REVISTA DE MATEMÁTICA: TEORÍA Y APLICACIONES 2018 **25**(2) : 185–214
CIMPA – UCR ISSN: 1409-2433 (PRINT), 2215-3373 (ONLINE)
DOI: https://doi.org/10.15517/rmta.v25i2.33608

MODELACIÓN DE POBLACIONES VÍA CADENAS DE MARKOV TRIDIMENSIONALES

DEMOGRAPHIC MODELING VIA 3-DIMENSIONAL MARKOV CHAINS

JUAN JOSÉ VÍQUEZ* JORGE AURELIO VÍQUEZ†
ALEXANDER CAMPOS‡ JORGE LORÍA§
LUIS ALFREDO MENDOZA¶

Received: 25/May/2017; Revised: 5/Mar/2018;
Accepted: 26/Apr/2018

*CIMPA & Escuela de Matemática, Universidad de Costa Rica, San José, Costa Rica. E-Mail: viquezejin@gmail.com
†Escuela de Matemática, Universidad de Costa Rica, San José, Costa Rica. E-Mail: javiquez42@gmail.com
‡Misma dirección que/*Same address as:* J.A. Víquez.
E-Mail: alexander.camp353@gmail.com
§Misma dirección que/*Same address as:* J.A. Víquez. E-Mail: jelorias95@gmail.com
¶Misma dirección que/*Same address as:* J.A. Víquez. E-Mail: luis.mf08@gmail.com

Resumen

En este artículo se presenta un nuevo modelo de generación poblacional que puede ser utilizado para proyectar cantidades de personas en fondos de pensiones (tanto cotizantes como jubilados) así como trabajadores en instituciones públicas. Aunado a esto, el modelo presenta oportunidades para cuantificar los flujos derivados de estas poblaciones futuras, tales como gastos en salarios, cotizaciones, pluses salariales, aportes patronales a ahorros/pensiones, entre otros. Claramente la implementación de este modelo probabilístico será de gran utilidad dentro de la caja de herramientas actuariales, aumentando la confiabilidad de las proyecciones, así como permitiendo análisis más profundos por cuanto el desglose poblacional y financiero del modelo es extenso. Aquí se construye un modelo matemático-probabilístico que permite capturar las singularidades de las transiciones entre estados, con suficiente flexibilidad como para aplicarse a varios escenarios. Se estiman exitosamente sus primeros momentos, así como el ajuste de las probabilidades que lo alimenta. Para verificar la idoneidad del modelo propuesto, se implementa con datos reales de una institución pública, y se calcula el error de estimación, presentando niveles inferiores al 2%.

Palabras clave: cadenas de Markov; generación demográfica; matemática financiera.

Abstract

This article presents a new model for demographic simulation which can be used to forecast and estimate the number of people in pension funds (contributors and retirees) as well as workers in a public institution. Furthermore, the model introduces opportunities to quantify the financial flows coming from future populations such as salaries, contributions, salary supplements, employer contribution to savings/pensions, among others. The implementation of this probabilistic model will be of great value in the actuarial toolbox, increasing the reliability of the estimations as well as allowing deeper demographic and financial analysis given the reach of the model. We build a mathematical and probabilistic model that allows us to capture the singularities of the transitions between states with enough flexibility that it can be applied to several scenarios. We successfully estimate its first moments, and show how to adjust the required probabilities. In order to verify the exactness of the proposed model we applied it to real data from a public institution, showing that the estimation error is below the 2%.

Keywords: Markov chains; demographic simulation; financial engineering.

Mathematics Subject Classification: 31C25, 60J60, 47D07, 58J65.

1 Introducción

A menudo en el mundo actuarial se presenta la imperiosa necesidad de contar con proyecciones poblacionales. Para regímenes de pensiones del tipo "Reparto" ("Pay-As-You-Go")[1], es necesario pronosticar el número de cotizantes que alimenten los ingresos del fondo, así como la cantidad de pensionados (quienes representan los gastos). Del mismo modo, en una institución pública, es de gran relevancia conocer la dinámica de la movilidad laboral, analizando el crecimiento/decrecimiento de trabajadores en ciertos puestos, salidas por jubilaciones y relevos generacionales, al tiempo que se cuantifican los gastos derivados de estos trabajadores y sus pluses salariales.

En la literatura existen varios modelos que permiten generar una población a lo largo del tiempo. El diagrama de Lexis es un ejemplo de este tipo de modelación[2], donde se produce la población esperada, segregada por edades, para cada año futuro. En forma análoga, las cadenas de Markov han sido utilizadas para calcular costos esperados de un determinado inventario.[3] Sin embargo, en las ciencias actuariales es necesario intermezclar las generaciones demográficas con los costos derivados de dicha población. Una aplicación en actuariado se puede leer en [7], donde utilizan las cadenas de Markov para realizar un estudio actuarial dentro de un fondo de pensiones costarricense.

La debilidad de estas aplicaciones, anteriormente mencionadas, radica en su naturaleza bidimensional, pudiendo generar solamente una característica dentro de su modelación. En los fondos de pensiones reales, se requiere incorporar más características que afectan directamente los costos y los ingresos. Por ejemplo, la pensión que se va a pagar a un afiliado depende del número de años que cotizó (antigüedad), por lo que no es suficiente el conocer que existen N personas con x edad, sino que se necesita saber cuántas cotizaciones posee, e incluso incluir el monto del salario con el que cotizó. Por otra parte, en una empresa pública se necesita conocer el número de empleados, pero además, cada empleado cuenta con pluses salariales que deben ser incorporados dentro de la generación poblacional.

En este sentido, existe la natural necesidad de trabajar con un modelo que sea multidimensional, y que permita incorporar todas estas características dentro de la población generada. El modelo aquí desarrollado logra, exitosamente, mode-

[1]Es un sistema de financiamiento de las pensiones en donde todo el dinero que se recauda sirve para el pago de las pensiones que se deben de cubrir en ese año, si queda un remanente se destina a la creación de una reserva de contingencias. Tomado de `https://www.supen.fi.cr/glosario`

[2]Se puede leer el capítulo 19 de [1] para ahondar en el tema.

[3]En el capítulo 2 de [5] se puede ver un ejemplo de cadenas de Markov aplicado a costos por inventarios.

lar una población con varias características, respetando sus transiciones entre los distintos grupos. El modelo se sustenta en considerar una tripleta markoviana, es decir, un vector en el espacio de estados $\mathcal{E} \subset \mathbb{R}^3$, tal que para todo conjunto $\{w_i \,/\, i = 1, \ldots, n\} \subset \mathcal{E}$ se debe cumplir

$$\mathbb{P}\left[\boldsymbol{W}_n = w_n \mid \boldsymbol{W}_1 = w_1, \ldots, \boldsymbol{W}_{n-1} = w_{n-1}\right] = \mathbb{P}\left[\boldsymbol{W}_n = w_n \mid \boldsymbol{W}_{n-1} = w_{n-1}\right].$$

Se descompone el espacio $\mathcal{E}$ en tres componentes importantes para la modelación: un espacio de "Categoría", uno de "Edad" y otro de "Antigüedad". La idea detrás del mismo consiste en considerar que esta tripleta determina el comportamiento del individuo, y los estados a los que "salta".

Por ejemplo, considere la situación de modelar el comportamiento de una universidad, específicamente para la categoría académica de catedrático. Personas en esta categoría pueden tener distintas edades, donde claramente un catedrático con 35 años va a acceder a pluses salariales distintos que a los que accede un catedrático de 50 años, el cual puede ser incluso Rector de la universidad. Aún más, suponiendo que estamos observando a un catedrático de 35 años, existen distintas formas que esto suceda. Por ejemplo, el trabajador podría haber logrado el estatus de catedrático en otra entidad y llevar solo un año trabajando en esta universidad, o podría llevar 15 años "asociado" a esta institución. Se esperaría que el primero presente un comportamiento muy distinto al segundo, el cual ha construido una carrera profesional dentro de la institución y le sería más difícil salirse. Para capturar estas especificaciones, se genera dentro de cada tripleta un distribución de características, los que permite conocer el número de personas que poseen determinada $N-$tupla de características, con lo que se puede determinar sus salarios, ya sea para determinar su cotización, su pensión, o el costo de este trabajador dentro de la institución.

El artículo está dividido de la siguiente manera: en la segunda sección se presenta la definición probabilística del modelo, sus estados y sus propiedades, presentando resultados sobre poblaciones esperadas y cómo calcularlas. Para la tercera sección se tratará el tema del ajuste estadístico del modelo, apalancándose en datos mensuales para computar los estimadores de las probabilidades de transición, de ingreso al sistema, y de la distribución inicial. La cuarta sección se concentra en presentar un algoritmo eficiente de generación poblacional, utilizando el hecho de que se puede considerar cada grupo poblacional como una realización de una multinomial con parámetros determinados por el modelo. Finalmente, en la sexta sección se presentará la implementación del modelo con datos reales, y se mostrará el nivel de ajuste al contrastarlo con datos observados, utilizando el conocido método del "backtesting".[4]

[4]Es una prueba de análisis inverso para comprobar la capacidad de predicción y la consistencia estadística del modelo en estudio.

2 Modelo probabilístico

En las siguientes secciones se definirán los estados que componen al espacio $\mathcal{E}$, se efectuarán los cálculos sobre la cadena una vez implementados ciertos supuestos, y se concluirá con el ajuste estadístico de los parámetros del modelo. Además, se crearán particiones en $N-$tuplas de cada tripleta, agregando características poblacionales importantes dentro de la modelación.

2.1 Definición de los estados

Se considera un espacio de estados $\mathcal{E} = C \times E \times A$, compuesto por un espacio de "Categoría" C, uno de "Edad" E y otro de "Antigüedad" A, con estados determinados por la forma $(c, e, a) \in C \times E \times A$, siendo c la categoría a la que pertenece, e la edad que posee, y a la antigüedad que tiene asignada.

