_0 estudiantes.*
2. *Cada miembro del grupo tiene una probabilidad q_x de retiro al comenzar el semestre x. Este evento de retiro incluye la deserción y el egreso por terminación de los estudios.*
3. *Al grupo no ingresan otros estudiantes fuera de los l_0 iniciales.*
4. *Los tiempos de supervivencia en cada cohorte se asumen variables aleatorias independientes.*

Para $\tau = \tau_1, \ldots, \tau_{n_c}$ se denota por $l_{x,\tau}$ el total de estudiantes de la cohorte τ que empiezan el semestre x, con $x = 1, ..., n_a$. La definición anterior implica $l_{\omega_\tau+1,\tau} = 0$. Cada valor ω_τ es diferente, y n_c denota el total de cohortes completas observadas. Los valores $l_{x,\tau}, \tau = \tau_1, \ldots, \tau_{n_c}$ se interpretan como el censo de estudiantes del semestre x.

Las funciones de la sección anterior se consideran sin referencia a una cohorte particular. De ahora en adelante se asume el supuesto que las funciones definidas hasta el momento depeden de una cohorte que se identifica con el símbolo τ, el cual indica el semestre al cual ingresan $l_{0,\tau}$ estudiantes. Por ejemplo, la identidad (2) se reescribe como:

$$l_{x+1,\tau} = l_{x,\tau} \exp(-\int_0^1 \mu_{x+s,\tau} ds). \tag{4}$$

3. El modelo Lee-Carter para la proyección de la población

Una población se define estacionaria cuando se cumple dos condiciones:

1. Las probabilidades de supervivencia en el semestre x no dependen de la cohorte, o sea:
$$p_{x,\tau} = p_x \text{ para } x = 1, \ldots, n_a$$

2. El número de estudiantes que ingresan en cada cohorte es constante. O sea, las variables $l_{0,\tau}$ se asumen constantes e iguales.

El modelo Lee-Carter (ver Lee & Carter (1992)) es un modelo para poblaciones no estacionarias, en las cuales las funciones biométricas dependen de la cohorte τ. Concretamente, el modelo asume que la fuerza de mortalidad a la edad x, medida con la tasa central de mortalidad, en la cohorte τ, es de la forma:

$$\log(m_{x,\tau}) = \alpha_x + \beta_x \kappa_\tau + \epsilon_{x,\tau}, \tag{5}$$

Donde $x = 1, \ldots, n_\tau$, $\tau = \tau_1, \ldots, \tau_{n_y}$ y $\epsilon_{x,\tau} \sim i.i.dN(0, \sigma^2)$. Las funciones $\alpha_x, \beta_x, \kappa_\tau$ son los parámetros del modelo. El modelo tiene un término bilineal $\beta_x \kappa_\tau$. Y no es identificable ya que es invariante por transformaciones como $\beta_x \to c\beta_x$,$\kappa_\tau \to \kappa_\tau / c$, $\forall c \neq 0$; pero resulta identificable con estas restricciones:

$$\sum_\tau \kappa_\tau = 0, \qquad \sum_x \beta_x = 1 \tag{6}$$

El modelo no se puede ajustar por regresión debido a que no tiene variables explicativas observadas en el término de la derecha de (5). El parámetro α_x describe el promedio en el tiempo del logaritmo de la fuerza de mortalidad o deserción-egreso. El parámetro κ_τ define la tendencia en el tiempo de la fuerza de mortalidad, y β_x es un parámetro que mide la sensibilidad del patrón dado en α_x a los cambios en el tiempo dados por κ_τ. Nótese (ver Brouhns & Denuit (2002): 50) que se cumple (ignorando el efecto del error aleatorio):

$$\frac{\partial \log(m_{x,\tau})}{\partial \tau} = \beta_x \frac{\partial \kappa_\tau}{\partial \tau}$$

Por ello, las edades x para las cuales β_x tenga valores más grandes son más sensibles a los cambios en la variación temporal de κ_τ.

Asumimos en este trabajo el supuesto de fuerza de mortalidad constante, es decir,

$$\mu_{x+s,\tau} = \mu_{x,\tau}, \ 0 \leq s < 1. \tag{7}$$

Por tanto:

$$l_{x+1,\tau} = l_{x,\tau} \exp(-\mu_{x,\tau}) = l_{x,\tau} \exp(-\exp(\alpha_x + \beta_x \kappa_\tau)).$$

Y también se cumple:

$$q_{x,\tau} = 1 - \exp(-\exp((\alpha_x + \beta_x \kappa_\tau))$$

4. Métodología original de estimación del modelo de Lee-Carter

La estrategia original de estimación del modelo de Lee-Carter (5) se basa en la minimización de esta función:

$$D(\underline{\alpha},\underline{\beta},\underline{\kappa}) = \sum_{x,\tau}(\log(m_{x,\tau}) - (\alpha_x + \beta_x\kappa_\tau))^2$$

