UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Ajuste y predicción de la mortalidad. Aplicación a Colombia.

Noviembre de 2021

Autora: Gisou Díaz Rojo

Directora: Ana Debón Aucejo

Agradecimientos

Agradezco a mi directora de tesis Ana Debón por su dedicación, paciencia y confianza durante este largo proceso de aprendizaje.

Agradezco a mi amigo y colega Jaime, por sus múltiples aportes, por ayudarme a finalizar este proyecto.

Agradezco a los profesores de la UPV y a los amigos que dejé en Valencia por los conocimientos y buenos momentos compartidos.

Agradezco a mi familia y a mis amigos del mundo por el cariño y la buena energía que me dan, por estar en mi vida.

Agradezco a la Universidad del Tolima por la oportunidad brindada y el apoyo económico que hizo posible este proyecto.

a mi hija Aryana

Resumen

En la actualidad resulta de gran importancia el análisis de los fenómenos como el crecimiento poblacional y la reducción de la mortalidad por la repercusión económica y social que dichos procesos tienen en el desarrollo de los países. En este sentido las tablas de vida constituyen una herramienta para comprender, a través de las probabilidades de muerte, la esperanza de vida y otros indicadores, la dinámica poblacional. Lee y Carter (1992), plantearon un modelo, cuyo ajuste permite a los analistas obtener una visión dinámica del comportamiento de la mortalidad durante un periodo de análisis.

Esta tesis doctoral busca contribuir en la comprensión de los cambios que ha experimentado la población colombiana en cuanto a mortalidad. Para lograrlo se plantearon cuatro objetivos. El primero, construir modelos estocásticos de mortalidad como Lee-Carter para datos de Colombia y hacer un estudio comparativo de dichos modelos para evaluar su coherencia a partir de la calidad de los resultados obtenidos. El segundo, calcular y analizar algunos indicadores relacionados con la mortalidad tales como la mortalidad infantil, la esperanza de vida al nacer, la esperanza de vida a los 65 años, el índice de Gini al nacer y el índice de Gini a los 65 años. El tercero, aplicar gráficos de control para identificar los momentos en el tiempo y los intervalos de edad en los que la probabilidad de muerte observada es sustancialmente diferente de la pauta de mortalidad en el período estudiado. Para esto, los residuos de los modelos seleccionados se vigilaron mediante el gráfico de control multivariado T^2 de Hotelling para detectar cambios sustanciales en la mortalidad que no fueron identificados por los modelos. El cuarto, analizar el comportamiento de la mortalidad para los departamentos de Colombia mediante técnicas de análisis

multivariado como el análisis de componentes principales, el clúster jerárquico y el fuzzy clúster, para posteriormente identificar grupos de departamentos con comportamientos similares y caracterizarlos mediante los indicadores de mortalidad estudiados.

La metodología descrita relacionada con los tres primeros objetivos se aplicó a datos de las tablas de vida abreviadas por sexo para Colombia para el período 1973-2005, utilizando la información disponible en *The Latin America Human Mortality Database*. Para el análisis de la mortalidad por departamentos se construyeron nuevas tablas de vida abreviadas por sexo con la información de los departamentos para el período 1985-2014, ajustándonos a la información disponible para los departamentos de Colombia en cuanto a defunciones y población. La metodología fue implementada a través del software estadístico libre R, lo que permite la replicabilidad y reproducibilidad de los resultados.

Resum

En l'actualitat resulta de gran importància l'anàlisi dels fenòmens com el creixement poblacional i la reducció de la mortalitat per la repercussió econòmica i social que aquests processos tenen en el desenvolupament dels països. En aquest sentit les taules de vida constitueixen una eina per a comprendre, a través de les probabilitats de mort, l'esperança de vida i altres indicadors, la dinàmica poblacional. Lee i Carter (1992), van plantejar un model, l'ajust del qual permet als analistes obtenir una visió dinàmica del comportament de la mortalitat durant un període d'anàlisi.

Aquesta tesi doctoral cerca contribuir en la comprensió dels canvis que ha experimentat la població colombiana quant a mortalitat. Per a aconseguir-ho es van plantejar quatre objectius. El primer, construir models estocàstics de mortalitat com Lee-Carter per a dades de Colòmbia i fer un estudi comparatiu d'aquests models per a avaluar la seua coherència a partir de la qualitat dels resultats obtinguts. El segon, calcular i analitzar alguns indicadors relacionats amb la mortalitat tals com la mortalitat infantil, l'esperança de vida en nàixer, l'esperança de vida als 65 anys, l'índex de Gini en nàixer i l'índex de Gini als 65 anys. El tercer, aplicar gràfics de control per a identificar els moments en el temps i els intervals d'edat en els quals la probabilitat de mort observada és substancialment diferent de la pauta de mortalitat en el període estudiat. Per a això, els residus dels models seleccionats es van vigilar mitjançant el gràfic de control multivariat T^2 de Hotelling per a detectar canvis substancials en la mortalitat que no van ser identificats pels models. El quart, analitzar el comportament de la mortalitat per als departaments de Colòmbia mitjançant tècniques d'anàlisi multivariada com l'anàlisi de components principals,

el clúster jeràrquic i el fuzzy clúster, per a posteriorment identificar grups de departaments amb comportaments similars i caracteritzar-los mitjançant els indicadors de mortalitat estudiats.

La metodologia descrita relacionada amb els tres primers objectius es va aplicar a dades de les taules de vida abreujades per sexe per a Colòmbia per al període 1973-2005, utilitzant la informació disponible en *The Latin America Human Mortality Database*. Per a l'anàlisi de la mortalitat per departaments es van construir noves taules de vida abreujades per sexe amb la informació dels departaments per al període 1985-2014, ajustant-nos a la informació disponible per als departaments de Colòmbia quant a defuncions i població. La metodologia va ser implementada a través del programari estadístic lliure R, la qual cosa permet la replicabilidad i reproducibilidad dels resultats.

Abstract

The analysis of phenomena such as population growth and mortality reduction is currently of great importance because of the economic and social impact that these processes have on the development of countries. In this sense, life tables are a tool for understanding population dynamics through death probabilities, life expectancy and other indicators. Lee and Carter (1992) proposed a model whose adjustment allows analysts to obtain a dynamic view of the behavior of mortality during a period of analysis.

This doctoral thesis seeks to contribute to the understanding of the changes experienced by the Colombian population in terms of mortality. To achieve this, four objectives were proposed. The first, to construct stochastic mortality models such as Lee-Carter for Colombian data and to make a comparative study of these models to evaluate their coherence based on the quality of the results obtained. The second is to calculate and analyze some mortality-related indicators such as infant mortality, life expectancy at birth, life expectancy at age 65, the Gini index at birth and the Gini index at age 65. The third is to apply control charts to identify moments in time and age intervals in which the observed probability of death is substantially different from the mortality pattern in the period studied. For this, the residuals of the selected models were monitored using Hotelling's T^2 multivariate control chart to detect substantial changes in mortality that were not identified by the models. Fourth, to analyze the behavior of mortality for the departments of Colombia using multivariate analysis techniques such as principal component analysis, hierarchical clustering and fuzzy clustering, in order to subsequently identify groups

of departments with similar behavior and characterize them by means of the mortality indicators studied.

The methodology described in relation to the first three objectives was applied to data from the abbreviated life tables by sex for Colombia for the period 1973-2005, using the information available in *The Latin America Human Mortality Database*. For the analysis of mortality by department, new abbreviated life tables by sex were constructed with information from the departments for the period 1985-2014, adjusting to the information available for the departments of Colombia in terms of deaths and population. The methodology was implemented through the free statistical software R, which allows the replicability and reproducibility of the results.

Índice general

Índice de figuras

Índice de tablas

Introducción

Predecir con exactitud el proceso de envejecimiento de la población es una preocupación permanente en todos los países del mundo debido a las repercusiones sociales y económicas que dicho proceso tiene. En este contexto, la aplicación de los modelos ya aplicados en el mundo desarrollado para la construcción de tablas dinámicas de mortalidad en países como Colombia se presenta como un punto clave de investigación. Aunque disponemos de algunas propuestas de modelos, es necesario profundizar en la fiabilidad de los datos y de su impacto en los productos actuariales en países con diferentes evoluciones demográficas.

En este sentido, los objetivos de esta tesis son los siguientes:

1. Construir modelos estocásticos de mortalidad como Lee-Carter para datos de Colombia mediante el uso de la base de datos internacional *Latin American Human Mortality Database*. Recopilar artículos de investigación relacionados con el tema de la modelización de la mortalidad en diferentes países y realizar un estudio comparativo que permita decidir la calidad de los resultados obtenidos.

2. Seleccionar indicadores de mortalidad como la mortalidad infantil, esperanza de vida al nacer, la esperanza de vida a los 65 años, el índice de Gini al nacer y el índice de Gini a los 65 años para cuantificar adecuadamente los datos de Colombia y caracterizarlos.

3. Aplicar gráficos de control para identificar los periodos y los intervalos de edad en los que la probabilidad de muerte observada es sustancialmente diferente del patrón de mortalidad de un periodo determinado. Utilizar

los residuos de los modelos estudiados en el gráfico de control multivariado T^2 de Hotelling para detectar cambios sustanciales en la mortalidad que no fueron identificados por los modelos.

4. Analizar el comportamiento de la mortalidad por departamentos para Colombia mediante técnicas de análisis multivariado como el análisis de componentes principales, el clúster jerárquico y el fuzzy clúster para identificar grupos de departamentos con comportamientos similares y luego caracterizarlos a través de los indicadores de mortalidad estudiados.

A partir de estos objetivos, el trabajo se organizó de la siguiente manera:

Capítulo 1.- Brinda una descripción global de la situación actual de la población colombiana en términos demográficos, su crecimiento, cambios y tendencias. Además, ofrece los principales conceptos relacionados con tablas de vida y un resumen de las fuentes de datos demográficos con información para Latinoamérica.

Capítulo 2.- Ofrece una recopilación de los artículos más recientes de modelización de la mortalidad en países de Latinoamérica. Presenta un estudio de algunos modelos estocásticos de mortalidad como el de Lee-Carter y varias de sus extensiones, con el fin de poder aplicarlos a datos de Colombia. En él se realiza un estudio comparativo que permita decidir la calidad de los resultados obtenidos con los modelos en base a su coherencia y se seleccionaron indicadores de mortalidad que resumieran y caracterizaran adecuadamente la mortalidad y sus cambios.

Capítulo 3.- Presenta la aplicación del gráfico de control multivariado T^2 de Hotelling para detectar aquellos cambios sustanciales en la mortalidad que no fueron identificados previamente por los modelos Lee-Carter y Lee-Carter con dos términos. Se identifican a través del gráfico T^2 de Hotelling los tiempos (periodos) e intervalos de edad con probabilidades de muerte que difieren significativamente del patrón de tendencia determinado por los modelos de mortalidad seleccionados y se interpretan según las características de la población colombiana en el contexto histórico.

Capítulo 4.- Analiza la mortalidad de las diferentes departamentos de Colombia y explora la formación de grupos a partir de técnicas multivariadas como el análisis de componentes principales, clúster jerárquico y fuzzy clúster. Se caracterizan los grupos identificados utilizando la información recogida para los indicadores de mortalidad mediante las técnicas de CART y random forest.

Capítulo 1

Análisis de la mortalidad y la tabla de vida en Colombia

Parte del contenido de este capítulo se ha incluido en la publicación: Tendencias y comportamiento de la mortalidad en Colombia entre 1973 y 2005, en Estadística Española, 58 (191), 277-300 (Díaz y Debón 2016).

Estudiar los cambios y tendencias de la mortalidad resulta de gran importancia dado que se registran fenómenos como el aumento progresivo de la población, el envejecimiento poblacional y la reducción de la mortalidad, los cuales tienen un impacto en términos económicos en el desarrollo general de los países. En este capítulo se estudió el comportamiento de la mortalidad en Colombia para el período 1973-2005 utilizando tablas de mortalidad construidas a partir de Latin American Human Mortality Database. El objetivo de este capítulo es proporcionar las herramientas teóricas y prácticas que permitan la depuración de los datos y la obtención de tablas dinámicas, particularmente para los datos de mortalidad colombiana, la actualización permanente de las mismas y de cualquier indicador relacionado con ellas.

1.1 Introducción

Comprender los cambios demográficos de los países es de gran interés en muchas áreas del conocimiento tales como la demografía, las ciencias económicas, la biología o las ciencias actuariales y financieras. Para conocer esta dinámica es necesario analizar tanto el crecimiento de la población, como la incidencia de la mortalidad.

1.1.1 Transición demográfica

La transición demográfica (TD) es un proceso que define el paso de altas tasas de fecundidad y mortalidad a bajas tasas, provocando cambios en el tamaño de la población y en su estructura por edades. Todos los países de alguna manera han transitado por sus etapas en diferentes grado y tiempo, razón por lo que se ha estudiado tanto en países desarrollados como en desarrollo, incluyendo los países de América Latina.

La TD se define como un proceso de larga duración, partiendo de una situación inicial donde existen altas tasas de mortalidad y fecundidad para llegar a una situación final de bajas tasas de mortalidad y fecundidad, para ambas situaciones la tasa de crecimiento demográfico sería baja. Durante el paso de la situación inicial a la final se establecen dos etapas: en la primera etapa, un descenso en la tasa de mortalidad provoca un aumento de la tasa de crecimiento de la población; en la segunda, el descenso en la tasa de fecundidad trae como consecuencia una disminución en la tasa de crecimiento poblacional. Además, en el proceso de TD intervienen factores socioeconómicos y culturales, específicamente con cambios a nivel de condiciones de salud, educación, nutrición, urbanización, estándares de vida, acceso a servicios de salud reproductiva, así como cambios culturales sobre el valor de la familia y los hijos entre otros (Bertranou 2008).

Según el Centro Latinoamericano y Caribeño de Demografía (CELADE 1996), aunque existe heterogeneidad entre los países latinoamericanos en cuanto a su situación demográfica, también se observan similitudes en su comportamiento en la fecundidad, la mortalidad y migraciones internacionales. Teniendo en cuenta la tipología propuesta en CEPAL (2014) que señala cuatro etapas de la transición demográfica para los países de América Latina de acuerdo a la esperanza de vida y a las tasas de fecundidad observadas en el periodo 2005-2010 (Moderada, Plena, Avanzada y Muy avanzada), Colombia se ubica en el grupo de países en transición demográfica avanzada-países con bajas tasas de natalidad y mortalidad y bajo crecimiento poblacional-, entre los que se encuentran además Brasil, Costa Rica, Ecuador, México, Panamá, Perú, República Dominicana y Venezuela.

Durante la segunda mitad del siglo XX en Colombia se dieron grandes cambios demográficos y socioeconómicos como consecuencia del proceso de urbanización y de industrialización. Eventos como el aumento del nivel educativo de la población, el desarrollo científico y tecnológico, la reducción de la mortalidad y la mejora de la calidad de vida de la población, se relacionan con la transformación demográfica y el envejecimiento poblacional que se percibió en el

país (MinSalud 2013). Estos cambios demográficos estuvieron en concordancia a los datos registrados para América Latina en la segunda mitad del siglo pasado: incremento constante en el crecimiento de las generaciones, descenso acentuado en la tasa de mortalidad infantil, aumento de la esperanza de vida al nacer, disminución de la tasa de fecundidad y disminución de la tasa anual de crecimiento poblacional (Saad, Miller y Martínez 2009).

En América Latina y el Caribe, la tasa de crecimiento anual de la población era del 2,6% a mediados del siglo XX, mientras que en la actualidad es del 0,9%. (CEPAL2020)

La Tabla 1.1 muestra los valores de la esperanza de vida al nacer para algunos países de América Latina y el Caribe.

Tabla 1.1: Esperanza de vida al nacer.

País	1990	2000	2008	2018
Argentina	72	74	75	77
Bolivia	56	62	67	71
Brasil	66	70	73	76
Chile	74	76	78	80
Colombia	70	73	75	77
Cuba	75	77	78	79
México	71	74	75	75
Uruguay	73	75	76	78

Fuente: Banco-Mundial (2018)

Entre las consecuencias de los cambios demográficos en Colombia está la variación en la estructura por edad de la población. Debido al descenso sostenido de la fecundidad en las últimas décadas, las generaciones son cada vez menos numerosas (ENDS 2015). Por lo anterior, se obtiene una progresiva reducción relativa de la población de menores edad (0 a 15 años) y el engrosamiento relativo de la población activa (15 a 64 años). En las etapas siguientes, el fenómeno predominante es el acelerado crecimiento de la población mayor. Por lo anterior, junto a la TD se estudia el fenómeno de envejecimiento de la población.

Autores como Bertranou (2008) y CEPAL (2014), indican que el envejecimiento de la población se puede analizar desde tres perspectivas que se relacionan entre sí pero se manifiestan en la sociedad de manera diferente:

- *El envejecimiento demográfico*, es el incremento sistemático de la proporción de personas adultas y mayores en la población total debido principalmente por un descenso drástico de la fecundidad.
- *El envejecimiento doméstico*, es el aumento de la proporción de personas con 60 años o más en los hogares o al aumento del promedio de personas mayores por hogar, asociado sobre todo a factores socioculturales.
- *El envejecimiento individual*, se refiere al incremento de la edad cronológica de las personas, determinado sobre todo por el incremento de la esperanza de vida, el contexto sociocultural y las características de las personas.

1.1.2 Cambios en la Población

El cambio de población se mide como la diferencia entre el tamaño de la población en diferentes fechas (Siegel y Swanson 2004). Usualmente se utilizan los informes censales para analizar estos cambios y detectar si ha ocurrido un aumento o un descenso en el tamaño de la población.

Para analizar los cambios en la población, generalmente se utilizan dos medidas: la *cantidad absoluta de cambio* en la población para diferentes fechas, que se calcula restando la población en la fecha anterior a la población en la fecha posterior; y el *cambio porcentual* para un período, el cual se obtiene dividiendo el cambio absoluto por la población en la fecha anterior y multiplicando por 100.

Por otra parte, para describir la evolución de una población, se examina el cambio ocurrido durante más de un período utilizando la información de tres o más fechas. De esta manera, se describen los cambios durante una serie de periodos sucesivos de la población. Igualmente, se podría describir el cambio para un periodo determinado calculando el cambio por año para analizar cómo se distribuye el cambio dentro del período. A menudo se aplican diferentes modelos para calcular el cambio en una población y luego se comparan con el cambio real de la población.

Resumiendo, la descripción del cambio es un ejemplo de interpolación o de ajuste de curvas a los datos de las series temporales, y por tanto, las técnicas utilizadas son similares a las empleadas para realizar estimaciones y proyecciones intercensales.

Al calcular el incremento medio durante un período de tiempo, implícitamente se hace una suposición sobre cómo se distribuye el crecimiento para ese periodo. Si suponemos un crecimiento es lineal en la población, asumimos que hay una cantidad constante de aumento por unidad de tiempo. Aunque son pocas las condiciones demográficas que hacen que la población aumente o disminuya de forma lineal, la línea recta se utiliza con frecuencia no sólo para describir el crecimiento de la población en períodos pasados, sino también para proyectar el crecimiento de la población en el futuro para periodos cortos de tiempo (Perz 2004).

La Tabla 1.2 muestra el cambio lineal para Colombia teniendo en cuenta los datos censales del Departamento Administrativo Nacional de Estadística de Colombia (DANE). Para el cálculo se dividió el cambio total entre el número de años entre los censos.

Tabla 1.2: Cambio lineal para la población de Colombia para los últimos censos.

Censo	Población	Incremento desde la fecha anterior	Años desde la fecha anterior	Incremento medio anual
1973	20.666.920	–	–	–
1985	27.837.932	7.171.012	12	597.584
1993	33.109.839	5.271.907	8	658.988
2005	41.468.384	8.358.545	12	696.545
2018	48.258.494	6.790.110	13	522.316

Fuente: Elaboración propia a partir datos censales del DANE.

La ecuación para calcular la variación lineal de la población para un período de tiempo se expresa como

$$P_n = P_0 + b \cdot n \tag{1.1}$$

donde P_0 es la población inicial, P_n es la población del periodo final, b es la cantidad anual de cambio en la población, y n es el tiempo en años.

La Tabla 1.3 muestra las tasas de crecimiento de la población teniendo en cuenta diferentes aproximaciones: aritmética, geométrica y exponencial para Colombia en los últimos períodos censales. El método geométrico tiende a producir las estimaciones más altas de crecimiento, seguido por el método exponencial y luego por la aproximación aritmética. No obstante, en el caso de Colombia con los datos utilizados, las tres aproximaciones arrojan valores similares para la tasa de crecimiento de la población.

Tabla 1.3: Tasas de crecimiento aritmético, geométrico y exponencial de la población en Colombia para los últimos censos.

	Tasa de crecimiento			
Fórmula	1973-1985	1985-1993	1993-2005	2005-2018
Aritmética	2,46	2,16	1,87	1,16
Geométrica	2,51	2,19	1,89	1,17
Exponencial	2,48	2,17	1,88	1,17

Fuente: Elaboración propia a partir datos censales del DANE.

1.1.3 Indicadores demográficos

Los indicadores demográficos permiten analizar el comportamiento de los fenómenos demográficos de un país o región. Según el Instituto Nacional de Estadística de España (INE) estos indicadores se utilizan para proporcionar información sobre las principales características y la evolución en el tiempo de ciertos fenómenos demográficos sobre una población como la mortalidad, la natalidad, la fecundidad, la esperanza de vida y la migración (INE 2020).

Aunque a nivel demográfico existen numerosos indicadores, sólo mencionaremos los básicos que generalmente son partes de los boletines e informes en Colombia.

- Tasa bruta de mortalidad: Se define como el total de defunciones a lo largo del año t de personas pertenecientes a un determinado entorno por cada 1.000 habitantes de ese entorno.
- Tasa de mortalidad infantil: Se define como el total de defunciones de menores de un año de vida, pertenecientes a un determinado entorno, por cada 1.000 nacidos vivos en ese entorno.
- Esperanza de vida al nacer: Se define como el número medio de años que vivirían los individuos de una generación sometidos en cada edad al patrón de mortalidad observada sobre las personas de un determinado entorno a lo largo del año t.
- Tasa bruta de natalidad: Se define como el total de nacimientos de madre perteneciente a un determinado ámbito en el año t por cada 1.000 habitantes.
- Tasa global de fecundidad: Se define como el total de nacimientos, de madre de un determinado entorno ocurridos en un año t, por cada 1.000 mujeres en edad fértil (de 15 a 49 años de edad) de dicho entorno.

De acuerdo con el Ministerio de salud y protección social de Colombia (MinSalud 2013), en las últimas décadas Colombia ha experimentado un aumento progresivo de la población y de su expectativa de vida (producto de la caída en de la tasa de mortalidad), con una reducción de la base de la pirámide poblacional (o envejecimiento de la población), los cuales tienen consecuentes cambios sobre su perfil demográfico.

Tabla 1.4: Evolución de los indicadores demográficos en Colombia.

Indicadores	1973	1985	1993	2005	2012	2018
Tasa de mortalidad infantil (‰)	75.0	53.3	33.9	25.6	17.2	10.7
Tasa bruta de mortalidad (‰)	9.4	7.4	7.2	5.9	5.6	4,6
Tasa bruta de natalidad (‰)	37.6	32.6	27.2	22.1	19.1	18.2
Tasa global de fecundidad (‰)	4.5	3.2	3.0	2.4	2.3	2.2
Esperanza de vida (años)	60.9	68.0	70.9	72.6	73.8	76.2

Fuente: DANE 2007a, UNICEF 2016 y Carmona-Fonseca 2005.

La Tabla 1.4 presenta la evolución de algunos indicadores demográficos para Colombia. Como se puede observar, la población ha crecido en cifras absolutas desde 1973 aunque la velocidad de crecimiento va variando en los diferentes periodos. La tasa de mortalidad infantil decreció de manera importante, la tasa bruta de mortalidad también disminuye notablemente, la esperanza de vida aumenta y la tasa bruta de natalidad decrece así como la tasa total de fecundidad. Estos cambios -propios de países en una transición demográfica avanzada- indican que, la población colombiana presenta características demográficas similares a la de países desarrollados.

Otros investigadores como Reyes (2010), indican que en Colombia la mortalidad tiene dos características fundamentales; la primera es que la mortalidad está subregistrada como en la mayoría de los países subdesarrollados y la segunda, que presenta sobremortalidad masculina por causa del conflicto interno y la violencia juvenil. La autora menciona el hecho de que tanto los registros de defunciones como los de nacimientos en Colombia presentan omisiones, señalando un incremento de las muertes violentas en la población masculina así como una posible influencia del conflicto armado sobre factores de producción.

Apoyando la idea anterior sobre los problemas de subregistro, Acosta y Romero (2014b) señalan que el subregistro en Colombia en 2011 se estimaba en 20.3%, lo que ubicaba al país entre los de mayor subregistro en América Latina. En este mismo sentido Carmona-Fonseca (2005), resalta las deficiencias que presenta Colombia en sus registros poblacionales o la falta de los mismos durante varios períodos de tiempo, cuestión que impide hacer en varias ocasiones un análisis demográfico adecuado debido a los diferentes valores que presentan

las fuentes y los diferentes métodos para ajustar los datos provenientes de los censos o hacer proyecciones.

1.1.4 Causas de la mortalidad en Colombia

Por otra parte, Acosta y Romero (2014a) plantean que Colombia presenta características de ciclos avanzados y primarios de la transición epidemiológica. A inicios del siglo XX, el país tenía entre sus principales causas de muertes enfermedades de tipo infecciosa y parasitaria, mientras que en los últimos años la principal causa de muerte se relacionan con enfermedades del sistema circulatorio y cánceres, enfermedades acordes con edades avanzadas. De igual forma, las causas externas como los homicidios y accidentes de transporte terrestre aún se encuentran dentro de las principales causas de muerte.

Según Horiuchi (1999), las características mencionadas anteriormente ubican a Colombia en un revés de la transición epidemiológica, espacio donde coexisten enfermedades propias de las etapas avanzadas de la transición y también un número importante de muertes por homicidios y producto de alienaciones sociales. Podemos decir además, que la reducción de las tasas de mortalidad puede atribuirse -como en otros países- a un incremento de los ingresos, lo cual permite unas mejores condiciones de vida y nutrición para la población; así como un mejor acceso a la salud y a sus avances científicos y tecnológicos.

Tabla 1.5: Distribución porcentual de la mortalidad general por grandes causas para Colombia.

Causas/Año	2005	2017
Enfermedades en sistema circulatorio	30.9	30.5
Otras enfermedades agrupadas	30.6	25.3
Neoplasias	19.2	20.2
Causas externas	14.0	15.5

Fuente: Herrera 2019

La Tabla 1.5 muestra la distribución porcentual de la mortalidad por grandes causas para Colombia en los años 2005 y 2017. La principal causa de muerte en la población colombiana fueron las enfermedades del sistema circulatorio, mostrando una tendencia de estabilización de la mortalidad por esta causa. En el grupo de " Otras enfermedades" que consolida los diagnósticos por enfermedades del sistema respiratorio, enfermedades infecciosas y parasitarias, y deficiencias nutricionales entre otros, se presenta una reducción en la tasa de mortalidad del 5.3%. Las muertes por neoplasias y por causas externas muestran una tendencia creciente para los años analizados.

En relación con las causas externas para 2017, el 42.2% del total de muertes fue por agresiones (homicidios) y el 25.0% fue por accidentes de transporte terrestre. Se debe destacar en este aspecto la gran diferencia que existe entre sexos, la tasa de mortalidad por homicidios es 10,5 veces más alta en hombres que en mujeres, y la tasa de mortalidad por accidentes de transporte terrestre es 4.7 veces mayor para los hombres con respecto a las mujeres (Herrera 2019).

1.2 La tabla de vida

La tabla de vida, también conocida como tabla de mortalidad, es la herramienta que usualmente se utiliza para medir las probabilidades de vida o de muerte de una población o un subgrupo de ésta, en función de la edad para un período establecido (Debón, Montes y Sala 2009). Su estructura está compuesta por estimaciones basadas en observaciones reales -datos de registro de mortalidad y censos de población- las cuales pueden ser estudiadas desde la perspectiva de la edad a la que ocurre el fallecimiento del individuo o desde los años que le restan por vivir después de cierta edad. Es habitual que una tabla de vida presente información relacionada con la mortalidad, la supervivencia y la esperanza de vida, elementos básicos para el análisis de la situación de un país y de sus regiones en el tema de la mortalidad.

1.2.1 Tipos tabla de vida

Las tablas de vida pueden clasificarse de varias formas teniendo en cuenta diferentes aspectos (Zarruk, Villegas y Ortiz 2011). Cuando en la construcción de tablas se observa la mortalidad en un período de tiempo específico y por tanto la probabilidad de muerte depende solo de la edad alcanzada por el individuo, se denominan tablas de mortalidad de *período*. En cambio, si se asume que las tasas de mortalidad además de depender de la edad, dependen del año calendario en que se alcanza dicha edad, se denomina tablas de mortalidad *dinámicas*.

Además podemos clasificar la tabla de vida como *completa*, cuando se construye presentando la información para cada una de las edades, es decir, con las edades año a año, desde el nacimiento hasta la última edad disponible. En caso de que la información se presente de manera agrupada por rangos de edades, entonces la tabla de vida se clasifica como *abreviada*. En las tablas de vida abreviadas, la agrupación que habitualmente se utiliza para la edad es: menores de 1 año, de 1 a 4 años y el resto, en grupos quinquenales de edad hasta el intervalo abierto final. Las tablas abreviadas se utilizan cuando se dispone de los datos

de mortalidad en tasas referidas a grupos quinquenales de edad y no las tasas de mortalidad de cada año de edad. Cuando se trabaja con este tipo de tablas se asume que las muertes se distribuyen uniformemente en cada intervalo de edad (Ayuso 2007).

1.2.2 La estructura de la tabla de vida

En cuanto a la estructura de la tabla de vida, el INE indica que una tabla de vida se compone por un conjunto de funciones biométricas que se definen para una cohorte ficticia de individuos, y destaca la importancia de la comprensión de cada una de ellas para su interpretación (INE 2009) .

Los datos básicos requeridos para la construcción de una tabla de vida son:

- el número de muertes por edad durante un período de tiempo,
- el número de personas por edad que estaban vivas a mediados del mismo año (población).

Generalmente, la información sobre las muertes se obtiene de los registros de defunción existentes para cada país o región, mientras que las cifras de población se toman de los censos de población o de los registros de población (Yusuf y col. 2014).

Las tablas de vida de periodo anual generalmente se construyen con el objetivo de describir el comportamiento de la mortalidad de la población residente en un país, ya sea por sexo o de forma conjunta para ambos sexos. Para esto, se somete a una cohorte ficticia de 100000 individuos al patrón de mortalidad por edad definido, es decir, por las tasas específicas de mortalidad observadas sobre la población en estudio en el año de referencia y luego se calculan sobre la misma el resto de funciones biométricas de la tabla (Zarruk, Villegas y Ortiz 2011).

Las funciones básicas de la tabla de vida son m_x, q_x, l_x, d_x, L_x, T_x, y e_x. Sin embargo, la tabla de vida no siempre publica todas estas funciones. La definición y cálculo de las funciones de la tabla de vida en una tabla completa sería la siguiente:

- La tasa de mortalidad m_x, es la ocurrencia de muertes expresadas por persona-año en cada edad x,

$$m_x = \frac{d_x}{P_x}.$$

Para una cohorte ficticia con una incidencia de mortalidad según las tasas de mortalidad que se han definido anteriormente:

- La probabilidad de muerte q_x, es la probabilidad de morir en un determinado período para cada edad x.

$$q_x = \frac{d_x}{P_x + \frac{1}{2}d_x}.$$

- El número de defunciones teóricas d_x, es el número de muertes dentro de la cohorte ficticia para cada edad x,

$$d_x = l_x \cdot q_x.$$

- El número de supervivientes l_x, es el número de individuos de la cohorte ficticia que alcanza la edad x; l_0 es el número de nacidos que componen la generación y usualmente se le asigna un valor de 100 000 por convenio,

$$l_{x+1} = l_x - d_x.$$

- La población estacionaria L_x, es el tiempo total vivido por todos los individuos de la generación ficticia que tienen x años,

$$L_x = l_{x+1} + \frac{1}{2}d_x,$$

$L_w = \frac{d_w}{m_w}$, donde w representa la edad más avanzada en la tabla.

- El total de años vividos T_x, es el total de años vividos por todos los individuos de la generación ficticia que tienen x o más años de edad,

$$T_x = L_x + L_{x+1} + L_{x+2} + ... + L_w,$$

$T_w = L_w$, donde w representa la edad más avanzada en la tabla.

- La esperanza de vida e_x, es el promedio de años que le quedan para vivir a los supervivientes de edad x. La esperanza de vida al nacer (e_0) es el número medio de años vividos por una generación de nacidos bajo condiciones de mortalidad dadas.

$$e_x = \frac{T_x}{l_x}.$$

1.2.3 La tabla de vida abreviada

La tabla de vida abreviada muestra funciones basadas en los datos de mortalidad de las estadísticas vitales y el tamaño de la población obtenidos de los censos de población. En algunos países los censos se realizan aproximadamente cada 10 años, como Argentina, Brasil y México, entre otros, y en otro países los censos se llevan a cabo con intervalos superiores a 10 años, como es el caso de Colombia. Algunos países en desarrollo, debido a los errores en los registros vitales relacionados con la edad durante la recopilación de la mortalidad, suelen elaborar la tabla de mortalidad para intervalos de edad.

Como en el caso de las tablas de vida completas, las tablas de vida abreviadas se calculan para hombres y mujeres por separado, aunque también pueden calcularse para ambos sexos combinados.

En una tabla de vida abreviada, la interpretación de las funciones es similar al caso de la tabla de vida completa, los valores ${}_nm_x$, ${}_nq_x$, ${}_nd_x$, y ${}_nL_x$ se calculan para el intervalo de edad $[x, x+n)$ como se muestra en la Tabla 1.6:

Tabla 1.6: Funciones de la tabla de vida según la clasificación.

Tabla de vida completa	Tabla de vida abreviada	Interpretación
m_x	${}_nm_x$	Tasa de mortalidad entre la edad x y $x+n$
q_x	${}_nq_x$	Probabilidad de morir entre la edad x y $x+n$
l_x	l_x	Supervivientes a la edad x
d_x	${}_nd_x$	Defunciones entre la edad x y $x+n$
L_x	${}_nL_x$	Población estacionaria entre la edad x y $x+n$
T_x	T_x	Total de años vividos por personas de edad x o más
e_x	e_x	Esperanza de vida a la edad x

Fuente: Yusuf y col. 2014

A continuación se presentan las fórmulas de las funciones para la tabla de vida abreviada que son diferentes de las tabla de vida completa.

- La probabilidad de muerte ${}_nq_x$, se calcula a partir de la tasa de mortalidad ${}_nm_x$,

$$ {}_nq_x = \frac{n \cdot {}_nm_x}{1 + (n - {}_na_x) \cdot {}_nm_x}. $$

  donde a_x es el promedio de años vividos por los individuos que mueren en el intervalo de edad $[x, x+n)$.