- **Categoría:** Asumimos que existen $N_C + 1$ categorías, es decir, $C = \left\{C^{(0)}, C^{(1)}, \ldots, C^{(N_C)}\right\}$. Aquí cada $C^{(i)}$ representa algún tipo de indicador sobre el estatus del individuo, tales como categoría salarial en caso de trabajador público, o tipo de sector y género (hombre/mujer - independiente/público/privado). Por otra parte, $C^{(0)}$ es la categoría que representa estar "fuera" del sistema, es decir, son las personas que no pertenecen a la organización, y que no le están generando ningún tipo de gastos o ingresos (directamente), pero que con probabilidad positiva pueden llegar a hacerlo en los siguientes años. No se asume ninguna estructura sobre estos, únicamente se busca la probabilidad de que entren a cada categoría. Por ejemplo, en una institución pública se consideraría como "no estar contratado" por dicha entidad, o como personas no cotizantes si fuera un fondo de pensiones.

- **Edad:** Se toman edades enteras (aunque se pueden desagregar más) como un rango $E = [E_l, E_u) \cap \mathbb{Z}$, donde E_l es la edad más pequeña (posiblemente negativa) y E_u la edad máxima. Se consideran $N_E + 1$ grupos de edad, es decir, $E = \bigcup_{i=0}^{N_E} E^{(i)}$. La idea es buscar grupos de edad que presenten similares comportamientos de transición entre categorías. El grupo de edad $E^{(0)}$ representa las edades "de reserva", es decir, aquellas personas que con el paso de los años vendrán a alimentar el modelo. Por ejemplo, si estamos en una institución pública, la cual contrata solamente a personas mayores de 18 años, y si se va a proyectar la población por 25 años, entonces este grupo de reserva sería $E^{(0)} = [-7, 18) \cap \mathbb{Z}$. Nótese que las edades pueden ser negativas, esto debido a que hay que

tomar en cuenta a aquellas personas que aún no han nacido pero que eventualmente podrían ingresar al sistema. De ahí el nombre de "población de reserva".

- **Antigüedad:** Del mismo modo, se consideran solamente antigüedades enteras $A = [0, A_u) \cap \mathbb{Z}$, las cuales representan los años de "ligamen" del individuo con el "sistema".[5] Tomamos N_A grupos de antigüedades, con $A = \bigcup_{i=1}^{N_A} A^{(i)}$. En un fondo de pensiones, los grupos de antigüedades serían "paquetes" de años cotizados, indicando su proximidad/lejanía con el estado de la categoría "pensión". Por ejemplo, $A^{(1)}$ podría estar compuesto por las personas que tienen entre 1 a 10 años de cotizar. En una institución pública, se podría tomar a las antigüedades como años laborados en dicha institución, indicando la "carrera" profesional que la persona haya construido en dicha entidad.

Sea S_i la $i-$ésima característica, con $S_i := \left\{S_i^{(1)}, S_i^{(2)}, \ldots, S_i^{(N_i)}\right\}$, donde $S_i^{(j)}$ es el $j-$ésimo estado de la $i-$ésima característica. Tomamos el vector $S = (S_1, S_2, \ldots, S_{N_S})$ de dichas características, el cual representa aquellos factores que inciden en los cálculos, financieros o demográficos, de la población a la cual se le está aplicando el modelo. Se denota por $I_n^{C^{(r)},E^{(i)},A^{(k)}}\left(S_1^{j_1}, \ldots, S_{N_S}^{j_{N_S}}\right)$ al número de personas que en el año n poseen una tripleta $(c, e, a) \in \left\{C^{(r)}\right\} \times E^{(i)} \times A^{(k)} \subset \mathcal{E}$, y N_S- tupla de características $\left(S_1^{j_1}, \ldots, S_{N_S}^{j_{N_S}}\right)$.

Ejemplo: Tomemos el caso de un profesor de la Universidad de Costa Rica, para el cual se tienen los rubros salariales usuales, (se incluyen los montos de las garantías sociales). El cálculo del salario de cada uno de los grupos de la población se haría de acuerdo con los valores de $S_i^{j_i}$, correspondientes a los componentes descritos en la Tabla 1.

[5] Se puede desagregar más, pero no se consideran antigüedades así en este artículo debido a que se complica mucho la notación (ya suficientemente compleja).

Tabla 1: Características salariales.

$S_i^{j_1}$		Descripción	Monto
$S_1^{j_1}$	=	Salario Base Docente	644 831
$S_2^{j_2}$	=	Porcentaje Categoría Académica	354 657
$S_3^{j_3}$	=	Anualidad	776 190
$S_4^{j_4}$	=	Escalafón Docente	119 939
$S_5^{j_5}$	=	Fondo Consolidado	18 854
$S_6^{j_6}$	=	Pasos Académicos	59 970
$S_7^{j_7}$	=	Reconocimiento por Elección	279 857
$S_8^{j_8}$	=	Magisterio	176 822
$S_9^{j_9}$	=	Seguro de Enfermedad y Maternidad	225 600
$S_{10}^{j_{10}}$	=	Banco Popular	12 195
$S_{11}^{j_{11}}$	=	Fondo de Capitalización Laboral	73 168
$S_{12}^{j_{12}}$	=	Fondo de Pensión Complementaria	36 584
$S_{13}^{j_{13}}$	=	Aguinaldo	203 235
$S_{14}^{j_{14}}$	=	Salario Escolar	184 627
$S_{15}^{j_{15}}$	=	JAFAP	60 973
		TOTAL GASTADO	3 227 500

Como se observa de los datos, el costo total para la Universidad por este profesor es de $3.227.500$ colones. Asuma ahora que para $n = 2$,

$$I_2^{C^{(r)},E^{(i)},A^{(k)}}\left(S_1^{j_1},\ldots,S_{15}^{j_{15}}\right) = 30,$$

es decir, que dentro de 2 años habrían 30 individuos en la categoría $C^{(r)}$, con rango de edad $E^{(i)}$, rango de antigüedad $A^{(k)}$, y con las características salariales de la Tabla 1. Entonces, el gasto total para la Universidad por este grupo de personas será de $96.824.998$ colones. Repitiendo este proceso con todas las categorías, rangos de edad y de antigüedad, y todas las características salariales, se logra obtener el monto total gastado en todos los empleados de la Universidad de Costa Rica.[6]

[6]El salario base se debería incrementar (semestralmente) de acuerdo con la inflación estimada, para cada año de proyección.

2.2 Cadena de Markov

Considere, (X_n, Y_n, Z_n) la tripleta aleatoria de la cadena, donde X_n, Y_n y Z_n representan el estado "Categoría", "Edad" y "Antigüedad", respectivamente, en el $n-$ésimo año. Se asume que se cumple la propiedad de Markov, es decir, para $(c_k, e_k, a_k) \in C \times E \times A$, $k = 0, \ldots, n$,

$$\begin{aligned}\mathbb{P}\big[(X_n, Y_n, Z_n) = (c_n, e_n, a_n) \,\big|\, &(X_{n-1}, Y_{n-1}, Z_{n-1}) \\ &= (c_{n-1}, e_{n-1}, a_{n-1}), ..., (X_0, Y_0, Z_0) = (c_0, e_0, a_0)\big] = \\ \mathbb{P}\big[(X_n, Y_n, Z_n) = (c_n, e_n, a_n) \,\big|\, &(X_{n-1}, Y_{n-1}, Z_{n-1}) = (c_{n-1}, e_{n-1}, a_{n-1})\big].\end{aligned}$$

Como se asume que la cadena es homogénea, se tiene que

$$\begin{aligned}&\mathbb{P}\big[(X_n, Y_n, Z_n) = (c_n, e_n, a_n) \,\big|\, (X_{n-1}, Y_{n-1}, Z_{n-1}) = (c_{n-1}, e_{n-1}, a_{n-1})\big] = \\ &\mathbb{P}\big[(X_1, Y_1, Z_1) = (c_1, e_1, a_1) \,\big|\, (X_0, Y_0, Z_0) = (c_0, e_0, a_0)\big],\end{aligned}$$

para todo n. Se nota además que

$$\begin{aligned}\mathbb{P}\big[(X_1, Y_1, Z_1) = (c_1, e_1, a_1) \,\big|\, &(X_0, Y_0, Z_0) = (c_0, e_0, a_0)\big] = \\ &\mathbb{P}\big[X_1 = c_1 \mid Y_1 = e_1, Z_1 = a_1, X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big] \\ &\times\mathbb{P}\big[Z_1 = a_1 \mid Y_1 = e_1, X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big] \\ &\times\mathbb{P}\big[Y_1 = e_1 \mid X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big].\end{aligned}$$

Observaciones, hipótesis y cálculos:

- Asumimos que el incremento de la antigüedad es homogénea (idénticamente distribuida); eso es, existe una variable aleatoria ξ con valores en $\{0, 1\}$ (que simboliza el haber "trabajado/cotizado" durante ese año o no haber estado en el sistema en dicho año), tal que $Z_1 - Z_0 \overset{d}{=} \xi$, para todo n.

  $$\begin{aligned}&\mathbb{P}\big[X_1 = c_1 \mid Y_1 = e_1, Z_1 = a_1, X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big] = \\ &\mathbb{P}\big[X_1 = c_1 \mid Y_1 = e_1, \xi = a_1 - a_0, X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big]\end{aligned}$$

  y

  $$\begin{aligned}&\mathbb{P}\big[Z_1 = a_1 \mid Y_1 = e_1, X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big] = \\ &\mathbb{P}\big[\xi = a_1 - a_0 \mid Y_1 = e_1, X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big].\end{aligned}$$

- Se asume que los aumentos de edad suceden en enero de cada año. Nótese que si $e_1 \neq e_0 + 1$, entonces $\{Y_1 = e_1\} \cap \{Y_0 = e_0\} = \emptyset$. Como