Esta se establece respecto a $(\underline{\alpha},\underline{\beta},\underline{\kappa})$, $\underline{\alpha},\underline{\beta} \in \mathbb{R}^{n_a}$, $\underline{\kappa} \in \mathbb{R}^{n_c}$. En forma matricial, si $M = [\log(m_{x,\tau})] \in \mathbb{R}^{n_a \times n_c}$, $\underline{1}_{n_c} = (1,\ldots,1)' \in \mathbb{R}^{n_c}$:

$$D(\underline{\alpha},\underline{\beta},\underline{\kappa}) = ||M - \underline{\alpha}\underline{1}'_{n_c} - \underline{\beta}\underline{\kappa}'||_2^2. \tag{8}$$

Donde $||A||_2 = \sqrt{\sum_i \sum_j |A_{i,j}|^2}$ es la norma Frobenius para matrices. A partir de la ecuación $\partial D/\partial\underline{\alpha} = \underline{0}$ las restricciones (6) permiten obtener como solución el vector dado por los promedios de las filas de M, $\widehat{\underline{\alpha}} = \frac{1}{n_c} M\underline{1}_{n_c}$. Reemplazando esta solución en (8), es problema ahora consiste en minimizar con respecto a $(\underline{\beta},\underline{\kappa})$ esta función

$$D(\underline{\beta},\underline{\kappa}) = ||M - \widehat{\underline{\alpha}}\underline{1}'_{n_c} - \underline{\beta}\underline{\kappa}'||_2^2. \tag{9}$$

La minimización en (9) es un caso del problema general de aproximar una matriz dada H por otra matriz de rango igual o menor, X, es decir, se trata de encontrar $\widehat{X}$ tal que

$$\widehat{X} = \underset{rank(X)\leq r}{\operatorname{argmin}} ||H - X||_2, \tag{10}$$

Donde $r = rank(H)$. La solución $\widehat{X}$ se puede encontrar mediante la descomposición singular de la matriz H.

El teorema de descomposición singular de una matriz (ver Shores (2004), sec. 5.6) establece que dada una matriz $H \in \mathbb{R}^{m\times n}$ se pueden definir dos matrices ortogonales (sus columnas son vectores ortogonales) $U \in \mathbb{R}^{m\times m}$, y $V \in \mathbb{R}^{n\times n}$, tales que $H = U\Sigma V'$, donde $\Sigma \in \mathbb{R}^{m\times n}$ es una matriz diagonal $\Sigma = \text{diag}(\sigma_1,\ldots,\sigma_p)$, con $p = min(m,n)$ y $\sigma_1 \geq \ldots \geq \sigma_p \geq 0$. Los valores σ_j se denominan valores singulares de H, las columnas de U los vectores singulares izquierdos y las de V vectores sigulares derechos. Si $r \leq min(m,n)$ es el rango de H se cumple $\sigma_r > 0$ y $\sigma_{r+1} = 0$. El siguiente resultado es un corolario de este teorema (ver Shores (2004), Cor.5.7,(6)). Si $U = [u_1,\ldots,u_m]$, $V = [v_1,\ldots,v_n]$, entonces se cumple la descomposición siguiente:

$$H = \sum_{j=1}^{r} \sigma_j u_j v'_j. \tag{11}$$

La solución $\widehat{X}$ al problema (10) para $rank(X) = h < r = min(m,n)$ está dada por la siguiente expresión, establecida en Eckart & Young (1936):

$$\widehat{X} = \sum_{j=1}^{h} \sigma_j u_j v'_j. \tag{12}$$

Aplicando este resultado a la matriz $H = M - \underline{\widehat{\alpha}}\,\underline{1}'_{n_c}$, asumiendo $n_a = m < n = n_c$, con $r = n_a$:

$$M - \underline{\widehat{\alpha}}\underline{1}'_{n_c} = \sum_{j=1}^{h} \sigma_j u_j v_j'. \tag{13}$$

En el trabajo original Lee & Carter (1992) utilizan una aproximación de rango $h = 1$ y, por tanto, se estiman $\underline{\widehat{\beta}} = \sqrt{\sigma_1}u_1$ y $\underline{\widehat{\kappa}} = \sqrt{\sigma_1}v_1$, respectivamente. Este resultado se puede colocar de manera informal así:

$$M \approx \underline{\widehat{\alpha}}\,\underline{1}'_{n_c} + \underline{\widehat{\beta}}\,\underline{\widehat{\kappa}}' \tag{14}$$

En Koissi & Shapiro (2008) y Brouhns et al. (2002) se analiza con detalle esta metodología. El procedimiento se puede realizar directamente utilizando la función de R, `svd()` que calcula U, V, o mediante la librería `demography`, Hyndman et al. (2014), con la función `lca()`, de acuerdo con la instrucción siguiente:

```
modlc = lca(datos.un, interpolate = TRUE, adjust = 'none')
```

Donde `datos.un` es un objeto creado con la función `demogdata` de la librería `demography`, con esta instrucción:

```
datos.un = demogdata(data=dx, pop=Ex,years=t,ages=x,
type="mortalidad",label="UNAL",name="cohortes")
```

Una modificación del modelo Lee-Carter (5) consiste en aplicar un suavizamiento a los parámetros dependientes de la edad x. El objetivo es mejorar las proyecciones eliminando la volatilidad en estos parámetros. La técnica utililizada consiste en utilizar splines penalizados (P-splines), ver Delwarde et al. (2007)); sin embargo, no se utilizó en este trabajo.