- El número de muertes ${}_nd_x$, es el número de individuos de la generación ficticia que murieron durante el intervalo de edad $[x, x+n)$,

$$ {}_nd_x = l_x \cdot {}_nq_x. $$

- La población estacionaria ${}_nL_x$, es el tiempo total vivido por todos los individuos de la generación ficticia de $[x, x+n)$ años,

$$ {}_nL_x = n \cdot l_{x+n} + {}_na_x \cdot {}_nd_x. $$

Los intervalos comúnmente utilizados para agrupar las edades en una tabla de vida abreviada son $[0;1); [1;5); [5;10); [10;15); \ldots$; hasta el intervalo abierto final, porque normalmente las edades preferidas son las que terminan en múltiplos de cinco en una declaración de muerte. Para asegurar una visión más amplia de la dinámica de la mortalidad en una población, es necesario además visualizar la tendencia temporal de la incidencia de la mortalidad. Para ello se utilizan las tablas dinámicas de vida, que corresponden a la colección de tablas de vida de períodos, completas o abreviadas, obtenidas para cada año de un intervalo de tiempo. En adelante, el número total de intervalos de edad de cada período se indicará con p, y el número total de períodos analizados se indicará con m.

1.2.4 Gráficos de las columnas (funciones) de la tabla de vida abreviada

A partir de las funciones q_x, l_x, d_x y e_x de la tabla de vida se construyen gráficos que muestran la forma de estos datos tanto para hombres como para mujeres. Estos gráficos son aplicables a cualquier tabla de vida independientemente del genero, país o periodo de tiempo. La forma general de L_x es similar a la de l_x, mientras que la de T_x es similar a la de e_x, por lo que usualmente los gráficos para L_x y T_x no se construyen.

Para muchos países, las columnas de la tabla de vida para las mujeres tienen una forma diferente a las de las tablas de vida para hombres. Generalmente, los valores de q_x y d_x en las mujeres son más bajos que en los hombres. En consecuencia, los valores L_x, T_x y e_x de las tablas de vida femeninas son generalmente más altos (Yusuf y col. 2014).

1.3 Fuentes de datos demográficos

Diferentes fuentes de información se pueden usar para obtener datos demográficos provenientes de encuestas y censos nacionales. Según el Departamento Admistrativo Nacional de Estadística (DANE), Colombia cuenta con una larga tradición de censos de población desde el siglo XVIII, sin embargo la historia institucional de estadísticas comienza en 1906 cuando se creó la Dirección General de Estadísticas, vinculada al Ministerio de Hacienda, institución que fue convertida en 1923 en la Dirección General de Estadística, la cual tuvo como misión desde 1935 recopilar, analizar y producir las cifras oficiales del país Carmona-Fonseca (2005). Los últimos censos han sido realizados por el DANE, entidad independiente creada en 1953 y adscrita a la presidencia de la República.

La cifras para Colombia están disponibles en algunas web que mencionamos a continuación.

- Departamento Administrativo Nacional de Estadística (DANE).
  Esta Institución colombiana produce y difunde información estadística de calidad para la toma de decisiones y la investigación del país. Es la encargada de desarrollar el Sistema Estadístico Nacional y tiene como objetivos garantizar la producción, disponibilidad y calidad de la información estadística estratégica, dirigir, planear, ejecutar, coordinar, regular y evaluar la producción y difusión de información oficial básica (`www.dane.gov.co`).

- Latin American Human Mortality Database (LAHMD).
  Esta web es una plataforma donde se encuentran disponibles datos para cinco países de Latinoamérica: Argentina, Brasil, Colombia, México y Perú relacionados con el total de defunciones y causas de muerte, tanto a nivel nacional como regional para cada país. En esta página de fácil acceso, se muestran los datos bajo un mismo formato y de manera organizada, tanto por años como por temas, en los periodos disponibles para cada país, lo que ofrece cierta ventaja sobre otras fuentes (`www.lamortalidad.org`).

- Latin American Mortality Database (LAMBdA).
  Esta base de datos contiene información relacionada con censos y estadísticas vitales de la población y el número de muertes para 18 países: Argentina, Brasil, Chile, Colombia, Costa Rica, Cuba, República Dominicana, Ecuador, El Salvador, Guatemala, Honduras, México, Nicaragua, Panamá, Paraguay, Perú, Uruguay y Venezuela. Conjuntamente muestra tablas de vida cada cinco y diez años, tasas de mortalidad por causas de defunción y tablas de vida de cohortes reconstruidas (estimadas) para cohortes de individuos nacidos a principios del siglo pasado (`www.ssc.wisc.edu/cdha/latinmortality`).

- Banco Mundial.
  Esta página de acceso abierto y gratuito, contiene datos sobre desarrollo para muchos países en el mundo. Permite crear gráficos, descargar información en diferentes formatos y consultar los diferentes indicadores económicos, sociales y de salud, así como resultados de investigaciones relacionadas con los temas de estudio (`www.bancomundial.org`).

- Bases de datos y publicaciones estadísticas de la Comisión Económica para América Latina y el Caribe (CEPALSTAT).
  CEPALSTAT permite acceder a toda la información estadística para los países de América Latina y el Caribe recolectada, sistematizada y publicada por la Comisión Económica para América Latina y el Caribe (CEPAL). Además, facilita estadísticas e indicadores periódicos a través de la consulta de tablas y gráficos pre-definidos mediante una consulta en línea en tiempo real (`www.cepal.org/es/datos-y-estadisticas`).

1.3.1 Datos

En la construcción de la tabla de vida para Colombia, se utilizó la información proveniente de Latin American Human Mortality Database (LAHMD 2015), donde las defunciones se publican para el periodo 1970-2012, organizadas para grupos de edades desde 0 hasta 84 años. En cuanto a los datos referentes a la población, encontramos información sólo para los años censales 1973, 1985, 1993 y 2005.

Dado que solo se dispone de datos de población para esos cuatro años censales, se complementó la información haciendo uso de la interpolación lineal a trozos, utilizando la expresión (1.2) propuesta por Delwarde y Denuit (2003) para calcular la población entre censos (1974 a 1984, 1986 a 1992 y de 1994 a 2004),

$$P_{xt} = P_{xt_1} + \left(\frac{t - t_1}{t_2 - t_1}\right)(P_{xt_2} - P_{xt_1}), \quad t \in [t_1, t_2] \tag{1.2}$$

donde t_1 y t_2 son los dos periodos censales consecutivos.

La interpolación permitió obtener tablas de vida para el periodo 1973-2005 sustituyendo la cifras de defunciones y población en la expresión (1.2).

1.4 Resultados

El análisis de la información se realizó a partir de la tabla de vida abreviada para Colombia para el periodo 1973-2005. La Tabla 1.7 muestra un extracto de la tabla de vida abreviada para los hombres colombianos, año 2005.

Tabla 1.7: Extracto de la tabla de vida abreviada para los hombres colombianos para el año 2005.

Edad	n	mx	qx	lx	ndx	Lx	Tx	ex
0	1	0.01	0.01	100000.00	1449.11	98612.48	7136044.95	71.36
1	4	0.00	0.00	98550.89	322.28	393447.17	7037432.47	71.41
5	5	0.00	0.00	98228.61	171.26	490714.90	6643985.31	67.64
10	5	0.00	0.00	98057.35	183.59	489827.77	6153270.41	62.75
15	5	0.00	0.01	97873.76	821.30	487315.55	5663442.64	57.86
20	5	0.00	0.02	97052.46	1558.83	481365.25	5176127.10	53.33
25	5	0.00	0.02	95493.64	1620.52	473416.90	4694761.85	49.16
30	5	0.00	0.02	93873.12	1525.11	465552.84	4221344.95	44.97
35	5	0.00	0.02	92348.01	1451.14	458112.22	3755792.11	40.67
40	5	0.00	0.02	90896.87	1489.81	450759.83	3297679.90	36.28
45	5	0.00	0.02	89407.06	1789.78	442560.84	2846920.07	31.84
50	5	0.01	0.03	87617.28	2480.51	431885.10	2404359.22	27.44
55	5	0.01	0.04	85136.76	3232.39	417602.82	1972474.13	23.17
60	5	0.01	0.06	81904.37	4875.62	397332.80	1554871.31	18.98
65	5	0.02	0.09	77028.75	7058.18	367498.31	1157538.51	15.03
70	5	0.03	0.14	69970.57	10122.74	324546.01	790040.20	11.29
75	5	0.05	0.21	59847.83	12702.11	267483.87	465494.19	7.78
80	5	0.08	0.32	47145.72	15087.31	198010.32	198010.32	4.20

Para analizar la distribución poblacional de Colombia según edad y sexo, se utilizaron dos métodos que resultan de gran utilizad para representar gráficamente la composición por edad de una población: el gráfico de series temporales y la pirámide de población.

El gráfico de series temporales, también denominado gráfico de áreas apiladas al cien por cien, puede emplearse para representar los cambios temporales en la composición porcentual por edades. Por otra parte, la pirámide de población se construye a través de un histograma doble, uno para cada sexo, a ambos lados del eje de las ordenadas, donde se representan los grupos de edad (número de personas o proporciones) en orden ascendente desde el más bajo hasta el más alto (Siegel y Swanson 2004).

La Figura 1.1 muestra el cambio en la distribución porcentual de la población en amplios grupos de edad para Colombia desde 1973 hasta 2018.

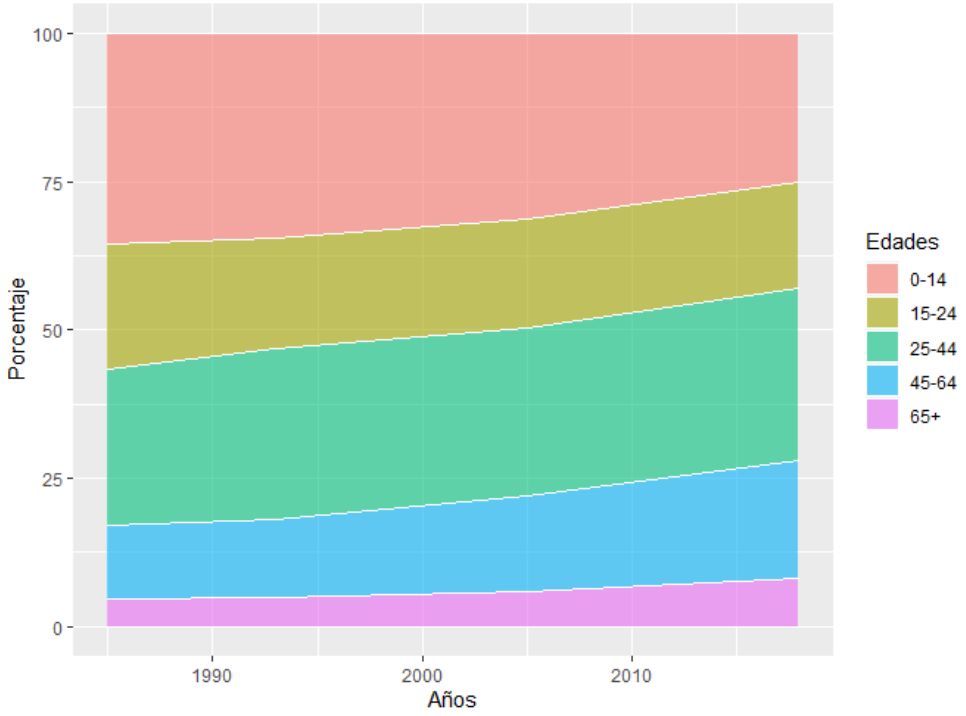
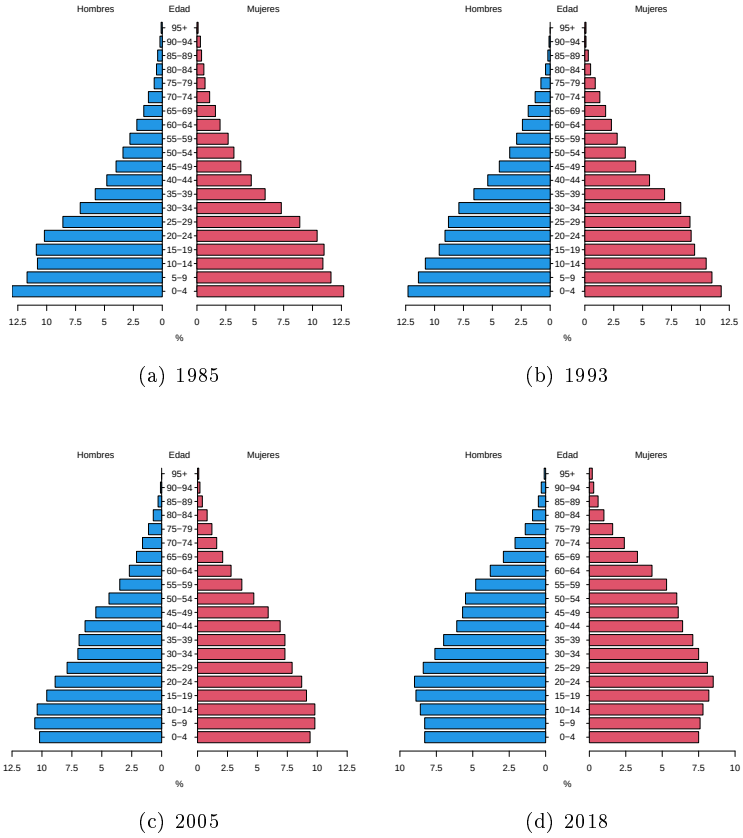


Figura 1.1: Gráfico de áreas apiladas al 100% para la distribución porcentual de la población por amplios grupos de edad para Colombia, 1985 a 2018.

La Figura 1.2 muestra la pirámide poblacional para Colombia en los últimos años censales: 1985, 1993, 2005 y 2018. Es evidente que la población ha venido creciendo gradualmente en estos 40 años (aproximadamente un 55%: de 31 millones de personas pasó a 48 millones), conjuntamente con un cambio

en la composición por edad y sexo de la población. Este cambio se evidencia en un cambio en la forma de la pirámide poblacional, que ha pasado de un perfil expansivo (típico de poblaciones jóvenes, con elevadas tasas de natalidad) a un perfil constrictivo (disminución de la natalidad y envejecimiento poblacional), donde los porcentajes de la población adulta se incrementan (como se observa en muchos países de Europa).



(a) 1985 (b) 1993

(c) 2005 (d) 2018

Figura 1.2: Distribución de la población por sexo y grupos quinquenales de edad para Colombia, 1985, 1993, 2005 y 2018

Según el DANE entre 1985 y 2018, la población colombiana aumentó en aproximadamente un 55%: de 31 millones de personas pasó a 48 millones. Este crecimiento se ha visto acompañado por un envejecimiento de la población, lo que se puede observar a través de los cambios en la estructura por grupos de

edad en la población, claramente reflejado en la pirámide de población (DANE 2018).

La Figura 1.3 muestra el comportamiento de las probabilidades de muerte, q_x, para cada sexo. Se puede observar que la probabilidad de muerte es muy alta en el grupo de los recién nacidos y luego decrece de manera abrupta a partir del primer año de vida cumplido del individuo. Posteriormente se estabiliza para edades intermedias, y vuelve a incrementarse a partir de los 60 años para ambos sexos. Además, podemos comprobar un descenso de la probabilidad de muerte en el tiempo, es decir, con cada nuevo censo.

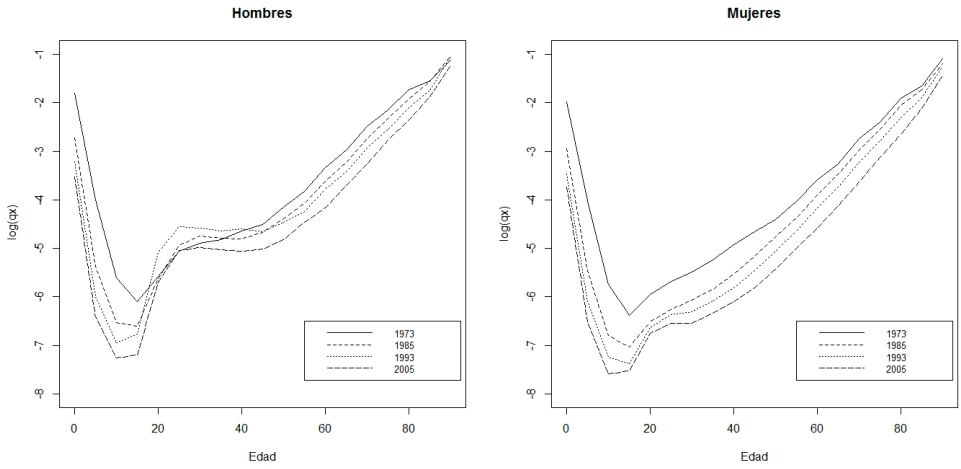


Figura 1.3: Probabilidades de muerte q_x en Colombia para los años censales.

Además, analizando la Figura 1.4, podemos mencionar que el fenómeno de rectangularización se nota de manera leve en la curva de supervivientes. La rectangularización se relaciona con el desplazamiento de la curva de supervivientes hacia edades avanzadas (Olivieri 2001), y se considera un fenómeno demográfico que se presenta en los países más desarrollados. En en el caso de Colombia, este fenómeno se observa más marcado en las mujeres.

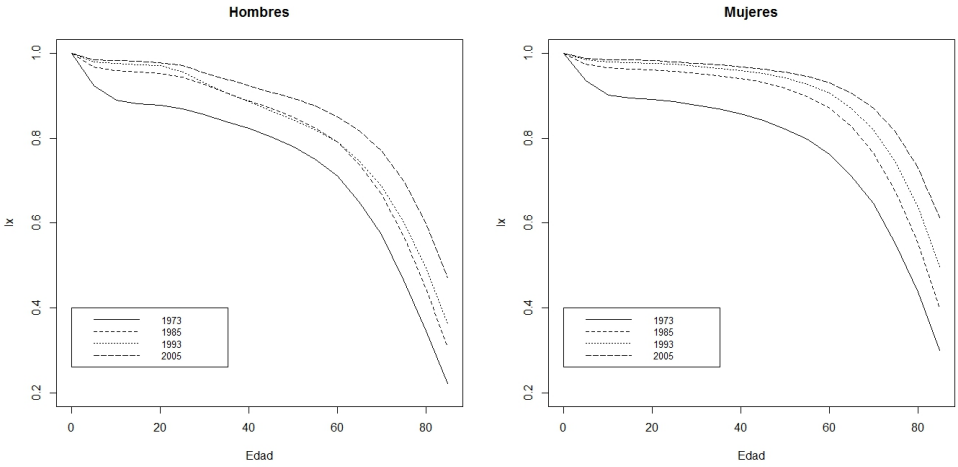


Figura 1.4: Supervivientes l_x en Colombia para los años censales.

La Figura 1.5 muestra la función de defunciones, d_x para cada sexo. En la representación de la defunciones suele observarse la edad modal de muerte, ubicada donde la función d_x alcanza el valor máximo. La edad modal es donde se produce la mayor cantidad de muertes en una población y generalmente toma valores entre los 65 y los 85 años.

En la población colombiana, aunque la edad modal de muerte a aumentado en el tiempo para ambos sexos, se nota como persiste en el tiempo un número importante de defunciones entre los 20 y 40 años en los hombres.

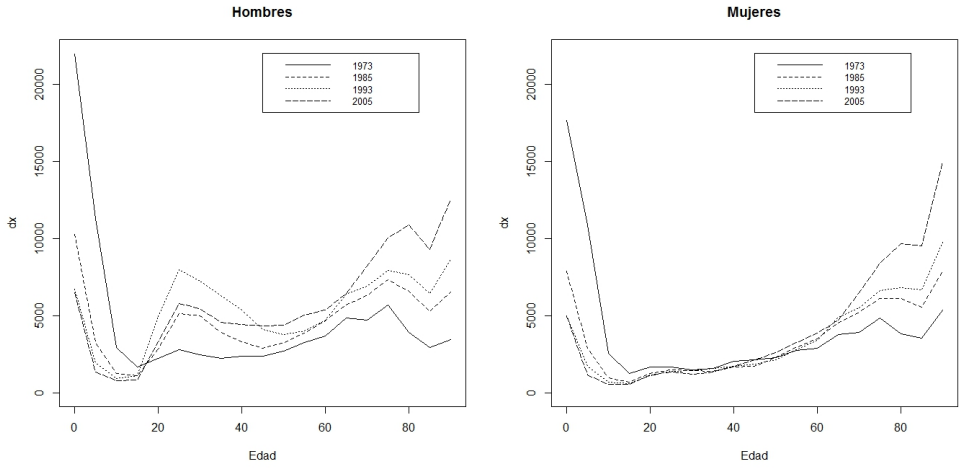


Figura 1.5: Defunciones d_x en Colombia para los años censales.

En la Figura 1.6 se representa la esperanza de vida (en años) la cual se incrementó de manera paulatina a través de los años tanto en hombres como en mujeres, pero presentando mayores valores para las mujeres en todas las edades y en los cuatro años considerados.

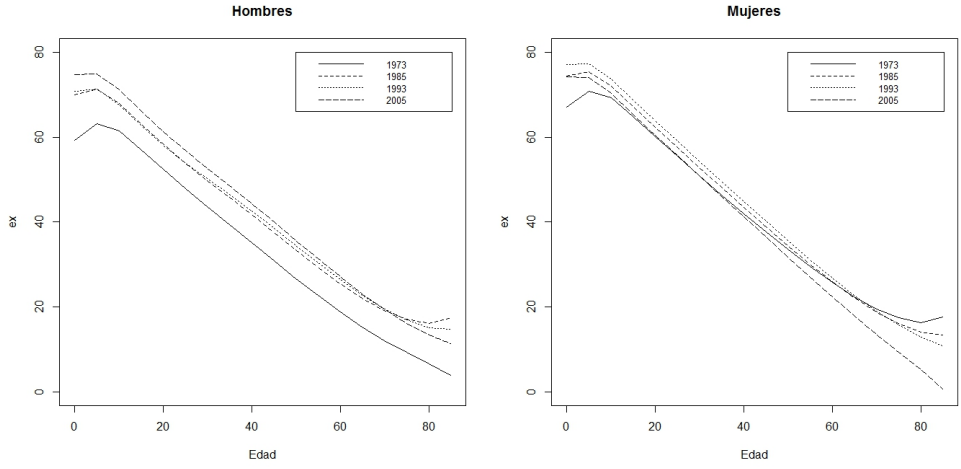


Figura 1.6: Esperanza de vida e_x en Colombia para los años censales.

1.5 Conclusiones

En este capítulo se exponen las principales características de la población colombiana en cuanto a mortalidad, para lo cual se analizó en primer lugar la distribución poblacional según edad y sexo. Se pudo evidenciar que la población colombiana ha crecido gradualmente en los últimos años, y ha experimentado un cambio en la composición por edad y sexo de la población, con un incipiente envejecimiento poblacional.

La construcción de la tabla de vida abreviada y sus funciones, posibilitó el análisis del comportamiento de la mortalidad y describir fenómenos demográficos.

Los patrones de mortalidad observados indican que:

- La mortalidad es mayor en las edades infantiles, luego se estabiliza a partir de los 15 años y posteriormente se incrementa conforme el individuo envejece.
- La mortalidad presenta una disminución con el tiempo independientemente de la edad.
- La mortalidad de las mujeres es inferior a la de los hombres.
- La mortalidad de los hombres colombianos es mayor entre los 15 y 40 años, lo que usualmente se conoce como "joroba social" asociada a una mayor exposición al riesgo.
- La edad modal de muerte y la esperanza de vida al nacer aumentaron a través de los años en ambos sexos.

Teniendo en cuenta algunas de las características descritas, podemos decir que la mortalidad en Colombia muestra similitudes con el comportamiento de la mortalidad de países desarrollados.

Capítulo 2

Pronóstico de la mortalidad en Colombia a partir de tablas de vida abreviadas por sexo

Parte del contenido de este capítulo se ha incluido en la publicación: Mortality forecasting in Colombia from abridged life tables by sex, en Genus, 74:15 (Díaz, Debón y Giner-Bosch 2018).

Un tema de interés para los países en transición demográfica es contar con modelo capaz de predicir la mortalidad y que permita a su vez analizar los diferentes cambios de la mortalidad en un periodo de tiempo para su población. Fenómenos como la reducción de la mortalidad, el envejecimiento y el aumento de la esperanza de vida, son de gran utilidad en la planificación de políticas públicas que buscan promover el desarrollo económico y social de los países.

En este capítulo se evaluó el desempeño de los modelos de predicción de la mortalidad aplicados a tablas de vida abreviadas. En este estudio se presentan diferentes modelos de mortalidad para tratar la modelización y el pronóstico de la probabilidad de muerte. Para la comparación de los modelos de mortalidad se analizaron dos criterios: el análisis de residuos gráficos y el método de hold-out para evaluar el rendimiento predictivo de los modelos, aplicando diferentes medidas de bondad de ajuste. Sólo tres modelos no tuvieron problemas de convergencia: Lee-Carter (LC), Lee-Carter con dos términos ($LC2$) y modelos de edad-período-costo (APC). Todos los modelos se ajustan mejor para las mujeres, la mejora de $LC2$ respecto a LC es sobre todo para las edades centrales para los hombres, y el ajuste del modelo APC es peor que los otros dos. El análisis de los residuos deviance estandarizados nos permite deducir que los modelos que se ajustan razonablemente a los datos de mortalidad colombiana son LC y $LC2$. Los modelos LC y $LC2$ presentan una mayor bondad de ajuste, identificando las principales características de la mortalidad para Colombia.

2.1 Introducción

El estudio de la mortalidad, sus características y las previsiones nos permiten comprender la dinámica de la población y sus tendencias. Fenómenos como el crecimiento de la población y la reducción de la mortalidad son de gran interés por el impacto económico y social que tienen en el desarrollo de los países.

En los últimos años se han desarrollado diferentes modelos para describir la mortalidad (Booth y Tickle 2008; O'hare y Li 2017). Los modelos para la estimación de tablas dinámicas de vida se utilizan para graduar las tasas de mortalidad crudas y para analizar el comportamiento de la mortalidad (Cairns y col. 2011; Villegas, Millossovich y Kaishev 2018). El modelo original de Lee-Carter (LC) (Lee y Carter 1992) es uno de los métodos más conocidos y aplicados en el área demográfica y actuarial en todo el mundo. Se han presentado numerosas extensiones y modificaciones de este modelo añadiendo más términos al modelo original.

Este modelo se ha utilizado para estudiar la mortalidad en países de América Central y del Sur. En México, García-Guerrero y Mellado (2012) y Aburto y García-Guerrero (2015) proyectan la mortalidad utilizando el modelo de Lee-Carter, mientras que Ornelas (2015) ajusta los modelos de Lee-Carter, Renshaw-Haberman y Age-Period-Cohort (APC) para obtener las tasas ajustadas para el mercado de seguros corregidas por la mortalidad general de México. En Argentina, la mortalidad ha sido estudiada por Belliard y Williams (2013), Andreozzi y Blaconá (2011), Andreozzi (2012) y Blaconá y Andreozzi (2014). En este último trabajo se presenta una descripción de la metodología de datos funcionales propuesta por Hyndman y Ullah (2007), que representa un avance sobre el modelo original de Lee-Carter ya que utiliza un suavizado no paramétrico para reducir la aleatoriedad inherente en los datos observados, y la descomposición de los componentes demográficos permite el uso de componentes principales clásicos (Blaconá y Andreozzi 2014).

Por otro lado, para Chile, Lee y Rofman (1994) extiende el modelo de Lee-Carter para resolver los problemas de datos de censo incompletos. Para Costa Rica, Aguilar (2013) utiliza dos variantes del modelo Lee-Carter para la estimación de la esperanza de vida: las dos proyecciones muestran un comportamiento muy similar y revelan valores más altos que los oficiales.

Además, cuando analizamos la mortalidad en América Latina, es importante mencionar el crecimiento de la delincuencia y las muertes violentas por homicidio en algunos países de la región. Según Levitt y Rubio (2000), las tasas de homicidio en Colombia están entre las más altas del mundo, siendo la tasa de

homicidio en Colombia tres veces más alta que la de Brasil o México, y diez veces más alta que la de Argentina o Uruguay. Garfield y Llanten (2004) analizan el hecho de que Colombia tiene el mayor nivel de muertes por homicidio y conflicto armado. Durante los años 1980-2003, muchas de las muertes fueron resultado directo del conflicto armado; otras estuvieron relacionadas con venganzas personales, vigilancia, ataques de venganza, fácil acceso a armas de fuego, competencia en el comercio de drogas ilícitas, y la impunidad de los servicios de represión.

Para analizar las características de la mortalidad y los fenómenos demográficos, hicimos previsiones de la mortalidad que proporcionaron varios indicadores demográficos. Estos se utilizaron para describir fenómenos como el envejecimiento, la transición demográfica, el nivel de vida o las desigualdades en materia de salud. Los indicadores que se incluyen en los estudios de mortalidad suelen provenir de indicadores de población, indicadores sociales o indicadores de nivel de vida, desigualdad y pobreza (Lora 2008). Entre los indicadores que se relacionan con la mortalidad y las tendencias actuales de la población se encuentran: la esperanza de vida al nacer, la esperanza de vida a los 65 años, la edad modal de la muerte, la curva de mortalidad de Lorenz y el índice de mortalidad de Gini.

2.2 Modelos de mortalidad

Los modelos para la estimación de las tablas dinámicas de mortalidad se dividen en dos grupos: modelos paramétricos (que pueden ser estructurales o no, dependiendo de si se asume que los parámetros han sido influenciados por el tiempo del calendario o si se incorpora el tiempo cronológico como una variable), o no paramétricos (generalizaciones de técnicas smoothing dependientes tanto de la edad como del tiempo). Ambos tipos de modelos ofrecen diversas herramientas para la construcción y graduación de tablas de mortalidad. Sin embargo, los modelos paramétricos suelen ser más utilizados en la actualidad pues es habitual que los parámetros permitan realizar predicciones sobre la mortalidad futura de una manera más sencilla que los métodos no paramétricos (Debón, Montes y Sala 2009).

El modelo Lee-Carter ha tenido una gran aceptación en el ámbito de las ciencias actuariales desde que consiguiera explicar en 1992 el 93% de la variación de los datos de mortalidad de EEUU, para el período 1933-1987 (Lee y Carter 1992).

El modelo Lee y Carter 1992 expresa la tasa de mortalidad, m_{xt}, como una medida que depende de la edad del individuo y del periodo de análisis correspondiente a través de una función exponencial de estas variables,

$$m_{xt} = \exp(a_x + b_x k_t + \epsilon_{xt}). \tag{2.1}$$

Una modificación de este modelo fue propuesta por Debón, Montes y Puig 2008 en la que se utiliza la transformación logit para la probabilidad de muerte, q_{xt}, ya que el modelo Lee-Carter original no garantizaba estimaciones para q_{xt} que no superaran el valor 1.

El modelo Lee-Carter modificado tiene la siguiente expresión

$$\text{logit}(q_{xt}) = \ln\left(\frac{q_{xt}}{1-q_{xt}}\right) = a_x + b_x k_t + \epsilon_{xt}, \tag{2.2}$$

donde a_x es el parámetro dependiente de la edad que describe el perfil general de la mortalidad a lo largo de la edad; b_x es el parámetro de sensibilidad dependiente de la edad que representa el cambio en la mortalidad a la edad x cuando la mortalidad cambia con el tiempo, y k_t es el índice de mortalidad, un parámetro que representa la tendencia de la mortalidad a lo largo del tiempo.

El modelo Lee-Carter con dos términos (LC2) representa un caso particular del modelo Lee-Carter generalizado con un término bilineal adicional, $b_x^2 k_t^2$, para modificar las tendencias de la mortalidad en el tiempo (Booth, Maindonald y Smith 2002). Este modelo se ha aplicado a datos de mortalidad de diferentes países europeos como es el caso de España (Debón, Montes y Puig 2008) e Italia (Carfora, Cutillo y Orlando 2017).

La expresión del modelo LC2 es:

$$\text{logit}(q_{xt}) = a_x + b_x^1 k_t^1 + b_x^2 k_t^2 + \epsilon_{xt}, \tag{2.3}$$

donde b_x^2 es un segundo parámetro dependiente de la edad que representa el cambio en la mortalidad a la edad x cuando la mortalidad cambia con el tiempo y k_t^2 es un segundo parámetro dependiente del tiempo que representa la tendencia de la mortalidad.

Por otra parte, Richards (2008) destaca la extraordinaria importancia de incluir el efecto cohorte en los modelos de mortalidad. La cohorte se define como el

año de nacimiento ($c = t - x$) en el estudio de los patrones de mortalidad para los actuarios. En Richards (2008) se encuentra una valiosa revisión de las técnicas utilizadas para identificar y modelar este efecto.

Otros modelos considerados en este trabajo incluyen el efecto de cohorte propuesto por Renshaw y Haberman (2006). En este sentido presentamos el modelo Lee-Carter con efecto de cohorte (LCC) que tiene la siguiente expresión:

$$\text{logit}(q_{xt}) = a_x + b_x^1 k_t + b_x^2 \gamma_c + \epsilon_{xt}, \tag{2.4}$$

y el modelo Edad-Período-Cohorte (APC) presentado en Tabeau (2001) que se obtiene cuando sustituimos $b_x^1 = 1$ y $b_x^2 = 1$ en la ecuación (2.4):

$$\text{logit}(q_{xt}) = a_x + k_t + \gamma_c + \epsilon_{xt}. \tag{2.5}$$

El modelo (2.4) es una extensión del modelo Lee-Carter (2.1) donde se añade un término bilineal, $b_x^2 \gamma_c$, para indicar un efecto de cohorte que muestra el comportamiento de la mortalidad por año de nacimiento (Renshaw y Haberman 2006). En este caso, b_x^2 es un parámetro de sensibilidad dependiente de la edad que representa el cambio en la mortalidad a la edad x en referencia a la mortalidad de la cohorte, y γ_c es un parámetro que representa la tendencia de la mortalidad a través de las cohortes. Cuando en este modelo el término $b_x^2 = 1$, entonces se obtiene el modelo de Renshaw-Haberman (RH). El modelo APC consiste en analizar de forma independiente el efecto de la edad, el periodo y la cohorte sobre la probabilidad de muerte (Currie, Durban y Eilers 2006).

El modelo de mortalidad Cairns-Blake-Dowd (CBD) sugerido por Cairns, Blake y Dowd (2006), propone una estructura de predicción con dos términos de edad-período, sin función de edad estática y sin efecto de cohorte:

$$\text{logit}(q_{xt}) = k_t^1 + (x - \overline{x}) k_t^2 + \epsilon_{xt}, \tag{2.6}$$

donde $\overline{x}$ es la edad media de los datos.

En Cairns y col. (2009) encontramos una generalización del modelo CBD donde se sugiere que el impacto del efecto de cohorte en una cohorte específica disminuye con el tiempo y por lo tanto se expresa como:

$$\text{logit}(q_{xt}) = k_t^1 + (x - \overline{x})k_t^2 + (x_c - x)\gamma_c + \epsilon_{xt}. \tag{2.7}$$

donde x_c es un parámetro constante a estimar. Este modelo se identifica generalmente como el modelo M8.

Las expresiones de los modelos descritos anteriormente se resumen en la Tabla 2.1 con sus respectivas restricciones para garantizar la identificabilidad de los modelos. Los modelos de las Tablas 2.1 y 2.2 fueron ajustados con los paquetes gnm (Turner y Firth 2015) y StMoMo (Villegas, Millossovich y Kaishev 2018) de R, respectivamente.

Tabla 2.1: Modelos de mortalidad ajustados con la librería gnm de R.

Modelo	Fórmula	Restricciones
LC	$\text{logit}(q_{xt}) = a_x + b_x k_t$	$\sum_x b_x = 1,\ k_{t_0} = 0$
LC2	$\text{logit}(q_{xt}) = a_x + b_x^1 k_t^1 + b_x^2 k_t^2$	$\sum_x b_x^i = 1, k_{t_0}^i = 0, i = 1, 2$
LCC	$\text{logit}(q_{xt}) = a_x + b_x^1 k_t + b_x^2 \gamma_c$	$\sum_x b_x^i = 1,\ k_{t_0} = 0,\ \gamma_{c_0} = 0$
APC	$\text{logit}(q_{xt}) = a_x + k_t + \gamma_c$	$k_{t_0} = 0,\ \gamma_{c_0} = 0$

Tabla 2.2: Modelos de mortalidad ajustados con la librería StMoMo de R.