$\mathbb{P}\big[Y_1 = e_1 \mid X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big] = 0$, en este caso no importaría si no se condiciona por $\{Y_1 = e_1\}$. Igualmente, $\{Y_1 = e_1\} \cap \{Y_0 = e_0\} = \{Y_0 = e_0\}$ si $e_1 = e_0 + 1$. Y en el segundo caso, condicionar por $\{Y_1 = e_1\} \cap \{Y_0 = e_0\}$ es lo mismo que condicionar solamente por $\{Y_0 = e_0\}$. En conclusión, se tiene que

$$\mathbb{P}\big[X_1 = c_1 \mid Y_1 = e_1, \xi = a_1 - a_0, X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big] = \mathbb{P}\big[X_1 = c_1 \mid \xi = a_1 - a_0, X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big],$$

y

$$\mathbb{P}\big[\xi = a_1 - a_0 \mid Y_1 = e_1, X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big] = \mathbb{P}\big[\xi = a_1 - a_0 \mid X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big],$$

siempre y cuando se multiplique el factor $\mathbb{P}\big[Y_1 = e_1 \mid X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big]$. Aún más,

$$\mathbb{P}\big[Y_1 = e_1 \mid X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big] = \begin{cases} 1 & \text{si } e_1 = e_0 + 1 \\ 0 & \text{si } e_1 \neq e_0 + 1 \end{cases}.$$

- Note que la probabilidad de estar en la categoría $c_1 \neq C^{(0)}$ dado que no aumentó su antigüedad ($\xi = 0$) es 0 (no puede cambiar de categoría dentro del sistema si no está en el sistema). Del mismo modo, la probabilidad de irse a la categoría $C^{(0)}$ si no permanece en el sistema es 1. Inversamente, si entra al sistema ($\xi = 1$), entonces la probabilidad de pasar por la categoría $C^{(0)}$ sería 0, y solo podría pasar por a las otras categorías $\{C^{(1)}, \ldots, C^{(N_C)}\}$.[7] De este modo tenemos

$$\mathbb{P}\big[X_1 = c_1 \mid \xi = a_1 - a_0, X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big] = \begin{cases} \mathbb{P}\big[X_1 = c_1 \mid X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big] & \text{si } \xi = 1 \text{ y } c_1 \neq C^{(0)} \\ 0 & \text{si } \xi = 1 \text{ y } c_1 = C^{(0)} \\ 1 & \text{si } \xi = 0 \text{ y } c_1 = C^{(0)} \\ 0 & \text{si } \xi = 0 \text{ y } c_1 \neq C^{(0)}. \end{cases}$$

En esencia, una persona puede acceder a la categoría $C^{(0)}$ solamente si no es "contratado/cotizante" ($\xi = 0$), y puede emigrar de ella solo si fue "contratado/cotizante" ($\xi = 1$).

[7] Recuerde si sale del sistema entonces no puede tener ninguna categoría dentro del sistema, y si entra al sistema no podría tener la categoria que significa estar afuera del sistema.

- Se asume que la probabilidad de pasar de la categoría c_0 a la categoría c_1, dado que fue contratado o no (ξ), $a_0 \in A_0 \in \left\{A^{(1)}, A^{(2)}, A^{(3)}, A^{(4)}\right\}$, y $e_0 \in E_0 \in \left\{E^{(0)}, E^{(1)}, E^{(2)}, E^{(3)}, E^{(4)}\right\}$, es la misma. Es decir, que la probabilidad de cambiar de categoría solo se ve afectada cuando se pasa de grupos de edad y antigüedad. De la misma manera, $\mathbb{P}\big[\xi = \cdot \mid X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big]$ no cambia para todo $a_0 \in A_0 \in \left\{A^{(1)}, A^{(2)}, A^{(3)}, A^{(4)}\right\}$, y $e_0 \in E_0 \in \left\{E^{(0)}, E^{(1)}, E^{(2)}, E^{(3)}, E^{(4)}\right\}$. Por el Lema 1, se concluye que

$$\begin{aligned}
&\mathbb{P}\big[X_1 = c_1 \mid \xi = a_1 - a_0, X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big] = \\
&\mathbb{P}\left[X_1 = c_1 \mid \xi = a_1 - a_0, X_0 = c_0, Y_0 \in E_0, Z_0 \in A_0\right] \\
&y \\
&\mathbb{P}\big[\xi = a_1 - a_0 \mid X_0 = c_0, Y_0 = e_0, Z_0 = a_0\big] = \\
&\mathbb{P}\left[\xi = a_1 - a_0 \mid X_0 = c_0, Y_0 \in E_0, Z_0 \in A_0\right].
\end{aligned}$$

Denote

$$P^{E_0,A_0}\big(c_0, c_1\big) := \mathbb{P}\left[X_1 = c_1 \mid X_0 = c_0, Y_0 \in E_0, Z_0 \in A_0\right],$$

la probabilidad de transición entre las categorías $\{C^{(1)}, \ldots, C^{(N_C)}\}$, dado los grupos de edad E_0 y de antigüedad A_0. Igualmente, denote

$$Q^{E_0,A_0,c_0}(\cdot) := \mathbb{P}\left[\xi = \cdot \mid X_0 = c_0, Y_0 \in E_0, Z_0 \in A_0\right],$$

la probabilidad de que una persona en el grupo de edad E_0, con antigüedad en el rango A_0, dentro de la categoría c_0, sea incluido o no al sistema en el año siguiente. Nótese que $Q^{E_0,A_0,c_0}(r) = 0$ para todo $r \notin \{0, 1\}$.

Tomando en cuenta todo lo anterior, se concluye que si $e_0 \in E_0$ y $a_0 \in A_0$, entonces

$$\begin{aligned}
&\mathbb{P}\big[(X_1, Y_1, Z_1) = (c_1, e_1, a_1) \,\big|\, (X_0, Y_0, Z_0) = (c_0, e_0, a_0)\big] \\
&= \begin{cases}
P^{E_0,A_0}\big(c_0, c_1\big) \cdot Q^{E_0,A_0,c_0}(1) \cdot \mathbf{1}_{\{e_1 = e_0 + 1\}} & \text{si } a_1 - a_0 = 1 \text{ y } c_1 \neq C^{(0)} \\
0 & \text{si } a_1 - a_0 = 1 \text{ y } c_1 = C^{(0)} \\
Q^{E_0,A_0,c_0}(0) \cdot \mathbf{1}_{\{e_1 = e_0 + 1\}} & \text{si } a_1 - a_0 = 0 \text{ y } c_1 = C^{(0)} \\
0 & \text{si } a_1 - a_0 = 0 \text{ y } c_1 \neq C^{(0)}.
\end{cases}
\end{aligned} \tag{1}$$

2.2.1 Transiciones mensuales

Aquí se está abusando del lenguaje, pues un año consiste de 12 meses, y en cada mes se podría observar una categoría diferente. En este sentido, es necesario considerar el sentido que tiene la frase "en el $n-$ésimo año la persona tuvo la tripleta

(c_n, e_n, a_n)". Una alternativa es considerar que $(X_n, Y_n, Z_n) = (c_n, e_n, a_n)$ representa que al inicio del año n, la persona se encontraba en esa tripleta, y asumir que no existen cambios de categoría, edad y antigüedad a lo largo del año. Otra opción sería pensar que $(X_n, Y_n, Z_n) = (c_n, e_n, a_n)$ significa que la categoría fue c_n, la edad e_n y la antigüedad a_n, en al menos un mes del $n-$ésimo año.

- Asuma que la transición entre las categorías $\{C^{(1)}, \ldots, C^{(N_C)}\}$ (dado el grupo de edad y de antigüedad) dentro de un mismo año es una cadena de Markov en escala "mensual". Es decir, sea $\widetilde{X}_m$ la variable que representa el estado de categoría en que se encuentra en el $m-$ésimo mes, y sea la probabilidad $\mathbb{P}^{E_0,A_0}[\cdot] = \mathbb{P}[\,\cdot \mid Y_0 \in E_0, Z_0 \in A_0]$, entonces,

$$\mathbb{P}^{E_0,A_0}\left[\widetilde{X}_m = \widetilde{c}_m \;\middle|\; \widetilde{X}_{m-1} = \widetilde{c}_{m-1}, \ldots, \widetilde{X}_0 = \widetilde{c}_0\right] = \mathbb{P}^{E_0,A_0}\left[\widetilde{X}_m = \widetilde{c}_m \;\middle|\; \widetilde{X}_{m-1} = \widetilde{c}_{m-1}\right].$$

- Como los aumentos de edad y antigüedad suceden en enero de cada año, de manera que la probabilidad de cambiar de categoría mensualmente es la misma durante todo el año, i.e., la cadena $\widetilde{X}_m$ es homogénea para $1 \leq m \leq 12$. Denote

$$\widetilde{P}^{E_0,A_0}\big(\widetilde{c}_0, \widetilde{c}_1\big) := \mathbb{P}\left[\widetilde{X}_1 = \widetilde{c}_1 \;\middle|\; \widetilde{X}_0 = \widetilde{c}_0, Y_0 \in E_0, Z_0 \in A_0\right],$$

  la probabilidad de transición **mensual** entre categorías, dado los grupos de edad E_0 y antigüedad A_0.

- Sea T_i el tiempo de parada donde se alcanza por primera vez la categoría $C^{(i)}$ dentro de un año determinado, con función de masa de probabilidad $p_t^{(i)} = \mathbb{P}[T_i = t]$. Como la tripleta (X_n, Y_n, Z_n) representa el estado de categoría, edad y antigüedad en el $n-$ésimo año, y de éstos solo la categoría cambia dentro de un mismo año, definimos la siguiente relación, con $c_1 = C^{(i)}$,

$$\begin{aligned} P^{E_0,A_0}\big(c_0, c_1\big) &:= \mathbb{E}^{T_i}\left[\mathbb{P}^{E_0,A_0}\left[\widetilde{X}_{T_i} = \widetilde{c}_1 \;\middle|\; \widetilde{X}_0 = \widetilde{c}_0, T_i\right]\right] \\ &= \sum_{t=1}^{12} \mathbb{P}^{E_0,A_0}\left[\widetilde{X}_t = \widetilde{c}_1 \;\middle|\; \widetilde{X}_0 = \widetilde{c}_0\right] \cdot p_t^{(i)}. \end{aligned}$$

  Aún más, aplicando la propiedad de Markov junto con el Lema 3 (dado que $\cup_{\widetilde{c}_i \in C}\{\widetilde{X}_i = \widetilde{c}_i\} = \Omega$) iteradamente, con la convención de

que $c_t = c_1$, se tiene

$$\begin{aligned}
&P^{E_0,A_0}\big(c_0, c_1\big) \\
&= \sum_{t=1}^{12} \mathbb{P}^{E_0,A_0}\left[\widetilde{X}_t = \widetilde{c}_1 \;\middle|\; \widetilde{X}_0 = \widetilde{c}_0\right] \cdot p_t^{(i)} \\
&= \sum_{t=1}^{12} \sum_{\widetilde{c}_1,\ldots,\widetilde{c}_{t-1} \in C} \prod_{l=1}^{t} \mathbb{P}^{E_0,A_0}\left[\widetilde{X}_l = \widetilde{c}_l \;\middle|\; \widetilde{X}_{l-1} = \widetilde{c}_{l-1}\right] \cdot p_t^{(i)} \\
&= \sum_{t=1}^{12} \sum_{\widetilde{c}_1,\ldots,\widetilde{c}_{t-1} \in C} \prod_{l=1}^{t} \widetilde{P}^{E_0,A_0}\big(\widetilde{c}_l, \widetilde{c}_{l-1}\big) \cdot p_t^{(i)}. \qquad (2)
\end{aligned}$$

Este término se sustituiría en (1) para calcular las probabilidades de transición.