5. Estimación con base en modelos lineales y no lineales generalizados

El modelo original Lee-Carter (5) presenta varias desventajas. Una de ellas es que asume homocedasticidad en los errores; pero este supuesto no se cumple con datos de tablas de vida en los cuales en las edades avanzadas al tener pocos datos para la exposición, los estimadores presentan alta varianza en contraste con los estimadores de edades tempranas, ya que $Var(\log(m_{x,\tau}) \approx 1/d_{x,\tau}$ (ver Wilmoth (1993): 2).

Para corregir este hecho Brouhns et al. (2002) modificaron el modelo original, asumiendo $\mathcal{D}(x,\tau) \sim Poisson(E_{x,\tau}m_{x,\tau})$ en lugar de $\mathcal{D}(x) \sim Bin(l_0, S(x) - S(x+1))$, tal que

$$\log(m_{x,\tau}) = \alpha_x + \beta_x\kappa_\tau \tag{15}$$

Donde $d_{x,\tau}/E_{x,\tau} = m_{x,\tau}$ y $d_{x,\tau}$ es un valor observado de $\mathcal{D}(x,\tau)$. Una expresión equivalente del modelo es:

$$\log(d_{x,\tau}) = \log(E_{x,\tau}) + \alpha_x + \beta_x \kappa_\tau. \tag{16}$$

El término $\log(E_{x,\tau})$ se denomina *offset*.

El método de máxima verosimilitud para estimación de la versión log-bilineal Poisson del modelo Lee-Carter (16) está desarrollado en Wilmoth (1993) y Brouhns et al. (2002) y complementado por Renshaw & Haberman (2009). Como $\mathcal{D}(x,\tau) \sim Poisson(L_{x,\tau} m_{x,\tau})$, entonces la función log-verosimilitud basada en los datos $d_{x\tau}$, $E_{x,\tau}$, y colocando $\lambda_{x,\tau} = E_{x,\tau} m_{x,\tau}$, está dada por:

$$l(d,\lambda) = \sum_{x,\tau} d_{x,\tau} \ln(\lambda_{x,\tau}) - \lambda_{x,\tau} - \ln(d_{x,\tau}!). \tag{17}$$

Se busca encontrar los valores en la matriz $\lambda = [\lambda_{x,\tau}]$ que maximicen (17). Como se cumple que:

$$\lambda_{x,\tau} = E_{x,\tau} m_{x,\tau} = E_{x,\tau} e^{\alpha_x + \beta_x \kappa_\tau}$$

entonces maximizar (17) con respecto a $\lambda_{x,\tau}$ equivale a maximizar:

$$l(\alpha, \beta\, \kappa) = \sum_{x,\tau} d_{x,\tau} \ln(E_{x,\tau} e^{\alpha_x + \beta_x \kappa_\tau}) - E_{x,\tau} e^{\alpha_x + \beta_x \kappa_\tau} \tag{18}$$

con respecto a $\alpha_x, \beta_x, \kappa_\tau, x = 1, \ldots, n_a, \tau = \tau_1, \ldots, \tau_{n_c}$

El estimador de máxima verosimilitud en el contexto de modelos lineales generalizados se obtiene minimizando la función deviance. Si $d_{x,\tau}$ es el conteo de bajas observado y $\widehat{d}_{x,\tau} = E_{x,\tau} e^{\widehat{\alpha} + \widehat{\beta}\widehat{\kappa}}$, el estadístico chi-cuadrado Deviance se define así:

$$D = 2 \sum_{x,\tau} \omega_{x,\tau} \left(d_{x,\tau} \log \left(\frac{d_{x,\tau}}{\widehat{d}_{x,\tau}} \right) - (d_{x,\tau} - \widehat{d}_{x,\tau}) \right), \tag{19}$$

donde $\omega_{x,\tau} = 0$ si $d_{x,\tau} = 0$ y $\omega_{x,\tau} = 1$ si $d_{x,\tau} > 0$. La minimización del estadístico D se hace mediante el método de Newton-Raphson unidimensional, aplicado a cada parámetro $\theta = \alpha_x, \beta_x, \kappa_t$, el cual actualiza el valor de θ según la expresión:

$$\theta_{j+1} = \theta_j - \frac{\frac{\partial D}{\partial \theta}}{\frac{\partial^2 D}{\partial \theta^2}} \tag{20}$$

Este método está implementado en la librería `ilc` de Butt & Haberman (2009), mediante la función `lca.rh()`, aplicada a un objeto que contiene la información d_x, m_x, x, τ; el código siguiente muestra la forma de utilizarla:

```
datos.un = demogdata(data=dx,pop=Ex,years=t,ages=x,
type="mortality",label="UNAL",name="cohortes")
m1 = lca.rh(datos.un, mod='lc', interpolate=TRUE,
verbose = FALSE)
```

La librería `ilc` está integrada con la librería `demography`, por lo que tiene la posibilidad de realizar pronósticos a partir de un modelo de marcha aleatoria, ARIMA(0,1,0), para κ_τ, mediante este comando
`prons = forecast(m1, h = 5, jump = 'fit', level = 90, shift = FALSE)`.