Modelo	Fórmula	Restricciones
RH	$\text{logit}(q_{xt}) = a_x + b_x^1 k_t^1 + \gamma_c$	$\sum_x b_x^1 = 1, \sum_t k_t = 0,\ \sum_c \gamma_c = 0$
CBD	$\text{logit}(q_{xt}) = k_t^1 + (x - \bar{x})k_t^2$	No tiene
M8	$\text{logit}(q_{xt}) = k_t^1 + (x - \bar{x})k_t^2 + (x_c - x)\gamma_c$	$\sum_c \gamma_c = 0$

2.3 Comparación de modelos

Para la comparación de los modelos de mortalidad se analizaron dos criterios: el análisis gráfico de residuos y el método hold-out para evaluar el rendimiento predictivo de los modelos. Para este último método se aplicaron diferentes medidas de bondad de ajuste.

Para validar los resultados de las predicciones de los modelos generalmente se utiliza alguna de las siguientes estrategias:

- Evaluar el modelo en una muestra de prueba diferente a la muestra de ajuste,

- Desarrollar el modelo con el 75 % de la muestra y calcular el poder de predicción con el 25 % restante de los datos, o
- Utilizar la misma muestra que en el ajuste del modelo, pero calcular los indicadores de predicción mediante técnicas de bootstrap.

En este trabajo, utilizamos la segunda estrategia. En concreto, utilizamos el método hold-out, que separa los datos en dos subconjuntos, uno utilizado para entrenar el modelo y otro para realizar la prueba de validación (Blum, Kalai y Langford 1999). Utilizamos el 75% de los períodos originales para ajustar los modelos (conjunto de entrenamiento) y calculamos el poder predictivo con el 25% restante de los períodos (conjunto de validación).

Los pasos que se siguieron en el método hold-out fueron los siguientes:

- Se ajustaron los modelos de mortalidad para el conjunto de datos de entrenamiento.
- Se hicieron predicciones para los índices k_t, k_t^2 y γ_c, utilizando un modelo de serie temporal (ARIMA) para cada índice para el periodo de validación.
- Se generaron predicciones de probabilidad de muerte ($\hat{q}_{xt}$) con los índices predichos (obtenidos en el paso anterior) para el conjunto de datos de validación.
- Las predicciones del modelo ($\hat{q}_{xt}$) se compararon con las probabilidades de mortalidad observadas (q_{xt}) en el periodo de validación obteniendo medidas de bondad de ajuste.

Las medidas de bondad de ajuste utilizadas fueron la raíz del Error Cuadrático Medio (en inglés, RMSE) y el Error Porcentual Medio Absoluto (en inglés, MAPE), cuyas expresiones son:

$$\text{RMSE}(\hat{q}_{xt}) = \sqrt{\sum_x \sum_t \frac{(q_{xt} - \hat{q}_{xt})^2}{m \cdot p}}, \quad x = 1, ..., p; \quad t = 1, ..., m. \tag{2.8}$$

y

$$\text{MAPE}(\hat{q}_{xt}) = \frac{\sum_x \sum_t \frac{|q_{xt} - \hat{q}_{xt}|}{q_{xt}}}{m \cdot p} 100\%, \quad x = 1, ..., p; \quad t = 1, ..., m \tag{2.9}$$

donde p el número total de intervalos de edad para cada período y m el número total de períodos analizados.

Además, se realizaron comprobaciones de diagnóstico del modelo ajustado de manera gráfica, utilizando los residuos deviance estandarizados, ya que el uso exclusivo de medidas de bondad de ajuste no es un indicador de diagnóstico satisfactorio según se señala en los trabajos de Debón, Montes y Puig (2008) y Debón, Martínez-Ruiz y Montes (2012).

En el análisis gráfico de los residuos deviance estandarizados se evaluó su comportamiento con respecto a la edad, el periodo y la cohorte mediante gráficos de dispersión. Este análisis permitió analizar la variación de los residuos y se pudo percibir las mejoras producidas por algunos modelos en edades y años específicos. Dado que se supone que los residuos deviance estandarizados son independientes e idénticamente distribuidos según una distribución normal estándar de $N(0, 1)$, en esos gráficos se debe observar que los residuos están distribuidos aleatoriamente.

La expresión para los residuos deviance basados en una distribución binomial para el número de muertes es

$$r_{dev_{xt}} = sign(d_{xt} - \hat{d}_{xt}) \sqrt{2 \left[d_{xt} \log \left(\frac{d_{xt}}{\hat{d_{xt}}} \right) + (E_{xt} - d_{xt}) \log \left(\frac{E_{xt} - d_{xt}}{E_{xt} - \hat{d}_{xt}} \right) \right]},$$

donde d_{xt} denota el número observado de muertes y E_{xt} es el número inicialmente expuesto al riesgo a la edad x en el año t.

El intervalo de referencia $[-2, 2]$ para el 95,5% de los residuos de desviación estandarizados permite identificar los valores atípicos, aunque a veces se utiliza $[-2,5, 2,52]$ para captar el 99%.

2.4 Indicadores de mortalidad

El análisis de los indicadores de mortalidad es esencial para evaluar la situación social, económica y sanitaria de un país. Dentro de los indicadores demográficos básicos, encontramos los llamados indicadores de población que permiten describir las características estructurales y el comportamiento de una población. Este grupo incluye indicadores de nacimiento y fertilidad, tasas de mortalidad por edad y esperanza de vida, entre otros. Otro grupo de indicadores de mortalidad resume las asociaciones entre las desigualdades en salud y los indicadores socioeconómicos, como la Edad modal de la muerte, la curva de mortalidad de Lorenz y el Índice de mortalidad de Gini (Debón, Martínez-Ruiz y Montes 2012).

A continuación se definen los indicadores utilizados en este trabajo relacionados con la mortalidad.

- Esperanza de vida a la edad x (e_x).
  La esperanza de vida se puede calcular a partir de la tasa de mortalidad. Representa el número medio de años que les quedan por vivir a los supervivientes a la edad x en caso de prevalecer las condiciones de mortalidad existentes (INE 2020), su expresión es:

$$e_{xt} = \frac{T_{xt}}{l_{xt}}, \quad t = 1, ..., T \tag{2.10}$$

  donde T_{xt} corresponde al tiempo que le queda por vivir a los individuos de una generación desde los x años de edad hasta su completa extinción y l_{xt} el número de supervivientes de la misma a la edad x.

  En este trabajo se calculó la esperanza de vida al nacer, e_{0t}, y la esperanza de vida a los 65 años , e_{65t}, que se obtiene sustituyendo en la expresión (2.10) $x = 0$ y $x = 65$ respectivamente.

  La esperanza de vida al nacer se define como el número medio de años que vivirían los recién nacidos de una generación sometidos en cada edad a las condiciones de vida observadas en un determinado ámbito en el año t.De igual forma, la esperanza de vida a los 65 años se define como el número medio de años que viviría con 65 años cumplidos los componentes de una generación de individuos sometidos en cada edad a las condiciones de vida observadas en un determinado ámbito, a lo largo del año t.

- Edad modal de muerte.
  La edad modal de muerte (M_t) constituye un indicador de longevidad. Representa la edad a la cual se produce el máximo de defunciones de una población. En una tabla de mortalidad, este valor indica la edad a la cual fallecen la mayoría de los individuos de la cohorte ficticia inicial. Según Canudas-Romo (2008) la edad modal de muerte está influenciada en gran medida por la tasa de mortalidad en edades más avanzadas y por la mortalidad infantil. Por tanto, la edad modal de muerte puede reflejar cambios en la probabilidad de muerte que no se detectan con la esperanza de vida.

- Curva de Lorenz de mortalidad.
  La curva de Lorenz tiene su origen en un contexto económico y se considera fundamental a la hora de hacer un diagnóstico sobre la situación económica de un país y de sus políticas económicas y sociales (Lee 1997). Generalmente se utiliza para representar la distribución de la renta o el bienestar entre la población. Puede aceptarse que la renta se encuentra distribuida equitativamente entre los miembros de la población cuando a cada uno corresponde la misma fracción de la renta total (Lora 2008). En el contexto de la mortalidad, tenemos que la curva de Lorenz de mortalidad representa la distribución de la edad de muerte de los individuos de una población.

  Para obtener la curva de Lorenz de mortalidad, se ubica la proporción de fallecidos antes de la edad x en las abscisas frente a la proporción acumulada de años que esos individuos han vivido en las ordenadas. Luego se unen los puntos, quedando la curva siempre por debajo de la diagonal. Cuando el número de años vividos está repartido por igual entre los individos de la población, la curva de Lorenz coincide con la diagonal. En cambio, si el número de años vividos se concentra en un solo individuo, la curva de Lorenz recorre los ejes horizontal inferior y vertical derecho (Llorca, Prieto y Delgado-Rodriguez 2000).

  Las siguientes expresiones representan los años, g_{xt}, y la población de la curva,f_{xt}, respectivamente.

$$g_{xt} = \frac{T_{x_0 t} - T_{xt} - (x - x_0) l_{xt}}{T_{x_0 t}}$$

$$f_{xt} = \frac{l_{x_0 t} - l_{xt}}{l_{x_0 t}}.$$

- Índice de Gini de mortalidad.
  Según Singh y col. (2017) el índice de Gini se considera la medida más útil para analizar la desigualdad en la esperanza de vida y resume la curva de Lorenz de mortalidad. Se calcula como una función adicional de la tabla de mortalidad, evaluando así la desigualdad entre los individuos correspondiente a los años vividos por una persona hasta la muerte. Si el coeficiente de Gini de mortalidad es cercano a cero indica que todos los individuos mueren aproximadamente a la misma edad, mientras que si es cercano a uno indica que hay grandes diferencias en la edad de muerte entre los individuos de esa población, por consiguiente, una gran cantidad de individuos mueren a una edad muy temprana y, muy pocos individuos consiguen sobrevivir más que la media (Llorca y col. 1998).

  Existen diferentes alternativas para el cálculo del índice de Gini las cuales dependen de si los datos están agrupados o no. En una tabla de mortalidad completa, para su cálculo se requieren las tasas de mortalidad específicas por edad x (m_x), el número de supervivientes a la edad x (l_x) y el número total de años vividos a partir de la edad x (T_x). La expresión del índice de Gini a la edad x_0 en un año determinado t viene dada en Shkolnikov, Andreev y Begun (2003):

$$\mathrm{G_{xt}} = \frac{\sum_{x=x_0}^{w-1} (f_{xt} - g_{xt})}{\sum_{x=x_0}^{w-1} f_{xt}}, \quad t = 1, ..., T$$

  donde w representa la edad más avanzada en la tabla de vida.

  Otra expresión del índice de Gini suele utilizarse para las tablas de mortalidad abreviadas. Se presenta la propuesta de Rodríguez (2007), donde se calcula este indicador de mortalidad para datos de Colombia para el año 2000, lo que resulta un buen referente a la hora de evaluar nuestros resultados. Su expresión es la siguiente:

$$\mathrm{G_{xt}} = \left| 1 - \sum_{i=x}^{w} (N_i - N_{i-1})(Y_{i-1} + Y_i) \right|, \quad t = 1, ..., T \tag{2.11}$$

donde

$$N_i = \frac{\sum_{x=0}^{i} d_x}{\sum_{x=0}^{w} d_x}$$

es la proporción acumulada de fallecidos a la edad i, y

$$Y_i = \frac{\sum_{x=0}^{i} d_x \bar{x}}{\sum_{x=0}^{w} d_x \bar{x}}$$

es la proporción acumulada de años que estos individuos han vivido, w representa la edad más avanzada en la tabla de vida, $\bar{x}$ es la edad media de los individuos que mueren entre las edades exactas x y $x + n$, y d_x es el número de muertes hasta la edad i.

En este trabajo calculamos el índice de Gini al nacer (G_{0t}) y el índice de Gini a los 65 años (G_{65t}), los cuales se obtienen sustituyendo en la expresión (2.11) $x = 0$ y $x = 65$, respectivamente.

2.5 Resultados

2.5.1 Comparación de los modelos ajustados

Los diferentes modelos de mortalidad de las Tablas 2.1 y 2.2 fueron ajustados a los datos de Colombia de forma separada para hombres y para mujeres. Algunos de los modelos presentaron dificultades, las cuales se describen a continuación.

Según Holford (2006), los modelos con efecto cohorte pueden presentar problemas de estimación de los parámetros, sobre todo cuando los intervalos para

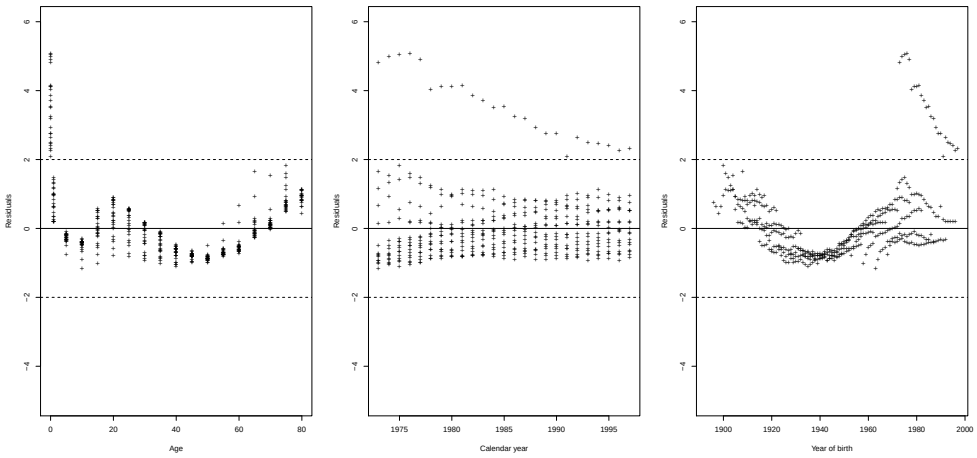
la edad o los períodos son de diferente amplitud. En nuestro trabajo se presentaron problemas de convergencia para el modelo LCC utilizando la librería gnm, y para los modelos RH y M8 con la librería StMomo con los datos para hombres. Este problema de convergencia para modelos de mortalidad con efecto cohorte ha sido señalado anteriormente por otros autores como Debón, Martínez-Ruiz y Montes (2010), Hunt y Villegas (2015) y Kennes (2017).

Por otro lado, el modelo CBD asume que la mortalidad es lineal en la escala logit, por lo que solo funciona bien para edades avanzadas, provocando residuos muy elevados a edades tempranas y mal comportamiento en general de los residuos (ver Figura 2.1).

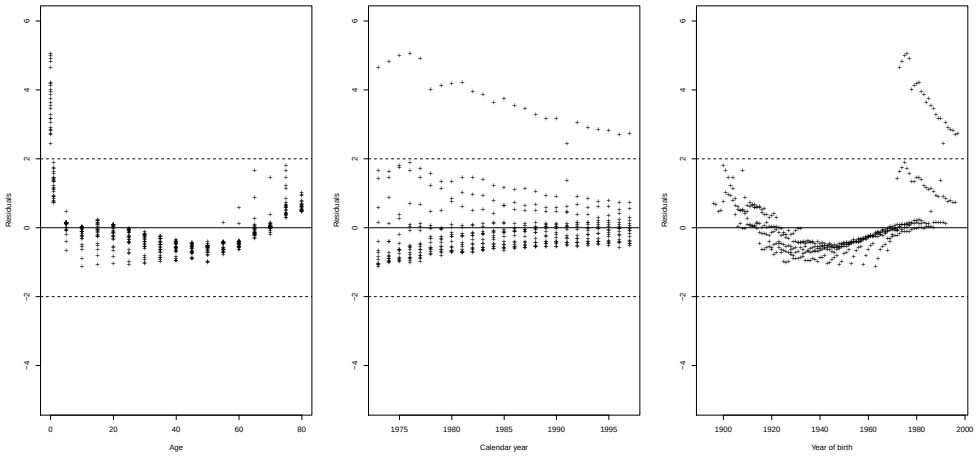
Resumiendo hasta aquí, podemos decir que para las tablas de vida abreviadas con datos de Colombia, los modelos LCC, RH y M8 presentaron problemas de convergencia para los hombres, y que el modelo CBD presenta un mal comportamiento lo que se evidencia en los gráficos de residuos. Estos modelos quedaron descartados en posteriores análisis.

Utilizando el método de hold-out descrito en la sección 2.3, se llevó a cabo una evaluación del ajuste y el rendimiento predictivo de los tres modelos de mortalidad que no presentaban problemas de convergencia: LC, LC2 y APC. Para estos tres modelos, tanto los valores ajustados como los proyectados se compararon con las probabilidades de muerte observadas en cada período mediante las medidas de bondad de ajuste RMSE y MAPE de las expresiones (2.8) y (2.9) respectivamente. Para este proceso, se utilizó el 75% de los períodos originales (años 1973-1997) para ajustar los modelos (conjunto de entrenamiento) y calculamos el poder predictivo con el 25% restante (conjunto de validación) de los períodos (años 1998-2005).

La Tabla 2.3 muestra que según las medidas de bondad de ajuste calculadas en el período de entrenamiento, el modelo LC2 es el que mejor ajusta los datos porque tiene los valores más bajos de RMSE y MAPE en ambos sexos. En cuanto al rendimiento predictivo de los modelos evaluados, podemos decir que LC2 tiene valores de MAPE más bajos (el mismo valor 12,63 en ambos sexos). Sin embargo, según los valores de RMSE, LC predice mejor para ambos sexos. En cuanto al modelo APC, podemos decir que presenta altos valores de RMSE y MAPE para ambos sexos en los dos conjuntos evaluados, por lo que fue descartado para el cálculo de los indicadores de mortalidad y para la evaluación gráfica de los residuos. Aunque el modelo APC presenta un peor ajuste, esto no implica necesariamente que el efecto de cohorte no sea importante, sino que es difícil de ajustar con tablas de vida abreviadas.



(a) Hombres



(b) Mujeres

Figura 2.1: Gráficos de dispersión de los residuos deviance estandarizados para el modelo CBD, librería StMomo, 1973-2005. Las líneas discontinuas representan el intervalo $[-2, 2]$.

Aunque los modelos LCC, RH y M8 se eliminaron de este análisis debido a los problemas de convergencia para los hombres, se muestran los resultados de estos modelos para las mujeres en la Tabla 2.3. Se puede observar que los

Tabla 2.3: Medidas de bondad de ajuste para los modelos de mortalidad ajustados.

	Conjunto de entrenamiento				Conjunto de validación			
	RMSE		MAPE		RMSE		MAPE	
Modelo	H	M	H	M	H	M	H	M
LC	0.11	0.09	5.89	8.28	0.01	0.01	15.66	18.70
LC2	0.08	0.07	4.01	6.49	0.01	0.01	12.63	12.63
APC	0.12	0.11	10.08	15.50	0.02	0.02	46.60	27.79
LCC	–	0.49	–	76.59	–	0.02	–	62.00
RH	–	0.10	–	9.87	–	0.31	–	19.24
M8	–	0.14	–	30.05	–	0.46	–	49.74

H (Hombres), M (Mujeres)

Tabla 2.4: Medidas de bondad de ajuste para los indicadores de mortalidad.

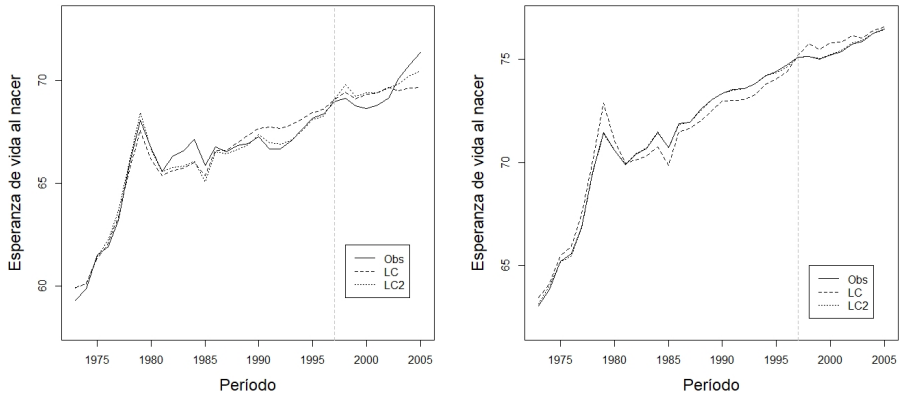
	Conjunto de entrenamiento				Conjunto de validación			
	RMSE		MAPE		RMSE		MAPE	
Indicador	H	M	H	M	H	M	H	M
e_{0t} - LC	0.56	0.39	0.68	0.37	0.92	1.08	0.97	1.41
e_{0t} - LC2	0.39	0.08	0.43	0.09	0.79	0.32	1.03	0.37
G_{0t} - LC	0.01	0.01	2.23	3.77	0.01	0.02	5.97	17.09
G_{0t} - LC2	0.01	0.00	1.95	1.77	0.01	0.01	6.07	10.73
e_{65t} - LC	0.15	0.15	0.91	0.89	0.11	0.08	0.71	0.43
e_{65t} - LC2	0.14	0.12	0.84	0.72	0.06	0.18	0.32	0.98
G_{65t} - LC	0.00	0.00	0.60	0.75	0.00	0.00	0.59	0.90
G_{65t} - LC2	0.00	0.00	0.59	0.63	0.00	0.00	0.63	0.91

H (Hombres), M (Mujeres)

valores de RMSE y MAPE de estos tres modelos son mayores que para los de los modelos LC y LC2 tanto en el conjunto de entrenamiento como en el conjunto de validación.

Además, se decidió evaluar el efecto del ajuste y la predicción con estos dos modelos (LC y LC2) en los indicadores de mortalidad. En la Tabla 2.4, e_{0t} es la esperanza de vida al nacer; e_{65t} es la esperanza de vida a los 65 años; G_{0t} es el índice de Gini al nacer y G_{65t} es el índice de Gini a los 65 años. De manera general podemos decir que en el conjunto de entrenamiento, el modelo LC2 presentó valores más bajos de RMSE y MAPE en ambos sexos. Sin embargo, en el conjunto de validación, es decir para las predicciones, el modelo LC tuvo mejor comportamiento para predecir en las mujeres en edades avanzadas.

Las Figura 2.2 muestra la comparación de la esperanza de vida al nacer utilizando los modelos LC y LC2 para hombres y mujeres, respectivamente. Para los hombres, en el periodo de validación, el modelo LC2 presenta valores más altos que el modelo LC, mientras que los datos observados muestran una trayectoria más errática, aumentando rápidamente su valor en los últimos años, algo que no recogen los modelos. Para las mujeres, en el periodo de validación, LC presenta una sobreestimación de la esperanza de vida al nacer, mientras que LC2 se acerca a los valores observados.



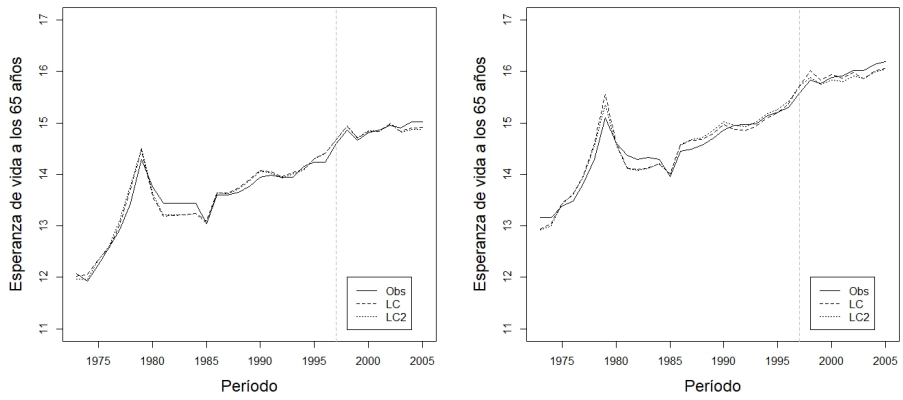
(a) Hombres (b) Mujeres

Figura 2.2: Comparación de la Esperanza de vida al nacer en el periodo 1973 a 2005.

Por otra parte, la Figura 2.3 muestran la comparación de la esperanza de vida a los 65 años utilizando los modelos LC y LC2 para hombres y mujeres, respectivamente. Las predicciones de ambos modelos se acercan a los datos observados para los hombres, mientras que para las mujeres LC2 subestima los valores en el periodo de validación.

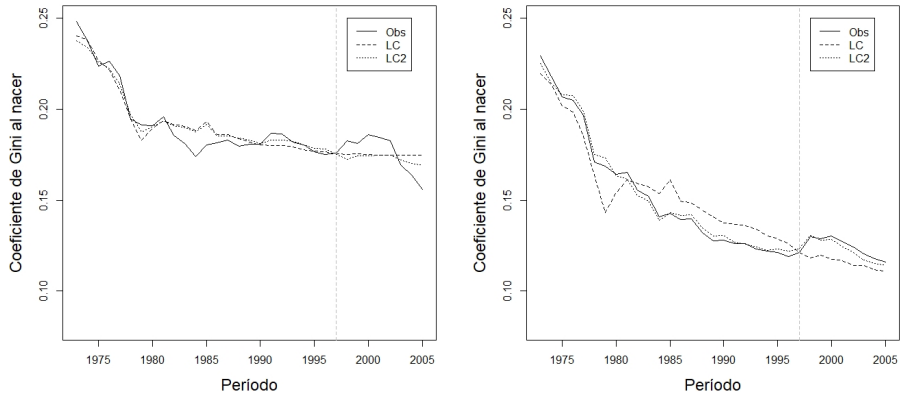
La Figura 2.4 muestran los valores del índice de Gini al nacer en hombres y mujeres. Para los hombres en el periodo de validación, los modelos no captan la tendencia decreciente presente en los datos observados. Para las mujeres, en ese periodo, ambos modelos presentan una subestimación, aunque muestran la tendencia decreciente presente en los datos observados.

La comparación del índice de Gini a los 65 años de los modelos en hombres y mujeres se muestra en la Figura 2.5. Para los hombres, en el periodo de



(a) Hombres (b) Mujeres

Figura 2.3: Comparación de la Esperanza de vida a los 65 años para en el periodo 1973 a 2005.

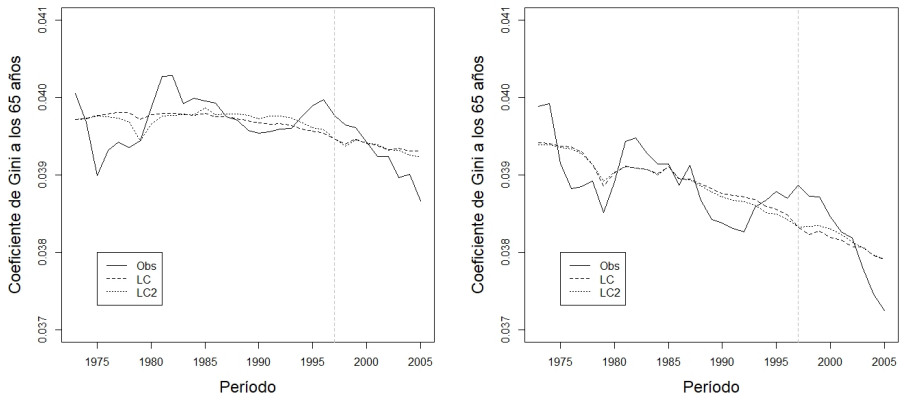


(a) Hombres (b) Mujeres

Figura 2.4: Comparación del Índice de Gini al nacer en el periodo 1973 a 2005.

validación, ambos modelos muestran una sobreestimación. Para las mujeres, en el periodo de validación, los modelos muestran la tendencia a la disminución

aunque no capturan la rápida caída en los últimos años presente en los datos observados.



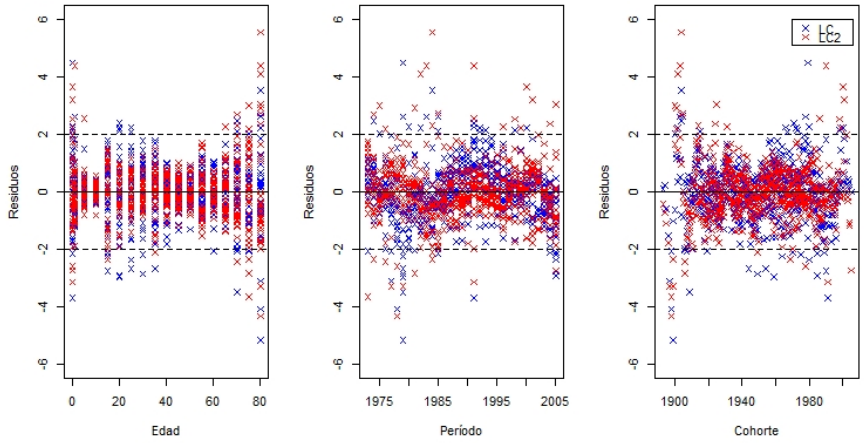
(a) Hombres (b) Mujeres

Figura 2.5: Comparación del Índice de Gini a los 65 años en el periodo 1973 a 2005.

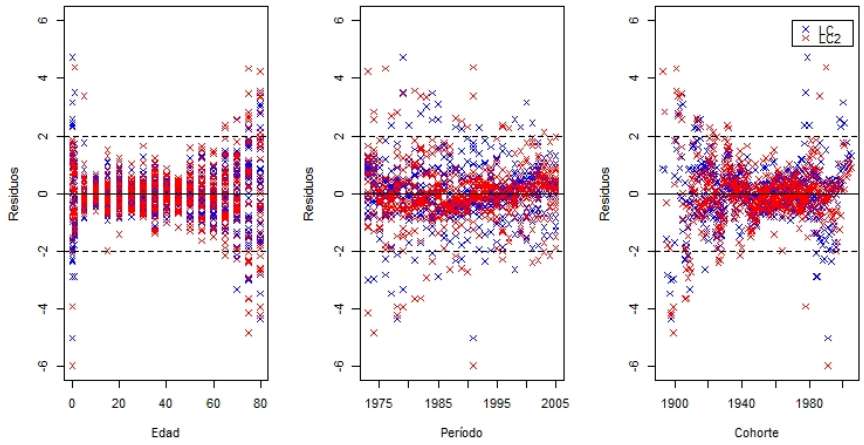
En general, LC2 no mejora las predicciones en los indicadores de mortalidad respecto a LC como podemos ver en las Figuras 2.2 y 2.3 para la esperanza de vida y en las Figuras 2.4 y 2.5 para el coeficiente de Gini, especialmente a los 65 años.

La Figura 2.6 muestra el comportamiento de los residuos de los modelos LC y LC2 frente a la edad, el periodo y la cohorte para hombres y mujeres. Se observa una mayor variabilidad en los residuos en las edades infantiles y en las edades avanzadas para ambos sexos.

En el caso de los hombres, se perciben valores elevados para las edades comprendidas entre los 15 y los 40 años, siendo el comportamiento del modelo LC2 mejor que el de LC sólo para estas edades. Además, los residuos que dependen del periodo y de la cohorte, presentan un comportamiento similar para ambos modelos. El análisis de los residuos permite afirmar que ambos modelos ajustan razonablemente los datos de la mortalidad colombiana.



(a) Hombres

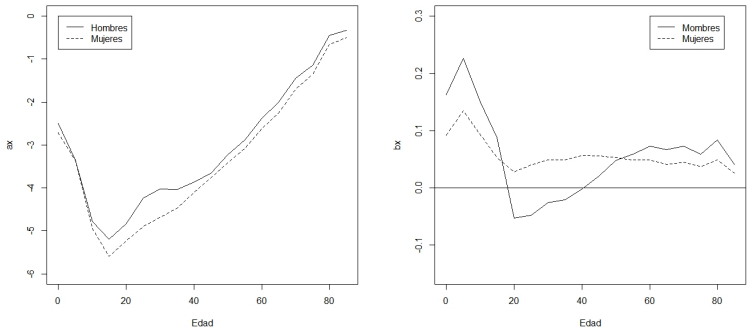


(b) Mujeres

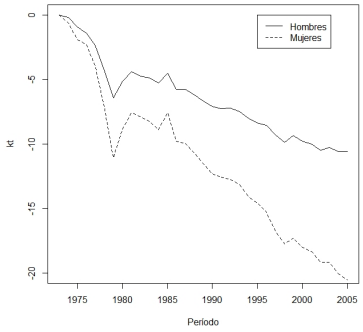
Figura 2.6: Gráficos de dispersiónpara los residuos deviance estandarizados de los modelos LC y LC2 para los años 1973 a 2005. Las líneas discontinuas representan el intervalo $[-2, 2]$.

2.5.2 Estimación de parámetros para los modelos seleccionados

El primer modelo ajustado a los datos de Colombia para el periodo 1973-2005 fue el modelo de LC. La Figura 2.7 presenta de forma gráfica las estimaciones de los parámetros del modelo LC, lo que permite obtener diferentes perspectivas del comportamiento de la mortalidad y evaluar las posibles diferencias entre las poblaciones de hombres y mujeres.



(a) a_x (b) b_x



(c) k_t

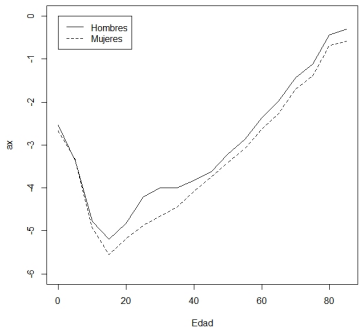
Figura 2.7: Parámetros del modelo LC ajustado a los datos colombianos para los años 1973 a 2005.

En la Figura 2.7 se observan las estimaciones de a_x, donde se puede apreciar las fases habituales de la mortalidad poblacional. Específicamente se observa que el riesgo de mortalidad desciende lentamente durante los primeros años de

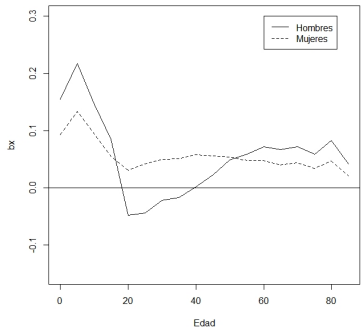
edad, sin que existen mayores diferencias entre hombres y mujeres. A partir de los 15 años de edad aproximadamente, el riesgo de mortalidad de las mujeres empieza a notarse inferior al de los hombres, ampliándose esta diferencia entre los 15 a 39 años de edad, y para las edades avanzadas el riesgo de muerte tiende a ser similar. A través de este parámetro se observa el fenómeno de "joroba" para la mortalidad, siendo más marcada en hombres con edades entre los 15 y 39 años. Este fenómeno que forma parte de la tendencia de la mortalidad en todos los países, conocido como "young adult mortality hump", se define como el exceso de mortalidad en un periodo generalmente corto en jóvenes adultos. Este fenómeno que históricamente se ha asociado con los accidentes de tráfico, en los últimos años se ha visto influenciado por enfermedades como el HIV, suicidios y homicidios (Remund, Camarda, Riffe y col. 2017). En Colombia la presencia de exceso de mortalidad masculina en los jóvenes se explica sobre todo por los homicidios o agresiones derivadas de hechos violentos aunque también se relacionan con los accidentes terrestres (Acosta y Romero 2014a).