Este tipo de herramienta se torna útil cuando se poseen pocos datos anuales, y se requieren utilizar datos mensuales para incrementar las observaciones y mejorar la convergencia de los estimadores de los parámetros.

2.3 Distribución de (X_n, Y_n, Z_n)

Se define la distribución inicial π_{c_0,e_0,a_0}, para una tripleta $(c_0, e_0, a_0) \in C \times E \times A$, como la probabilidad de que un individuo tenga categoría c_0, edad e_0 y antigüedad a_0 en el año inicial.

Proposición 1 *La probabilidad de que una persona, después de un año, se encuentre en el estado $(c_1, e_1, a_1) \in C \times E \times A$, es decir, que al año siguiente un individuo tenga e_1 años de edad, con antigüedad a_1 y en la categoría c_1, vendría dado por*

$$\begin{aligned}
&\mathbb{P}\big[(X_1, Y_1, Z_1) = (c_1, e_1, a_1)\big] = \\
&\sum_{c_0 \in C} \left(P^{E_{e_1-1},A_{a_1-\delta_{c_1}}}\big(c_0, c_1\big)\right)^{\delta_{c_1}} \cdot Q^{E_{e_1-1},A_{a_1-\delta_{c_1}},c_0}\big(\delta_{c_1}\big) \cdot \pi_{c_0,e_1-1,a_1-\delta_{c_1}}
\end{aligned}$$

donde $E_e \in \big\{E^{(0)}, E^{(1)}, E^{(2)}, E^{(3)}, E^{(4)}\big\}$ y $A_a \in \big\{A^{(1)}, A^{(2)}, A^{(3)}, A^{(4)}\big\}$ son tales que $e \in E_e$ y $a \in A_a$, y con

$$\delta_{c_1} = \begin{cases} 1 & \textit{si } c_1 \neq C^{(0)} \\ 0 & \textit{si } c_1 = C^{(0)}. \end{cases}$$

Demostración. Como $\bigcup_{c_0\in C,e_0\in E,a_0\in A}\big\{(X_0,Y_0,Z_0)=(c_0,e_0,a_0)\big\}=\Omega$, entonces
$\mathbb{P}\big[(X_1,Y_1,Z_1)=(c_1,e_1,a_1)\big]$

$$
\begin{aligned}
&=\mathbb{P}\left[\big\{(X_1,Y_1,Z_1)=(c_1,e_1,a_1)\big\}\bigcap\bigcup_{c_0\in C,e_0\in E,a_0\in A}\big\{(X_0,Y_0,Z_0)=(c_0,e_0,a_0)\big\}\right]\\
&=\sum_{c_0\in C,e_0\in E,a_0\in A}\mathbb{P}\big[(X_1,Y_1,Z_1)=(c_1,e_1,a_1)\,\big|\,(X_0,Y_0,Z_0)=(c_0,e_0,a_0)\big]\\
&\quad\times\mathbb{P}\big[(X_0,Y_0,Z_0)=(c_0,e_0,a_0)\big]\\
&=\sum_{c_0\in C,e_0\in E,a_0\in A}\mathbb{P}\big[(X_1,Y_1,Z_1)=(c_1,e_1,a_1)\,\big|\,(X_0,Y_0,Z_0)=(c_0,e_0,a_0)\big]\\
&\quad\times\mathbf{1}_{\{e_1=e_0+1\}}\cdot\pi_{c_0,e_0,a_0}\\
&=\sum_{c_0\in C,a_0\in\{a_1-1,a_1\}}\mathbb{P}\big[(X_1,Y_1,Z_1)=(c_1,e_1,a_1)\,\big|\,(X_0,Y_0,Z_0)=(c_0,e_0,a_0)\big]\cdot\pi_{c_0,e_1-1,a_0}\\
&=\begin{cases}\sum_{c_0\in C}Q^{E_{e_1-1},A_{a_1},c_0}(0)\cdot\pi_{c_0,e_1-1,a_1} & \text{si } c_1=C^{(0)}\\[2ex] \sum_{c_0\in C}P^{E_{e_1-1},A_{a_1-1}}\big(c_0,c_1\big)\cdot Q^{E_{e_1-1},A_{a_1-1},c_0}(1)\cdot\pi_{c_0,e_1-1,a_1-1} & \text{si } c_1\neq C^{(0)}\end{cases}
\end{aligned}
$$

■

Teorema 1 *La probabilidad de que una persona se encuentre en el estado $(c_n,e_n,a_n)\in C\times E\times A$ en el año n, es decir, que en n años un individuo tenga e_n años de edad, con antigüedad a_n y en la categoría c_n, vendría dado por*

$$
\begin{aligned}
&\mathbb{P}\big[(X_n,Y_n,Z_n)=(c_n,e_n,a_n)\big]\\
&=\sum_{c_0,c_1,\ldots,c_{n-1}\in C}\prod_{r=1}^{n}\Big(P^{E_{e_n-r},A_{a_n-\sum_{l=1}^{r}\delta_{c_{n+1-l}}}\left(c_{n-r},c_{n+1-r}\right)}\Big)^{\delta_{c_{n+1-r}}}\\
&\qquad\cdot Q^{E_{e_n-r},A_{a_n-\sum_{l=1}^{r}\delta_{c_{n+1-l}}},c_{n-r}}(\delta_{c_{n+1-r}})\times\pi_{c_0,e_n-n,a_n-\sum_{l=0}^{n-1}\delta_{c_{l+1}}}.
\end{aligned}
$$

Demostración. Se probará por inducción.
El caso $n=1$ es la Proposición 1.
Asuma que se cumple para $n-1$, y se probará para n. Repitiendo los pasos de la prueba de la Proposición 1, pero tomando n en lugar de 1 y $n-1$ en lugar de 0, se obtiene

$$
\begin{aligned}
&\mathbb{P}\big[(X_n,Y_n,Z_n)=(c_n,e_n,a_n)\big] \qquad (3)\\
&=\sum_{c_{n-1}\in C}\Big(P^{E_{e_n-1},A_{a_n-\delta_{c_n}}\left(c_{n-1},c_n\right)}\Big)^{\delta_{c_n}}\cdot Q^{E_{e_n-1},A_{a_n-\delta_{c_n}},c_{n-1}}\big(\delta_{c_n}\big)\\
&\qquad\times\mathbb{P}\big[(X_{n-1},Y_{n-1},Z_{n-1})=(c_{n-1},e_n-1,a_n-\delta_{c_n})\big].
\end{aligned}
$$

Por hipótesis de inducción,

$$\begin{aligned}
&\mathbb{P}\big[(X_{n-1},Y_{n-1},Z_{n-1})=(c_{n-1},e_{n-1},a_{n-1})\big] \qquad (4)\\
&=\sum_{c_0,c_1,\ldots,c_{n-2}\in C}\prod_{r=1}^{n-1}\Big(P^{E_{e_{n-1}-r},A_{a_{n-1}-\sum_{l=1}^{r}\delta_{c_{n-l}}}\left(c_{n-1-r},c_{n-r}\right)}\Big)^{\delta_{c_{n-r}}}\\
&\quad\times Q^{E_{e_{n-1}-r},A_{a_{n-1}-\sum_{l=1}^{r}\delta_{c_{n-l}}},c_{n-1-r}}(\delta_{c_{n-r}})\\
&\quad\times\pi_{c_0,e_{n-1}-(n-1),a_{n-1}-\sum_{l=0}^{n-2}\delta_{c_{l+1}}}.
\end{aligned}$$

Se sustituye 4 en 3, tomando en cuenta que $e_{n-1}=e_n-1$ y $a_{n-1}=a_n-\delta_{c_n}$, y pasando el índice a $r-1$, se concluye que

$$\begin{aligned}
&\mathbb{P}\big[(X_n,Y_n,Z_n)=(c_n,e_n,a_n)\big]\\
&=\sum_{c_{n-1}\in C}\Big(P^{E_{e_n-1},A_{a_n-\delta_{c_n}}\left(c_{n-1},c_n\right)}\Big)^{\delta_{c_n}}\cdot Q^{E_{e_n-1},A_{a_n-\delta_{c_n}},c_{n-1}}\big(\delta_{c_n}\big)\\
&\quad\times\sum_{c_0,c_1,\ldots,c_{n-2}\in C}\prod_{r=2}^{n}\Big(P^{E_{e_n-r},A_{a_n-\sum_{l=1}^{r}\delta_{c_{n+1-l}}}\left(c_{n-r},c_{n+1-r}\right)}\Big)^{\delta_{c_{n+1-r}}}\\
&\quad\times Q^{E_{e_n-r},A_{a_n-\sum_{l=1}^{r}\delta_{c_{n+1-l}}},c_{n-r}}(\delta_{c_{n+1-r}})\times\pi_{c_0,e_n-n,a_n-\sum_{l=0}^{n-1}\delta_{c_{l+1}}}\\
&=\sum_{c_0,c_1,\ldots,c_{n-1}\in C}\prod_{r=1}^{n}\Big(P^{E_{e_n-r},A_{a_n-\sum_{l=1}^{r}\delta_{c_{n+1-l}}}\left(c_{n-r},c_{n+1-r}\right)}\Big)^{\delta_{c_{n+1-r}}}\\
&\quad\times Q^{E_{e_n-r},A_{a_n-\sum_{l=1}^{r}\delta_{c_{n+1-l}}},c_{n-r}}(\delta_{c_{n+1-r}})\times\pi_{c_0,e_n-n,a_n-\sum_{l=0}^{n-1}\delta_{c_{l+1}}}.
\end{aligned}$$