Este método está también implementado en la librería `LifeMetrics`, debida a Cairns (2007), la cual lo aplica no solamente el modelo Lee-Carter log-bilineal Poisson (16), sino otros siete modelos desarrollados en Cairns et al. (2009), entre los cuales están el modelo de Renshaw-Haberman, y el modelo CBD de Cairns, Blake y Dowd. Un ejemplo de la sintaxis con esta librería para estimar el modelo Lee-Carter está a continuación:

```
x=qdata.usa$x; y=qdata.usa$y; etx=qdata.usa$etx; dtx=qdata.usa$dtx;
wa=qdata.usa$wa
res=fit707(x,y,etx,dtx,wa)
```

El modelo log-bilineal Poisson (16) se puede definir como un ejemplo de la clase de modelos no-lineales generalizados, definidos a continuación (ver Davidian (2009): capítulo 4).

La librería `gnm`, creada por Turner & Firth (2007), para estimación en modelos no lineales generalizados, tiene como uno de sus ejemplos el modelo Lee-Carter, el cual se puede ver como un modelo no lineal generalizado debido a la presencia del término bilineal. En Currie (2014) se analiza la aplicación de la función para estimar modelos lineales generalizados `glm()` para algunos de los modelos definidos en Cairns et al. (2009), y la función `gnm()` para los modelos nolineales generalizados. El objetivo de Currie (2014) es mostrar que estas funciones proveen una mayor flexibilidad frente a las posibilidades de las librerías `ilc` y `LifeMetrics`. En particular, estas funciones permiten estimar modelos con base en las tasas de mortalidad $q_{x,\tau}$, utilizar distribuciones con efecto de sobredispersión. En Turner & Firth (2007) se encuentra una exposición muy detallada de las opciones de la función `gnm()` y su aplicación a la estimación del modelo Lee-Carter.

La sintaxis para estimar un modelo consiste en definir las variables `t,x,dx,Ex`, contenidas en un `data.frame`, por ejemplo, D, con esta instrucción

```
m1 = gnm(dx ~ x + Mult(x, t), offset = log(Ex),
family = "poisson", data = D)
```

Esta opción permite utilizar otras familias de modelos exponenciales, como "quasi-poisson", con la que se estima un modelo Poisson con parámetro de sobre-dispersión ϕ.

6. Resultados de la estimación

Nota sobre los datos

En este trabajo se consideraron datos de las cohortes entre los semestres 02/1989 y 01/2006, para la Universidad Nacional de Colombia, sede Medellín, proporcionados por la Oficina de Planeación. Los datos presentan algunas inconsistencias, que se analizan a continuación. Por su definición, la función l_x debe ser monótona decreciente, $l_{x+1} \leq l_x$; pero esto no sucede siempre, posiblemente debido a reingresos y translados. Por tanto, se aplicó un suavizamiento monótono a los datos, mediante la función `loess()` de R. Adicionalmente, los datos de cada cohorte en los últimos semestres tiene valores pequeños o tienen valores cero. Esto genera valores L_x muy pequeños que producen estimadores d_x/L_x con mucha varianza. La acción que se tomó fue considerar en cada cohorte solamente los valores x tales que $x \leq 17$, es decir, se utilizaron datos se supervivientes hasta $n_a = 17$ semestres; quienes presentaron una permanencia de más de 17 semestres se distribuyen de manera irregular, con poco datos. En la figura 1 aparece una gráfica de las $n_c = 34$ cohortes entre 02/1989 y 01/2006; cada trayectoria muestra cuántos estudiantes de cada cohorte van sobreviviendo hasta que egresa el último.

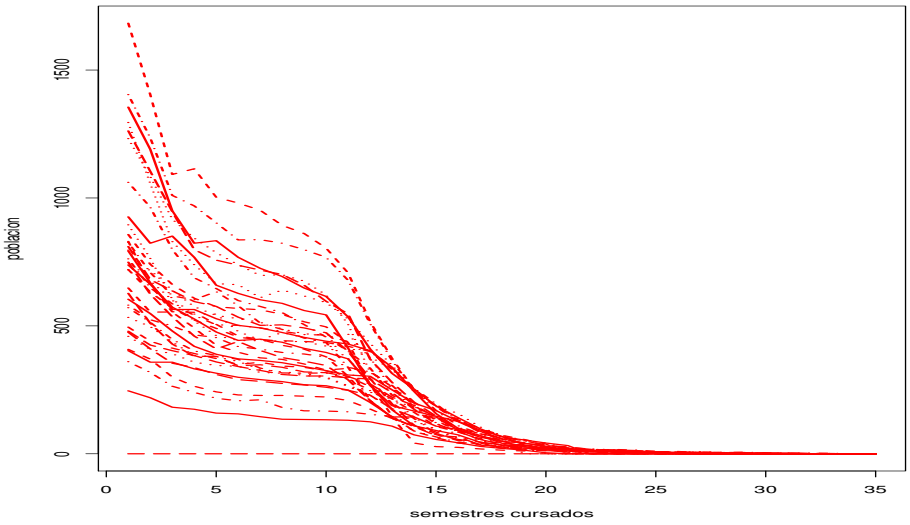

Figura 1: *Evolución de cohortes: 1989/02 - 2006/01. Fuente: Elaboración propia.*

Comparación de la estimación por SVD versus GNM

Comparamos los resultados de la estimación por descomposición singular (14) versus la estimación por minimización del estadístico Deviance (19), equivalente a máxima verosimilitud, en el modelo log-bilineal Poisson, mediante el valor del estadístico chi-cuadrado Deviance (19) y el estadístico chi-cuadrado Pearson,

definido así

$$\chi^2_P = \sum_{x,\tau} \frac{(d_{x,\tau} - \widehat{d}_{x,\tau})^2}{\widehat{d}_{x,\tau}}. \tag{21}$$

Se obtienen los resultados de la tabla 4 siguiente.