Las estimaciones del parámetro b_x nos indica la forma en que responde la mortalidad de cada edad, x, a los cambios en k_t, es decir, a lo largo de los años. En las mujeres toma valores positivos para todas las edades, lo que indica que la mortalidad ha disminuido para todas las edades. En los hombres, la estimaciones de este parámetro tienen valores negativos entre los 15 y 39 años, indicando que la mortalidad aumenta para estas edades a través del tiempo. Las estimaciones del índice k_t muestran un comportamiento decreciente. Tanto en hombres como en mujeres la mortalidad disminuye, siendo este descenso mucho más notable en las mujeres. La mayor diferencia entre sexos se presenta en los últimos años del periodo analizado.

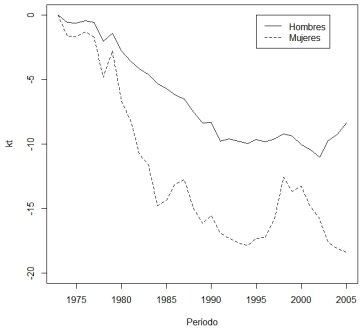
La Figura 2.8 presenta las estimaciones obtenidas de los distintos parámetros del modelo LC2. Los parámetros a_x y b_x tienen un comportamiento similar a lo observado con el modelo LC. El comportamiento decreciente del índice k_t hace evidente la tendencia de disminución de la mortalidad especialmente de las mujeres colombianas, pues para lo hombres se observa un pequeño repunte al final.



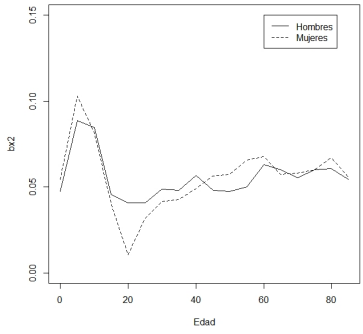
(a) a_x



(b) b_x



(c) k_t



(d) b_x^2

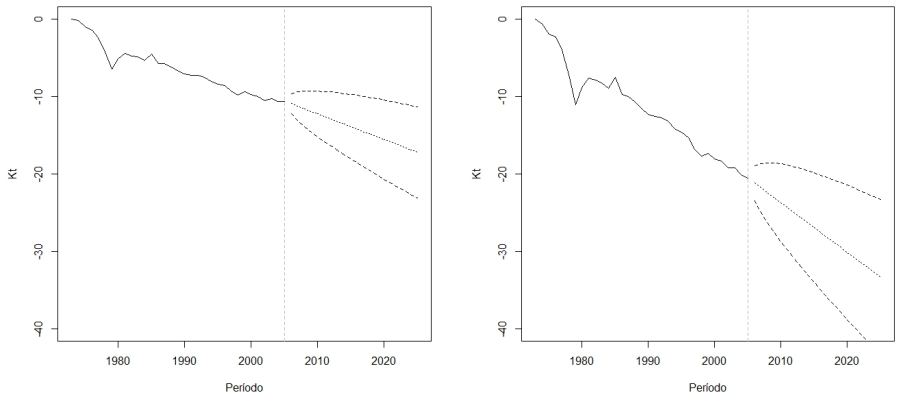
(e) k_t^2

Figura 2.8: Parámetros del modelo LC2 ajustado a los datos colombianos para los años 1973 a 2005.

Adicionalmente, se puede observar que el parámetro b_x^2 presenta valores más altos en las primeras edades hasta los 15 años y un valor constante para el resto de edades. Por otra parte, el parámetro k_t^2 presenta valores cercanos a cero, aunque entre 1975 y 1980 los valores decrecen de manera considerable en ambos sexos. Podemos decir que este segundo término (k_t^2) tiene efecto solo para unos pocos años en todas las edades, mejorando el ajuste de LC solo para esas observaciones.

2.5.3 Cálculo y proyecciones de Indicadores de mortalidad

Las proyecciones de los índices k_t, k_t^1 y k_t^2 de los modelos LC y LC2 para el periodo 2006-2025 se realizaron ajustando un modelo ARIMA a todo el periodo 1973-2005, utilizando la ecuación de predicción respectiva para cada caso como proyector de los valores futuros de estos índices. Los intervalos de confianza se obtuvieron según la propuesta original de Lee y Carter 1992, es decir, a partir de los errores de predicción de los índices k_t, k_t^1 y k_t^2 proyectados por los modelos ARIMA. Para la implementación se utilizaron las funciones auto.arima y forecast de la biblioteca forecast de R de Hyndman 2016.

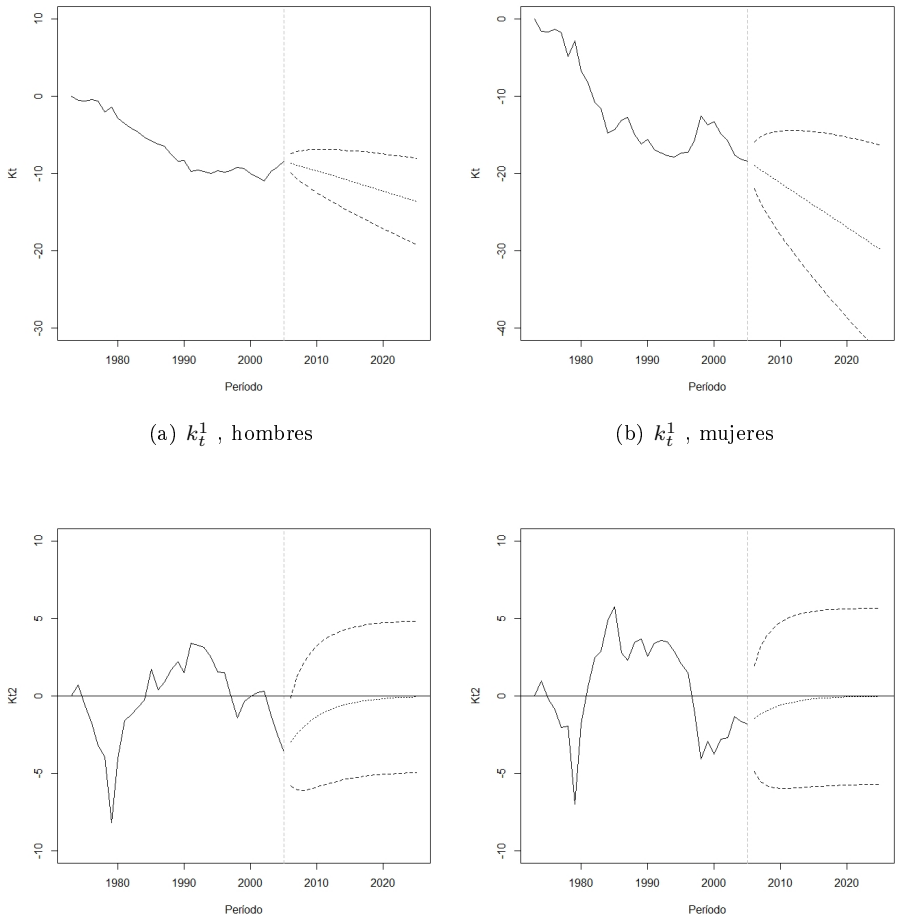


(a) Hombres (b) Mujeres

Figura 2.9: Proyección del índice k_t del modelo LC para el periodo 2006 a 2025 a Colombia. Las líneas discontinuas representan el pronóstico central y las líneas punteadas representan los intervalos de predicción del 95%.

La Figura 2.9 muestra los resultados de las predicciones del índice k_t utilizando el modelo LC para hombres y mujeres, con sus intervalos de confianza. Se

proyecta una tendencia a continuar con la disminución de la mortalidad tanto para hombres como para mujeres, aunque en las mujeres esta proyección tiene una reducción más marcada.



(a) k_t^1 , hombres

(b) k_t^1 , mujeres

(c) k_t^2, hombres

(d) k_t^2, mujeres

Figura 2.10: Proyección de los índices k_t^1 y k_t^2 del modelo LC2 para el periodo 2006 a 2025 a Colombia. Las líneas discontinuas representan el pronóstico central y las líneas punteadas representan los intervalos de predicción del 95%.

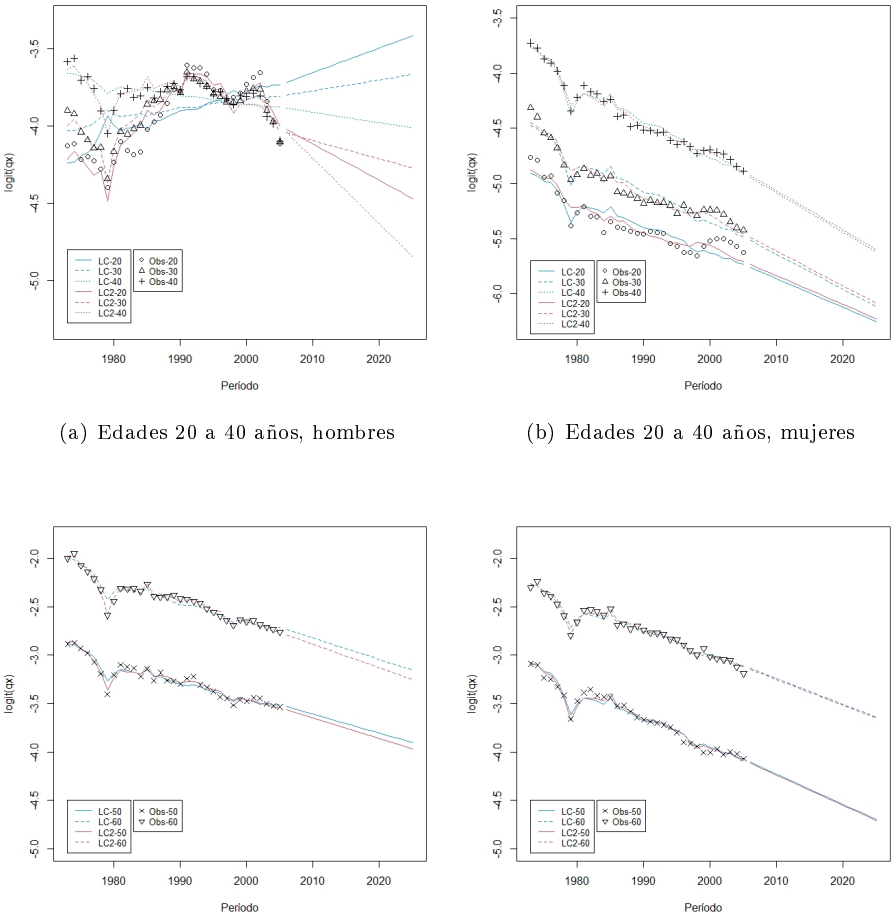
Los resultados de las predicciones de los índices k_t^1 y k_t^2 del modelo LC2 con sus intervalos de confianza se muestran en la Figura 2.10. Para el índice k_t^1, aunque sus valores aumentaron en los últimos años para los hombres, según el $ARIMA(0,1,0)$ ajustado hay una tendencia a disminuir. El índice k_t^2 tiende a cero en ambos sexos ($ARIMA(1,0,0)$). De este modo, los valores previstos indican una tendencia a la disminución de la mortalidad tanto para los hombres como para las mujeres.

Las probabilidades de muerte previstas para las edades de 20, 30 40, 50 y 60 años se muestran en la Figura 2.11 para hombres y mujeres. En el caso de los hombres, las probabilidades de muerte ajustadas para las edades de 20 y 30 años muestran grandes diferencias entre los modelos. Los valores pronosticados para el modelo LC2 muestran una disminución de las probabilidades de muerte, como ya hemos mencionado. Además, se confirma que el modelo LC2 se ajusta y predice mejor para los hombres, lo que indica que la inclusión del segundo término adapta mejor el modelo a los cambios de tendencia para las edades intermedias. Por otra parte, las predicciones de las probabilidades de muerte para las mujeres muestran una clara tendencia a la baja que se atenúa a partir de los 20 años. Para edades superiores, 50 y 60 años, casi no hay diferencias en el ajuste y la predicción de los modelos.

Igualmente, se calculó la esperanza de vida al nacer y el índice de Gini para el periodo analizado y se hicieron proyecciones hasta el año 2025. Esto con el fin de analizar las tendencias en los próximos años y su relación con los cambios demográficos que se están dando en Colombia.

La Figura 2.12 muestra que la esperanza de vida prevista para la población colombiana aumenta para ambos sexos para el periodo 1973-2025. Para los hombres el aumento fue de unos 10 años y para las mujeres de 13 años durante el periodo estudiado 1973-2005. Además, podemos decir que la esperanza de vida aumentará para ambos durante el periodo previsto 2006-2025. Los hombres tendrán un aumento de 7 años desde los 71 años y las mujeres experimentarán un aumento de 8 años desde los 76 años. Las mujeres tendrán una mayor esperanza de vida que los hombres (6 años más), manteniendo así la tendencia a vivir más tiempo.

Según la Figura 2.13, en ambos sexos se observa una ligera tendencia hacia la diagonal de las curvas de Lorenz para 2005, siendo más notable en el caso de las mujeres. Además, se observa que los niños pequeños y los jóvenes tienen una pequeña contribución en la distribución de los años que viven, lo que muestra una desigualdad en la edad de muerte (o esperanza de vida) de la población colombiana.



(a) Edades 20 a 40 años, hombres

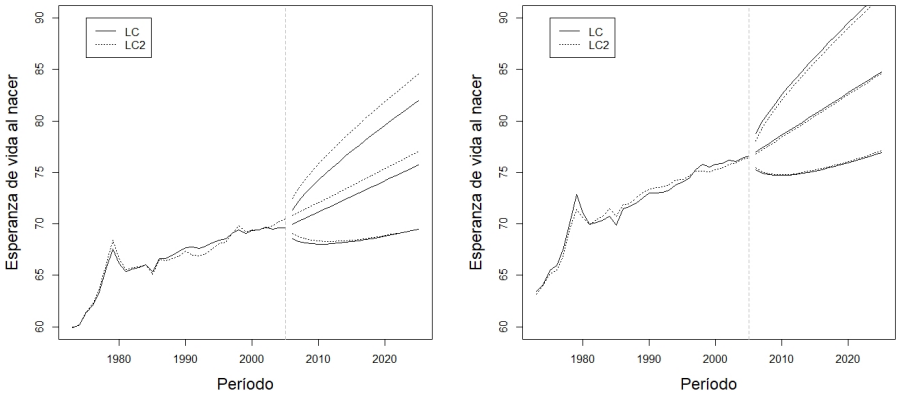
(b) Edades 20 a 40 años, mujeres

(c) Edades 50-60 años, hombres

(d) Edades 50-60 años, mujeres

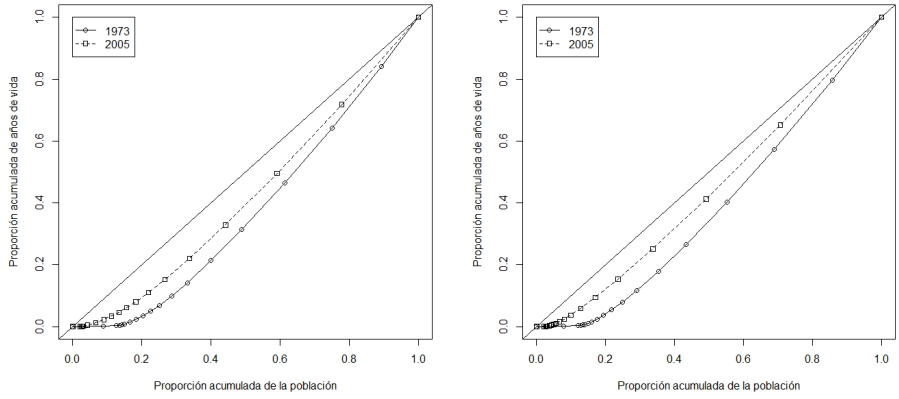
Figura 2.11: Probabilidades de muerte entre 1973 y 2005 y predicciones hasta 2025 para algunas edades a Colombia.

La Figura 2.14 muestra el comportamiento del índice de Gini, que disminuye para ambos sexos durante el período analizado, siendo la disminución mucho más marcada para las mujeres. Los valores disminuyen para los hombres de 0,24 en 1973 a 0,17 en 2005, y para las mujeres de 0,22 a 0,11. Por lo tanto, se observa que las desigualdades en la edad de fallecimiento son mayores para los



(a) Hombres (b) Mujeres

Figura 2.12: Evolución de la esperanza de vida al nacer, 1973 y 2025 a Colombia.

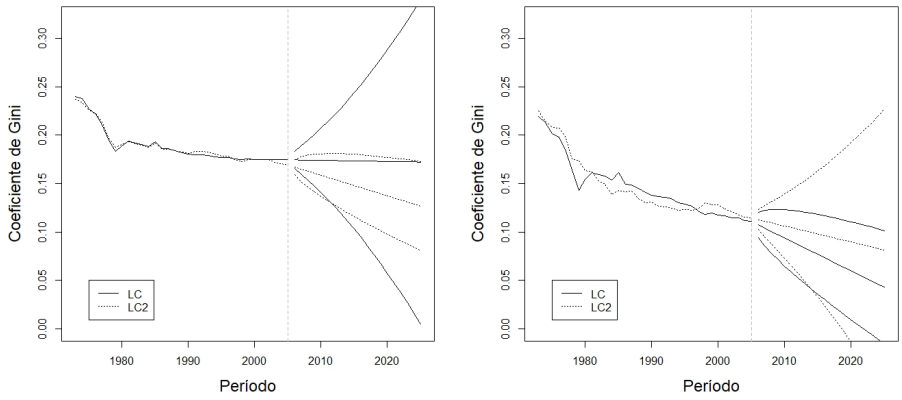


(a) Hombres (b) Mujeres

Figura 2.13: Evolución de la Curva de Lorenz para Colombia en los años censales seleccionados 1973 y 2005.

hombres que para las mujeres durante todo el periodo analizado. Además, se observa que la proyección es que la tendencia a disminuir se mantenga hasta el año 2025. Los resultados encontrados están en concordancia con el valor

reportado de 0,11 para Colombia en el año 2000 en Rodríguez (2007) y es consistente con el proceso de mejoramiento de la calidad de vida y la salud del país.



(a) Hombres (b) Mujeres

Figura 2.14: Evolución del índice de Gini entre 1973 y 2025 para Colombia.

Tabla 2.5: Evolución de la Edad modal de muerte para los años 1973 a 2005 en Colombia.

Año	Hombres	Mujeres	Año	Hombres	Mujeres
1973	75-80	75-80	1990	80-85	80-85
1974	75-80	75-80	1991	80-85	80-85
1975	75-80	75-80	1992	80-85	80-85
1976	75-80	75-80	1993	80-85	80-85
1977	75-80	75-80	1994	80-85	80-85
1978	75-80	75-80	1995	80-85	80-85
1979	75-80	75-80	1997	80-85	80-85
1981	75-80	75-80	1998	80-85	80-85
1982	75-80	75-80	1999	80-85	80-85
1983	75-80	75-80	2000	80-85	80-85
1984	75-80	80-85	2001	80-85	80-85
1985	75-80	75-80	2002	80-85	80-85
1986	75-80	80-85	2003	80-85	80-85
1987	75-80	80-85	2004	80-85	80-85
1988	75-80	80-85	2005	80-85	80-85
1989	75-80	80-85	-	-	-

Considerando la edad modal de muerte, podemos decir que para los hombres durante el periodo 1973 a 1989 el intervalo fue de [75,80) años, y entre 1990 a 2005, la edad modal de muerte aumentó al intervalo [80,85) años. En las mujeres, la edad modal de muerte se situó en el intervalo [75,80) años para el periodo de 1973 a 1983, mientras que en el período de 1984 a 2005, la edad modal de fallecimiento aumentó hasta el intervalo [80,85) años. Lo anterior refuerza la idea de que las mujeres han vivido más tiempo durante muchos más periodos en Colombia. Esta evolución se puede observar en la Tabla 2.5.

2.6 Conclusiones

La estimación de la mortalidad a partir de un buen modelo de previsión es importante teniendo en cuenta el impacto que sus resultados tienen en los diferentes procesos de planificación social y económica de un país. En algunos países en desarrollo, los datos de defunciones suelen darse en grupos de edad. Se trata de un fenómeno habitual en los registros vitales relacionado con los errores en la edad registrada (generalmente preferencias por las edades que terminan en múltiplos de cinco y algunas otras dificultades de registro). Por ello, una cuestión de interés en el ámbito demográfico y actuarial es la estimación y previsión del patrón de mortalidad mediante tablas de vida abreviadas.

En este capítulo se realizaron pronósticos de mortalidad en Colombia que muestran el comportamiento de la mortalidad para tablas de vida abreviadas. A diferencia de estudios anteriores para Colombia y otros países latinoamericanos, utilizamos una amplia variedad de extensiones del modelo Lee-Carter que nos permitieron seleccionar el modelo con mejor bondad de ajuste, y a partir de éste realizar pronósticos de la mortalidad y estimación de algunos indicadores. Es importante señalar que, hasta donde sabemos, la librería StMoMo de R no ha sido utilizado hasta la fecha para la graduación de los datos de mortalidad de Colombia, y que la librería gnm de R permitió ajustar algunas extensiones del modelo de Lee-Carter.

Como en muchos otros países del mundo, todos los modelos predicen mejor la mortalidad de las mujeres, ya que la experiencia de mortalidad de las mujeres tiene menos variabilidad. Además, es importante destacar el uso de estos 7 modelos en tablas de vida abreviadas y los resultados encontrados a pesar de la no convergencia de algunos modelos. En este estudio, los modelos presentaron problemas de convergencia con el efecto cohorte con ambos paquetes de R para los hombres, excepto el modelo APC. El problema de convergencia para los modelos de mortalidad con efecto cohorte ha sido señalado por otros autores

como Debón, Martínez-Ruiz y Montes (2010), Hunt y Villegas (2015) y Kennes (2017). Es importante señalar que el efecto de cohorte presenta problemas de estimación de los parámetros en las tablas de vida abreviadas, ya que las cohortes representan subconjuntos de cinco cohortes con diferentes números de observaciones. Por otra parte, el modelo CBD demostró un comportamiento muy malo para las edades infantiles y las edades avanzadas. Por lo tanto, la comparación se llevó a cabo ajustando los modelos LC, LC2 y APC. En resumen, podemos concluir que el modelo LC2 proporciona un mejor ajuste para ambos sexos, aunque la mejora de LC2 sobre LC es sobre todo para las edades intermedias.

Mediante el ajuste de los modelos de mortalidad LC y LC2 se identificaron algunas características de la mortalidad en Colombia:

- Analizando el comportamiento habitual de la probabilidad de muerte según la edad, se observa como la alta mortalidad en edades infantiles disminuye gradualmente hasta los 15 años y luego aumenta a medida que la población envejece. La mortalidad disminuyó significativamente en el periodo 1973 a 2005 para la mayoría de las edades con una pequeña tendencia al aumento en los últimos años para los hombres.

- El fenómeno de la joroba se observa para la mortalidad principalmente en los hombres de 15 a 39 años que se visualiza claramente por el modelo LC pero se descompone en dos términos para el modelo LC2. Esta sobremortalidad se explica principalmente por los homicidios o las agresiones derivadas de actos violentos, aunque también se relacionan con los accidentes de tráfico. Este patrón de mortalidad es más notable en Colombia que en otros países latinos. Según nuestros resultados, las probabilidades de muerte previstas son más factibles con LC2 que con LC, especialmente para los hombres. Sin embargo, podría ser necesario analizar los datos de un período más reciente para obtener estimaciones de los parámetros que den un pronóstico razonable a todas las edades.

- Fenómenos como la sobremortalidad de los hombres jóvenes (fenómeno de la joroba), que hacen que el comportamiento de la mortalidad sea diferente entre los sexos, son importantes para las compañías de seguros. Las tablas de vida son la herramienta que utilizan las aseguradoras para calcular el riesgo y valorar los productos que emiten en el mercado. En Colombia, las aseguradoras no tienen en cuenta el fenómeno de la joroba en los hombres jóvenes por dos razones. En primer lugar, para evitar que las personas tomen la decisión de posponer la contratación de un seguro hasta cumplir cierta edad para ahorrar dinero, hecho que podría suponer

que estas personas se queden sin asegurar durante muchos años, y en segundo lugar para evitar que se obtengan valores negativos a la hora de calcular el valor de la reserva que deben establecer las aseguradoras. La aplicación de esta medida en la tarificación puede ser importante para países con condiciones de desarrollo similares a las de Colombia para evitar el efecto de este fenómeno en la tarificación.

- La previsión de los indicadores demográficos y de mortalidad permite concluir que la población colombiana está inmersa en un fenómeno de mejora gradual de sus condiciones de vida. La esperanza de vida sigue siendo la medida de longevidad más conocida por los demógrafos, y aunque refleja los cambios en la mortalidad con el tiempo, lo hace de manera suave debido a su robustez. Por ello, en el presente trabajo se han estudiado otros indicadores: la edad modal a la muerte, la curva de Lorenz y el índice de Gini. La evolución de la edad modal al morir, la curva de Lorenz y el índice de Gini también confirmaron los cambios demográficos en Colombia. Se confirma una mayor longevidad en las mujeres que en los hombres, mostrando una mayor esperanza de vida y un menor índice de Gini. Por lo tanto, podemos concluir que LC2 no mejora las predicciones para los indicadores de mortalidad con respecto a LC para la esperanza de vida y el coeficiente de Gini especialmente a los 65 años, aunque LC2 es mejor para la predicción de probabilidades. Se puede apreciar que LC es bastante pobre en términos de predicción, especialmente en la clase de edad 20-40 años.

- Las diferencias que observamos en la disminución de la mortalidad y el aumento de la esperanza de vida entre sexos, deben ser tenidas en cuenta por las compañías de seguros colombianas para la elaboración de las tablas de vida y el cálculo de sus productos. Según la resolución 1555 de 2010 de la Superintendencia Financiera de Colombia, las tablas de vida que utilizan las entidades administradoras del Sistema General de Pensiones, el Sistema General de Riesgos Profesionales y las compañías de seguros de vida para la elaboración de sus productos y para los cálculos actuariales están discriminadas por sexo. Algo diferente ocurre en la Unión Europea donde según la junta 2004/113/CE del Tribunal de Justicia de la Unión Europea (UE), no se puede establecer una discriminación por razón de sexo en los bienes y servicios que impliquen la utilización de tablas de mortalidad unisex en el sector asegurador.

Por último, queremos señalar que aunque en este capítulo sólo aplica la graduación a las tablas de vida abreviadas colombianas, la metodología puede extenderse a las tablas de vida abreviadas de cualquier zona geográfica.

Capítulo 3

Gráfico de control multivariante y modelos Lee-Carter para estudiar los cambios de la mortalidad en Colombia

Parte del contenido de este capítulo se ha incluido en la publicación: Multivariate Control Chart and Lee-Carter Models to Study Mortality Changes, en Mathematics, 8, 2093 (Díaz-Rojo, Debón y Mosquera 2020).

La estructura de la mortalidad de una población suele reflejar el desarrollo económico y social del país. El propósito de este estudio fue identificar los momentos en el tiempo y los intervalos de edad en los que la probabilidad de muerte observada es sustancialmente diferente de la pauta de mortalidad de un período estudiado. Por consiguiente, se ajustó un modelo de mortalidad para descomponer el patrón histórico de mortalidad. Los residuos del modelo se vigilaron mediante el gráfico de control multivariado T^2 de Hotelling para detectar cambios sustanciales en la mortalidad que no fueron identificados por el modelo. El modelo Lee-Carter recoge información sobre la violencia en Colombia. Por lo tanto, los años identificados como fuera de control en las tablas se asocian con edades de muerte muy tempranas o bastante avanzadas y están inversamente relacionadas con la violencia que no se cobró tantas víctimas a esas edades. Los cambios en la mortalidad identificados en los gráficos de control corresponden a cambios en las condiciones de salud de la población o a nuevas causas de muerte como COVID-19 en los próximos años. La metodología propuesta se puede generalizar a otros países, especialmente a los países en desarrollo.

3.1 Introducción

El análisis de la mortalidad y sus tendencias temporales permite a un país comprender la dinámica de su población y constituye una guía fundamental para establecer la política económica y social. Hay una gran variedad de modelos de mortalidad para entender esa dinámica. Según Alexopoulos, Dellaportas y Forster (2019), el modelo de mortalidad más conocido, y el más exitoso en términos de generación de extensiones, es el modelo Lee-Carter (LC). Este modelo fue construido para descomponer el patrón histórico, obteniendo las tendencias de la mortalidad y su relación con la edad de la población.

El modelo LC propuesto en 1992 por Lee y Carter (Lee y Carter 1992) y sus diferentes extensiones o variantes se han aplicado para modelar y predecir la mortalidad en estudios de seguros y de población. En este sentido, la mayoría de las aplicaciones se han realizado en países desarrollados. Callot, Haldrup y Kallestrup-Lamb (2016) proponen una modificación del modelo LC que facilita la separación de la dinámica determinista y la estocástica; y proporcionan ilustraciones empíricas de datos de mortalidad para Estados Unidos, Japón y Francia para demostrar los avances del modelo modificado. Carfora, Cutillo y Orlando (2017) proponen una comparación cuantitativa de los principales modelos de mortalidad (incluido el modelo básico de LC) para evaluar tanto su bondad de ajuste como su rendimiento de previsión en los datos de la población italiana. Booth y col. (2006) comparan cinco variantes o extensiones del método Lee-Carter para la previsión de la mortalidad en poblaciones de 10 países desarrollados (Australia, Canadá, Dinamarca, Inglaterra, Finlandia, Francia, Italia, Noruega, Suecia y Suiza). Salhi y Loisel (2017) proponen un enfoque multivariante para predecir las tasas de mortalidad por pares de poblaciones relacionadas y realizaron una comparación con el modelo clásico de LC para los datos de Inglaterra y Gales. Recientemente, algunas investigaciones propusieron procedimientos alternativos al modelo clásico de LC para obtener la tasa de mortalidad. Postigo-Boix, Agüero y Melús-Moreno (2019) presentan funciones polinómicas en las que se reduce considerablemente la cantidad de datos necesarios para establecer la tasa de mortalidad.

También existen aplicaciones exitosas del modelo LC y sus diferentes versiones para los datos de mortalidad de diferentes países de América Latina. Algunos ejemplos son los trabajos de Andreozzi (2012) y Belliard y Williams (2013) para Argentina; García-Guerrero y Mellado (2012) para México; Lee y Rofman (1994) para Chile; Aguilar (2013) para Costa Rica; y Díaz, Debón y Giner-Bosch (2018) para Colombia. Estos trabajos muestran la utilidad de

los modelos de LC para analizar y modelar la mortalidad en los países en vía de desarrollo.

Como se ha señalado anteriormente, la mayoría de los trabajos publicados que se han consultado se han centrado en la modelización de la dinámica de la mortalidad. Sin embargo, el modelo Lee-Carter capta el patrón general de comportamiento de la mortalidad de la población, según la edad y a lo largo del tiempo, con un ajuste excelente o regular. Los patrones y cambios en la mortalidad se pueden analizar a través de las estimaciones obtenidas para los parámetros del modelo LC.

En ocasiones el modelo LC no reproduce correctamente la mortalidad observada, y parte de la información sobre ese fenómeno puede permanecer en el vector de residuos. Es precisamente aquí donde el gráfico de control desempeña un papel importante intentando descubrir otros cambios sustanciales en el comportamiento de la mortalidad que no hayan sido recogidos anteriormente por el modelo LC.

Por lo anterior, este capítulo va más allá de la modelización de la mortalidad y se propone el uso de los residuos de los modelos de LC para controlar e identificar situaciones de cambio sustancial en la tendencia de la mortalidad. Se quiere entonces, determinar si la probabilidad de muerte cambia significativamente a lo largo del período estudiado. Por esto, se identificaron los tiempos (años) e intervalos de edad con probabilidades de muerte que difieren significativamente del patrón de tendencia determinado por los modelos LC. Para cumplir con este objetivo, se utilizó el gráfico de control T^2 implementado sobre los residuos de los modelos LC complementado con la descomposición MTY (Mason, Tracy y Young 1995) para detectar el intervalo de edad en el que se produjo el cambio.

Un gráfico de control es una herramienta gráfica sencilla, propuesta inicialmente por Shewhart en 1927 (Shewhart 1927), para verificar la estabilidad temporal de un parámetro de interés en la distribución de probabilidad de una variable aleatoria. De esta manera, el gráfico de control univariante controla una sola variable. Luego, en 1947, Hotelling (1947) amplió la aplicación de los gráficos de control para el control simultáneo de dos o más variables aleatorias, creando el gráfico de control multivariado T^2.

Aunque los gráficos de control se propusieron inicialmente para vigilar los procesos industriales (control estadístico de la calidad), muchos trabajos evidenciaron su aplicación a otras áreas del conocimiento. Por ejemplo, en medicina, Woodall (2006) mencionó que los gráficos de control están vinculados a la vi-

gilancia de la atención sanitaria. En el reciente trabajo de revisión de Vetter y Morrice (2019) se propuso el uso de los gráficos de control como una herramienta que permite a los profesionales sanitarios comprender y comunicar los datos de rendimiento y mejorar la calidad de la atención al paciente, la anestesiología, la medicina perioperatoria, los cuidados críticos y el manejo del dolor. Algunos trabajos informan de su uso para monitorizar indicadores de rendimiento hospitalario (Benneyan, Lloyd y Plsek 2003), variables clínicas en pacientes (Alemi y Neuhauser 2004), y enfermedades crónicas e infecciosas (Imam y col. 2019; Williamson y Hudson 1999; Thacker y col. 1995), y para monitorizar la eficacia de los procedimientos quirúrgicos (Yue y col. 2017).

Además, la literatura ha puesto de manifiesto el uso de gráficos de control para los datos de mortalidad. Chamberlin y col. (1993) propusieron utilizar gráficos de control para determinar si la gravedad de las enfermedades de los pacientes y las tasas de mortalidad cambiaban significativamente en los cinco años comprendidos entre 1986 y 1990. Marshall y Mohammed (2007) utilizaron gráficos de control para supervisar las tasas de mortalidad tras la realización de bypass coronarios. Más recientemente, Urdinola y Rojas-Perilla (2013) proponen utilizar este enfoque para identificar el subregistro de la mortalidad en adultos en Colombia.

Los estudios sobre mortalidad son importantes para países como Colombia, teniendo en cuenta que se encuentra en el grupo de países latinoamericanos con las tasas de mortalidad más altas (Briceño-León, Villaveces y Concha-Eastman 2008). Además, hay que tener en cuenta su rápido aumento de la delincuencia violenta en el último siglo. Según Gaviria (2000), la tasa de homicidios comenzó su tendencia al alza a finales de la década de 1970 y se había más que triplicado a principios de la década de 1990. Para este período, la tasa de homicidio en Colombia era tres veces mayor que la de Brasil y México, siete veces mayor que la de Estados Unidos y 50 veces mayor que la de un país europeo típico.

En este contexto, los modelos de mortalidad ajustados para las tablas de vida abreviadas están diseñados para predecir simultáneamente un vector de tasas de mortalidad. Por lo tanto, cada vez que se realiza una predicción, se obtiene un vector de tasas de mortalidad estimadas, una para cada intervalo de edad, y consecuentemente, se genera un vector de residuos. Los residuos miden las desviaciones entre la tasa de mortalidad actual y la tasa esperada según el modelo. Un residuo elevado indica que la mortalidad actual no se corresponde con la tendencia observada. En consecuencia, se sospecha que existe un cambio en la mortalidad de la población para un rango de edad concreto.

Según esto, el seguimiento de la tabla de vida es un problema de control multivariante que consiste en controlar simultáneamente p variables aleatorias (cada residuo observado es una variable aleatoria). Por lo tanto, proponemos supervisar los residuos del modelo de mortalidad con un gráfico de control T^2, y sólo para los años identificados como fuera de control, utilizar el primer término de la descomposición MTY para identificar el rango de edad implicado en la señal de fuera de control.