■

Corolario 1 *La probabilidad de que una persona, en el año n, se encuentre en la categoría c_n, dentro de los rangos de "edad" y "antigüedad", $E_n\in\{E^{(0)},E^{(1)},E^{(2)},E^{(3)},E^{(4)}\}$ y $A_n\in\{A^{(1)},A^{(2)},A^{(3)},A^{(4)}\}$, respectivamente, sería*

$$\mathbb{P}\left[(X_n,Y_n,Z_n)\in\{c_n\}\times E_n\times A_n\right]=\sum_{\substack{e_n\in E_n\\ a_n\in A_n}}\mathbb{P}\left[(X_n,Y_n,Z_n)=(c_n,e_n,a_n)\right].$$

Corolario 2 *De una población inicial I_0, la cantidad esperada de personas $\mathbb{E}\big[I_n^{c_n,E_n,A_n}\big]$, en la categoría $c_n\in C$, dentro de los rangos de edad $E_n\in\{E^{(0)},E^{(1)},E^{(2)},E^{(3)},E^{(4)}\}$ y de antigüedad $A_n\in\{A^{(1)},A^{(2)},A^{(3)},A^{(4)}\}$, para el $n-$ésimo año sería*

$$\mathbb{E}\big[I_n^{c_n,E_n,A_n}\big]=I_0\cdot\mathbb{P}\big[(X_n,Y_n,Z_n)\in\{c_n\}\times E_n\times A_n\big].$$

2.4 Distribución de las características

Sea $(W_1, W_2, \ldots, W_{N_S})$ el vector aleatorio de características.

Hipótesis: Asumimos el vector aleatorio $(W_1, W_2, \ldots, W_{N_S})$ es estacionario, por lo que podemos considerarlo independiente del tiempo. Aún más, se asume que, para $\{c_0\} \times E_0 \times A_0 \subset C \times E \times A$,

$$\begin{aligned}\mathbb{P}\big[W_i = S_i^{j_i}, i = 1, \ldots, N_S \,\big/\, (X_n, Y_n, Z_n) \in \{c_0\} \times E_0 \times A_0\big] \\ = \mathbb{P}\big[W_i = S_i^{j_i}, i = 1, \ldots, N_S \,\big/\, (X_0, Y_0, Z_0) \in \{c_0\} \times E_0 \times A_0\big].\end{aligned}$$

Defina la probabilidad estacionaria de la distribución de la población respecto a las características, como

$$\begin{aligned}&R^{c_0,E_0,A_0}\big(S_1^{j_1}, \ldots, S_{N_S}^{j_{N_S}}\big) \\ &\quad := \quad \mathbb{P}\big[W_i = S_i^{j_i}, i = 1, \ldots, N_S \,\big/\, (X_0, Y_0, Z_0) \in \{c_0\} \times E_0 \times A_0\big]. \qquad (5)\end{aligned}$$

Corolario 3 *De una población inicial I_0, la cantidad esperada de personas $\mathbb{E}\big[I_n^{c_n,E_n,A_n}\big(S_1^{j_1}, \ldots, S_{N_S}^{j_{N_S}}\big)\big]$, en la categoría $c_n \in C$, dentro de los rangos de edad y antigüedad, $E_n \in \big\{E^{(0)}, E^{(1)}, E^{(2)}, E^{(3)}, E^{(4)}\big\}$ y $A_n \in \big\{A^{(1)}, A^{(2)}, A^{(3)}, A^{(4)}\big\}$, respectivamente, para el $n-$ésimo año, y con las características $S_i^{(j_i)}$, para $i = 1, \ldots, N_S$, sería*

$$\begin{aligned}&\mathbb{E}\big[I_n^{c_n,E_n,A_n}\big(S_1^{j_1}, \ldots, S_{N_S}^{j_{N_S}}\big)\big] = \\ &\quad I_0 \cdot R^{c_n,E_n,A_n}\big(S_1^{j_1}, \ldots, S_{N_S}^{j_{N_S}}\big) \cdot \mathbb{P}\big[(X_n, Y_n, Z_n) \in \{c_n\} \times E_n \times A_n\big].\end{aligned}$$

3 Ajuste del modelo

Se considera una base histórica donde cada individuo es registrado mes a mes, con su categoría, su edad, y su antigüedad; así como las características del mismo. Se trabajará el escenario de ajuste con categorías mensual, puesto que es más probable que falten datos a que sobren, sin embargo, el ajuste con datos anuales se sigue de manera sencilla. Además, se asume que solo se tiene información de los individuos dentro del sistema, mientras que la información de las personas fuera del sistema no se posee.

3.1 Población total/reserva

Se toma una población total de I_0 personas[8], distribuidas de la siguiente manera:

- $N_m^{C^{(r)},e,a}$ es la cantidad de personas que en el mes m se encontraban en la categoría $C^{(r)}$, con e años de edad, y a años de antigüedad, para $r = 1, \ldots, N_C$, y $e \in E$, $a \in A$.

- Se asume que existen N_e personas de edad $e \in [E_l, E_u]$, es decir, hay N_{E_l} personas con edad E_l años, N_{E_l+1} personas con edad $E_l + 1$ años, y así sucesivamente hasta llegar a N_{E_u} personas con edad E_u años. De este modo, el número de personas con edad e años que están en la categoría $C^{(0)}$ durante el mes m serían

$$N_m^{C^{(0)},e} = N_e - \sum_{r=1}^{N_C} \sum_{a \in A} N_m^{C^{(r)},e,a}.$$

- Esta población "reserva" de e años de edad, que se encuentra en la categoría $C^{(0)}$, se asume uniformemente distribuida entre antigüedades, es decir, se divide la población $N_m^{C^{(0)},e}$ en partes iguales, y dicha cantidad correspondería a $N_m^{C^{(0)},e,a}$ para los valores de a posibles dada la edad e. Por el contrario, $N_m^{C^{(0)},e,a} = 0$ para las antigüedades a que son imposibles con la edad e.

3.2 Distribución inicial

Sea $\mathcal{M}$ el conjunto de meses observados, donde $m = 0$ para el último mes observado, $m = -1$ para el penúltimo mes observado, y así sucesivamente, hasta llegar a $m = -M$ que correspondería al mes más antiguo observado. Se emplea la contabilización anterior, tomando $N_0^{C^{(r)},e,a}$ como el número promedio de personas en la categoría $C^{(r)}$, con e años de edad, y a años de antigüedad, definido por

$$N_0^{C^{(r)},e,a} = \frac{1}{12} \sum_{m=-11}^{0} N_m^{C^{(r)},e,a}.$$

[8]Se puede tomar como la Población Económicamente Activa (PEA) para un régimen de pensiones como el de Invalidez, Vejez y Muerte (IVM) de la Caja Costarricense del Seguro Social (CCSS), o como un número definido en el caso de una institución que no posee un número representativo de la PEA. En este caso I_0 puede cambiarse por I_t al hacer esa modificación, ya que la población total sería variable en el tiempo.

El promedio empleado es sobre el último año observado, el cual es la base para la proyección, pero puede hacerse para más periodos.

Se toma la distribución inicial de la población como

$$\pi_{c_0,e_0,a_0} = \frac{N_0^{c_0,e_0,a_0}}{I_0} = \frac{\text{Población de categoría } c_0\text{, edad } e_0\text{, y antigüedad } a_0}{\text{Población Total}},$$

para una tripleta $(c_0, e_0, a_0) \in C \times E \times A$.

3.3 Probabilidades de transición mensual

Para el $m-$ésimo mes, tome

$$N_m^{C^{(r)},E^{(i)},A^{(k)}} = \sum_{e\in E^{(i)},a\in A^{(k)}} N_m^{C^{(r)},e,a},$$

como el número total de personas que estaban en la categoría $C^{(r)}$, con edad y antigüedad en los rangos $E^{(i)}$ y $A^{(k)}$, respectivamente, en el mes m.

Por otro lado, sea $N_m^{E^{(i)},A^{(k)}}\left(C^{(r)}, C^{(l)}\right)$, el número de personas con edad y antigüedad en los rangos $E^{(i)}$ y $A^{(k)}$, respectivamente, que se encontraba en el mes m en la categoría $C^{(r)}$, después de un mes (en el mes $m+1$), se encontraba en la categoría $C^{(l)}$.

La probabilidad de transición mensual del $m-$ésimo mes estaría dada por[9]

$$\widetilde{P}_m^{E^{(i)},A^{(k)}}\left(C^{(r)}, C^{(l)}\right) = \frac{N_m^{E^{(i)},A^{(k)}}\left(C^{(r)}, C^{(l)}\right)}{N_m^{C^{(r)},E^{(i)},A^{(k)}} - N_m^{E^{(i)},A^{(k)}}\left(C^{(r)}, C^{(0)}\right)}.$$

El estimador de esta probabilidad sería

$$\widetilde{P}^{E^{(i)},A^{(k)}}\left(C^{(r)}, C^{(l)}\right) = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \widetilde{P}_m^{E^{(i)},A^{(k)}}\left(C^{(r)}, C^{(l)}\right),$$

donde $|\mathcal{M}|$ es el número de elementos que tiene $\mathcal{M}$. Además, según [6, p. 21], para una cadena de Markov este es un estimador insesgado de las probabilidades de transición.

[9] Observe que se resta del denominador el número de personas que llegaron a $C^{(0)}$, esto debido a que las transiciones se cuentan solamente para aquellos movimientos dentro del sistema. Sin embargo, no hay restricción en que la persona pase de $C^{(0)}$ a una categoría dentro del sistema.