Tabla 1: *Comparación de Deviance y Pearson*

	SVD	Max.Ver
Deviance	2115.68	1719.19
Pearson	2219.40	1666.68

Como el ajuste mediante máxima verosimilitud con el modelo log-bilineal Poisson resulta superior, es este el elegido para realizar pronósticos de mortalidad.

Resultados de la estimación del modelo Lee-Carter

Utilizando la función `gnm()` se procede de acuerdo con lo indicado en Currie (2014), en donde se advierte que esta función no aplica las restricciones para que el modelo sea identificable dadas en (6), por lo que es necesario aplicarlas al resultado inicial de la estimación.

```
#------estimacion y ajuste de parametros cf. Currie(2014)
m.gnm.un = gnm(dx.v ~ -1 + edad.f + Mult(edad.f, semestres.f)+
offset(log(Ex.v)),
family = poisson(link="log"))
coeff = m.gnm.un$coefficients
#-------------aplicar restricciones cf. Currie(2014)
alfa = coeff[1:na]
beta = coeff[(na+1):(2*na)]
k = coeff[(2*na+1):(2*na+ny)]
m.kR = sum(k)/ny
m.betaR = sum(beta)/na
alfa.est = alfa +m.kR*beta
k.est = na*m.betaR*(k-m.kR)
beta.est = beta/(na*m.betaR)
```

Los resultados de la estimación son los vectores $\widehat{\underline{\kappa}}$, $\widehat{\underline{\alpha}}$,$\widehat{\underline{\beta}}$, que aparecen graficados en la figura 2. La gráfica de $\widehat{\underline{\kappa}}$ (panel (c)) muestra una tendencia aleatoria. El nivel de $\widehat{\underline{\kappa}}$ parece estar revertiendo a mediados de 2006 al nivel de finales de la década de 1990. Un modelo adecuado es definitivo. En este caso se encontró un modelo ARIMA(1,1,0); sin embargo, también es posible utilizar una regresión local Loess. El patrón descrito en α_x (panel (a)) muestra valores mínimos en los semestres 6,7,8 para luego aumentar de manera persistende hasta llegar a un punto de equilibrio. Este patrón determina los patrones de la tasa central de mortalidad (de retiro) m_x; dos casos de m_x se muestran en el panel (d) de la figura 2: la tasa m_x para

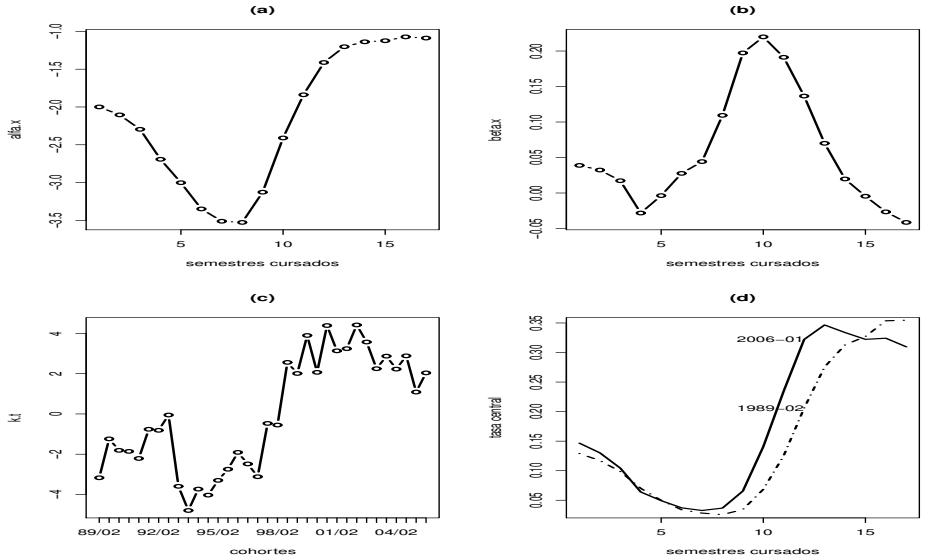

Figura 2: *Modelo Lee-Carter. Parámetros estimados por Max.Ver. Fuente: Elaboración propia.*

la cohorte inicial 1989-02, y la final 2006-01. Se puede observar al inicio de ambas curvas que la tasa es mayor en el caso 2006-01, con relación a 1989-01. Como se anotó anteriormente, el modelo utilizado en este estudio no da información acerca del riesgo de deserción; solamente conjeturamos que tal aumento podría ser debido a que la deserción en los 3 primeros semestres ha aumentado y también, que los egresos en estos semestres son casi nulos. En los semestres 4,5,6 las tasas en ambas curvas son similares, pero en los semestres 7,8,9,10 la tasa es mayor para 2006-01. De nuevo, como no hay información sobre el riesgo de egreso, nos limitamos a conjeturar que la tasa de egreso en 2006 podría ser mayor que en 1989, y también, que las tasas de deserción serían bajas en tales semestres. Una conclusión definitiva, reiteramos, solamente es posible con la medición concreta de las tasas de deserción y de egreso. Nótese que en semestres posteriores al 10 la tendencia en la fuerza de retiros se intercambia.