Principalmente, Colombia ha presentado una serie de fenómenos demográficos en los últimos 60 años, como el aumento progresivo de la población, que se refleja en un cambio de la pirámide poblacional, y un aumento de la esperanza de vida, que se asocia principalmente a la disminución de la tasa de mortalidad y al envejecimiento de la población. Aunque la metodología se ilustra para los datos de Colombia, esta propuesta es estándar y generalizable a otros contextos similares que requieran explicaciones sociodemográficas de las "anomalías" detectadas con los gráficos de control.

3.2 Metodología

En esta sección, se describe brevemente las tablas de vida y los modelos Lee-Carter (LC) y Lee-Carter con dos términos (LC2). Luego, con más detenimiento, se presenta el gráfico de T^2 de Hotelling para las observaciones individuales y la descomposición MTY. Estos elementos teóricos apoyan nuestra propuesta de estudiar los cambios sustanciales en la mortalidad de los países en desarrollo.

3.2.1 Tablas de vida

Las tablas de vida periódicas, también conocidas como tablas de mortalidad, son una herramienta de análisis demográfico que resume la información de la incidencia de la mortalidad de una población para un periodo determinado. Las tablas de vida se clasifican según la longitud del intervalo de edad en el que se presentan los datos: "completas" cuando contienen datos para cada una de las edades desde el nacimiento hasta la última edad aplicable, y "abreviadas" cuando contienen datos en intervalos de edad, generalmente de 5 años para la mayor parte del rango de edad Siegel y Swanson 2004. Las funciones básicas de la tabla de vida son m_x, q_x, l_x, d_x, L_x, T_x y e_x. Sin embargo, la tabla de vida no siempre publica todas estas funciones. La interpretación de las funciones de la tabla de vida en una tabla completa sería la siguiente (Riva, Cantalapiedra y Lopéz 2010):

- La tasa de mortalidad o tasa de mortalidad m_x; la ocurrencia de muertes expresada por persona-año a cada edad x.

Para una cohorte ficticia con una incidencia de mortalidad según las tasas de mortalidad que se han definido:

- La probabilidad de muerte q_x es la probabilidad de que se produzcan muertes en un determinado periodo a cada edad x.
- El número de supervivientes l_x es el número de individuos de la cohorte ficticia que alcanzan la edad x.
- El número de defunciones d_x es el número de defunciones dentro de la cohorte ficticia a cada edad x. *La población estacionaria L_x es el tiempo total vivido para todos los individuos de la generación ficticia que tienen x años.*
- *El total de años vividos T_x es el total de años vividos para todos los individuos de la generación ficticia de x años o más.*
- *La esperanza de vida al nacer e_x es el número medio de años que les queda por vivir a los supervivientes a la edad x.*

La tabla de vida abreviada muestra las estimaciones basadas en los datos de mortalidad de las estadísticas vitales y el tamaño de la población obtenido de los censos de población. Los censos se realizan aproximadamente cada 10 años en algunos países, como Argentina, Brasil y México, entre otros, y algunos censos se realizan con intervalos mayores a 10 años, como es el caso de Colombia (Comisión Económica para América Latina y el Caribe (CEPAL) 2017). Los países en desarrollo, debido a los errores en los registros vitales relacionados con la edad durante la recolección de la mortalidad, suelen construir la tabla de mortalidad con intervalos de edad .

En una tabla de vida abreviada, la interpretación de las funciones es similar al caso de la tabla de vida completa, salvo que los valores ${}_nm_x$, ${}_nq_x$, ${}_nd_x$ y ${}_nL_x$ se refieren al intervalo de edad $[x, x+n)$:

- La probabilidad de muerte ${}_nq_x$ se calcula a partir de la tasa de mortalidad ${}_nm_x$:

  ${}_nq_x = \frac{n \cdot {}_nm_x}{1 + (n - {}_na_x) \cdot {}_nm_x}$, donde a_x es el número medio de años vividos por los individuos que mueren en el intervalo de edad y n es la amplitud del intervalo de edad.

- El número de muertes ${}_nd_x$, es el número de individuos de la generación ficticia que murieron durante el intervalo de edad $[x, x+n)$.
- La población estacionaria ${}_nL_x$, es el tiempo total vivido por todos los individuos de la generación ficticia de $[x, x+n)$ años.

Los intervalos comúnmente utilizados para agrupar las edades en una tabla de vida abreviada son $[0;1); [1;5); [5;10); [10;15); \ldots$; hasta el último intervalo abierto, ya que se prefieren las edades que terminan en múltiplos de cinco en una declaración de muerte. Para asegurar una visión más amplia de la dinámica de la mortalidad en una población, es necesario además visualizar la tendencia temporal de la incidencia de la mortalidad. Para ello, se utilizan tablas de vida dinámicas, que corresponden a la colección de tablas de vida periódicas, completas o abreviadas, obtenidas para cada año de un intervalo de tiempo. En adelante, el número total de intervalos de edad para cada período se denotará por p, y el número total de períodos analizados se denotará por m.

3.2.2 Modelos de Lee-Carter

Lee y Carter (1992) propusieron un método sencillo para modelar y pronosticar la mortalidad: un modelo de tasas de mortalidad por edad con un componente temporal y un componente fijo de edad relativa, y un modelo de series temporales (una media móvil integrada autorregresiva (ARIMA)) del componente temporal. Este método ofrece tres ventajas significativas: es un modelo demográfico parsimonioso combinado con métodos estadísticos estándar de series temporales, la previsión se basa en tendencias persistentes a largo plazo y se proporcionan intervalos de confianza probabilísticos para las previsiones (Booth, Maindonald y Smith 2002).

El modelo Lee-Carter (LC) expresa la tasa de mortalidad específica por edad como una medida que depende tanto de la edad de los individuos, x, como del periodo de tiempo, t. El modelo LC clásico se expresa como sigue:

$$\ln(m_{xt}) = a_x + b_x k_t + \epsilon_{xt}, \tag{3.1}$$

donde a_x es un parámetro específico de la edad que es independiente del tiempo (describe el perfil general de mortalidad según la edad), b_x es un parámetro específico de la edad que representa la rapidez o lentitud con que varía la mortalidad en cada edad cuando cambia el nivel general de mortalidad, y k_t es el índice general de mortalidad que depende del tiempo y refleja el nivel

general de mortalidad. Se supone que los errores ϵ_{xt} son variables aleatorias independientes e idénticamente distribuidas $N(0, \sigma^2)$.

El modelo LC tiene una estructura que es invariante bajo algunas transformaciones lineales de los parámetros. Por ejemplo, para cualquier valor de la constante c, se verifica que

$$\begin{aligned}(a_x, b_x, k_t) &\mapsto (a_x, b_x/c, ck_t)\\ (a_x, b_x, k_t) &\mapsto (a_x + cb_x, b_x, k_t - c).\end{aligned}$$

Para asegurar la identificabilidad del modelo, Lee y Carter (1992) propusieron incluir las siguientes restricciones en el modelo: $\sum_x b_x = 1$ y $\sum_t k_t = 0$.

Renshaw y Haberman (2003) desarrollaron una modificación del modelo Lee-Carter, denominada modelo Lee-Carter con dos términos (LC2). Indicaron que la interacción entre la edad y el tiempo puede captarse mejor añadiendo términos al modelo LC. El modelo LC2 se expresa de la siguiente manera:

$$\ln(m_{xt}) = a_x + b_x^1 k_t^1 + b_x^2 k_t^2 + \epsilon_{xt}. \tag{3.2}$$

En este capítulo, se ajustaron las ecuaciones (3.1) y (3.2) con la adecuación propuesta por Debón, Montes y Puig (2008), quienes sugieren modelar la probabilidad de muerte logit q_{xt} considerando una distribución binomial para la tasa de mortalidad. Así, el modelo LC se expresa como:

$$\text{logit}(q_{xt}) = \ln\left(\frac{q_{xt}}{1 - q_{xt}}\right) = a_x + b_x k_t + \epsilon_{xt}$$

con las restricciones $\sum_x b_x = 1$, $k_{t_0} = 0$; y el modelo LC2

$$\text{logit}(q_{xt}) = a_x + b_x^1 k_t^1 + b_x^2 k_t^2 + \epsilon_{xt}$$

con las restricciones$\sum_x b_x^i = 1$ $(i = 1, 2)$ y $k_{t_0}^1 = k_{t_0}^2 = 0$.

Los detalles sobre este ajuste utilizando R pueden encontrarse en Debón, Martínez-Ruiz y Montes (2010). Por último, los modelos de LC se basan en patrones históricos de mortalidad, y si las tendencias no se mantienen, entonces los modelos dejarán de ser válidos (Wang y Lu 2005).

3.2.3 Residuos deviance estandarizados

Los residuos son la base de la mayoría de los métodos de diagnóstico y suelen utilizarse para analizar la bondad del ajuste de los modelos de mortalidad. Sin embargo, como mencionamos anteriormente, los residuos pueden identificar momentos en el tiempo e intervalos de edad en los que la probabilidad de muerte observada es sustancialmente diferente del patrón de mortalidad para un periodo de tiempo. Con este objetivo, proponemos utilizar un gráfico de control multivariante T^2 de Hotelling.

En el análisis de bondad de ajuste de los modelos de mortalidad, se suelen utilizar los residuos deviance (Renshaw y Haberman 2008; Coelho y Nunes 2011; Debón, Montes y Puig 2008), teniendo en cuenta que los patrones en los residuos podrían indicar que el modelo no describe adecuadamente todas las características de los datos (Millossovich, Villegas y Kaishev 2018). Los residuos deviance basados en una distribución binomial para el número de muertes a la edad x son los siguientes:

$$r_{xt} = sign(d_{xt} - \hat{d}_{xt})\sqrt{\frac{2}{\hat{\phi}}\left[d_{xt}\log\left(\frac{d_{xt}}{\hat{d}_{xt}}\right) + (l_{xt} - d_{xt})\log\left(\frac{l_{xt} - d_{xt}}{l_{xt} - \hat{d}_{xt}}\right)\right]},$$

donde d_{xt} denota el número observado de muertes, $\hat{d}_{xt}$ son las muertes estimadas por el modelo, l_{xt} es el número de personas que viven al principio del intervalo de edad indicado, y $\hat{\phi}$ es un factor de escala empírico estimado por la expresión

$$\hat{\phi} = \frac{D(d_{xt}, \hat{d}_{xt})}{(m \cdot p) - \nu}.$$

Además, $D(d_{xt}, \hat{d}_{xt})$ es la desviación total del modelo, $m \cdot p$ es el número de observaciones en los datos, y ν es el número efectivo de parámetros (Millossovich, Villegas y Kaishev 2018).

Los residuos deviance suelen ser simétricos, pero su varianza y escala no son estándar. Por lo tanto, para corregir estas situaciones, los residuos deviance suelen estandarizarse.

Los residuos deviance estandarizados se definen por

$$str_{xt} = \frac{r_{xt}}{\sqrt{(1 - h_{xt})}}$$

donde h_{xt} es el leverage, la distancia entre una observación (x, t) y el centro de las observaciones.

Los residuos deviance estandarizados se distribuyen mediante una distribución normal estándar con varianza unitaria cuando el modelo ajustado es satisfactorio. Por esta razón, los valores de estos residuos estarán generalmente entre -2 y 2 (Collett 2003). Además, estos residuos cumplen los supuestos de los gráficos de control de Hotelling T^2. Teniendo en cuenta lo anterior, se utilizaron los residuos deviance estandarizados para controlar la tendencia de la mortalidad.

3.2.4 *Gráficos multivariado T^2 de Hotelling para la monitorización de residuos del modelo LC*

Los gráficos de control son útiles para determinar si un proceso ha estado en un estado de control estadístico mediante el examen de los datos históricos (Ryan 2011). Específicamente, los gráficos de control multivariados se utilizan para problemas de supervisión de procesos en los que varias variables relacionadas son de interés.

Hotelling (1947) propuso el gráfico de control T^2 para cumplir el objetivo de controlar simultáneamente $p \geq 2$ variables aleatorias, que generalmente tienen algún grado de asociación no despreciable. Bajo la suposición de que el vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)'$ sigue una distribución normal de $p$variables, $\boldsymbol{X} \backsim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con un vector de medias conocido $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$ y una matriz de covarianzas **Sigma**, la estadística:

$$T^2 = (\boldsymbol{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \tag{3.3}$$

sigue una distribución chi-cuadrado con grados de libertad p.

La carta T^2 es entonces un gráfico del estadístico T^2 frente al número de observaciones, con un límite superior de control (UCL) situado en $\chi^2_{(\alpha,p)}$, que representa el percentil superior α de la distribución chi-cuadrado. Aquí, α es la probabilidad deseada de error de tipo I.

Dado que $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ suelen ser desconocidos, estos parámetros deben estimarse a partir de una muestra de referencia compuesta por m observaciones de $\boldsymbol{X}$. El vector de la media de la muestra ($\bar{\boldsymbol{X}}$) y la matriz de covarianza ($\mathbf{S}$), obtenidos a partir de la muestra de referencia, son los estimadores de $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ respectivamente, para la Ecuación (3.3).

Según Tracy, Young y Mason (1992), cuando se estiman $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, el estadístico T^2 de Hotelling

$$T^2 = (\boldsymbol{X} - \bar{\boldsymbol{X}})'\mathbf{S}^{-1}(\boldsymbol{X} - \bar{\boldsymbol{X}}) \backsim \frac{(m-1)^2}{m} B_{(p/2,(m-p-1)/2)},$$

donde $B_{(p/2,(m-p-1)/2)}$ representa una distribución beta con parámetros $p/2$ y $(m-p-1)/2$. Esta distribución depende del número de variables p y del tamaño de la muestra m, que debe satisfacer $m > p+1$ (Mason y Young 2002).

Por lo tanto, el límite superior de control (UCL) para la carta T^2 debe situarse en:

$$\text{UCL} = \frac{(m-1)^2}{m} B_{(\alpha,p/2,(m-p-1)/2),}$$

donde $B_{(\alpha,p/2,(m-p-1)/2)}$ es el percentil superior α de una distribución beta con parámetros $p/2$ y $(m-p-1)/2$.

Cuando un punto sobrepasa el límite superior de control en un gráfico T^2, se interpreta como una señal de cambio en la distribución de $\boldsymbol{X}$. Entonces, se recomienda realizar una investigación para encontrar las causas que producen la señal o aplicar algún procedimiento para identificar la(s) variable(s) causal(es) de la señal, ya que el gráfico T^2 por sí mismo no puede hacerlo.

En nuestra propuesta, los cambios en la dinámica de la mortalidad se evalúan a través del seguimiento del vector p-dimensional de residuos deviance estandarizados $\boldsymbol{R_t} = (str_{1t}, str_{2t}, \ldots, str_{xt}, \ldots, str_{pt})'$, que se obtuvo a partir de un modelo Lee-Carter ajustado para una muestra de referencia de m períodos consecutivos ($t = 1, 2, 3, \ldots, m$). En esta aplicación particular, el vector medio $\bar{\boldsymbol{R}}_{\boldsymbol{t}}$ y la matriz de covarianza $\mathbf{S_R}$ se estiman a partir de la muestra de referencia de los vectores de residuos estandarizados históricos m. Con este enfoque, el estadístico T^2 de Hotelling adopta la forma

$$T^2 = (\boldsymbol{R} - \bar{\boldsymbol{R}})'\mathbf{S_R}^{-1}(\boldsymbol{R} - \bar{\boldsymbol{R}}).$$

En este contexto de aplicación, una observación que supera el límite de control del gráfico T^2 de Hotelling se interpreta como una desviación del patrón de tendencia de la mortalidad reproducido al ajustar cualquier modelo Lee-Carter. En consecuencia, este periodo se etiqueta como fuera de control, se sospecha un cambio sustancial en la dinámica de la mortalidad y la segunda fase del análisis investiga los intervalos de edad que pueden estar implicados en el diagnóstico fuera de control.

Como puede verse, la propuesta de control multivariante es más flexible en sus supuestos que el análisis habitual de los residuos, ya que ahora se relaja la condición de independencia completa impuesta al término de error. Bajo el enfoque de control multivariante, cada conjunto de p-residuos, asociados a los p-intervalos de edad para un año concreto, forman un vector de p-variables aleatorias que no son necesariamente independientes, ni están idénticamente distribuidas. En este sentido, el gráfico de control multivariante permite aplicar la metodología incluso cuando el modelo presenta problemas de ajuste local. Obsérvese que un caso particular de la estrategia multivariante se constituye cuando las p-residuos son independientes e idénticamente distribuidos.

Otra diferencia entre estas estrategias para identificar las anomalías está relacionada con el número de evaluaciones realizadas en la prueba de hipótesis. En el análisis de los residuos, cada residuo se comprueba individualmente con respecto a los límites de variación tolerables, lo que implica hacer un conjunto de $m \times p$ comparaciones. Por el contrario, en el enfoque multivariante, la comparación con el límite de control se realiza para cada observación de dimensión p: es decir, una comparación para cada año. Entonces, el gráfico de control multivariante reduce el número de comparaciones a m, lo que reduce la probabilidad de error global de tipo I. Como se sabe, la probabilidad de error global aumenta exponencialmente en función del número total de comparaciones realizadas simultáneamente (Mason y Young 2002).

Por último, cabe señalar que el rendimiento de un gráfico de control Hotelling T^2 está relacionado con el número de períodos m que componen el período de observación al que se ajusta el modelo de Lee-Carter. Para el gráfico de control Hotelling T^2, m define el tamaño de la muestra utilizada para estimar los parámetros de la distribución de probabilidad multivariante del vector de residuos estandarizados. En este sentido, a través de estudios de simulación, Champ y Jones-Farmer (2007) demostraron que cuando el tamaño de la muestra m es pequeño, la tasa de alarma verdadera-falsa de los gráficos de control multivariante suele ser sustancialmente mayor que la tasa nominal establecida. Una recomendación de estos autores es utilizar límites de control más amplios. El efecto del error de estimación sobre la tasa de falsas alarmas se absorbe sin afectar sustancialmente al rendimiento del gráfico en la detección de cambios.

3.2.5 Descomposición MTY

Se presentan varios enfoques para el problema de la interpretación de una señal multivariante. Mason, Tracy y Young (1995) propusieron el método MTY de descomposición para encontrar las causas que producen la señal. El método MTY descompone la estadística T^2 de Hotelling en p componentes ortogonales aditivos, cada uno de los cuales revela la contribución de la variable de proceso individual y la contribución conjunta relativa de la misma variable de proceso.

Para el caso de p variables, hay $p!$ diferentes descomposiciones MTY posibles. Una de estas descomposiciones viene dada por

$$T^2 = T_1^2 + T_{2.1}^2 + T_{3.1,2}^2 + \ldots + T_{p.1,2,\ldots,p-1}^2 = T_1^2 + \sum_{j=2}^{p} T_{j.1,2,\ldots,j-1}^2.$$

El primer término de la descomposición se llama término incondicional y corresponde al estadístico T^2 calculado para la variable $j = 1, 2, 3, \ldots, p$. La expresión es la siguiente:

$$T_j^2 = \left(\frac{str_j - \hat{\mu}_j}{\hat{\sigma}_j} \right)^2$$

donde $\hat{\mu}_j$ y $\hat{\sigma}_j$ son las estimaciones de la media y la desviación estándar de los residuos deviance estandarizados str_j obtenidos con las m observaciones históricas de str_j. En nuestro contexto, el término incondicional mide la distancia estandarizada entre la mortalidad observada en un intervalo de edad y el patrón esperado según el modelo. Así, cuando el estadístico T^2 emite una señal de cambio, el valor elevado del término incondicional indica que la señal de cambio puede estar relacionada con el intervalo de edad j.

Los otros términos, llamados términos condicionales, se calculan como

$$T_{j.1,2,\ldots,k}^2 = \left(\frac{str_j - \hat{\mu}_{j|1.2,\ldots,k}}{\hat{\sigma}_{j|1.2,\ldots,k}} \right)^2$$

donde $\hat{\mu}_{j|1,2,\ldots,k}$ y $\hat{\sigma}_{j|1,2,\ldots,k}$ son las estimaciones de la media y la desviación estándar condicional de str_j respectivamente.

Estos parámetros pueden estimarse mediante la estimación de un modelo de regresión lineal (str_j como variable de respuesta y $str_1, str_2, \ldots, str_k$ como predictores) con las m observaciones históricas del vector de residuos deviance

estandarizados. Cuando el cálculo del término incondicional arroja un valor alto, es una indicación de que la señal puede estar asociada a un cambio en la estructura de correlación de las variables que se están controlando. En nuestro contexto de aplicación, donde las variables de seguimiento corresponden a los residuos deviance estandarizados de un modelo Lee–Carter, la interpretación de este tipo de cambio no tiene sentido práctico. Por esta razón, su análisis no se considera en nuestra propuesta.

Una distribución F apropiada puede describir la distribución de probabilidad para el término incondicional en las descomposiciones MTY:

$$T_j^2 \sim \left(\frac{m+1}{m}\right) F_{(1,m-1)}.$$

Utilizando estas distribuciones, para un nivel α especificado y una muestra de referencia de tamaño m, los límites superiores de control para el término incondicional se obtienen como sigue:

$$\text{UCL} = \left(\frac{m+1}{m}\right) F_{(\alpha,1,m-1)}$$

donde $F_{(\alpha,1,m-1)}$ es el percentil superior α de la distribución F con grado de libertad $(1, m-1)$.

Como resultado, se puede utilizar la distribución F para determinar cuando un término incondicional individual de la descomposición es significativamente amplio y contribuye a la señal. Por lo tanto, un valor significativo para un término incondicional implica que la variable designada está fuera de control (Mason y Young 2002). Para nuestra propuesta, utilizamos sólo el término incondicional de la descomposición MTY para identificar el rango de edad implicado en la señal de cambio de mortalidad.

Por último, es esencial señalar que nuestra propuesta de vigilancia de la mortalidad basada en la aplicación de un gráfico de control multivariante presenta algunas ventajas con respecto al simple ejercicio de verificación del ajuste del modelo mediante el análisis de los residuos. Habitualmente, el análisis de residuos se realiza para validar los supuestos distribucionales establecidos a priori sobre los errores del modelo de mortalidad: media cero, homocedasticidad, independencia y igualdad de distribución de probabilidades. Bajo los supuestos establecidos, los residuos se tratan como observaciones independientes $m \times p$ de una variable aleatoria de media cero y varianza constante. Implícitamente,

esto equivale a suponer que la bondad de ajuste del modelo es la misma para cualquier intervalo de edad y cualquier instante de tiempo, lo cual no suele cumplirse en la práctica. En una aplicación con datos reales, el modelo suele ajustarse muy bien para ciertos intervalos de edad, pero sobreestima o subestima las tasas de mortalidad de otros intervalos. En estos casos, habrá una diferencia natural en la distribución de errores para ciertos intervalos de edad. Bajo un análisis de residuos, esta situación se identificará como un problema de ajuste del modelo, lo que limita su utilización.

Por otro lado, nuestra propuesta de control multivariante está orientada a identificar momentos de tiempo e intervalos de edad, en los que la probabilidad de muerte observada es sustancialmente diferente del patrón de mortalidad que ha recogido el modelo ajustado. Con esta estrategia, los cambios en la mortalidad pueden identificarse tanto en el modelo como en el gráfico de control.

3.3 Resultados

3.3.1 Datos

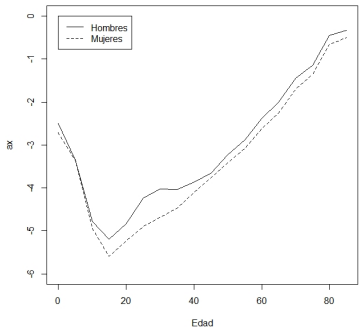
Los datos utilizados en este trabajo fueron tomados de las tablas de vida abreviadas que construimos para Colombia para el período 1973-2005, tanto para hombres como para mujeres, utilizando la información disponible en "The Latin America Human Mortality Database" (Urdinola, Torres y Velasco 2015). Los datos poblacionales corresponden a los últimos cuatro censos (1973, 1985, 1993 y 2005), por lo que la información se completó mediante interpolación lineal. A partir de los datos de mortalidad y población, fue posible calcular las funciones de las tablas de vida para Colombia desde 1973 hasta 2005. Las edades se agrupan de la siguiente manera: $[0;1);[1;5);[5;10);[10;15);\ldots;[80;85)$ para cada sexo. En el Capítulo 1 se describe con detalle el método de obtención de estas tablas de vida abreviadas para Colombia (Díaz y Debón 2016) .

3.3.2 Ajuste de los modelos LC y LC2

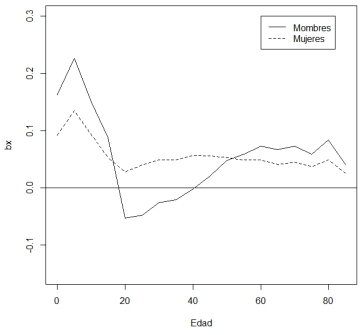
En el Capítulo 2 se ajustaron diferentes modelos de mortalidad, siendo los modelos Lee-Carter (LC) y Lee-Carter con dos términos (LC2) los que mejor se ajustaron. Los modelos LC y LC2 se ajustaron utilizando el paquete gnm Turner y Firth 2015 del software estadístico R (R Core Team 2018).

Considerando el gran número de parámetros estimados en los modelos LC y LC2 para hombres y mujeres, las estimaciones obtenidas para los parámetros

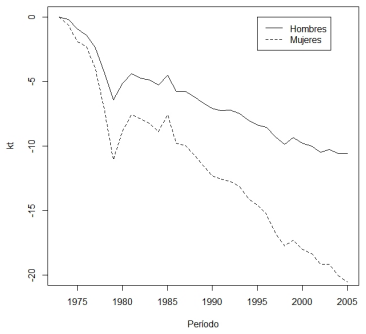
de los dos modelos se presentan en las Figuras 3.1 y 3.2. Se notan algunas características de la mortalidad y sus diferencias entre hombres y mujeres.



(a) a_x



(b) b_x



(c) k_t

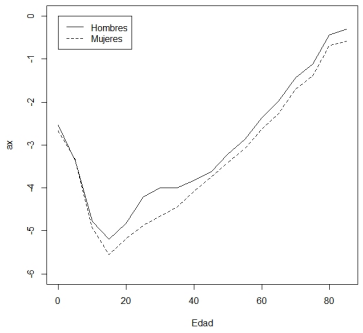
Figura 3.1: Parámetros del modelo LC ajustado a los datos colombianos en el periodo 1973-2005.

En la Figura 3.1 , el parámetro a_x muestra las fases habituales de la mortalidad de la población: para ambos sexos, la probabilidad de muerte es alta en los primeros años y disminuye lentamente hasta los 15 años, para luego aumentar a medida que la población envejece. Este parámetro también muestra el fenómeno conocido como "joroba de mortalidad de adultos jóvenes", que describe el exceso de mortalidad para un período de tiempo para los adultos jóvenes. Para los datos colombianos, este fenómeno de joroba se observa de manera muy marcada en los hombres entre 20 y 40 años de edad.

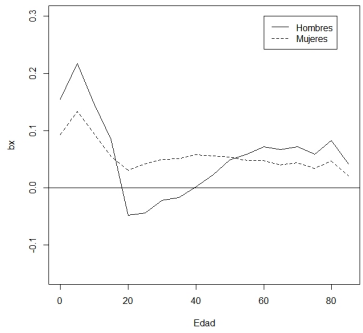
Por otra parte, el índice de mortalidad k_t presenta diferentes estructuras en su comportamiento durante el periodo analizado. Hay un descenso de 1973 a 1978 y luego un aumento notable para 1981. También hay un pico alrededor de 1985. La disminución de la mortalidad es más pronunciada en el caso de las mujeres que en el de los hombres, lo que puede explicarse por la mejora de la salud y las condiciones de vida. En el caso de los hombres, esta pendiente es menor porque estuvieron más expuestos a los acontecimientos bélicos de los años 80 y 90. La diferencia más considerable entre los sexos se produce en los últimos años del periodo analizado.

El parámetro b_x indica cómo responde la mortalidad de cada edad x a los cambios en k_t. En las mujeres, b_x toma valores positivos para todas las edades, lo que indica que la mortalidad ha disminuido para todas las edades. En los hombres, el parámetro toma valores negativos entre los 20 y los 40 años, lo que significa que a pesar de la tendencia a la reducción de la mortalidad observada a nivel general durante el periodo de estudio, la subpoblación de hombres entre los 20 y los 40 años presentó el fenómeno inverso: su mortalidad aumentó, acentuando el efecto joroba en la mortalidad masculina. Este fenómeno de joroba se ha asociado principalmente a los homicidios por el conflicto armado (Díaz, Debón y Giner-Bosch 2018).

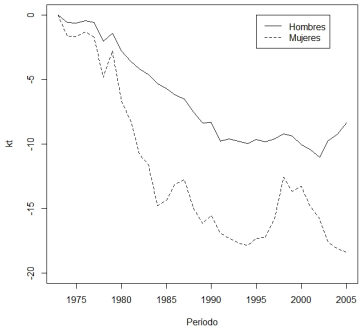
En la Figura 3.2, se muestran los parámetros del modelo LC2. Comparando las Figuras 3.1 y 3.2 del a) al c), se observa que, en general, los parámetros a_x, k_t, y b_x presentan un comportamiento similar en ambos modelos. El parámetro k_t^2 presenta valores cercanos a cero, aunque los valores disminuyen considerablemente en ambos sexos entre 1975 y 1980. Este segundo término sólo tiene efecto durante unos pocos años en todas las edades, mejorando el ajuste de LC. El parámetro b_x^2 presenta valores más altos a edades tempranas de hasta 15 años, para luego disminuir rápidamente, con un pico que se acerca a cero para las mujeres alrededor de los 20 años; para las edades posteriores, el valor se mantiene casi constante para ambos sexos.



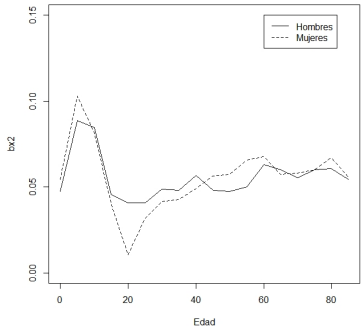
(a) a_x



(b) b_x



(c) k_t



(d) b_x^2

(e) k_t^2

Figura 3.2: Parámetros del modelo LC2 ajustado a los datos colombianos en el periodo 1973-2005.

3.3.3 Uso del gráfico T^2 de Hotteling para identificar cambios sustanciales en la tendencia de la mortalidad

El gráfico de control T^2 de Hotelling se utilizó retrospectivamente para identificar cambios sustanciales en la mortalidad durante el periodo de estudio que no fueron recogidos por el modelo Lee-Carter. La información sobre estos cambios se encuentra en los residuos del modelo. En esta aplicación, los residuos asociados a cada intervalo de edad específico se ve como una variable aleatoria. De esta manera, los residuos del modelo Lee-Carter forman un vector de variables aleatorias, y en la práctica no son necesariamente independientes.

Para el monitoreo se utilizaron los residuos deviance estandarizados para garantizar la simetría de la distribución de los residuos. Se construyeron gráficos de control Hotelling T^2 utilizando el paquete `qcc` de R (Scrucca 2017).

Para estimar el límite superior de control para el gráfico de control Hotelling T^2, se ajustó α^* (probabilidad de falsa alarma para cada punto). Dado que representamos un conjunto fijo de m puntos en el gráfico, en este caso con los años ($m = 33$), α^* se calculó como sigue:

$$\alpha^* = 1 - (1 - \alpha)^{\frac{1}{m}}$$

donde α es la probabilidad total de falsa alarma (fijada a priori en 0,05).

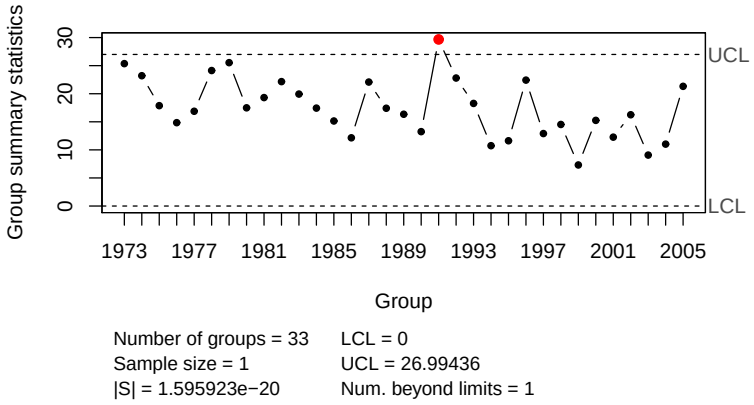
En las Figuras 3.3 y 3.4, se muestran los gráficos de control T^2 de Hotelling para los residuos de los modelos LC y LC2. Se pueden observar los años cuyos residuos están fuera de los límites de control.

La Figura 3.3 muestra que el modelo LC identifica el año 1991 como fuera de control para los hombres, mientras que para las mujeres se identifican dos años: 1979 y 1991.

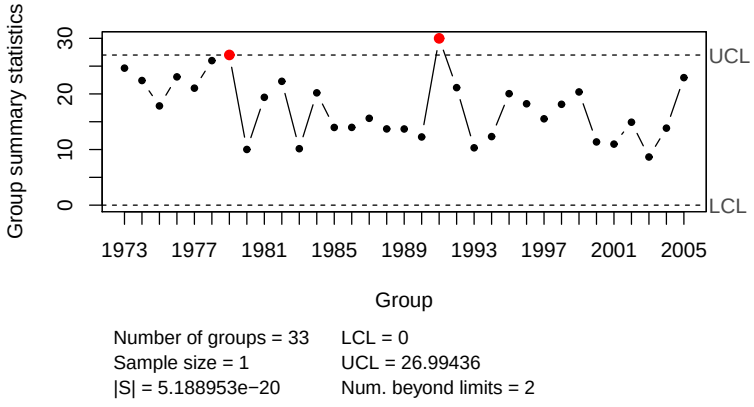
Para los residuos del modelo LC2, el año 1991 está fuera de control para los hombres, y para las mujeres, se detectaron los años 1976 y 1991 (véase la Figura 3.4).

En resumen, el gráfico de control T^2 de Hotelling con los modelos LC y LC2 identifica el año 1991 como fuera de control para ambos sexos. Además, el año 1979 se identifica para las mujeres con el modelo LC y el año 1976 se identifica para las mujeres con el modelo LC2 (ver Tabla 3.1).

En este contexto, la señal de fuera de control debe interpretarse como un cambio en la mortalidad no captado por el modelo Lee-Carter. Por lo tanto,



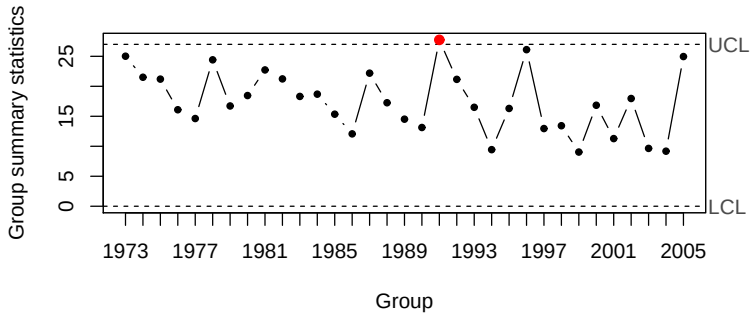
(a) Hombres



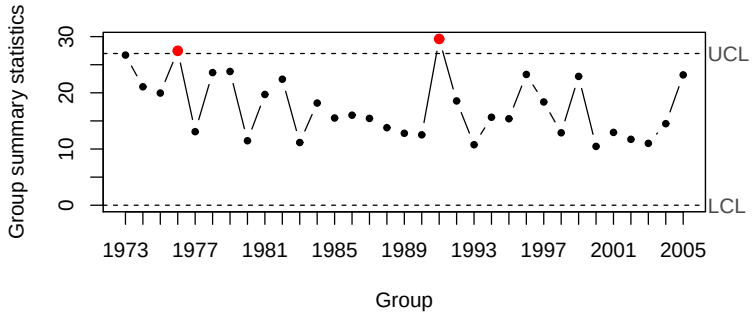
(b) Mujeres

Figura 3.3: Gráfico de control T^2 de Hotelling para los residuos del modelo LC.

cuando el gráfico no genera señales fuera de control, no debe interpretarse como que no se producen cambios en la mortalidad, sino que lo más probable es que estos cambios ya hayan sido captados por el modelo.