3.4 Probabilidades de transición anual

Una vez estimadas las probabilidades de transición mensuales, se obtienen las anuales con la fórmula (2)

$$P^{E^{(i)},A^{(k)}}\left(C^{(r)},C^{(l)}\right)=\sum_{t=1}^{12}\sum_{\widetilde{c}_1,\ldots,\widetilde{c}_{t-1}\in C}\prod_{k=1}^{t}\widetilde{P}^{E^{(i)},A^{(k)}}\left(\widetilde{c}_k,\widetilde{c}_{k-1}\right)\cdot p_t^{(i)},$$

con $\widetilde{c}_0=C^{(r)}$ y $\widetilde{c}_t=C^{(l)}$.

3.5 Probabilidades de ingreso al sistema

Sea $\mathcal{N}$ el conjunto de años observados. $N_n^{E^{(i)},A^{(k)},C^{(r)}}(1)$ el número de personas de la categoría $C^{(r)}$, con edad y antigüedad en los rangos $E^{(i)}$ y $A^{(k)}$, respectivamente, que en el $n-$ésimo año fueron contratados, y $N_n^{E^{(i)},A^{(k)},C^{(r)}}(0)$ aquellos con las mismas características que no lo fueron.

Las probabilidades de ingreso al sistema del $n-$ésimo año estarían dadas por

$$Q_n^{E^{(i)},A^{(k)},C^{(r)}}(l)=\frac{N_n^{E^{(i)},A^{(k)},C^{(r)}}(l)}{N_n^{E^{(i)},A^{(k)},C^{(r)}}(0)+N_n^{E^{(i)},A^{(k)},C^{(r)}}(1)}.$$

El estimador de esta probabilidad sería

$$Q^{E^{(i)},A^{(k)},C^{(r)}}(l)=\frac{1}{|\mathcal{N}|}\sum_{n\in\mathcal{N}}Q_n^{E^{(i)},A^{(k)},C^{(r)}}(l).$$

3.5.1 Distribución de la población por características

Tomamos el dato $N_m^{C^{(r)},E^{(i)},A^{(k)}}$ del número total de personas que en el $m-$ésimo mes pertenecían a la categoría $C^{(r)}$, dentro del grupo de edad $E^{(i)}$ y rango de antigüedad $A^{(k)}$. De este grupo, obtenga el número $N_m^{C^{(r)},E^{(i)},A^{(k)}}(S_1^{j_1},\ldots,S_{N_S}^{j_{N_S}})$ de personas que tenían las características salariales $(S_1^{j_1},\ldots,S_{N_S}^{j_{N_S}})$, durante el mes m. Utilizando la fórmula (5),

$$R_m^{C^{(r)},E^{(i)},A^{(k)}}\left(S_1^{j_1},\ldots,S_{N_S}^{j_{N_S}}\right)=\frac{N_m^{C^{(r)},E^{(i)},A^{(k)}}\left(S_1^{j_1},\ldots,S_{N_S}^{j_{N_S}}\right)}{N_m^{C^{(r)},E^{(i)},A^{(k)}}}.$$

El estimador de esta probabilidad sería

$$R^{C^{(r)},E^{(i)},A^{(k)}}\left(S_1^{j_1},\ldots,S_{N_S}^{j_{N_S}}\right)=\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}R_m^{C^{(r)},E^{(i)},A^{(k)}}\left(S_1^{j_1},\ldots,S_{N_S}^{j_{N_S}}\right).$$

4 Simulación de Montecarlo

Se desea proyectar la población utilizando una simulación de Montecarlo. Defina

$$P^n_{r,i,k} := \mathbb{P}\left[(X_n, Y_n, Z_n) \in \{C^{(r)}\} \times E^{(i)} \times A^{(k)}\right];$$

y

$$R^{r,i,k}_{j_1,\ldots,j_{N_S}} := R^{C^{(r)},E^{(i)},A^{(k)}}\big(S^{j_1}_1, \ldots, S^{j_{N_S}}_{N_S}\big).$$

Luego, defina:

$$V^{n,i,r,k}_{j_1,\ldots,j_{N_S}} := P^n_{r,i,k} R^{r,i,k}_{j_1,\ldots,j_{N_S}},$$

que representa la probabilidad de que una persona esté en la $\{N_S+3\}$-tupla dada por: $i, r, k, j_1, \ldots, j_{N_S}$ en el año n, por definición de probabilidad condicional.

4.1 Algoritmo

Suponga que se van a proyectar N años, utilizando la técnica de Montecarlo con 10.000 iteraciones por año.

Pseudocódigo:

- **Recibe:** Probabilidades $V^{n,i,r,k}_{j_1,\ldots,j_{N_S}}$, y la población inicial I_0.

- Inicie $I^{C^{(r)},E^{(i)},A^{(k)}}_1\big(S^{j_1}_1, \ldots, S^{j_{N_S}}_{N_S}\big) = 0, \ldots, I^{C^{(r)},E^{(i)},A^{(k)}}_N\big(S^{j_1}_1, \ldots, S^{j_{N_S}}_{N_S}\big) = 0$, para todo $r, i, k, j_1, \ldots, j_{N_S}$.

- Para $n = 1, \ldots N$ (proyección para N años) genere el siguiente vector aleatorio

$$\Big\{I^{C^{(r)},E^{(i)},A^{(k)}}_n\big(S^{j_1}_1, \ldots, S^{j_{N_S}}_{N_S}\big)\Big\}_{i,r,k,j_1,\ldots,j_{N_S}} \sim \; Multinom\big(I_0; \{V^{n,i,r,k}_{j_1,\ldots,j_{N_S}}\}_{i,r,k,j_1,\ldots,j_{N_S}}\big).$$

- **Devuelve:** La población para cada categoría, grupo de edad, rango de antigüedad, y cada año en el futuro, agregada y desagregada por características, para los siguientes N años.

5 Ejemplo de cálculo de gastos por remuneraciones

Se utilizaron los datos de planillas de una institución pública para los años 2004-2015. Las proyecciones se harán por un rango de 25 años. Además, se consideran los grupos de edad de la siguiente forma:

- Grupo $E^{(0)}$: Personas menores a 18 años, considerados como potenciales empleados al cumplir 18 años.
- Grupo $E^{(1)}$: Aquellas personas con edades entre los 18 años y los 30 años (no cumplidos).
- Grupo $E^{(2)}$: Individuos con edades entre los 30 años y hasta los 40 años (no cumplidos).
- Grupo $E^{(3)}$: Compuesto por personas con edades entre los 40 años y los 50 años (no cumplidos).
- Grupo $E^{(4)}$: Este grupo es considerado como el de las potenciales jubilaciones, pues está integrados por individuos con edades superiores a los 50 años.

Del mismo modo, se consideran cuatro grupos de antigüedades (indicador de su ligamen con la institución):

- Grupo $A^{(1)}$: Personas con menos de 15 años de trabajar en la institución (no necesariamente consecutivos).
- Grupo $A^{(2)}$: Individuos que han trabajado para la entidad entre 15 y 30 años (no necesariamente consecutivos).
- Grupo $A^{(3)}$: Integrado por trabajadores que han laborado entre 30 y 45 años (no necesariamente consecutivos).
- Grupo $A^{(4)}$: Compuesto por personas con un nivel elevado de relación con la institución, habiendo laborado por más de 45 años (no necesariamente consecutivos).

La tripleta, compuesta por categoría, grupo de edad y antigüedad, utiliza las categorías salariales dadas por:

$$
\begin{array}{lllll}
C^{(1)}=11; & C^{(2)}=12; & C^{(3)}=13; & C^{(4)}=14; & C^{(5)}=21;\\
C^{(6)}=22; & C^{(7)}=23; & C^{(8)}=24; & C^{(9)}=25; & C^{(10)}=31;\\
C^{(11)}=32; & C^{(12)}=34; & C^{(13)}=35; & C^{(14)}=36; & C^{(15)}=37;\\
C^{(16)}=38; & C^{(17)}=41; & C^{(18)}=42; & C^{(19)}=43; & C^{(20)}=44;\\
C^{(21)}=45; & C^{(22)}=47; & C^{(23)}=48; & C^{(24)}=49; & C^{(25)}=53;\\
C^{(26)}=54; & C^{(27)}=57; & C^{(28)}=63; & C^{(29)}=79; & C^{(30)}=82;\\
C^{(31)}=84; & C^{(32)}=86; & C^{(33)}=87; & C^{(34)}=88; & C^{(35)}=89;\\
C^{(36)}=90; & C^{(37)}=91. & & &
\end{array}
$$

Por su parte, se consideraron como "características" solamente las siguientes: anualidad, dedicación exclusiva, prohibición, disponibilidad, régimen de pensiones, y jornada laboral. Esto por cuanto los gastos derivados de los cuatro primeros representan el 90% de los gastos en pluses salariales, el quinto se emplea para calcular el correcto monto gastado por la universidad en aportes patronales a pensiones, así como para poder efectuar el respectivo análisis sobre el efecto del régimen sobre las jubilaciones, y el último debido a que no todos los empleados trabajan en jornadas de 40 horas semanales (algunos trabajan más y otros trabajan menos horas).

Para un trabajador de categoría salarial $C^{(i)}$ (asociada unívocamente a un salario base W_i según la escala salarial del segundo semestre del 2015), con jornada laboral J, y porcentajes[10] de anualidad A, dedicación exclusiva DX, prohibición P y disponibilidad D, se calcula el gasto anual en el año N por concepto de sus salarios sin garantías sociales G con la siguiente fórmula:

$$
\begin{aligned}
&G(N,J,W_i,A,DX,P,D)=\\
&\frac{J}{40}\cdot\left[6\cdot W_i\cdot\left((1+3,88\%)^{(N-2016)+1/2}+(1+3,88\%)^{N-2015}\right)\right]\cdot\frac{(1+A+DX+P+D)}{0.90}.
\end{aligned}
$$

El primer término $(J/40)$ nos indica el porcentaje del salario que recibe por la proporción de horas que trabaja respecto de la jornada completa. El segundo término nos contabiliza los 12 salarios base, tomando en cuenta que hay 6 salarios correspondientes a $2\cdot(N-2016)+1$ aumentos[11] semestrales desde diciembre del 2015, y otros 6 salarios con $(N-2015)$ aumentos anuales desde diciembre del 2015. El último término nos presenta lo gastado por estos 12 salarios, más su porcentaje de anualidad, dedicación exclusiva, prohibición, y disponibilidad.