En la figura 3 siguiente se muestran los residuos deviación

$$D_{x,\tau} = 2\,\omega_{x,\tau}\left(d_{x,\tau}\log\left(\frac{d_{x,\tau}}{\widehat{d}_{x,\tau}}\right) - (d_{x,\tau} - \widehat{d}_{x,\tau})\right) \tag{22}$$

En el panel (a) se grafican diagramas box-plot para $D_{x,\tau}$ para cada $x = 1, \ldots, n_a = 17$, y en el panel (b), para cada cohorte, $\tau = 89/02, \ldots, 06/01$. En el panel (a) parece haber varianza constante a pesar de que los primeros semestres 1,2,3 presentan mayor dispersión, posiblemente debida a que el retiro es más pronunciado en estos. En el panel (b) se observa una posible heterocedasticidad, debido a que

las cohorte del año 2000 en adelante muestran mayor dispersión. Los paneles (c) y (d) corresponden a los residuales de las poblaciones que egresaron a los 10 semestres y las que egresaron con permanencia mayor a 10 semestres, ambas a través de las 34 cohortes. Se observa que a partir del 2000 se ha generado una mayor variabilidad en ambos casos.

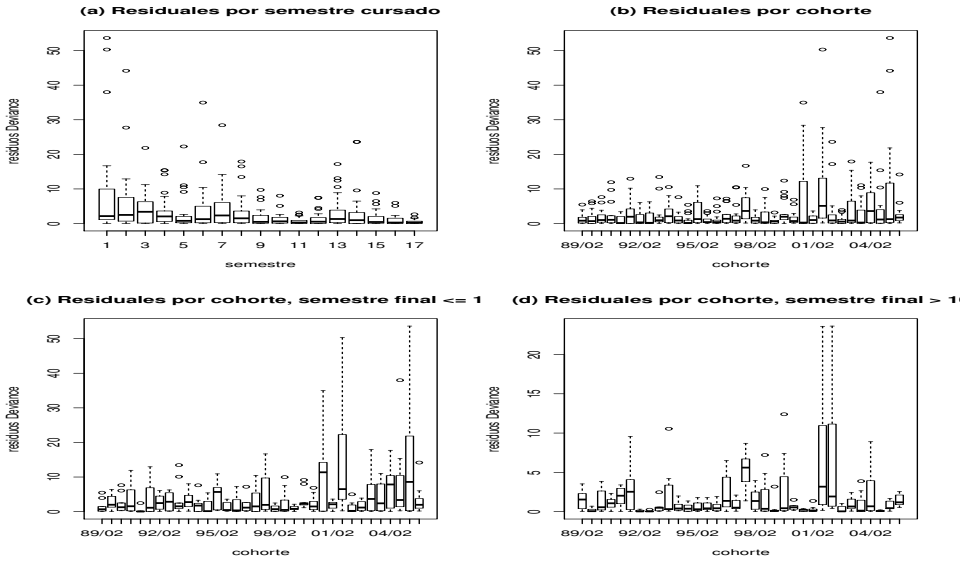

Figura 3: *Residuos (Deviance) del modelo Lee-Carter. Fuente: Elaboración propia.*

Pronóstico de mortalidad

El modelo Lee-Carter (5) es un modelo para pronosticar la tasa central de mortalidad $m_{x,\tau}$, con base en un pronóstico de la serie de tiempo $\widehat{\underline{\kappa}}$. Se ajustaron dos modelos de series de tiempo: un modelo paramétrico tipo ARIMA y un modelo semi-paramétrico de regresión local Loess.

El primero se eligió con base en la función `auto.arima()` de la librería `forecast` de Hyndman (2014). El resultado fué un modelo ARIMA(1,1,0) con parámetro autorregresivo $\widehat{\varphi} = -0.3419$ y error estándar 0.1666.

Para la regresión local Loess se utilizó la función `loess()` con un ancho de ventana `span = 0.42` (no se utilizó el valor por defecto de 0.75).

Los resultados de un pronóstico con ambos modelos, a 10 semestres, que corresponden a los semestres entre 2006/02 y 2011/01, se pueden observar en la figura 5 siguiente. Los pronósticos con ARIMA(1,1,0) son constantes, mientras que con Loess son decrecientes.