(a) Hombres



(b) Mujeres

Figura 3.4: Gráfico de control T^2 de Hotelling para los residuos del modelo LC2.

Tabla 3.1: Valores de T^2 para los puntos fuera de control.

Modelo	Sexo	Año	Valores T^2
LC	Hombre	1991	29.67
	Mujeres	1979	27.01
		1991	29.99
LC2	Hombres	1991	27.74
	Mujeres	1976	27.47
		1991	29.59

$\alpha = 0.05$, UCL $= 26.99$.

3.3.4 Descomposición MTY para interpretar las señales fuera de control

Las edades relacionadas con las señales se exploraron aplicando la descomposición MTY sobre el gráfico de control de Hotelling T^2 sobre las señales fuera de control .

Tabla 3.2: Términos incondicionales de la descomposición MTY para los puntos fuera de control

	LC			LC2		
	Hombres	Mujeres		Hombres	Mujeres	
Edad	1991	1979	1991	1991	1976	1991
[0–1)	4.80 *	4.78 *	5.27 *	3.68	0.15	12.58 *
[1–5)	3.79	7.21 *	1.39	17.38 *	0.28	14.47 *
[5–10)	15.42 *	0.85	9.95 *	18.80 *	0.01	22.06 *
[10–15)	0.83	0.76	0.39	2.20	0.01	0.27
[15–20)	1.60	0.19	0.55	0.02	0.17	0.00
[20–25)	3.44	0.11	0.04	0.76	0.15	0.33
[25–30)	3.69	0.82	0.01	1.39	0.83	1.41
[30–35)	2.88	0.37	0.49	0.31	0.55	0.01
[35–40)	2.37	0.24	0.02	0.10	0.04	0.79
[40–45)	2.13	0.44	1.03	0.08	1.38	0.07
[45–50)	0.69	0.83	0.38	0.05	1.85	0.11
[50–55)	2.86	1.01	0.00	1.13	1.49	0.15
[55–60)	1.47	0.27	0.76	0.13	0.21	0.22
[60–65)	2.03	4.10	0.08	0.24	0.51	0.10
[65–70)	0.04	0.49	0.69	0.46	0.21	1.11
[70–75)	0.45	6.05 *	0.63	0.41	3.36	0.02
[75–80)	0.24	0.01	0.50	0.00	3.34	0.11
[80–85)	0.06	9.25 *	0.52	0.69	0.64	0.20

$\alpha = 0.05$, UCL = 4.27, * Denota significancia

La Tabla 3.2 muestra un resumen de la información obtenida del término incondicional de la descomposición MTY de los puntos de señalización para el vector de residuos de los modelos LC y LC2. Para los hombres, los residuos de ambos modelos de mortalidad señalan a 1991 como fuera de los límites de control, donde las edades infantiles provocan esta alarma. En mujeres, los residuos del modelo LC señalan los años 1979 y 1991 como fuera de control; en ambos años, las edades infantiles tienen una fuerte influencia, además de las edades [70-75) y [80-85) para el año 1979.

Los residuos del modelo LC2 indicaron los años 1976 y 1991 como fuera de los límites de control, y las edades muy jóvenes están relacionadas con esta alarma. Cabe señalar que, aunque el año 1976 fue identificado por el gráfico de control de Hotelling T^2 como fuera de los límites de control, no se detectaron edades relacionadas con la descomposición MYT, lo que podría indicar una falsa alarma o un error de la prueba MYT (Mason y Young 2002).

También cabe señalar que el cambio identificado en el año 1976 no fue generado por una variación anormal de la mortalidad en ninguno de los rangos de edad específicos; la anomalía fue generada por un movimiento multivariado contrario a la estructura de correlación entre los residuos de este año. Este tipo de cambio no tiene ninguna interpretación práctica y, por lo tanto, su exploración se descarta en la metodología propuesta.

3.4 Discusión

Los resultados obtenidos se compararon con una serie de eventos registrados en Colombia desde una perspectiva sociodemográfica.

La alerta detectada con los residuos puede relacionarse con un cambio en las causas de muerte desde el punto de vista epidemiológico. Cristancho (2017) menciona que según el Departamento Administrativo Nacional de Estadística (DANE), el homicidio se convirtió en la primera causa de muerte para los hombres en la década de 1980. En esa década, también hubo un aumento de la mortalidad por enfermedades no infecciosas y por causas externas. Las causas externas, especialmente para los hombres, se asociaron a la violencia y a los accidentes.

Cabe destacar que en la década de los noventa, las llamadas enfermedades emergentes aumentaron en Colombia. En 1990 hubo una epidemia masiva de dengue con una tasa de mortalidad del 40%: su intensidad disminuyó en 1991-1995 (Padilla, Rojas y Sáenz 2012). Además, en 1991 hubo un brote epidémico de cólera.

Al analizar los hechos relacionados con la salud pública en Colombia, se identificaron dos eventos cruciales como fuera de control y relacionados con los años detectados. En 1975, se estableció el Sistema Nacional de Salud; en el período comprendido entre 1990 y 2000, hubo cambios en la mortalidad relacionados con esta Reforma del Sistema de Salud.

Por otra parte, con la Constitución Política de 1991, se creó una nueva división político administrativa en Colombia, creando nuevos departamentos, lo que benefició la recolección de información en la región sur del país. La exhaustiva labor de esta división en la recuperación de la información aumentó el número de muertes registradas.

3.5 Conclusiones

Este capítulo demuestra la utilidad de los gráficos de control como herramienta para detectar cambios sustanciales en el comportamiento de la mortalidad mediante el seguimiento de los residuos de los modelos de mortalidad. En algunos países en desarrollo, los datos se presentan en grupos de edad debido a errores relacionados con la edad; normalmente, se prefiere que la declaración de la edad de la muerte se produzca en múltiplos de cinco, y existen otras dificultades de registro. Por ello, una cuestión de interés en el ámbito demográfico y actuarial es la utilización de los residuos de un modelo para controlar la mortalidad de las tablas de vida abreviadas.

Trabajos anteriores como el de Urdinola y Rojas-Perilla (2013) utilizan gráficos de control con una sola medida por unidad de tiempo para monitorizar los datos de mortalidad, lo que implica construir numerosos gráficos de control y aumentar la tasa de falsas alarmas (puntos fuera de control). En cambio, en este trabajo, utilizamos un gráfico de control multivariante T^2 de Hotelling, que supervisa simultáneamente p variables aleatorias (intervalos de edad) por unidad de tiempo, lo que reduce drásticamente el número de gráficos a construir. Además, este gráfico de control multivariante tiene en cuenta las relaciones entre los residuos asociados a los distintos intervalos de edad. La metodología que proponemos considera que la mortalidad es un fenómeno que no es estable a lo largo del tiempo, sino que presenta tendencias recogidas a través de los modelos de Lee-Carter. En consecuencia, los gráficos de control aplicadas a los residuos de estos modelos pueden detectar los otros tipos de cambios en la mortalidad que no fueron recogidos previamente por los modelos.

Los modelos de mortalidad LC y LC2 identificaron las principales características de la mortalidad en Colombia. La mortalidad infantil es alta y disminuye lentamente hasta los 15 años, después de lo cual aumenta progresivamente a medida que la población envejece. El fenómeno de la sobremortalidad se registra en los adultos jóvenes, principalmente en los hombres. Este fenómeno ha sido descrito con anterioridad en otros trabajos (Jones y Ferguson 2006; Díaz, Debón y Giner-Bosch 2018) y se explica principalmente por los homici-

dios relacionados con el conflicto armado interno (Urdinola, Torres y Velasco 2017) y las actividades ilícitas como los mercados ilegales de drogas Gaviria 2000, así como la disponibilidad de armas de fuego (Briceño-León, Villaveces y Concha-Eastman 2008).

El gráfico de control T^2 de Hotelling es una opción interesante para controlar los residuos de los modelos de mortalidad y así identificar los años en los que la mortalidad difiere de los patrones que recoge el modelo. Con el uso combinado de Hotelling T^2 y la descomposición MTY, se identificaron los años y el rango de edad de ese patrón atípico. Se identificaron dos años como fuera de control: 1991 para los residuos de ambos modelos, tanto para hombres como para mujeres, con influencia de edades muy jóvenes, y el año 1979 para los residuos del modelo Lee–Carter para mujeres con influencia de edades muy jóvenes y muy avanzadas .

El modelo Lee-Carter recoge información sobre el fenómeno de la violencia en Colombia. Por lo tanto, los años identificados como fuera de control en los gráficos están asociados a edades muy tempranas o bastante avanzadas, las cuales están inversamente relacionadas con la violencia que no cobró tantas víctimas a esas edades. Además, los cambios de mortalidad identificados en los gráficos de control pertenecen a cambios en las condiciones de salud de la población, o nuevas causas de muerte como el COVID-19 en futuras investigaciones. Sería interesante realizar estudios futuros para evaluar esta metodología combinada agregando información de nuevos censos para Colombia.

No obstante, nuestra propuesta de vigilancia de la mortalidad consiste en dos herramientas de análisis que funcionan de manera secuencial, por lo que este estudio tiene limitaciones en cuanto a modelos y gráficos de control. En primer lugar, el modelo capta la tendencia temporal y el perfil etario de la mortalidad en la población. Posteriormente, el gráfico de control T^2 de Hotelling y la descomposición MTY identifican aquellos años y rangos de edad cuyas probabilidades de muerte difieren sustancialmente de la tendencia del modelo. Por lo tanto, las señales fuera de control del gráfico T^2 se interpretan como posibles cambios en la mortalidad según la tendencia del modelo. Por ejemplo, el modelo LC es más sencillo que el LC2, ya que el modelo LC2 incluye estructuras de cambio adicionales más particulares. La selección de uno u otro modelo conllevará una caracterización diferente de la tendencia de la mortalidad y, en consecuencia, una variación en el diagnóstico del gráfico de control T^2 de Hotelling.

Por último, queremos señalar que aunque en este trabajo sólo se aplicaron gráficos de control a las tablas de vida abreviadas colombianas, la metodología

puede extenderse a las tablas de vida abreviadas de cualquier país en desarrollo. Podría ser útil analizar otros conjuntos de datos y examinar si las conclusiones son consistentes para diferentes países.

Capítulo 4

Análisis de la mortalidad en Colombia por departamentos

El análisis de la mortalidad permite comprender la dinámica poblacional, así como explorar tendencias y examinar comportamientos a través del tiempo. Resulta interesante estudiar los diferentes indicadores de desarrollo relacionados con la mortalidad de un país dado que son el reflejo de la importancia que se da a la vida y a la salud. El objetivo de este trabajo fue analizar la mortalidad de las diferentes departamentos de Colombia y explorar la formación de grupos según su comportamiento en la mortalidad mediante técnicas multivariadas.

4.1 Introducción

El estudio de temas demográficos es de actual interés en muchas áreas del conocimiento tales como la demografía, las ciencias económicas, la biología o las ciencias actuariales y finanzas. La dinámica de la población en ocasiones se refleja en el crecimiento de su esperanza de vida, el envejecimiento de la población o la disminución de la fecundidad.

Numerosos estudios abordan el análisis de la mortalidad para grupos de países o regiones de un mismo país, teniendo en cuenta su disposición geográfica y las características comunes que los definen en cuanto a aspectos sociales, económicos o sanitarios. Es conocido que países o regiones con características

comunes suelen tener tendencias de la mortalidad que evolucionan de manera similar (Guibert y col. 2020). Algunos de estos trabajos modelan la mortalidad y hacen pronósticos aplicando previamente métodos de agrupación provenientes del análisis multivariado o la minería de datos, para identificar grupos o conglomerados de poblaciones más pequeños que presenten comportamientos similares.

Guibert y col. (2020) proponen un nuevo modelo a partir de los conocidos modelos como el de Lee-Carter y el de Li-Lee, para pronosticar la mortalidad en poblaciones de manera simultánea asumiendo que el principio de coherencia a largo plazo se verifica en subgrupos de países que tienen la propiedad de "coherencia local". Para determinar los grupos de coherencia local, aplicaron el método de análisis jerárquico de conglomerados (HCA, hierarchical clúster analysis).

En Schnürch, Kleinow y Korn (2021) proponen variantes del modelo de efecto de edad común (Kleinow 2015), que describe las tasas de mortalidad de múltiples poblaciones, aplicando previamente métodos de agrupación como el fuzzy clúster de máxima verosimilitud para identificar los subgrupos adecuados. Debón y col. (2017) presentan una metodología para agrupar países europeos y caracterizar los grupos mediante diferentes indicadores de mortalidad.

Por otra parte, en Fritzell y col. (2013) realizan un análisis comparativo de la relación entre la pobreza relativa y la mortalidad en 26 países a lo largo del tiempo teniendo en cuenta las diferencias entre los regímenes de bienestar, mediante un análisis de series temporales transversales agrupadas. Se comparan cuatro modelos anidados en los que la variable dependiente es la tasa de mortalidad y la variable explicativa es la tasa de pobreza. Las tasas de mortalidad se evaluaron por sexo y por intervalos de edad.

Además, en Christiansen, Spodarev, Unseld y col. (2015) se presenta un enfoque geométrico en el plano edad-período para modelar las diferencias de la tasa de mortalidad en países de Europa Occidental utilizando técnicas de geoestadística.

Otro enfoque en la comparación de la mortalidad para países europeos lo encontramos en Carracedo y col. (2018), donde se utiliza la metodología espaciotemporal (el tiempo y las relaciones de vecindad entre los países) para detectar grupos de países europeos.

Un estudio interesante sobre el las diferencias de mortalidad entre los departamentos metropolitanos franceses lo encontramos en Barbieri y Depledge (2013). La autora explica de manera detallada el comportamiento y las variaciones de

la mortalidad por departamentos y sexos. Utiliza el análisis de componentes principales (ACP) para determinar los distintos patrones de la mortalidad por grupos de edad en los departamentos franceses para el período 1976-2008 y determinar las causas de muerte explican las variaciones geográficas de la mortalidad.

Un estudio reciente de la CEPAL advierte que en Colombia los departamentos se han desarrollado a ritmos diferentes, logrando algunos aumentar su prosperidad económica y bienestar social, mientras que otros departamentos se han mantenido rezagados en los últimos años (Ramírez y Aguas 2017). Según este estudio los territorios alcanzan distintos niveles de prosperidad teniendo en cuenta la economía, infraestructura, capital humano, ciencia, instituciones, gestión y finanzas públicas, y tecnología e innovación, en general con base en patrones de especialización particulares.

Un indicador comúnmente utilizado al analizar las desigualdades socioeconómicas y de salud entre las regiones de un país es la Tasa de Mortalidad Infantil (la TMI en Colombia mide las muertes de infantes ocurridas antes del primer año de edad por cada 1000 nacidos vivos). En Alvis-Zakzuk y col. (2015) donde se analiza aspectos de la desigualdad de la mortalidad y pobreza en Colombia, se tiene en cuenta los datos de TMI para los años censales 1993 y 2005 y se clasificaron los departamentos de acuerdo a su TMI. Se clasificaron los departamentos según cuatro categorías para la TMI: baja, media, alta y muy alta. Comparando el comportamiento para estos dos años se puede observar que los departamentos que tuvieron un retroceso fueron: Norte de Santander, Cundinamarca, Guaviare, Vaupés, Arauca y Atlántico, Guainía y Cauca; y entre lo que mejoraron están Quindío, Huila, Magdalena, Putumayo, Cesar, La Guajira, Caquetá, San Andrés y Prov. Esta clasificación tendrá en cuenta en nuestro análisis aunque se trabajará con otros indicadores.

- Baja (23 o menos): Antioquia, Caldas, Quindío, Bogotá, Risaralda, Santander, Valle del Cauca, San Andrés y Prov.
- Media ($23 - 33$): Tolima, Atlántico, Boyacá, Norte de Santander, Cundinamarca, Huila, Magdalena, Sucre.
- Alta ($33 - 43$): Córdoba, Meta, Putumayo, Guaviare, Casanare, Cesar, Guajira, Bolívar, Caquetá
- Muy alta (43 o más): Amazonas, Vaupés, Guainía, Nariño, Cauca, Arauca, Vichada, Chocó.

Teniendo en cuenta este comportamiento desigual en los departamentos de Colombia es aspectos socieconómicos, el objetivo principal de este capítulo fue estudiar el comportamiento de la mortalidad en los departamentos de Colombia mediante técnicas de análisis multivariado como son el análisis de componentes principales, análisis de clúster jerárquico y fuzzy clúster, para identificar grupos de departamentos con comportamientos similares en cuanto a mortalidad.

4.2 Ideas generales del análisis multivariado

Las técnicas multivariadas son cada vez más utilizadas tanto en estudios científicos como en las áreas sociales. Los métodos de análisis exploratorio y confirmatorio de datos, que generalmente se utilizan de forma combinada, requieren de un conocimiento previo del problema y de los factores con que está relacionado. Se puede definir el análisis multivariado como aquella rama de la estadística interesada en el estudio de las relaciones entre series de variables (independientes o no) y de los individuos que la sustentan.

La definición clásica de Kendall (1975), uno de los primeros en impulsar el análisis multivariado, destaca como rasgo más característico de este tipo de análisis, la existencia de una serie de n individuos, en cada uno de los cuales se observan los valores de p variables, lo que implica que los datos se pueden expresar en forma matricial y pueden interpretarse tanto desde los indivious como desde las variables.

Entre los objetivos principales de un análisis multivariado se encuentran:

- Reducción de dimensión,
- Agrupamiento de variables,
- Búsqueda de relaciones entre las variables.

Según Kendall (1975), las técnicas de análisis multivariado se dividen en dos, unas que tienen en cuenta las relaciones de interdependencia entre las variables y otras que se basan en las relaciones de dependencia. Las técnicas basadas en relaciones de dependencia establecen una distinción entre las variables a explicar y las variables explicativas u observadas. Las variables a explicar son las llamadas dependientes y las variables explicativas, independientes. Tales técnicas tienen por objeto establecer la relación entre las variables como base para realizar una predicción. Las técnicas basadas en relaciones de independencia no establecen ninguna diferenciación entre las variables, recibiendo todas ellas

el mismo tratamiento. El objetivo que se persigue al utilizar estas técnicas es el de organizar los datos de forma que sean más manejables para el investigador y ofrezcan una mayor comprensión global.

4.3 Análisis de Componentes Principales

El análisis de componentes principales (ACP en español o PCA en inglés) desarrollado por Pearson (a finales del siglo XIX) y luego por Hotelling (alrededor de los años 30 del siglo XX), es un método exploratorio multivariado cuyo objetivo es detectar la estructura subyacente de un conjunto de datos almacenados en la matriz $\mathbf{X}$ (Johnson y Wichern 2014).

Supongamos que tenemos p variables observadas en n individuos. Las observaciones pueden ordenarse en forma de una matriz $\mathbf{X}(n,p)$, donde x_{ij} es la i-ésima observación en la j-ésima variable. El objetivo del análisis de componentes principales es disminuir el número de variables, es decir, construir nuevas variables y reemplazar las variables originales por la menor cantidad posible de nuevas variables ficticias (componentes principales).

Específicamente, el ACP consiste en descomponer la matriz $\mathbf{X}(n,p)$ como

$$\mathbf{X} = \mathbf{AB}' + \mathbf{E}. \tag{4.1}$$

En la ecuación 4.1, $\mathbf{X}$ se resume en la matriz $\mathbf{AB}'$ que tiene rango $(c < p)$, donde $\mathbf{A}$ $(n \times c)$ y $\mathbf{B}$ $(p \times c)$ son, respectivamente las matrices de puntuación y de carga de los componentes y c es el número de componentes extraídos. Normalmente, n y p denotan el número de unidades de observación y de variables respectivamente. Por último, $\mathbf{E}$ $(n \times p)$ denota la matriz residual (Giordani y Kiers 2007).

Las matrices óptimas de los componentes se obtienen minimizando

$$\left\|\mathbf{X} - \mathbf{AB}'\right\|^2,$$

donde $\|\cdot\|^2$ denota la norma euclidiana al cuadrado, es decir la suma de los cuadrados de los elementos de la matriz indicada.

El ACP es un análisis interno, concerniente a las varianzas y covarianzas de los elementos de un vector aleatorio, por lo que no involucra relaciones externas

con cualquier otro vector. Debe tenerse en cuenta que al transformar y reducir el número de variables originales, parte de la información contenida en la matriz de varianzas y covarianzas de las variables se sacrifica, pero en el ACP esta pérdida de información es mínima (Linares 1990).

Para determinar el número de componentes principales existen tres criterios fundamentales, que como no tienen una fundamentación matemática sólida, se recomienda trabajar con un sentido de compromiso entre ellos.

- Diagrama de la pendiente (o gráfico de sedimentación). En el eje y se expresa el porcentaje de varianza explicada por cada componente, y en el eje x el número de la componente principal. Estos puntos se unen por segmentos de rectas y se analiza el punto a partir del cual se observa una caída. Ese punto determina el número de componentes principales.
- Criterio del porcentaje. Se incluye el número de componentes principales que den un porcentaje de la varianza aceptable para el investigador.
- Criterio de Kaiser. Si se trabaja con la matriz de correlaciones $\mathbf{R}$, se toman las primeras k componentes tales que el autovalor asociado sea mayor que 1.

Para la interpretación de las componentes, es necesario tener en cuenta el grado de asociación (mediante un coeficiente de correlación) de una variable original con una componente. Según Linares (1990), para determinar la importancia de las variables originales se tiene en cuenta el valor absoluto del mayor elemento del vector propio dividido entre dos, considerando importantes todas las variables cuyos elementos sean mayores o iguales que el valor anterior. En un segundo momento se debe tener en cuenta que dos variables originales con elementos del vector propio con valores semejantes, se consideran que contribuyen en el mismo sentido si tienen el mismo signo y son opuestos si tienen signo contrario.

4.4 Análisis de clúster

El análisis de clúster de manera general pretende buscar grupos dentro de un conjunto de datos. Dentro de cada grupo (clúster) los datos muestran similitud, a su vez la similitud se define mediante una medida de distancia. Las particiones generadas por este enfoque definen para todos los elementos de datos a qué clase (clúster) pertenecen. Las particiones pueden definir una frontera dura (fija) entre las subparticiones; esto se llama clústering duro. Por

el contrario los límites entre subpartes generados por un algoritmo de clúster difuso (conocido como fuzzy clúster) son vagos, lo que significa que cada patrón de datos de objetos de una partición difusa pertenece a diferentes clases con diferentes valores de pertenencia. El análisis de clúster se ha desarrollado en áreas tan diferentes como la medicina, la biología, la química, la astronomía o la antropología, a su vez con aplicaciones o enfoques muy diversos.

En general, podemos decir que un análisis clúster se puede realizar para:

- Explorar los datos. Se aplica la técnica en cuestión para ver qué se obtiene y apoyándose en la salida, se da una interpretación de los datos.
- Reducir el número de datos o variables. se reemplazan los integrantes de cada grupo por uno de sus elementos.
- Generar hipótesis. Una vez identificado el número de grupos en que se dividen los datos, se construye alguna teoría que explique el agrupamiento en la población de estudio.
- Realizar predicciones. Se utilizan las agrupaciones obtenidas para realizar predicciones sobre futuros estudios.

Una forma para lograr obtener los grupos es hacer agrupaciones progresivas, cada vez mayores, o por el contrario, divisiones sucesivas, cada vez de menor tamaño, de modo que se tiene una especie de jerarquía entre agrupaciones (Linares 1990).

Para esto, se parte de matriz primaria de datos de observación $\mathbf{X}(n, p)$, con n individuos y p variables, donde los elementos de $\mathbf{X}$ se supone presentan una estructura de grupos o de jerarquía de grupos encajadas, aunque no se tiene en cuenta ningún criterio preestablecido. La aplicación de este método de clasificación se desarrolla en tres etapas:

1. Se crea una matriz $\mathbf{D}(n, n)$ o $\mathbf{D}(p, p)$ en dependencia de si se quiere trabajar con las observaciones o con las variables, que presenta el grado de semejanza de cada individuo j de $\mathbf{X}(n, p)$, tomando en cuenta las p características observadas.
2. Algoritmo de clasificación jerárquica.
    - Se comienza con una partición del conjunto de manera tal que cada uno sea el único elemento de cada una de las clases de una partición en un número de clases igual al número de individuos.

- Se reúnen en una clase única las dos clases más parecidas de la etapa anterior. El número de clases restantes disminuye en una unidad.
- Se prosigue así hasta no disponer más que de una sola clase que reúne todas las clases, y en consecuencia individuos.

3. Se describen los contenidos de los sub-conjuntos de clases obtenidas en cada etapa y se evalúa la calidad de la clasificación obtenida.

En la primera etapa, para definir la semejanza entre individuos de la matriz de datos $\mathbf{X}(n,p)$ se crean los llamados "índices de similitud". Para evaluar la similitud entre los individuos de $\mathbf{X}(n,p)$ se definen primero los índices de disimilitud, los cuales varían a la inversa de los índices de similitud. A partir de este índice de disimilitud se define el concepto de distancia. La selección de una distancia entre los objetos depende del nivel de medida de las características observadas, a partir de las cuales se quiere hacer la comparación entre los objetos. Para el cálculo de distancias o de disimilitudes existen varios procedimientos, entre las más utilizadas se encuentran la distancia Euclideana, la distancia de Mahalanobis, la distancia de Minkowski y la city-block o de Manhattan. Una vez obtenida la matriz de distancias (o dismilitudes) entre los puntos iniciales, se calculan las distancias entre grupos (Hennig y col. 2016).

Los métodos de clasificación jerárquica están destinados a producir una representación gráfica de la información contenida en la matriz de datos. Por tanto, las clasificaciones jerárquicas tienen como objetivo principal representar de manera sintética el resultado de las comparaciones entre los objetos de una matriz $\mathbf{X}(n,p)$ observada, y se pueden definir como una serie de particiones encajadas, su resultado es un árbol de clasificación o dendrograma (Johnson, Wichern y col. 2002).

Existen diferentes algoritmos (procesos iterativos) de agregación que son utilizados corrientemente:

1. El método del vecino más cercano (distancia mínima o similitud máxima). En este método la distancia entre dos grupos o conglomerados es determinada por la distancia de los dos objetos más cercanos en los diferentes grupos. Esta regla será en un sentido, se tendrán filas de objetos juntos que forman grupos. El grupo resultante tiende a representar largos eslabones.

2. El método del vecino más lejano (distancia máxima o similitud mínima). En este método la distancia entre dos grupos es determinada por la mayor

distancia entre dos objetos cualesquiera en los diferentes grupos. Usualmente este método funciona bastante bien en los casos donde los objetos actuales forman de manera natural agrupaciones distintas.

3. El método de los centroides (distancia media). En este método la semejanza entre dos clústers viene dada por la semejanza entre sus centroides, es decir, los vectores de medias de las variables medidas sobre los individuos del clúster. Existen dos variantes: el método del centroide ponderado (tiene en cuenta los tamaños de los clústers al efectuar los cálculos), y el método del centroide no ponderado (que no tiene en cuenta los tamaños de los clústers al efectuar los cálculos).

4. El método de Ward. Este método se distingue de los demás porque usa un análisis de varianza para aproximar la evaluación de la distancia entre dos grupos, es decir, este método intenta minimizar la suma de cuadrados de dos grupos hipotéticos cualesquiera, que pueden ser formados en cada paso. Es un procedimiento jerárquico y los detalles concernientes a este método se le atribuyen a Ward (1963). En general este método es considerado como muy eficiente, sin embargo, tiende a crear grupos de tamaño pequeño.

Teniendo en cuenta que el análisis de clúster se utiliza para crear grupos homogéneos, es natural considerar que estos procedimientos puedan ser usados para determinar justamente el número de grupos que está presente en los resultados de un estudio de clasificación. Por ejemplo, la estructura de árbol de un dendrograma que sugiere que varios grupos diferentes quizás están presentes en los datos, de allí sin embargo surge una duda en el sentido que no se conoce el punto óptimo en donde se debería cortar el árbol de tal modo que el número óptimo de grupos sea encontrado, que es precisamente lo que se recomienda para una etapa posterior de este análisis.

4.4.1 Fuzzy clúster

El método fuzzy c-mean o fuzzy clúster (FC) fue desarrollado por Bezdek, Ehrlich y Full (1984). El FC es un algoritmo de agrupamiento no supervisado que se ha aplicado con éxito a una serie de problemas relacionados con el análisis de características, el agrupamiento de individuos o variables y el diseño de clasificadores. El FC tiene numerosas aplicaciones en diferentes áreas como son las ingenierías, la astronomía, la química, la geología, la medicina entre otras (Rezaee, Lelieveldt y Reiber 1998).

Mientrás que en el análisis clásicos de clúster los individuos pertenecen a un solo grupo, al aplicar el algoritmo fuzzy clúster los individuos pueden pertenecer a más de un grupo. Se describe la asociación a cada grupo por medio de un nivel de pertenencia que indicará el grado de relación entre un individuo y un grupo. El FC intenta encontrar el punto más característico de cada clúster, denominado centro de un clúster; a continuación, calcula el grado de pertenencia de cada objeto a los clústeres. Este algoritmo también minimiza la varianza intraclúster. Sin embargo, hereda los problemas de K-means, ya que el mínimo es sólo local y los clústers finales dependen de la elección inicial de los pesos (Fahad y col. 2014).

El algoritmo FC sigue el mismo principio que el algoritmo K-means, es decir, busca iterativamente los centros de los clústers y actualiza la pertenencia de los individuos. La principal diferencia consiste que en lugar de tomar una decisión firme sobre a qué clúster debe pertenecer el individuo, asigna a un individuo un valor que va de 0 a 1 para medir la probabilidad de que el individuo pertenezca a ese clúster. Una regla difusa establece que la suma del valor de pertenencia de un individuo a todos los clústers debe ser 1. Cuanto mayor sea el valor de pertenencia, más probable será que el individuo pertenezca a ese clúster.

El FC se obtiene minimizando la función objetivo J_m que es la suma ponderada de los errores al cuadrado dentro de grupos Bezdek, Ehrlich y Full (1984):

$$J_m(\boldsymbol{U}, \boldsymbol{V}, X) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^m \left| x_k - v_i \right|^2, \tag{4.2}$$

donde $\boldsymbol{U} = [u_{ik}]$ es una partición c difusa de X, $\boldsymbol{V} = (v_1, v_2, ..., v_c)$ es el vector de centro (prototipo de clúster) $v_i \in \mathbb{R}^p$, n es el número de objetos, c es el número de clúster definidos, m es un factor de confusión $(1 < m < \infty)$, u_{ik} son los valores de probabilidad al asignar el objeto i al clúster k (grado de pertenencia), y $|x_k - v_i|$ es la distancia euclidiana entre el i-ésimo objeto p_i y el k-ésimo centro de clúster v_k definido por:

$$|x_k - v_i| = \sqrt{\sum_{i=1}^{n} (x_k - v_i)} \tag{4.3}$$

El centroide del k^{th} clúster se actualiza mediante la ecuación:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m}, \ 1 \leq i \leq c; \tag{4.4}$$

La tabla de pertenencia se calcula utilizando la ecuación:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{|x_k - v_i|}{|x_k - v_j|} \right)^{\frac{2}{m-1}}} \tag{4.5}$$

para $1 \leq i \leq c$, $1 \leq k \leq n$.

4.4.2 Índices de validación

Teniendo en cuenta que los algoritmos de clúster no están supervisados, independientemente del método de clúster que se utilice, duro o difuso, las particiones finales de los datos requieren algún tipo de validación. Los índices de validez de los clúster permiten evaluar y comparar las particiones del FC, además de identificar cuan compactos son los grupos creados y determinar el número de clúster a seleccionar del conjunto de datos (Rezaee, Lelieveldt y Reiber 1998).

Tabla 4.1: Índices de validación para el fuzzy clúster

Índices de validez	Descripción	Número óptimo de grupos
XB (Xie y Beni)	Es función del conjunto de datos y los centroides de los clústers. Describe la relación entre la variación total de la partición, los centroides y la separación de los vectores de los centroides.	Valores mínimos indican las mejores particiones.
FS (Fukuyama y Sugeno)	Mide la diferencia entre dos términos, el primero que combina la difusividad en la matriz de pertenencia con la compacidad geométrica de la representación del conjunto de datos a través de los prototipos, y el segundo la difusividad en la fila de la matriz de partición con la distancia del i-ésimo prototipo a la gran media de los datos.	Valores mínimos indican las mejores particiones.
PE (Partition entropy)	Utiliza sólo los valores de pertenencia u_{ik} de una partición difusa del conjunto de datos X. Proporciona información sobre la matriz de pertenencia sin considerar los datos en sí mismos.	Valores mínimos indican las mejores particiones.
PC (Partition coefficient)	Utiliza sólo los valores de pertenencia u_{ik} de una partición difusa del conjunto de datos X. Proporciona información sobre la matriz de pertenencia sin considerar los datos en sí mismos.	Valores máximos indican las mejores particiones.

Para decidir el número de grupos al aplicar el FC en este capítulo se utilizarón cuatro índices de validación: XB, (Xie y Beni 1991); FS, (Fukuyama y Sugeno 1989); PE (Partition entropy) y PC (Partition coefficient). Los índices XB, FS y PE indican que la agrupación es mejor cuanto más bajos sean sus valores, y el índice PC indica que a mayor valor mejor será la partición.

4.5 Árboles de clasificación

Los árboles de clasificación o decisión, son una técnica de la minería de datos que se ha venido trabajando desde hace varios años y es común que se utilice junto a técnicas del análisis multivariado. Se utilizan para revelar la estructura oculta de los datos y reducir el número de posibles predictores. Su objetivo principal es identificar qué combinaciones de variables (explicativas) son capaces de predecir mejor la variable de interés (Schiattino y Silva, 2008).

Los métodos basados en árboles de clasificación constituyen una herramienta flexible y eficaz que permite explorar estructuras de datos complejas, siendo útiles cuando necesitamos clasificar o hacer predicciones de resultados (Sistachs, 2011). Este procedimiento crea un modelo de clasificación basado en árboles, y clasifica casos en grupos o pronostica valores de una variable dependiente basada en valores de variables independientes, proporcionando además herramientas de validación para análisis de clasificación exploratorios y confirmatorios.

Los árboles de clasificación se puede utilizar principalmente para:

- Segmentación (identificar las personas que pueden ser miembros de un grupo específico).
- Estratificación (asignar los casos a una categoría de entre varias).
- Predicción (crear reglas y utilizarlas para predecir eventos futuros, dado que el árbol no utiliza las variables que no aportan información).
- Identificación de interacción (identificar las relaciones que pertenecen sólo a subgrupos específicos y las especifica en un modelo paramétrico formal).
- Fusión de categorías y discretización de variables continuas (recodificar las variables continuas y las categorías de los predictores del grupo, con una pérdida mínima de información).