[10]Estos porcentajes pueden ser 0%.

[11]Considerados para efectos de este estudio como iguales a la inflación esperada del 3,88%.

Como estos no son todos los pluses salariales, y dado que la suma de los pluses que no están siendo considerados representan el 10% de los gastos totales en salarios, se procede a normalizar el monto calculado por el factor $\frac{1}{0{,}90}$, estimando así el verdadero gasto anual para este trabajador.

Del mismo modo, para un trabajador con las características anteriormente indicadas, aunado al porcentaje de cotización de régimen de pensiones R, el porcentaje de salario escolar E, 14,25% de garantías sociales (Seguro de Enfermedad y Maternidad, Ley de Protección al Trabajador, Banco Popular)[12], 4,25% para el Fondo de Cesantía y Asociaciones, 8,33% de Aguinaldo, y 0,25% por Riesgos de Trabajo del INS, se utiliza la siguiente fórmula para estimar el gasto total anual GT en este empleado:

$$GT(N, J, W_i, A, DX, P, D, R, E) =$$
$$G(N, J, W_i, A, DX, P, D) \cdot (1 + E) \cdot \big((1 + R + 14,25\% + 4,25\%) + 8,33\%\big) * (1 + 0,25\%).$$

Según directiva presidencial, el salario escolar será incrementado en tractos durante los años 2016-2018 hasta alcanzar el valor de un salario completo (como el aguinaldo). Esto fue tomado en cuenta, de modo que los porcentajes que se utilizaron para E fueron de 8,19% hasta el 2015, pasando a 8,23% en el 2016, 8,28% en el 2017, y manteniéndose en un nivel de 8,33% a partir del 2018. Por su parte, para efectos de los aportes patronales a los regímenes de pensiones que inciden en los costos de la institución, se utilizó un porcentaje sobre el salario de 6,75% si el empleado pertenece al Régimen de Capitalización Colectiva (RCC) de la Junta de Pensiones y Jubilaciones del Magisterio Nacional (JUPEMA), de 5% si está en el de Régimen Transitorio de Reparto (RTR) de JUPEMA, y se siguió el Transitorio XI[13] de la CCSS para los empleados en el IVM, es decir, se tomaron aportes patronales de 4.92% en el 2014, de 5.08% entre los años 2015-2019, de 5.25% en el rango 2020-2024, de 5.42% entre el 2025 y el 2029, de 5.58% para el quinquenio 2030-2034, y de 5.75% a partir del 2035.

Se considera que los tiempos de parada T_i se distribuyen uniformemente, es decir, $p_t^{(i)} = \frac{1}{12}$ para todo i. Además, las probabilidades de transición se ponderaron por sus jornadas, de modo que cada trabajador aporta a las transiciones dependiendo de las horas que laboran en la institución.

Para medir la calidad del ajuste, así como intentar cuantificar la certeza de sus proyecciones, se implementó el método conocido como "backtesting", el cual consiste en utilizar los datos de los años 2004-2013 para ajustar los parámetros,

[12] La institución está excenta de pagar IMAS, Asignaciones Familiares e INA.

[13] Se refiere a los distintos porcentajes de cotización de los trabajadores al régimen de Invalidez, Vejez y Muerte (IVM).

y contrastar los valores observados en los años 2014-2015 contra los proyectados por el modelo. Inicialmente notamos que las predicciones globales son bastante buenas, donde el gasto total (de las cuentas descritas por este modelo) cuantificaron 64.215.040.730 colones en el 2014 y 70.044.868.080 colones en el 2015, mientras que lo esperado[14] por el modelo rondaba los 64.149.862.676 colones para el 2014 y 69.262.303.902 colones para el 2015. Se estaría contando con un error de 0,1% en el 2014 y de 1,12% para el 2015.

Se comprueba que el nivel de predicción del modelo respecto a los valores observados (tanto en número de trabajadores como en colones gastado en sus salarios) es muy alta, validando la confiabilidad de las conclusiones que de este modelo se deriven. Al final del documento (sección de anexos) se presentan los gráficos sobre población observada versus esperada, así como los gastos calculados contra los observados, tanto para las tres categorías más importantes, así como para los tres grupos de edad más relevantes. Se observa claramente las propiedades predictivas del modelo, ya que las proyecciones distan muy poco de los valores observados, aunado al hecho de que los mismos se encuentran dentro del intervalo de confianza al 95%.

5.1 Conclusiones

A lo largo del proceso se comprobó que el modelo probabilístico desarrollado en este artículo, logra capturar información importante de las transiciones entre tripletas, permitiendo proyectar exitosamente los valores futuros de las variables en cuestión. Aún más, al aplicarlo a datos reales se aprecia claramente cómo el modelo logra ajustar los datos observados, y así proyectar sus valores futuros de manera confiable. Para la base de datos utilizada, se logró con este modelo pronosticar los valores futuros tanto demográficos como financieros, permitiendo reconocer el impacto por efectos de contratación (transición entre categorías) así como su contribución financiera por medio de los respectivos valores de las $n-$tuplas, con las cuales se logró estimar los costos anuales de la institución. De esta forma, este modelo se concreta como una herramienta útil para trabajar el problema de proyectar poblaciones para fondos de pensiones o para instituciones públicas.

El modelo presenta un reto de programación, pero una vez hecha esta inversión, el mismo posee suficiente flexibilidad como para poder efectuar análisis demográficos (como poblaciones en vías de jubilación) así como financieros (montos cotizados por los trabajadores a lo largo de la proyección). Se pueden incluir comportamientos futuros, como aumentos en probabilidades de transición

[14]Nos referimos a la "Esperanza Matemática" de la proyección.

en ciertas categorías, o aumentos/reducciones en pluses salariales.

Se logra de este modo el objetivo de presentar un modelo probabilístico suficientemente riguroso, formal, sólido y robusto, como para agregarse a los análisis actuariales nacionales e internacionales, dando confiablidad a los resultados que se desprendan de esta modelación. Se enfatiza en el éxito de la estimación de los parámetros, de manera tal que quien desee utilizar esta modelación lo puede hacer sin tener que recurrir a otras fuentes, unificando de esta manera la implementación del modelo y su formalización, facilitando así el uso generalizado del mismo.

Referencias

[1] Bowers, N.; Gerber, H.; Hickman, J.; Jones, D.; Nesbitt, C. (1986) *Actuarial Mathematics*. Society of Actuaries, Estados Unidos.

[2] Ching, W.; Zhang, S.; Ng, M. (2007) "On multi-dimensional Markov chain models", *Pacific Journal of Optimization* **3**(2): 235–243.

[3] Chung, K. (1967) *Markov Chain with Stationary Transition Probabilities*, 2da edición. Springer-Verlag, Berlin, Heidelberg.

[4] Hoel, P.; Port, S.; Stone, C. (1987) *Introduction to Stochastic Processes*. Waveland Press, Boston.

[5] Kulkarni, V. (2011) *Introduction to Modeling and Analysis of Stochastic Systems*. Springer, New York.

[6] MacDonald, I.; Zucchini, W. (2009) *Hidden Markov Models for Time Series. An Introduction Using R*. Chapman & Hall, Boca Raton FL, EE.UU.

[7] Morales, I.; Castro, M. (2016) "Proyecciones demográficas y actuariales por medio del método de cadenas de Markov con Monte Carlo", *Revista de Matemática: Teoría y Aplicaciones* **23**(1):241–253.

[8] Norris, J.R. (1997) *Markov Chains*. Cambridge University Press, Reino Unido.

[9] Ross, S. (1997) *Introduction to Probability Models*, 6ta edición. Academic Press, California.

[10] Taylor, H.; Karlin, S. (1998) *An Introduction to Stochastic Modeling*, 3a edición. Academic Press, San Diego.

Anexos

Resultados probabilísticos útiles

Para el desarrollo del modelo de generación poblacional es necesario establecer unos resultados conocidos[15], que son herramientas útiles para efectuar los cálculos necesarios del modelo.

Lema 1 *Si D_i son disjuntos y $\mathbb{P}[C \mid D_i] = p$, independientemente de i, entonces*

$$\mathbb{P}[C \mid \cup_i D_i] = p.$$

Demostración.

$$\mathbb{P}[C \mid \cup_i D_i] = \frac{\mathbb{P}\left[C \cap \cup_i D_i\right]}{\mathbb{P}[\cup_i D_i]} = \frac{\sum_i \mathbb{P}[C \cap D_i]}{\mathbb{P}[\cup_k D_k]}$$
$$= \frac{1}{\mathbb{P}[\cup_k D_k]} \sum_i \overbrace{\mathbb{P}[C \mid D_i]}^{=p} \mathbb{P}[D_i] = p \frac{1}{\mathbb{P}[\cup_k D_k]} \overbrace{\sum_i \mathbb{P}[D_i]}^{=\mathbb{P}[\cup_i D_i]} = p.$$

■

Lema 2 *Si C_i son disjuntos, entonces*

$$\mathbb{P}[\cup_i C_i \mid D] = \sum_i \mathbb{P}[C_i \mid D].$$

Demostración. Por aditividad contable de $\mathbb{P}$ se tiene que

$$\mathbb{P}[\cup_i C_i \mid D] = \frac{\mathbb{P}[(\cup_i C_i) \cap D]}{\mathbb{P}[D]} = \frac{\mathbb{P}[\cup_i (C_i \cap D)]}{\mathbb{P}[D]} = \sum_i \frac{\mathbb{P}[C_i \cap D]}{\mathbb{P}[D]} = \sum_i \mathbb{P}[C_i \mid D].$$

■

Lema 3 *Si E_i son disjuntos y $\cup_i E_i = \Omega$, entonces*

$$\mathbb{P}[C \mid D] = \sum_i \mathbb{P}[E_i \mid D] \mathbb{P}[C \mid E_i \cap D].$$

Demostración. Note que

$$\mathbb{P}[E_i \mid D] \mathbb{P}[C \mid E_i \cap D] = \frac{\mathbb{P}[E_i \cap D]}{\mathbb{P}[D]} \cdot \frac{\mathbb{P}[C \cap E_i \cap D]}{\mathbb{P}[E_i \cap D]} = \mathbb{P}[C \cap E_i \mid D].$$

[15] Ver [4] para adentrarse más en el tema.