Pronóstico de poblaciones por cohortes

Pero está el problema de decidir cuál es el pronóstico adecuado. Una respuesta

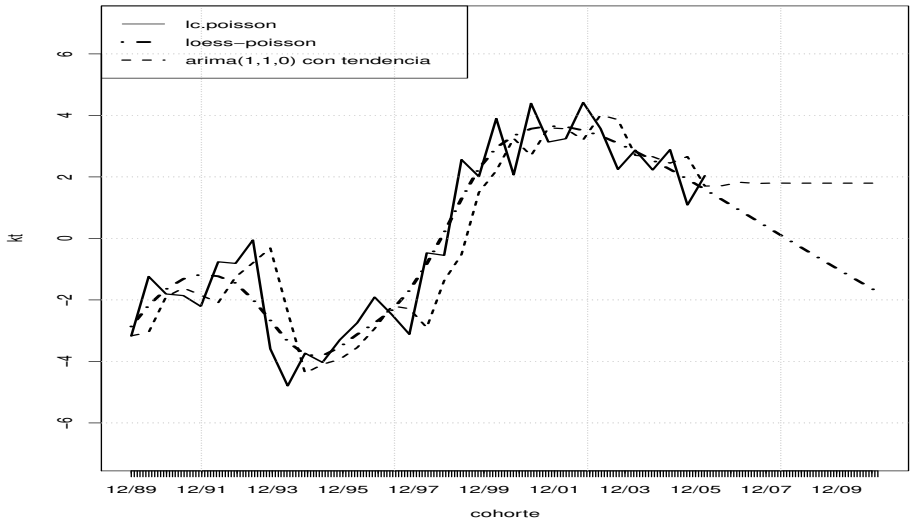

Figura 4: *Ajuste de κ_τ con ARIMA y Loess, y pronóstico a 10 semestres. Fuente: Elaboración propia.*

puede obtenerse utilizando valor final de $\widehat{\kappa}_{34+10}$, que corresponde a la cohorte $\tau_f = 2011 - 01$. Con este valor se reconstruye la tasa central de mortalidad m_{x,τ_f} y los valores $l_{x+1,\tau_f} = l_{x,\tau_f} e^{-\mu_{x,\tau_f}}$, para los semestres $x = 1, \ldots, 17$ de la cohorte τ_f; pero con estos valores son pronósticos la población futura de esta cohorte. Concretamente, para 2014-02 se conocían los primeros 8 valores del total de 17 semestres; son: `1465,1124,1107,1007,961,908,878,851`. A partir de $l_{1,\tau_f} = 1465$ y la estimación de μ_{x,τ_f} por Loess y ARIMA(1,1,0), se estiman l_{x,τ_f}. Los valores obtenidos están en la tabla 2 siguiente, y en la figura 5. En ambas se observa que, excepto los 4 primeros semestres, los pronósticos con Loess sobreestima, mientras que con ARIMA(1,1,0) subestima los valores observados en 2014 para los primeros 8 semestres de la cohorte 2011-01. Sin embargo, consideramos que los valores son aceptables e incluso podrían combinarse (promediarse) para obtener pronósticos más precisos.

7. Conclusiones

1. Los resultados sobre la evolución de la tasa central de mortalidad (de retiro) se pueden obtener a partir de la gráfica del parámetro κ_τ en la figura 2.c. Se evidencia una disminución inicial de κ_τ en la década 1990-1999, para luego aumentar y estabilizarse en la primera mitad de la década 2000-2009. Por lo que las respectivas tasas centrales de mortalidad (de retiro) aumentaron en esa primera mitad. Un ejemplo de la evolución de estas tasas está en la figura

Tabla 2: *Cohorte 2011-01, pronosticada con Loess y Arima*

	Observada	Loess	ARIMA
1	1465	1465	1465
2	1124	1291	1267
3	1107	1151	1113
4	1007	1043	1004
5	961	972	941
6	908	924	896
7	878	894	863
8	851	870	836
9		849	806
10		823	758
11		774	663
12		691	530
13		570	388
14		437	276
15		321	198
16		231	143
17		162	103

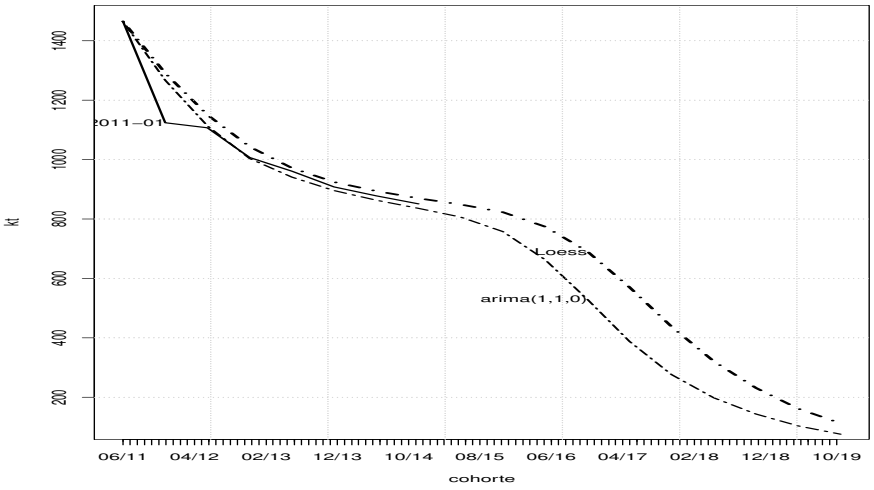

Figura 5: *Pronósticos en el modelo Lee-Carter. Fuente: Elaboración propia.*

2.d, en donde se compararon los casos de la cohorte 1989-02 y 2006-01. Una conclusión es que la tasa central de mortalidad (de retiro) ha aumentado a través del periodo analizado, con excepción de la década 1990-1999, para estabilizarse a mediados de la década 2000-2009.