Se define un árbol de clasificación como un conjunto de nodos y ramas, donde cada nodo representa un subconjunto de la población. Es necesario distinguir al nodo raíz que representa a toda la población y no tiene ramas entrantes, a los nodos terminales que representa la partición final, y los nodos intermedios cuyos ramas salientes apuntan a los nodos hijos (Schiattino y Silva, 2008). Se comienza con un nodo inicial y nos preguntamos cómo dividir el conjunto de datos disponibles en dos partes más homogéneas utilizando una de las variables.

Esta variable se escoge de modo que la partición de datos se haga en dos conjuntos lo más homogéneos posibles.

En el proceso de construcción de un árbol de clasificación se involucran las siguientes decisiones:

1. Seleccionar las variables y sus puntos de corte para hacer las divisiones.
2. Cuándo se considera que un nodo es terminal y cuándo se continúa dividiendo.
3. La asignación de las clases a los nodos terminales.

La principal diferencia entre los algoritmos para construir árboles radica en la estrategia para podar los árboles y el criterio para dividir los nodos. Dependiendo de la estrategia de podar los árboles y cómo se lleve a cabo la partición de los nodos, se distinguen diferentes algoritmos para realizar árboles de clasificación, entre los que se encuentran CART y CHAID:

CHAID.- Detección automática de interacciones mediante Chi-cuadrado (en inglés, CHi-square Automatic Interaction Detection). CHAID en cada paso elige la variable independiente (predictora) que presenta la interacción más fuerte con la variable dependiente. Las categorías de cada predictor se funden si no son significativamente distintas respecto a la variable dependiente. Generalmente se utiliza cuando la variable dependiente es categórica y las variables explicativas también, si no han de categorizarse.

CART.- Árboles de clasificación y regresión (en inglés, Classification and Regression Trees) Este método divide los datos en segmentos para que sean lo más homogéneos que sea posible respecto a la variable dependiente. Un nodo terminal en el que todos los casos toman el mismo valor en la variable dependiente es un nodo homogéneo y "puro". Se debe señalar que cuando se trabaja con una variable dependiente continua se utiliza un modelo de regresión y se obtiene un árbol de regresión, pero cuando la variable dependiente es categórica se ajusta a través de un modelo de clasificación, obteniendo un árbol de clasificación.

4.5.1 Random forest

Random Forest (RF) o Bosques aleatorios, es una técnica de minería de datos que resulta una alternativa a los tradicionales árboles de clasificación. Como herramienta combina árboles de decisión de manera que cada árbol depende de los valores de un vector aleatorio de la muestra de manera independiente y con la misma distribución de todos los árboles en el bosque (Breiman 2001).

Esta herramienta a menudo proporciona una mayor precisión en comparación con un modelo de árbol de decisión único, logrando mantener algunas de las cualidades beneficiosas de los modelos de árbol como son la capacidad de interpretar las relaciones entre los predictores y el resultado (Speiser y col. 2019). Además, según (Fernández-Delgado y col. 2014) RF ofrece de manera sistemática una de las mayores precisiones de predicción en comparación con otros modelos de clasificación en general.

En la construcción de los árboles se utiliza una muestra bootstrap (muestra aleatoria construida con reemplazamiento a partir de la muestra disponible) diferente de los datos, cada nodo se divide utilizando el mejor entre un subconjunto de predictores elegidos aleatoriamente en ese nodo en lugar de utilizar todas las variables como en CART. RF se encuentra entre las técnicas de aprendizaje automático más populares gracias a su precisión relativamente buena, robustez y facilidad de uso (Debón y col. 2017).

Está técnica es muy fácil de usar pues solo tiene dos parámetros: el número de variables que se selecciona en el subconjunto aleatorio en cada nodo, y el número de árboles que componen el bosque; y en general, no suele ser muy sensible a los valores de estos parámetros. La tasa de error del RF tiene relación con estos dos parámetros. Si se reduce el número de variables p, se reduce también la correlación entre los árboles por lo que la precisión del árbol se vera afectada. Al aplicar RF en una aplicación de clasificación (como es el caso de este trabajo), se recomienda utilizar $\sqrt{p}$ variables en cada nodo.

Una información valiosa que resulta del uso de la técnica de RF, es la "importancia de las variables" en el modelo de clasificación (Janitza, Celik y Boulesteix 2018). La idea es tener en cuenta como posibles cambios en las variables de entrada afectan al modelo. De esta manera las variables de entrada que provoquen mayor variabilidad en la salida, serán las que más influencian tengan y mejor explicaran el modelo, por lo tanto las más importantes. La importancia de las variables suele analizarse mediante el gráfico del criterio de Gini. Cuando se elige una variable para dividir un nodo, el índice de Gini se mantiene o se

reduce, la media de los decrecimientos para cada variable es lo que se conoce como "decrecimiento medio del índice de Gini".

Como medida de error, se suele utilizar el *Out of bag error* (OOB) que representa el error de predicción cometido por el random forest cuando se tiene en cuenta el conjunto de variables que no fue utilizado en los nodos al construir el árbol. Se calcula contando el número de observaciones que han sido clasificadas erróneamente y dividiendo este número por el total de observaciones.

4.6 Resultados

4.6.1 Datos

La división de Colombia en departamentos y municipios se da con la constitución de 1886 donde se decretó la República de Colombia, dejando atrás la división del territorio en provincias y estados. Con la Constitución de 1991 en Colombia se realizaron cambios importantes en la organización política territorial: se incluyen nuevos departamentos (marcados con * en la Tabla 4.2), se establecen los 32 departamentos como se conocen actualmente, y se establece el área metropolitana de Santafé de Bogotá como Distrito Capital (DANE 2007b).

Un mapa de Colombia por departamentos se muestra en la Figura 4.1 y en la Tabla 4.2 se puede ver la lista de departamentos y regiones de Colombia.

Para este capítulo se construyeron nuevas tablas de vida abreviadas para los departamentos de Colombia y para el período 1985-2014, tanto para hombres como para mujeres. El periodo de análisis se definió teniendo los datos disponibles. Las defunciones por departamentos de Colombia están a partir de 1970 hasta 2014, organizadas por grupos de edades en cada sexo en The Latin America Human Mortality Database (Urdinola, Torres y Velasco 2015); y las proyecciones de población a nivel departamental por área, sexo y edad se encontraron para el período 1985-2017 en la página digital del Departamento Administrativo Nacional de Estadística de Colombia (DANE). Combinando esta información, las tablas de vida abreviadas para los departamentos por sexo se construyeron con el máximo de datos posibles, es decir, de 1985-2014 por área, sexo y edad.

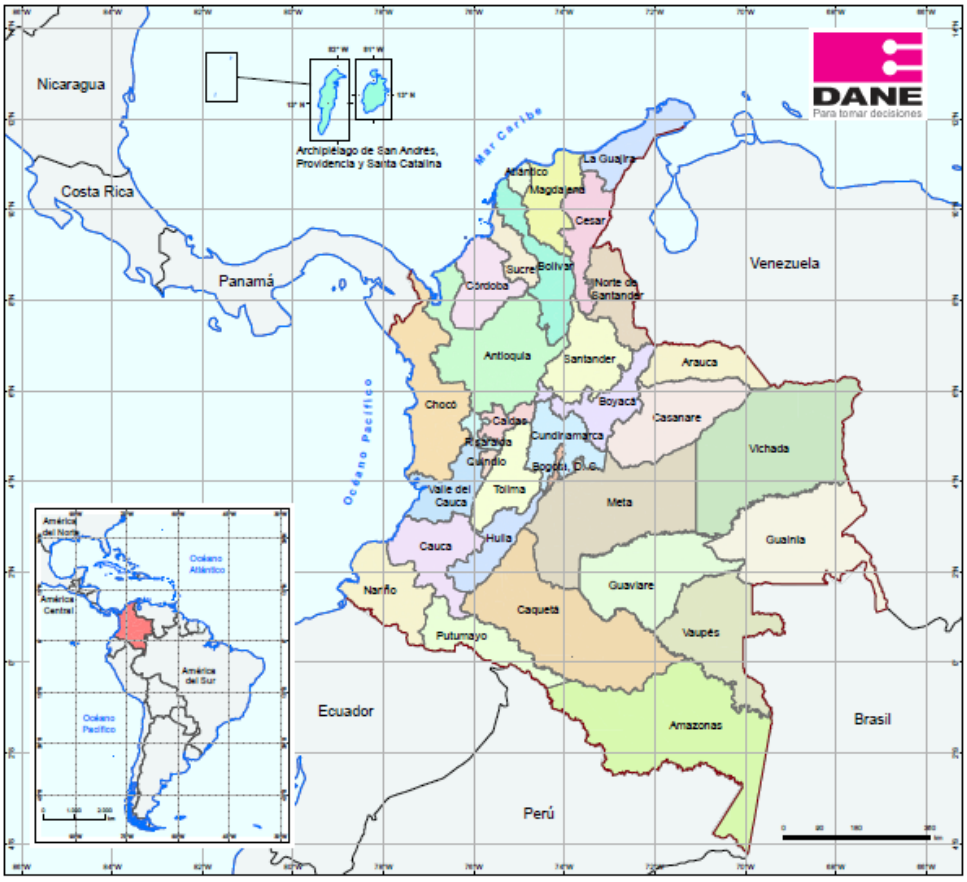


Figura 4.1: Ubicación geográfica del territorio colombiano y sus departamentos. Fuente: DANE, Dirección de Geoestadística.

Tabla 4.2: Listado de las regiones y departamentos de Colombia.

Región	Departamento	Municipios
Bogotá D.C.	Bogotá D.C.	1
Caribe	Atlántico	23
	Bolívar	46
	Cesár	25
	Córdoba	30
	Sucre	26
	Magdalena	30
	La Guajira	15
Oriental	Norte de Santander	40
	Santander	87
	Boyacá	123
	Cundinamarca	116
	Meta	29
Central	Antioquia	125
	Caldas	27
	Risaralda	14
	Quindío	12
	Tolima	47
	Huila	37
	Caquetá	16
Pacífica	Chocó	30
	Valle del Cauca	42
	Cauca	42
	Nariño	64
Nuevos departamentos *	Arauca	7
	Casanare	19
	Guanía	1
	Guaviare	4
	Amazonas	2
	Archipiélago de San Andrés, Providencia y Santa Catalina	1
	Putumayo	13
	Vichada	4
	Vaupés	3

* Los nuevos departamentos se adicionaron con la Constitución de 1991. Fuente: DANE.

Un resumen de los datos analizados se presentan en la Tabla 4.3:

Tabla 4.3: Resumen de datos

Variables	Categorías
Departamentos (32 + Capital)	33
Periodos (1985-2014)	30
Edades agrupadas (Intervalos de edad)	18
Sexo	2
Total de datos	35640

4.6.2 Agrupación de departamentos mediante análisis multivariado

Para el análisis de la mortalidad por departamentos se realizaron las siguientes fases:

1. Actualización de tabla de vida para departamentos de Colombia con datos de 1985 a 2014.
2. Reducción de la dimensionalidad mediante análisis de componentes principales -ACP- con el objetivo de disminuir la cantidad de variables con la menor pérdida de información posible.
3. Agrupación de los departamentos según probabilidad de muerte a través del análisis de clúster y el análisis fuzzy clúster.
4. Caracterización de los grupos obtenidos según indicadores de mortalidad.
5. Selección de los indicadores para los grupos de departamentos mediante random forest y árboles de clasificación. Análisis de la importancia de dichos indicadores.
6. Interpretación de los resultados.

La agrupación de los departamentos (etapa 3) se realizó con la transformación logit para la probabilidad de muerte descrita en el Capítulo 2:

$$\text{logit}(q_{xt}) = \ln\left(\frac{q_{xt}}{1-q_{xt}}\right).$$

Cada uno de los análisis se realizó separadamente para cada sexo (hombres y mujeres) dado que para Colombia se presentan diferencias en su mortalidad.

Análisis de componentes principales

En este estudio se tenían 540 variables para cada departamento, una cantidad bastante elevada, por lo que se aplicó un análisis de componentes principales (ACP) como paso previo a la identificación de grupos.

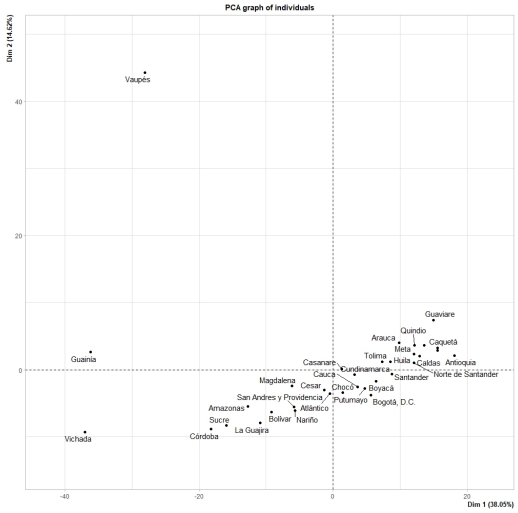
Entre los objetivos del ACP está analizar la posibilidad de representar adecuadamente una gran cantidad de información con un número menor de variables construidas como combinaciones lineales de las originales e identificadas como componentes principales, las cuales serán independientes.

La Tabla 4.4 muestra los valores propios y la varianza explicada de las componentes seleccionadas del ACP para hombres y mujeres (cuyos valores propios son mayores a uno). Al seleccionar 11 componentes para los hombres podemos explicar alrededor del 92 % de la varianza; y para las mujeres con las 12 componentes elegidas también podemos explicar alrededor del 92% de la varianza total.

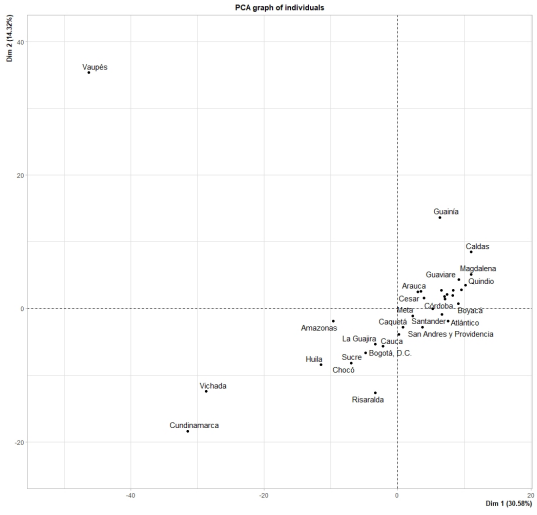
Para identificar una posible agrupación de los departamentos analizamos las 2 primeras componentes que son las que más varianza explican, alrededor del 53% en los hombres y 45% en las mujeres (ver Figura 4.2).

Tabla 4.4: Valores propios y varianza explicada por las componentes del ACP

Hombres				Mujeres			
Comp.	Valor propio	%Var.	%Var.Ac.	Comp.	Valor propio	%Var.	%Var.Ac.
1	205.45	38.04	38.04	1	165.11	30.57	30.58
2	78.95	14.62	52.66	2	77.33	14.32	44.90
3	54.76	10.14	62.81	3	70.39	6.00	57.93
4	38.46	7.12	69.93	4	45.38	5.46	66.34
5	32.29	5.97	75.91	5	39.21	5.06	73.60
6	26.84	4.97	80.88	6	30.38	4.42	79.23
7	19.27	3.57	84.45	7	22.39	4.17	83.37
8	16.93	3.13	87.58	8	14.63	3.16	86.08
9	11.88	2.20	89.79	9	10.96	2.86	88.11
10	7.12	1.31	91.11	10	7.57	2.41	89.52
11	6.00	1.11	92.22	11	7.14	1.32	90.84
–	–	–	–	12	5.87	1.09	91.92



(a) Hombres



(b) Mujeres

Figura 4.2: Distribución de departamentos según las componentes 1 y 2

En ambos sexos es clara la creación de grupo central que agrupa la mayoría de departamentos. En los hombres quedan separados de ese gran grupo central los departamentos Guanía y Vichada y muy alejado el departamento Vichada.

Por otra parte, en las mujeres los departamentos alejados son Cundinamarca y Vichada, y nuevamente queda muy alejado el departamento Vichada.

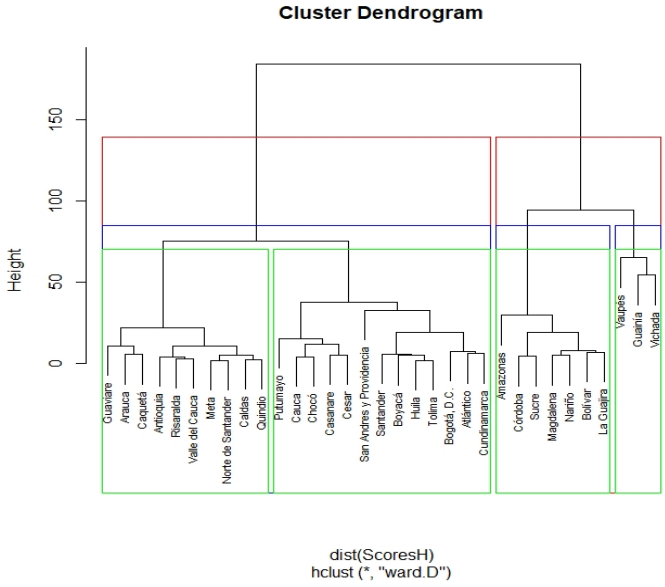
La información proveniente del ACP se utilizó posteriormente para agrupar los departamentos con comportamientos similares en cuanto a la mortalidad. La creación inicial de los grupos vistos en el ACP deben ser validados. Para lo anterior, se aplicó el análisis de clúster como se muestra en la siguiente sección.

Análisis de clúster jerárquico y fuzzy clúster

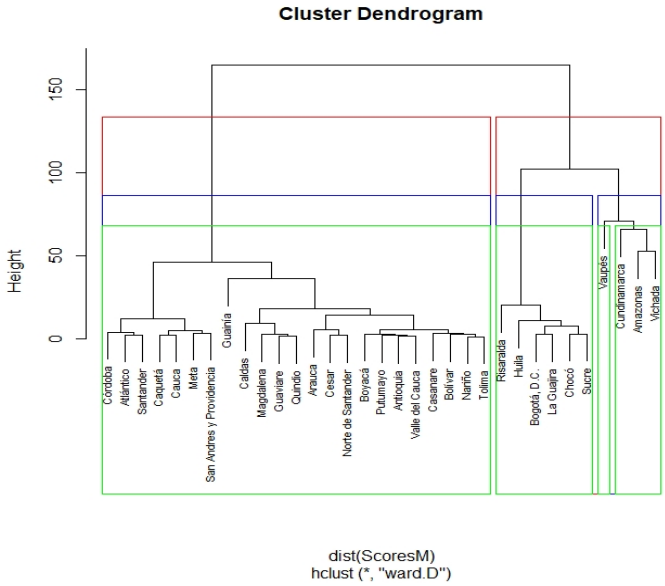
La identificación de los grupos de departamentos se realizó aplicando un análisis de clúster jerárquico primero. Este análisis clasificó los departamentos de Colombia en grupos relativamente homogéneos teniendo en cuenta las variables analizadas. Dentro de cada grupo o conglomerado, se identificaron los departamentos similares entre sí, logrando homogeneidad al interior del grupo y heterogeneidad entre los grupos.

Para el análisis de clúster jerárquico se utilizó el paquete FactorMineR de R, desarrollado por Husson y col. (2020) y para el fuzzy clúster, el paquete e1071 de R, diseñado por Meyer y col. (2021). Ambos análisis de clúster se realizaron con las nuevas variables resultantes del análisis de componentes principales.

El análisis de jerárquico se realizó junto con el método de Ward como algoritmo de resolución y la distancia euclídea. Como resultado se muestra el dendograma (Figura 4.3), herramienta que ayuda a visualizar las particiones y las distancias entre los grupos, lo cuál es de gran ayuda a la hora de decidir el número más idóneo de grupos. Los dendogramas obtenidos tras aplicar el análisis de clúster jerárquico señalan diferentes agrupaciones (2, 3 y 4 grupos) para hombres y mujeres. Analizando los resultados, tanto para los hombres como para las mujeres, se sugiere la creación de dos o tres grupos (señalados en rojo y azul respectivamente) para cada sexo.



(a) Hombres



(b) Mujeres

Figura 4.3: Dendogramas del análisis de clúster jerárquico

Tabla 4.5: Agrupación de departamentos según análisis de clúster con 2 grupos y 3 grupos, hombres

	2 Grupos		3 Grupos		
Sexo	**Grupo 1**	**Grupo 2**	**Grupo 1**	**Grupo 2**	**Grupo 3**
H	Antioquia	Amazonas	Antioquia	Amazonas	Guanía
	Arauca	Bolívar	Arauca	Bolívar	Vaupés
	Atlántico	Córdoba	Atlántico	Córdoba	Vichada
	Bogotá	Guanía	Bogotá	La Guajira	
	Boyacá	La Guajira	Boyacá	Magdalena	
	Caldas	Magdalena	Caldas	Nariño	
	Caquetá	Nariño	Caquetá	Sucre	
	Casanare	Sucre	Casanare		
	Cauca	Vaupés	Cauca		
	Cesar	Vichada	Cesar		
	Chocó		Chocó		
	Cundinamarca		Cundinamarca		
	Guaviare		Guaviare		
	Huila		Huila		
	Meta		Meta		
	N. de Santander		N. de Santander		
	Putumayo		Putumayo		
	Quindio		Quindio		
	Risaralda		Risaralda		
	S. Andrés y Prov.		S. Andrés y Prov.		
	Santander		Santander		
	Tolima		Tolima		
	Valle del Cauca		Valle del Cauca		
M	Antioquia	Amazonas	Antioquia	Bogotá	Amazonas
	Arauca	Bogotá	Arauca	Chocó	Cundinamarca
	Atlántico	Chocó	Atlántico	Huila	Vaupés
	Bolívar	Cundinamarca	Bolívar	La Guajira	Vichada
	Boyacá	Huila	Boyacá	Risaralda	
	Caldas	La Guajira	Caldas	Sucre	
	Caquetá	Risaralda	Caquetá		
	Casanare	Sucre	Casanare		
	Cauca	Vaupés	Cauca		
	Cesar	Vichada	Cesar		
	Córdoba		Córdoba		
	Guanía		Guanía		
	Guaviare		Guaviare		
	Magdalena		Magdalena		
	Meta		Meta		
	Nariño		Nariño		
	N. de Santander		N. de Santander		
	Putumayo		Putumayo		
	Quindio		Quindio		
	S. Andrés y Prov.		S. Andrés y Prov.		
	Santander		Santander		
	Tolima		Tolima		
	Valle del Cauca		Valle del Cauca		

En la Tabla 4.5 se muestran los departamentos que conformarían la división en 2 ó 3 grupos para cada sexo. Se puede observar que al pasar de 2 clústers a 3 clústers la diferencia entre los grupos radica en que se genera una partición del segundo grupo. En ambos sexos, ese tercer grupo tendría a los departamentos de Vaupés y Vichada. Comparando estas agrupaciones con lo que se había obtenido en el ACP (4.4), observamos que en esta ocasión los departamentos de que componen ese tercer grupo coinciden con los que quedaban más alejados e el ACP.

Posterior al análisis de clúster jerárquico, se realizó el análisis del fuzzy clúster donde se analizó la probabilidad de pertenencia a diferentes grupos. Lo anterior con el fin de estudiar qué departamentos son más difíciles de clasificar y validar los anteriores clúster (ver Tablas 4.6 y 4.7).

Tabla 4.6: Valores de pertenencia de los departamentos de Colombia a los grupos creados por el fuzzy clúster en hombres

	2 Grupos		3 Grupos		
Departamento	1	2	1	2	3
Amazonas	0.30	**0.70**	0.25	0.11	**0.64**
Antioquia	**0.91**	0.09	**0.90**	0.02	0.08
Arauca	**0.91**	0.09	**0.88**	0.02	0.10
Atlántico	**0.56**	0.44	0.41	0.04	**0.55**
Bogotá, D.C.	**0.74**	0.26	**0.64**	0.04	0.32
Bolívar	0.08	**0.92**	0.02	0.00	**0.97**
Boyacá	**0.93**	0.07	**0.88**	0.01	0.11
Caldas	**0.96**	0.04	**0.96**	0.01	0.04
Caquetá	**0.91**	0.09	**0.88**	0.02	0.09
Casanare	**0.81**	0.19	**0.68**	0.02	0.30
Cauca	**0.81**	0.19	**0.70**	0.03	0.27
Cesar	**0.54**	0.46	0.34	0.02	**0.63**
Chocó	**0.67**	0.33	**0.52**	0.03	0.44
Córdoba	0.07	**0.93**	0.08	0.04	**0.88**
Cundinamarca	**0.84**	0.16	**0.73**	0.02	0.25
Guainía	0.39	**0.61**	0.28	0.31	**0.41**
Guaviare	**0.86**	0.14	**0.82**	0.05	0.14
Huila	**0.99**	0.01	**0.99**	0.00	0.01
La Guajira	0.07	**0.93**	0.03	0.01	**0.96**
Magdalena	0.19	**0.81**	0.08	0.01	**0.90**
Meta	**0.98**	0.02	**0.98**	0.00	0.02
Nariño	0.21	**0.79**	0.08	0.01	**0.91**
N. de Santander	**0.98**	0.02	**0.99**	0.00	0.01
Putumayo	**0.71**	0.29	**0.62**	0.05	0.33
Quindio	**0.96**	0.04	**0.96**	0.01	0.03
Risaralda	**0.94**	0.06	**0.93**	0.01	0.05
S. Andrés y Prov.	0.42	**0.58**	0.35	0.10	**0.55**
Santander	**0.94**	0.06	**0.92**	0.01	0.07
Sucre	0.05	**0.95**	0.06	0.02	**0.92**
Tolima	**0.99**	0.01	**0.97**	0.00	0.03
Valle del Cauca	**0.93**	0.07	**0.93**	0.01	0.06
Vaupés	0.44	**0.56**	0.03	**0.94**	0.03
Vichada	0.31	**0.69**	0.24	0.26	**0.50**

Tabla 4.7: Valores de pertenencia de los departamentos de Colombia a los grupos creados por el fuzzy clúster en mujeres

	2 Grupos		3 Grupos		
Departamento	1	2	1	2	3
Amazonas	0.46	**0.54**	0.39	**0.44**	0.17
Antioquia	**0.98**	0.02	**0.98**	0.01	0.00
Arauca	**0.96**	0.04	**0.93**	0.07	0.01
Atlántico	**0.88**	0.12	**0.78**	0.21	0.01
Bogotá, D.C.	0.36	**0.64**	0.07	**0.93**	0.00
Bolívar	**0.98**	0.02	**0.97**	0.02	0.00
Boyacá	**0.96**	0.04	**0.96**	0.04	0.00
Caldas	**0.87**	0.13	**0.84**	0.13	0.03
Caquetá	**0.72**	0.28	0.44	**0.54**	0.01
Casanare	**0.99**	0.01	**0.99**	0.01	0.00
Cauca	**0.61**	0.39	0.28	**0.71**	0.01
Cesar	**0.97**	0.03	**0.92**	0.08	0.00
Chocó	0.12	**0.88**	0.03	**0.96**	0.00
Córdoba	**0.97**	0.03	**0.93**	0.07	0.00
Cundinamarca	0.43	**0.57**	0.33	**0.40**	0.27
Guainía	**0.71**	0.29	**0.64**	0.27	0.10
Guaviare	**0.96**	0.04	**0.96**	0.03	0.00
Huila	0.07	**0.93**	0.09	**0.89**	0.02
La Guajira	0.31	**0.69**	0.05	**0.94**	0.00
Magdalena	**0.93**	0.07	**0.92**	0.07	0.01
Meta	**0.85**	0.15	**0.66**	0.33	0.01
Nariño	**0.99**	0.01	**0.99**	0.01	0.00
N. de Santander	**0.96**	0.04	**0.91**	0.08	0.01
Putumayo	**0.98**	0.02	**0.98**	0.01	0.00
Quindio	**0.95**	0.05	**0.96**	0.04	0.00
Risaralda	0.35	**0.65**	0.23	**0.72**	0.05
S. Andrés y Prov.	**0.83**	0.17	**0.62**	0.36	0.01
Santander	**0.94**	0.06	**0.88**	0.12	0.01
Sucre	0.20	**0.80**	0.02	**0.98**	0.00
Tolima	**0.99**	0.01	**1.00**	0.00	0.00
Valle del Cauca	**0.96**	0.04	**0.96**	0.03	0.00
Vaupés	0.43	**0.57**	0.02	0.02	**0.96**
Vichada	0.36	**0.64**	0.30	**0.46**	0.24

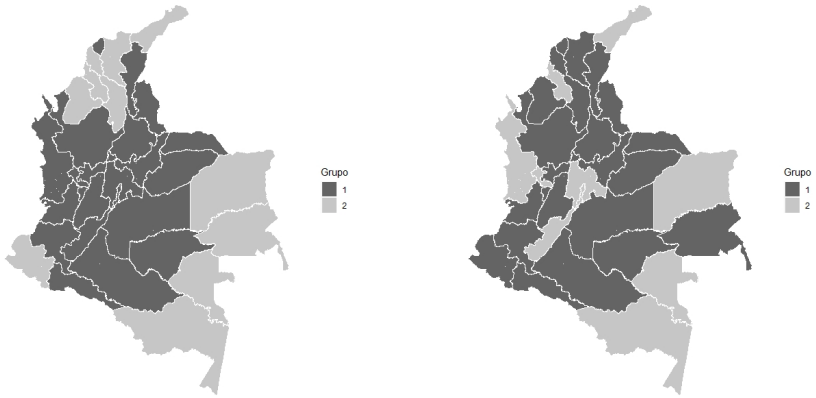
La Tabla 4.8 muestra los valores de los índices de validez para el fuzzy clúster y indican cuan compactos son los grupos creados. En los hombres, según los índices FS, PE y PC, la mejor agrupación sería con 2 grupos. En las mujeres, el índice XB indica que se deben utilizar 3 ó 4 grupos, sin embargo los índices FS, PE y PC indican que lo mejor sería utilizar 2 grupos. Por lo anterior,

se decide utilizar 2 grupos para la caracterización de los departamentos de manera agrupada teniendo en cuenta que los resultados del fuzzy clúster están en concordancia con los resultados del ACP y del clúster jerárquico.

Tabla 4.8: Índices de validación del análisis fuzzy clúster

	Hombres			Mujeres		
Índice	2 grupos	3 grupos	4 grupos	2 grupos	3 grupos	4 grupos
XB	0.01	0.01	0.01	0.02	**0.01**	0.02
FS	**-2974**	-3398	-7894	**-2106**	-11628	-8990
PE	**0.39**	0.65	0.70	**0.35**	0.40	0.60
PC	**0.76**	0.63	0.62	**0.78**	**0.78**	0.67

La ubicación geográfica de los departamentos que conforman los grupos creados por el fuzzy clúster se muestran en la Figura 4.4.



(a) Hombres (b) Mujeres

Figura 4.4: Grupos creados por el fuzzy clúster en el mapa de Colombia

4.6.3 Caracterización de grupos de departamentos según indicadores de mortalidad

Para caracterizar los grupos identificados en el análisis de fuzzy clúster se utilizaron los indicadores de mortalidad utilizados en el Capítulo 2: la mortalidad infantil (q_0), la esperanza de vida al nacer (e_0), la esperanza de vida a los 65 años (e_{65}), el coeficiente de Gini al nacer (G_0) y el coeficiente de Gini a los 65 años (G_{65}). Estos indicadores se calcularon para 1985 y 2014, el principio y el final del periodo de tiempo analizado, con el objetivo de analizar su evolución en el tiempo e identificar posibles patrones o tendencias.

La descripción de los grupos de departamentos con los indicadores de mortalidad en 1985 y 2014 se muestra en las Tablas 4.9 - 4.12, para hombres y mujeres respectivamente.

En los hombres, la esperanza de vida al nacer y la esperanza de vida a los 65 años han aumentado, de manera más notable en el Grupo 1 donde hay 6 años de diferencia para el periodo analizado, mientras que en el grupo 2 solo aumentó un año. Adicionalmente, el índice de Gini al nacer tuvo un descenso en ambos grupos, siendo mayor esta disminución en el grupo 2 (de 0.17 a 0.12). En cuanto a la mortalidad infantil, disminuyó en el grupo 1 pero en el grupo 2 permaneció igual. El índice de Gini a los 65 años mantuvo valores similares para ambos grupos en este periodo.

Para las mujeres, también ha aumentado la esperanza de vida al nacer en ambos grupos, aunque de una manera más marcada en el grupo 1, 6 años de diferencia en el periodo analizado, mientras que en el grupo 2 solo aumentó un año. En cuanto a la esperanza de vida a los 65 años, en el grupo 1, aumentó 2 años mientras que en grupo 2 permaneció igual. El índice de Gini al nacer disminuyó en ambos grupos, siendo más notable la disminución en el grupo 1 donde este indice presentaba un mayor valor al inicio del periodo analizado. La mortalidad infantil y el índice de Gini a los 65 años permanecieron con similares valores para ambos grupos en este periodo.

Comparando el compartimento de los índices de mortalidad entre los sexos en el periodo analizado, se observa que la esperanza de vida al nacer y la esperanza de vida a los 65 años tienen valores mayores en las mujeres. La mortalidad infantil, el índice de Gini al nacer y a los 65 años tuvieron valores similares en 2014 para los grupos creados para los departamentos.

En resumen, en ambos sexos se observan cambios en los indicadores de mortalidad para el periodo de 1985 hasta 2014. Para determinar si estos cambios podían considerarse estadísticamente significativos y, teniendo en cuenta que

los datos no presentaban homogeneidad de varianzas ni distribución normal de manera general, se aplicó la prueba no paramétrica de Wilcoxon al nivel del 5% para datos pareados entre los valores de 1985 y 2014 para cada indicador. La Tabla 4.12 muestra los resultados del análisis pareado.