Como los $C \cap E_i$ son disjuntos, entonces por el Lema 2,

$$\sum_i \mathbb{P}[E_i \mid D]\mathbb{P}[C \mid E_i \cap D] = \sum_i \mathbb{P}[C \cap E_i \mid D] =$$
$$\mathbb{P}[\cup_i (C \cap E_i) \mid D] = \mathbb{P}[C \cap (\cup_i E_i) \mid D] = \mathbb{P}[C \mid D].$$

■

Gráficos del ajuste: valor esperado vs valor observado

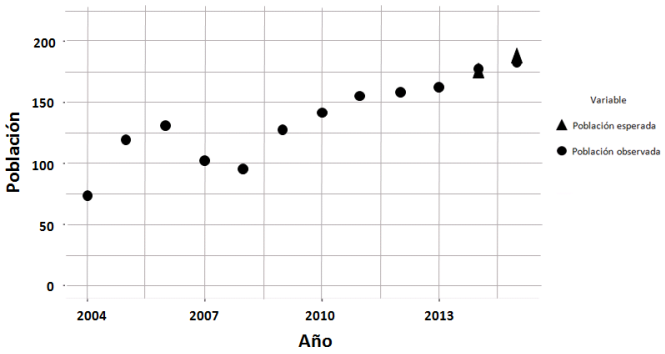

Figura 1: Backtesting poblacional: categoría 35.

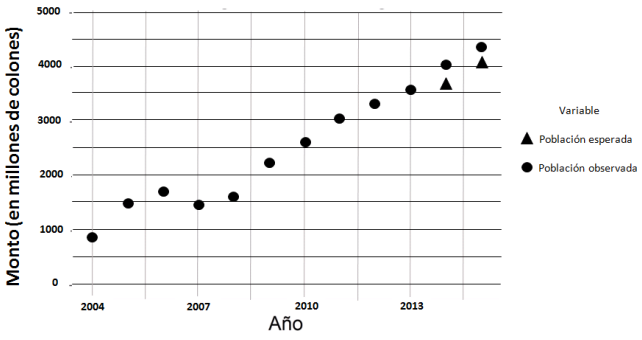

Figura 2: Backtesting salarial: categoría 34.

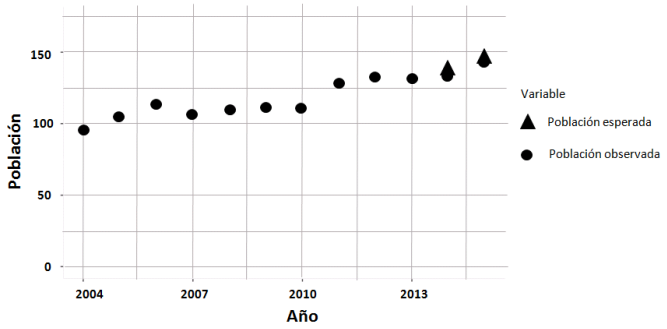

Figura 3: Backtesting poblacional: categoría 35.

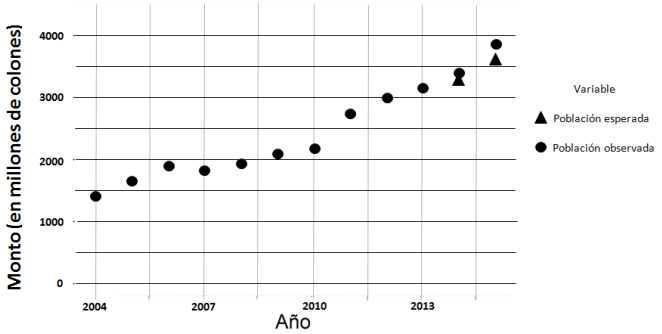

Figura 4: Backtesting salarial: categoría 35.

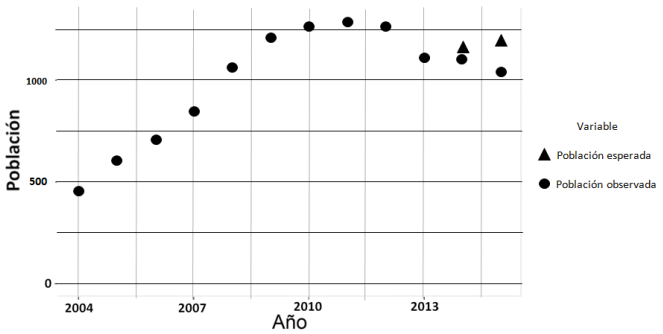

Figura 5: Backtesting poblacional: categoría 88.

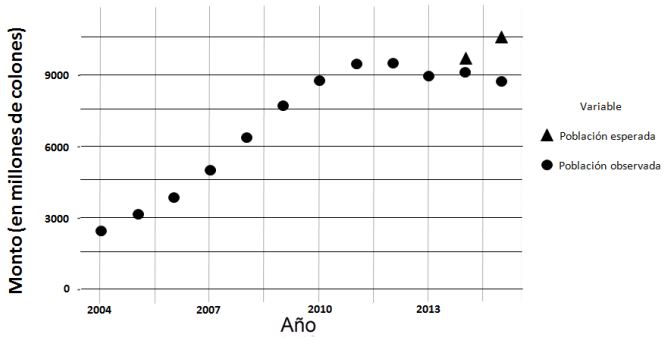

Figura 6: Backtesting salarial: categoría 88.

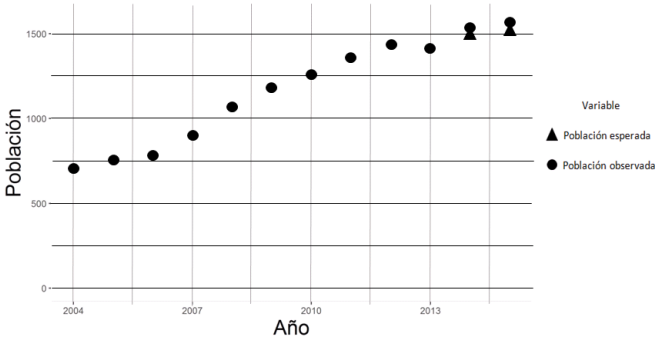

Figura 7: Backtesting poblacional: de 30 a 39 años.

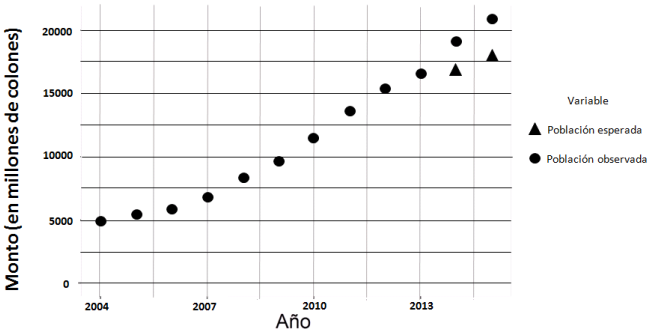

Figura 8: Backtesting salarial: de 30 a 39 años.

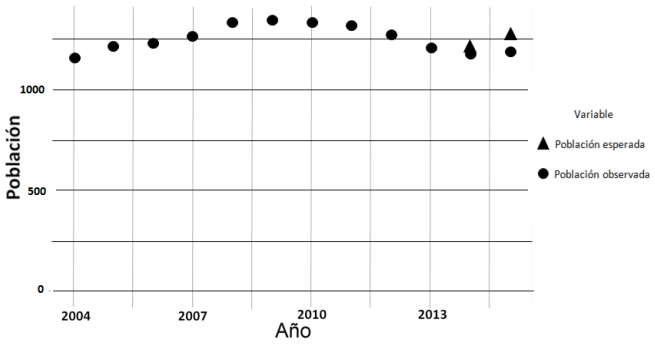

Figura 9: Backtesting poblacional: de 40 a 49 años.

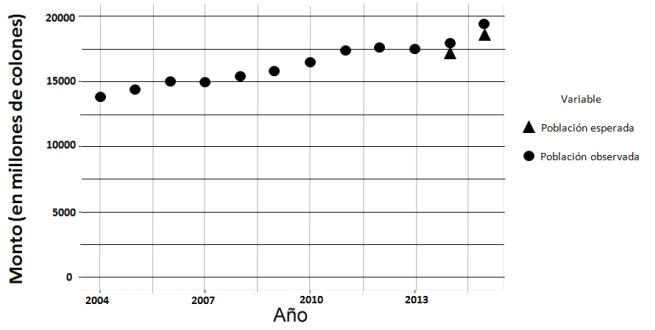

Figura 10: Backtesting salarial: de 40 a 49 años.

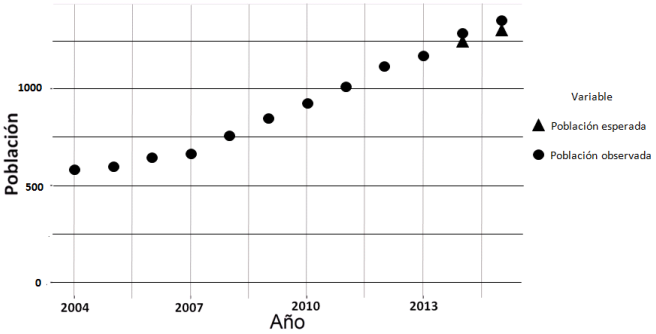

Figura 11: Backtesting poblacional: de 50 años en adelante.

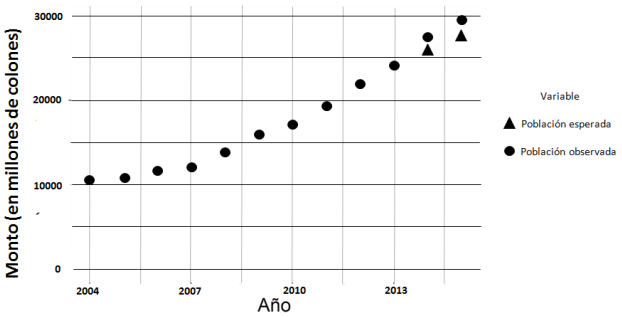

Figura 12: Backtesting salarial: de 50 en adelante.