2. Los resultados sobre el pronóstico de κ_τ muestran que las metodologías - ARIMA y Loess pueden no ser equivalentes. Con base en el caso analizado, en el pronóstico ARIMA la tasa central parece estabilizarse mientras que con Loess decrece, ver figura 4. Estos pronósticos permiten calcular pronósticos de los totales de estudiantes en cada cohorte. En el caso Loess, al decrecer la tasa central aumenta el número de estudiantes, contrario a lo obtenido con ARIMA. Una posible solución podría ser una combinación de pronósticos, por ejemplo, el promedio.

3. Como se mencionó en la Introducción, la utilidad de esta metodología puede verse desde la planeación académica. Concretamente, con la información hasta un semestre determinado, se pueden realizar los pronósticos con un horizonte de varios semestres, para las cohortes aún presentes en tal semestre, y la suma aritmética de estos es un pronóstico del total de la población estudiantil. Estos pronósticos podría ser de utilidad para calcular la demanda de diferentes servicios a cargo de la universidad y los costos asociados, entre otras posibilidades.

4. La metodología original de Lee-Carter con base en la descomposición en valores singulares no parece fácil de implementar en modelos alternos de Lee-Carter como los propuestos en Cairns et al. (2009); por ello, la implementación en R de la estimación con base en modelos no lineales generalizados resulta de suma utilidad y permite la extensión a otros modelos.

Agradecimientos

Agradecimientos a la Oficina de Planeación Académica de la Universidad Nacional de Colombia, sede Medellín, por proporcionar los datos de los censos de la población universitaria de la Sede, y el permiso para utilizarlos. Igualmente, agradecemos a uno de los árbitros por señalar algunas inconsistencias en el planteamiento inicial del estudio.

Recibido: 14 de abril del 2015
Aceptado: 12 de junio del 2015

Referencias

Booth, H., Hyndman, R. & Tickle, L. (2013), 'Prospective life tables', `url:robjhyndman.com/papers/prospect.pdf` .

Bowers, N., Gerber, H., Hickman, J., Jones, D. & Nesbitt, C. J. (1986), *Actuarial Mathematics*, The Society of Actuaries, Ithasca, Il.

Brouhns, N. & Denuit, M. (2002), 'Risque de longèevitè et rentes viagères II. Tables de mortalitè prospectives pour la population belge', *Belgian Actuarial Bulletin* **2**(1), 50–63.

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

Butt, Z. & Haberman, S. (2009), 'Ilc: A Collection of R Functions for Fitting a Class of Lee-Carter Mortality Models using Iterative Fitting Algorithms', *Actuarial Research Paper No. 190,Faculty of Actuarial Science and Insurance, Cass Business School* pp. 1–37.

Cairns, A. J. G. (2007), 'LifeMetrics: A toolkit for measuring and managing longevity and mortality risk', `url:www.lifemetrics.com` .

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A. & Balevich, I. (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.

Cleveland, W. S., Grosse, E. & Shyu, W. M. (1992), *Local regression models*, In: J.M. Chambers and T.J. Hastie, Statistical Models in S, Wadsworth and Brooks/Cole, Chapter 8.

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

Currie, I. (2014), 'On fitting generalized linear and nonlinear models of mortality', *Scandinavian Actuarial Journal* **14**, 1–28.

Davidian, M. (2009), *Nonlinear models for univariate and multivariate response*, Lecture Notes. Department of Statistics., North Caroline State University.

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

Eckart, C. & Young, G. (1936), 'The approximation of one matrix by another of lower rank', *Psychometrika* **1**, 211–218.

Huertas, J. A. (2001), *CÃ¡lculo Actuarial: contingencias de vida individual*, Editorial Unibiblos, Universidad Nacional de Colombia, BogotÃ¡, D.C.

Hyndman, R. (2014), *forecast: Forecasting functions for time series and linear models*. R package version 5.5.
*http://CRAN.R-project.org/package=forecast

Hyndman, R., Booth, H., Tickle, L. & Maindonald, J. (2014), *demography: Forecasting mortality, fertility, migration and population data*. R package version 1.17.
*http://CRAN.R-project.org/package=demography

Koissi, M.-C. & Shapiro, A. (2008), 'The Lee-Carter model under the condition of variables age-specific parameters', *43rd Actuarial Research Conference, Regina, Canada* .

Lee, R. & Carter, L. R. (1992), 'Modelling and Forecasting U.S.Mortality', *Journal of the American Statistical Association* **87**, 659–671.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Renshaw, A. E. & Haberman, S. (2009), 'A cohort-based extension to the Lee-Carter model for mortality reduction factors', *Insurance Mathematics and Economics* .

Shores, T. (2004), *Applied Linear Algebra and Matrix Analysis*, Springer Verlag, Heildelberg.

Steinsaltz, D. (2010), *Statistical Life-Time Models. Lecture Notes*, Department of Statistics, Oxford University.
*http://www.steinsaltz.me.uk/BS3b/BS3b.html

Turner, H. & Firth, D. (2007), *Generalized nonlinear models in R: An overview of the gnm package*, ESRC National Centre for Research Methods, NCRM Working Paper Series 6/07.
*http://www.ncrm.ac.uk

Wilmoth, J. (1993), 'Computational methods for fitting and extrapolating the Lee-Carter model of mortality change', *Technical Report. Department of Demography, University of California, Berkeley* .