Tabla 4.9: Caracterización de los grupos de departamentos de Colombia según indicadores de mortalidad en 1985 para los hombres

Departamento	q_0	e_0	e_{65}	G_0	G_{65}	Grupo
Antioquia	0.0288	64.81	13.87	0.1646	0.0438	1
Arauca	0.0122	71.38	17.36	0.2046	0.0385	1
Atlántico	0.0219	71.18	14.16	0.1435	0.0388	1
Bogotá, D.C.	0.0282	65.03	11.24	0.2097	0.0379	1
Boyacá	0.0351	66.29	13.99	0.1423	0.0401	1
Caldas	0.0330	66.56	13.36	0.1300	0.0386	1
Caquetá	0.0471	66.71	15.17	0.1561	0.0385	1
Casanare	0.0111	69.95	15.27	0.1199	0.0375	1
Cauca	0.0395	64.48	14.48	0.1391	0.0386	1
Cesar	0.0105	72.35	15.95	0.1464	0.0367	1
Chocó	0.0340	65.24	13.83	0.2009	0.0387	1
Cundinamarca	0.0206	71.52	14.62	0.1792	0.0392	1
Guaviare	0.0468	53.01	12.95	0.2379	0.0377	1
Huila	0.0364	65.70	13.25	0.1063	0.0355	1
Meta	0.0283	64.14	14.01	0.2305	0.0384	1
Norte de Santander	0.0327	67.97	14.32	0.1686	0.0376	1
Putumayo	0.0180	69.86	16.62	0.1826	0.0387	1
Quindio	0.0277	68.80	13.74	0.2035	0.0395	1
Risaralda	0.0281	67.12	13.58	0.1070	0.0368	1
Santander	0.0245	66.55	13.18	0.1282	0.0372	1
Tolima	0.0225	69.16	14.05	0.1353	0.0000	1
Valle del Cauca	0.0245	67.31	13.69	0.1412	0.0000	1
Media G1	**0.0278**	**67.05**	**14.21**	**0.1626**	**0.0349**	
Amazonas	0.0200	65.66	12.88	0.2052	0.0163	2
Bolívar	0.0133	75.26	16.20	0.1329	0.0343	2
Córdoba	0.0057	78.47	17.59	0.1327	0.0404	2
Guainía	0.0129	74.13	14.14	0.1975	0.0385	2
La Guajira	0.0070	75.68	17.32	0.1774	0.0440	2
Magdalena	0.0115	75.46	16.30	0.1882	0.0387	2
Nariño	0.0254	71.93	15.56	0.135	0.0373	2
San Andrés y Providencia	0.0074	73.27	14.04	0.1305	0.0386	2
Sucre	0.0085	77.43	17.05	0.1506	0.0386	2
Vaupés	0.0143	68.20	12.70	0.1731	0.0370	2
Vichada	0.0036	78.12	14.70	0.2883	0.0417	2
Media G2	**0.0117**	**73.96**	**15.32**	**0.1738**	**0.0368**	

Tabla 4.10: Caracterización de los grupos de departamentos de Colombia según indicadores de mortalidad en 2014 para los hombres

Departamento	q_0	e_0	e_{65}	G_0	G_{65}	Grupo
Antioquia	0.0091	73.60	13.87	0.1658	0.0390	1
Arauca	0.0083	70.45	17.36	0.1356	0.0374	1
Atlántico	0.0146	73.70	14.16	0.0984	0.0353	1
Bogotá, D.C.	0.0120	74.81	11.24	0.1605	0.0383	1
Boyacá	0.0074	74.96	13.99	0.1294	0.0370	1
Caldas	0.0078	73.57	13.36	0.0953	0.0341	1
Caquetá	0.0090	72.21	15.17	0.1194	0.0369	1
Casanare	0.0094	74.13	15.27	0.1082	0.0353	1
Cauca	0.0107	74.90	14.48	0.1371	0.0373	1
Cesar	0.0129	74.31	15.95	0.0980	0.0344	1
Chocó	0.0157	75.09	13.83	0.1224	0.0376	1
Cundinamarca	0.0101	74.78	14.62	0.1290	0.0362	1
Guaviare	0.0083	74.06	12.95	0.1523	0.0385	1
Huila	0.0113	73.63	13.25	0.1016	0.0344	1
Meta	0.0114	72.78	14.01	0.1568	0.0371	1
Norte de Santander	0.0091	73.00	14.32	0.1354	0.0376	1
Putumayo	0.0099	74.00	16.62	0.1401	0.0341	1
Quindio	0.0111	71.61	13.74	0.1743	0.0391	1
Risaralda	0.0096	72.61	13.58	0.1119	0.0352	1
Santander	0.0104	73.92	13.18	0.0951	0.0337	1
Tolima	0.0123	73.00	14.05	0.1255	0.0382	1
Valle del Cauca	0.0084	71.41	13.69	0.1645	0.0299	1
Media G1	**0.0103**	**73.48**	**15.58**	**0.1299**	**0.0362**	
Amazonas	0.0140	72.90	12.88	0.1600	0.0396	2
Bolívar	0.0127	75.15	16.20	0.1136	0.0359	2
Córdoba	0.0137	76.08	17.59	0.0961	0.0353	2
Guainía	0.0111	71.73	14.14	0.1123	0.0359	2
La Guajira	0.0141	76.14	17.32	0.1354	0.0363	2
Magdalena	0.0140	74.28	16.30	0.1164	0.0369	2
Nariño	0.0093	75.41	15.56	0.1039	0.0336	2
San Andrés y Providencia	0.0197	72.42	14.04	0.1385	0.0366	2
Sucre	0.0119	76.06	17.05	0.1201	0.0365	2
Vaupés	0.0029	74.02	12.70	0.0995	0.0303	2
Vichada	0.0066	77.51	14.70	0.1279	0.0323	2
Media G2	**0.0118**	**74.70**	**15.89**	**0.1203**	**0.0353**	

Tabla 4.11: Caracterización de los grupos de departamentos de Colombia según indicadores de mortalidad en 1985 para las mujeres

Departamento	q_0	e_0	e_{65}	G_0	G_{65}	Grupo
Antioquia	0.0237	70.42	13.83	0.1913	0.0397	1
Arauca	0.0069	75.98	16.26	0.1373	0.0372	1
Atlántico	0.0239	69.12	12.32	0.1752	0.0384	1
Bolívar	0.0282	70.89	14.13	0.1497	0.0397	1
Boyacá	0.0225	70.78	13.54	0.1641	0.0404	1
Caldas	0.0371	70.35	15.57	0.2183	0.0393	1
Caquetá	0.0086	73.81	14.02	0.1578	0.0390	1
Casanare	0.0312	69.12	14.77	0.1750	0.0376	1
Cauca	0.0083	75.81	16.49	0.1512	0.0391	1
Cesar	0.0227	71.14	15.72	0.1894	0.0391	1
Córdoba	0.0173	74.39	15.38	0.2329	0.0366	1
Guainía	0.0376	60.82	10.49	0.1641	0.0388	1
Guaviare	0.0268	69.60	13.50	0.1435	0.0393	1
Magdalena	0.0280	68.85	13.78	0.0981	0.0329	1
Meta	0.0199	73.73	15.90	0.1261	0.0370	1
Nariño	0.0236	72.48	15.05	0.1736	0.0401	1
Norte de Santander	0.0120	75.16	16.34	0.1338	0.0374	1
Putumayo	0.0229	71.66	14.17	0.1298	0.0390	1
Quindio	0.0228	70.55	13.95	0.1124	0.0352	1
San Andrés y Providencia	0.0141	74.55	15.66	0.1396	0.0383	1
Santander	0.0191	71.66	14.15	0.1779	0.0393	1
Tolima	0.0173	72.44	14.76	0.1991	0.0173	1
Valle del Cauca	0.0199	72.19	14.63	0.2613	0.0000	1
Media G1	**0.0214**	**71.54**	**14.54**	**0.1652**	**0.0358**	
Amazonas	0.0106	75.31	16.89	0.1495	0.0399	2
Bogotá, D.C.	0.0084	77.89	16.99	0.1194	0.0382	2
Chocó	0.0042	79.51	17.63	0.1503	0.0421	2
Cundinamarca	0.0069	73.67	13.64	0.1377	0.0381	2
Huila	0.0046	81.15	18.56	0.1758	0.0389	2
La Guajira	0.0097	77.64	16.84	0.1114	0.0455	2
Risaralda	0.0030	76.97	16.96	0.1649	0.0386	2
Sucre	0.0050	78.94	17.41	0.1383	0.0384	2
Vaupés	0.0191	70.61	13.33	0.1915	0.0000	2
Vichada	0.0013	80.52	17.97	0.0638	0.0160	2
Media G2	**0.0072**	**77.22**	**16.62**	**0.1402**	**0.0335**	

Tabla 4.12: Caracterización de los grupos de departamentos de Colombia según indicadores de mortalidad en 2014 para las mujeres

Departamento	q_0	e_0	e_{65}	G_0	G_{65}	Grupo
Antioquia	0.0075	77.96	13.83	0.1363	0.0373	1
Arauca	0.0067	76.46	16.26	0.1315	0.0366	1
Atlántico	0.0091	78.28	12.32	0.1620	0.0389	1
Bolívar	0.0068	78.60	14.13	0.1357	0.0360	1
Boyacá	0.0065	78.15	13.54	0.1088	0.0351	1
Caldas	0.0072	76.52	15.57	0.1377	0.0379	1
Caquetá	0.0059	77.63	14.02	0.1080	0.0344	1
Casanare	0.0079	78.74	14.77	0.1460	0.0357	1
Cauca	0.0105	77.74	16.49	0.1033	0.0366	1
Cesar	0.0125	78.79	15.72	0.1298	0.0375	1
Córdoba	0.0075	78.06	15.38	0.1965	0.0378	1
Guainía	0.0070	75.75	10.49	0.1498	0.0375	1
Guaviare	0.0099	76.71	13.50	0.1130	0.0354	1
Magdalena	0.0113	76.32	13.78	0.1352	0.0358	1
Meta	0.0082	78.67	15.90	0.1051	0.0345	1
Nariño	0.0073	77.38	15.05	0.1254	0.0371	1
Norte de Santander	0.0067	77.99	16.34	0.1021	0.0351	1
Putumayo	0.0089	76.83	14.17	0.1381	0.0366	1
Quindio	0.0088	76.96	13.95	0.1126	0.0353	1
San Andrés y Providencia	0.0128	77.10	15.66	0.105	0.0350	1
Santander	0.0082	78.05	14.15	0.1599	0.0374	1
Tolima	0.0090	77.31	14.76	0.1453	0.0372	1
Valle del Cauca	0.0062	77.63	14.63	0.139	0.0314	1
Media G1	**0.0083**	**77.55**	**16.57**	**0.1316**	**0.0361**	
Amazonas	0.0134	77.89	16.89	0.1071	0.0365	2
Bogotá, D.C.	0.0102	78.30	16.99	0.1054	0.0353	2
Chocó	0.0117	77.97	17.63	0.1171	0.0357	2
Cundinamarca	0.0062	75.72	13.64	0.1015	0.0359	2
Huila	0.0109	79.43	18.56	0.1375	0.0370	2
La Guajira	0.0102	77.45	16.84	0.0998	0.0351	2
Risaralda	0.0186	78.10	16.96	0.1449	0.0374	2
Sucre	0.0116	78.14	17.41	0.1025	0.0350	2
Vaupés	0.0153	74.29	13.33	0.1304	0.0304	2
Vichada	0.0074	79.79	17.97	0.1122	0.0287	2
Media G2	**0.0115**	**77.71**	**16.59**	**0.1158**	**0.0346**	

La Tabla 4.13 muestra los resultados de la prueba de Wilcoxon para los indicadores de mortalidad comparando sus valores entre los años 1985 y 2014, donde es notable que en ambos sexos se produjo una mejora estadísticamente significativa en todos los indicadores de mortalidad analizados.

Tabla 4.13: Prueba no paramétrica de Wilcoxon para comparar el comportamiento de los indicadores de mortalidad en los años 1985 y 2014

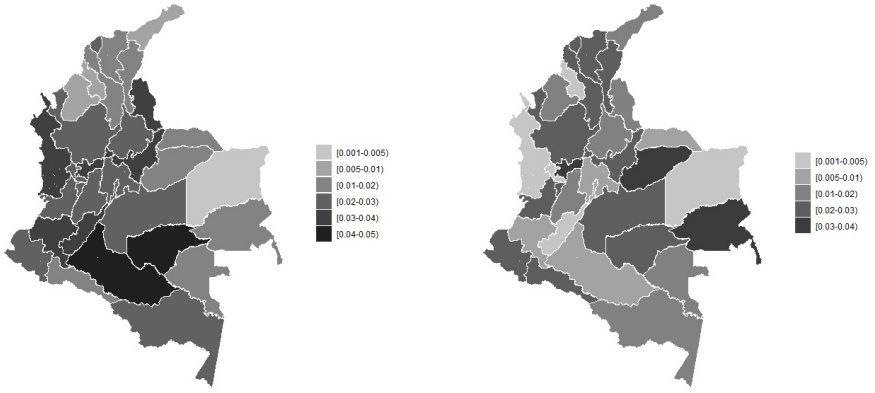
	Hombres		Mujeres	
Indicador	W test	p-value	W test	p-value
q_0	500	0.0000	466	0.0006
e_0	45	0.0000	25	0.0000
e_{65}	73	0.0000	54	0.0000
G_0	541	0.0000	512	0.0000
G_65	447	0.0022	409	0.0207

Las Figuras 4.5 y 4.6 presentan la mortalidad infantil por sexo para los departamentos de Colombia en 1985 y 2014 respectivamente. Aunque las diferencias son muy notables entre sexos en cada periodo, en 2014 hay un cierto acercamiento en su comportamiento.

En 1985, se muestran seis categorías para la mortalidad infantil en los hombres, donde el mínimo fue 0,0035 y el máximo 0,0470; y cinco categorías para las mujeres con en mínimo de 0,0013 y el máximo 0,0375. En los hombres, el 36% de los departamentos presentó una mortalidad infantil entre $[0,02; 0,03)$, el 24% presentó valores entre $[0,01; 0,02)$ y el el 18% estuvo en la categoría $[0,03; 0,04)$. Los departamentos con mayor mortalidad infantil (superior a 0,04) fueron Caquetá y Guaviare; el departamento con menor mortalidad infantil fue Vichada (inferior a 0.005). En las mujeres, el 36% de los departamentos estuvo en la categoría $[0,02; 0,03)$, el 24% en la categoría $[0,01; 0,02)$ y el 15% en la categoría $[0,005; 0,01)$. Los departamentos con mayor mortalidad infantil fueron Caldas, Casanare y Guanía (superior a 0,03); los departamentos con menor mortalidad infantil fueron Chocó, Hulia, Risaralda, Sucre y Vichada (inferior a 0,005).

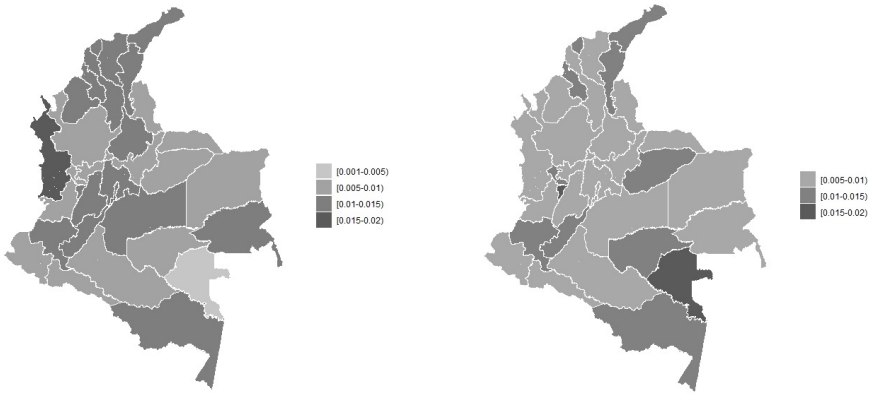
En 2014, la mortalidad infantil presentó valores mucho menores con respecto al periodo anterior, por lo que se modificó la escala de valores en los mapas. Para este año, se muestran cuatro categorías para los hombres, donde el mínimo fue 0,029 y el máximo 0,0196; y tres categorías para las mujeres con en mínimo de 0,0059 y un máximo de 0,0186. En los hombres, el 52% de los departamentos presentó una mortalidad infantil entre $[0,01; 0,015)$, el 39% estuvo en la categoría $[0,005; 0,01)$. Los departamentos con mayor mortalidad infantil (superior a 0,015) fueron Chocó y San Andrés y Providencia, el departamento con menor mortalidad infantil fue Vaupés (inferior a 0,005). En las mujeres, el 64% de los departamentos estuvo en la categoría más baja $[0,005; 0,01)$, el

30% en la siguiente categoría $[0,01; 0,015)$ y el 6% en la categoría de mayores valores $[0,015; 0,02)$ con solo dos departamentos: Risaralda y Vaupés.



(a) Hombres (b) Mujeres

Figura 4.5: Mortalidad infantil por departamentos de Colombia en 1985



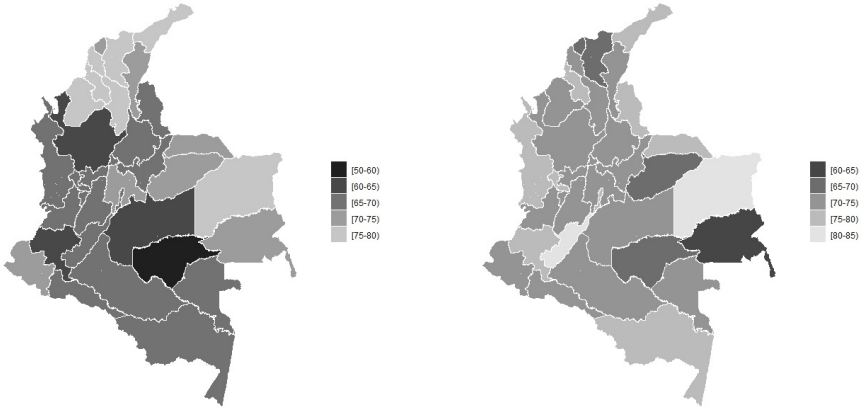
(a) Hombres (b) Mujeres

Figura 4.6: Mortalidad infantil por departamentos de Colombia en 2014

Aunque de manera general se observan mejoras de este indicador, algunos departamentos tuvieron un retroceso en este periodo como es el caso de San Andrés y Providencia, Cesar, Córdoba, la Guajira, Magdalena, Sucre y Vichada, en los hombres; y los departamentos de Amazonas, Cauca, Chocó, Cundinamarca, Huila, La Guajira, Risaralda, Sucre y Vichada, en las mujeres.

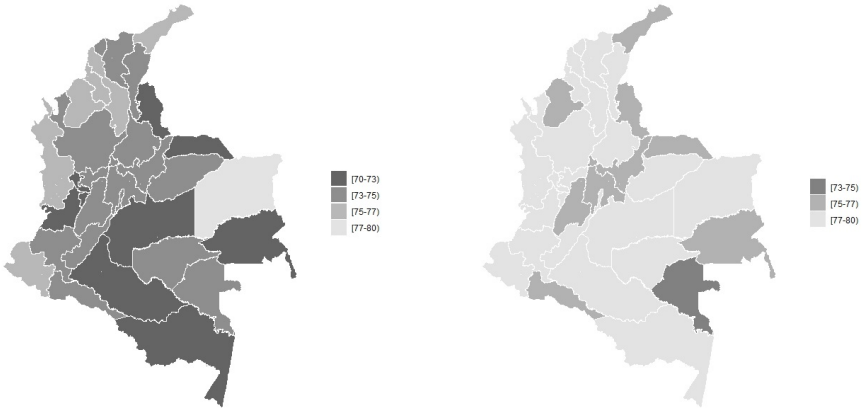
Las Figuras 4.7 y 4.8 presentan la esperanza de vida al nacer por sexo para Colombia en 1985 y 2014 respectivamente. Los mapas muestran como las diferencias geográficas entre sexos en la esperanza de vida al nacer permanecen a través del tiempo. En los hombres la media pasó de 69 a 74 años, y en las mujeres este indicador pasó fue de 73 a 78 años, lo que indica que en esos 30 años en ambos sexos ha mejorado pero entre los sexos sigue existiendo una diferencia importante.

En 1985, se muestran cinco categorías para los hombres, donde el mínimo fue 53 años y el máximo 78 años; y cinco categorías para las mujeres con en mínimo de 60 años y el máximo de 81 años. En los hombres el 45% de los departamentos se encuentra en la categoría $[65; 70)$ con valores alrededor de la media que fue 69 años, el 24% en la categoría $[70; 75)$ y el el 18% en la categoría $[75; 80)$. En las mujeres, el 52% de los departamentos se encuentra en la categoría $[70; 75)$ con valores alrededor de la media que fue 73 años, el 28% en la categoría $[75; 80)$ y el 12% en la categoría $[65; 70)$. Para este año, en los hombres los departamentos con menor esperanza de vida fueron Gauviare, Meta, Antioquia y Cauca que representan el 12 % del total de departamentos, ubicados en el Grupo 1 (donde la media fue de 67 años, 2 años por debajo de la media nacional); mientras que en las mujeres el departamento con menor esperanza de vida fue Guanía ubicado en la categoría $[60; 65)$ y en el Grupo 1 donde la media fue de 71 años.



(a) Hombres (b) Mujeres

Figura 4.7: Esperanza de vida al nacer por departamentos de Colombia en 1985



(a) Hombres (b) Mujeres

Figura 4.8: Esperanza de vida al nacer por departamentos de Colombia en 2014

En 2014, la esperanza de vida tuvo un mejoramiento significativo en ambos sexos lo que se refleja en su distribución. Para este año, se muestran cuatro

categorías para los hombres, donde el mínimo fue 70 años y el máximo 77 años; y tres categorías para las mujeres con en mínimo de 74 años y el máximo de 80 años. En los hombres, el 48% de los departamentos se encuentra en la categoría [73; 75) años con valores alrededor de la media que fue 74 años, el 30% en la categoría [70; 73) años, el el 18% en la categoría [75; 77) y el 3% en la categoría [77; 80) donde solo estuvo el departamento de Vichada. En las mujeres, el 70% de los departamentos se ubicó en la categoría [77; 80) años con valores alrededor de la media que fue 78 años, el 27% en la categoría [75; 77) años y el 3% en la categoría [73; 75) años. Para este año, los departamentos con menor esperanza de vida en los hombres (menor a 73 años) fueron: Amazonas, Arauca, Archipiélago de San Andrés y Providencia, Caquetá, Guanía, Meta, Norte de Santander, Quindío, Risaralda, Valle del Cauca, todos ubicados en el Grupo 1, mientras que en las mujeres, sólo el departamento del Cauca tuvo esperanza de vida menor a 75 años.

Al igual que para la mortalidad infantil, la esperanza de vida al nacer tuvo un descenso en el periodo analizado en algunos departamentos. En los hombres fueron los departamentos de Arauca, Córdoba, Guanía, Magdalena y Sucre; y en las mujeres los departamentos de Chocó, Huila, Sucre y Vichada.

Para complementar la caracterización de los grupos creados para los departamentos de Colombia por el análisis de fuzzy clúster y obtener una medida de la importancia de los indicadores de mortalidad, se aplicó la técnica de random forest utilizando el paquete randomForest de R (Liaw y Wiener 2002) con los indicadores de mortalidad para el año 2014 (el año más reciente del periodo analizado). Previo a este análisis se balancearon las muestras de cada grupo en ambos sexos haciendo uso de la función upSample del paquete caret desarrollado por Kuhn (2021), para evitar que al construir los árboles no se favoreciera excesivamente la clasificación en los grupos más grandes.

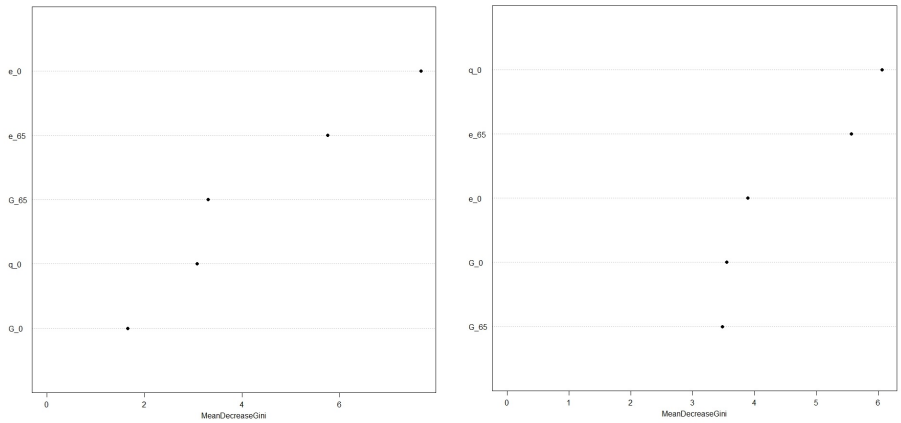
La Figura 4.9 muestra la importancia de cada indicador de mortalidad en la construcción de los árboles de clasificación, es decir en la clasificación de los departamentos por grupos. En el eje Y se observan los indicadores, y su importancia en el eje X. Los indicadores de mortalidad más importantes están en la parte superior y una estimación de su importancia viene dada por la posición del punto en el eje de abscisas.

Para los hombres el orden de importancia de los indicadores fue: esperanza de vida al nacer, esperanza de vida a los 65 años, mortalidad infantil, índice de Gini a los 65 años y índice de Gini al nacer. Por otra parte, en las mujeres el orden fue el siguiente: mortalidad infantil, esperanza de vida a los 65 años, esperanza de vida al nacer, índice de Gini al nacer, índice de Gini a los 65 año.

El hecho de que la importancia de las variables sea diferente en la clasificación de los grupos por sexo, refuerza las diferencias existentes en el comportamiento de estos indicadores entre los sexos.

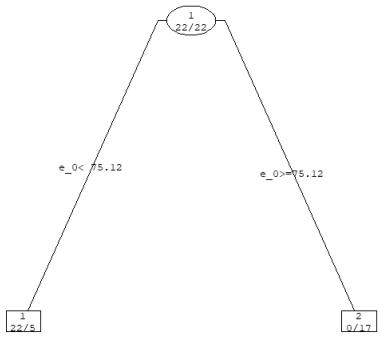
La estimación de la tasa de error OOB para la técnica del random forest tuvo valores similares en ambos sexos: 13.64% en los hombres, y 13.04% en las mujeres. Lo que indica similar precisión en ambos sexos.

En la Figura 4.10 se muestran los árboles de clasificación obtenidos mediante el paquete rpart de R, que ha sido implementado por Therneau y Atkinson (2019).

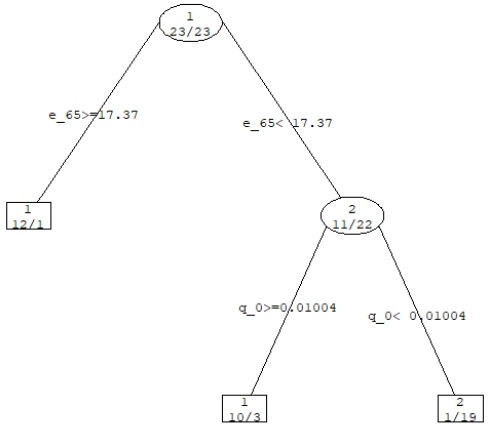


(a) Hombres (b) Mujeres

Figura 4.9: Importancia de los indicadores de mortalidad incluidos en las reglas de clasificación.



(a) Hombres



(b) Mujeres

Figura 4.10: Árboles de clasificación para los departamentos

En los hombres solo se utiliza para la clasificación la esperanza de vida al nacer, agrupando correctamente a todos los departamentos del Grupo 1. De esta manera en los hombres, los miembros del Grupo 1 tendrían una esperanza de vida al nacer menor a 75 años y los miembros del Grupo 2 mayor a 75 años. En las mujeres, se utilizó para las clasificación la esperanza de vida a los 65 años y la mortalidad infantil. En esta ocasión, con la esperanza de vida a los 65 años se clasifican correctamente la mayoría de miembros del Grupo 2 (22/23) con una esperanza de vida a los 65 años menor a 17 años, mientras que para el Grupo 1 solo 12 miembros están bien clasificados con una esperanza de vida a los 65 años mayor a 17 años. Para la segunda ramificación se utiliza la mortalidad infantil, clasificando la mayoría de miembros de manera correcta en sus correspondientes grupos. De esta manera, quedan en el Grupo 2 aquellos miembros con mortalidad infantil menor a 0,0104 y esperanza de vida a los 65 años menor a 17 años, y en el Grupo 1 quedarían ubicados aquellos miembros con mortalidad infantil mayor a 0,0104 y esperanza de vida a los 65 años menor a 17 años. Resumiendo,podemos decir que los indicadores que permiten caracterizar mejor los grupos de departamentos en los hombres es la esperanza de vida al nacer; y en las mujeres, la esperanza de vida a los 65 años y la mortalidad infantil.

4.7 Discusión

El Índice de Desarrollo Regional (IDERE), el cual recoge información de ocho dimensiones (Educación, Salud, Bienestar y Cohesión, Actividad Económica, Instituciones, Seguridad, Medio Ambiente, Género) para países de América Latina, ubica a Colombia como el país con más desigualdades territoriales de América Latina, seguido de Paraguay, Brasil y México, frente a Chile y Uruguay que tiene un desarrollo con mayor equilibrio Rodríguez y Vial (2020). Según este índice, en Colombia como en el resto de América Latina, se perciben dinámicas a nivel regional que identifican áreas geográficas de rezago y otras de situación privilegiada.

Según el Informe IDERE LATAM 2020 (Rodríguez y Vial 2020), los departamentos de Colombia con mayor Índice de Desarrollo Regional son Bogotá, Santander y Cundinamarca, que tienen niveles de desarrollo similares a regiones como Buenos Aires (Argentina), Artigas (Uruguay) o Araucanía (Chile); mientras que Vichada, Arauca y Guaviare, presentan valores más bajo para este índice, presentando similitudes con el Alto Paraguay (Paraguay) y Cabañas y La Unión (El Salvador). Las desigualdades en términos generales entre los departamentos de Colombia sustentan de alguna manera los altos valo-

res del índice de Gini de mortalidad al nacer y sus diferencias en los grupos identificados.

Los grupos obtenidos con el análisis de componentes principales responden a dos aspectos, uno de tipo geográfico y otro de tipo económico-social. Los departamentos de Vichada, Guanía y Vaupés (que se separran claramente del resto) son periféricos, los dos últimos están declarados por el DANE como "territorios especiales" biodiversos y fronterizos. El segundo factor que estaría definido por cuestiones económicas, sociales, condiciones de vida, y de gestión pública. Este segundo elemento se puede establecer a partir del *Escalafón de Competitividad de los Departamentos de Colombia* que mide factores como las condiciones económicas, la infraestructura, el bienestar social, la institucionalidad y gestión pública de los departamentos. Según este escalafón de Competitividad estos tres departamentos están en la categoría de "rezagados" (Ramírez y Aguas 2021). Asimismo, según el PIB por departamento, Guanía y Vaupés se encuentran entre los departamentos que en las últimas décadas tienen menor PIB (DANE 2021).

En cuanto a los grupos identificados por el análisis de clúster jerárquico y el fuzzy clúster, los grupos se caracterizaron según los valores de los indicadores de mortalidad en general. En 1985, el Grupo 1, mayoritario, tenía peores valores de mortalidad, y el grupo 2, minoritario, presentó mejores valores de mortalidad.

En ambos sexos, el Grupo 1 estuvo compuesto por departamentos como Valle del Cauca, Quindío, Meta, Antioquia y Casanare, considerados como lo de peor seguridad, con hechos relacionados con el conflicto armado interno pero también con otras cuestiones sociales (Rodríguez y Vial 2020). Lo anterior que podría explicar los bajos valores de esperanza de vida al nacer y a los 65 años, hecho que es más notable en los hombres.

Se debe resaltar que en las mujeres, el Grupo 2 de mejor mortalidad, recoge a aquellos departamentos denominados especiales y a otros como Bogotá y Cundinamarca que según el Índice de Desarrollo Regional, presentan un mayor desarrollo en binestar y actividad económica.

Aunque se evidencia una mejora de los indicadores de mortalidad a nivel general en los grupos identificados para los departamentos, es notable la heterogeneidad entre los departamentos en cuanto a mortalidad y su relación con el desarrollo económico, aspectos sociales y de seguridad.

4.8 Conclusiones

En este capítulo, se ha analizado la mortalidad de los departamentos de Colombia, encontrado grupos que tienen un comportamiento similar en cuanto a mortalidad. Se caracterizaron los grupos encontrados mediante algunos indicadores de mortalidad y se pudo establecer que mortalidad en la población colombiana muestra diferencias entre los departamentos tanto para hombres como para mujeres. Según lo analizado se plantean las siguientes conclusiones:

La técnica de ACP es una herramienta útil en estudios de mortalidad donde el número de variables es elevado, permitiendo reducir la dimensión del problema y una mejor comprensión del fenómeno de estudio.

El ACP nos permitió reducir la información de los hombres en 11 componentes y en las mujeres en 12 componentes, explicando aproximadamente el 92% de la varianza total en ambos sexos. La información comprimida por el ACP facilitó un manejo mejor de la información para posteriores análisis y agrupaciones de los departamentos.

Las técnicas de clúster jeránquico y fuzzy clúster resultaron de gran ayuda para identificar agrupaciones entre los departamentos de Colombia. Los agrupaciones creadas para los departamentos por sexo coinciden en estás técnicas. Con estás técnicas se identificaron dos grupos para los hombres (22/11) y 2 grupos para los mujeres (23/10).

Los grupos identificados permitieron caracterizar los departamentos según indicadores de mortalidad. Comparando los valores de los indicadores de mortalidad entre 1985 y 2014, se observó una mejora estadísticamente significativa entre estos años en ambos sexos. De manera general, la esperanza de vida al nacer, la esperanza de vida a los 65 años y el índice de Gini al nacer fueron los indicadores que más mejoraron durante el periodo 1985-2014. La mortalidad infantil y el índice de Gini a los 65 años también han mejorado pero en menor magnitud.

En 1985, en ambos sexos, el Grupo 1 estuvo compuesto por alrededor del 67% de los departamentos de Colombia, agrupando aquellos con mayor mortalidad infantil, menor esperanza de vida al nacer y a los 65 años, y mayor índice de Gini al nacer y a los 65 años. Por otra parte, el Grupo 2 se compuso por aquellos departamentos con menor mortalidad infantil, mayor esperanza de vida al nacer y a los 65 años, y menores valores del índice de Gini al nacer y a los 65 años. En 2014, el Grupo 1 presenta valores cercanos a los del Grupo 2 en cuanto a los indicadores, acercamiento que se da en ambos sexos.

Los árboles de clasificación permitieron caracterizar los grupos identificados y obtener una medida de la importancia de los indicadores de mortalidad. En los hombres solo se utilizó para la clasificación la esperanza de vida al nacer y en las mujeres, se utilizó para las clasificación la esperanza de vida a los 65 años y la mortalidad infantil.

Con esta técnica se logró describir el comportamiento de los grupos identificados según los indicadores de mortalidad más importantes en cada sexo para el año 2014. En los hombres, los grupos se separaron según si esperanza de vida al nacer era menor a 75 años (grupo 1) o mayor a 75 años (grupo 2). En las mujeres los grupos se separaron primero según la esperanza de vida a los 65 años, mayor a 17 años o menor a 17 años; y en un segundo momento según la mortalidad infantil, menor a 0,0104 o mayor a 0,0104. Este resultado ratifica que las diferencias en el comportamiento de la mortalidad entre sexos y departamentos en Colombia es una cuestión que sigue presentándose y que debe ser atendida por las instituciones y autoridades administrativas.

Por último decir que la metodología propuesta en este capítulo permite explorar el comportamiento de la mortalidad en regiones (ya sean de un país o continente), posibilitando la identificación de grupos con similitudes y diferencias utilizando diferentes indicadores de mortalidad, mediante técnicas como el análisis de componentes principales, clúster jerárquico, fuzzy clúster, CART y random forest.

Bibliografía

Aburto, J.M. y V.M. García-Guerrero (2015). "El modelo aditivo doble multiplicativo. Una aplicación a la mortalidad mexicana." En: *Papeles de Población* 21.84, págs. 9-44.

Acosta, K. y J. Romero (2014a). *Cambios recientes en las principales causas de mortalidad en Colombia*. 209. Banco de La República, Colombia.

— (2014b). *Estimación indirecta de la tasa de mortalidad infantil en Colombia, 1964-2008*. 199. Banco de La República, Colombia.

Aguilar, E. (2013). "Estimación y proyección de la mortalidad para Costa Rica con la aplicación del método Lee-Carter con dos variantes." En: *Población y Salud en Mesoamérica* 11.1, págs. 3-24.

Alemi, F. y D. Neuhauser (2004). "Time-between control charts for monitoring asthma attacks". En: *The Joint Commission Journal on Quality and Safety* 30.2, págs. 95-102.

Alexopoulos, A., P. Dellaportas y J.J. Forster (2019). "Bayesian forecasting of mortality rates by using latent Gaussian models". En: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182.2, págs. 689-711.

Alvis-Zakzuk, N. y col. (2015). "Desigualdades de la mortalidad infantil y pobreza en Colombia: Análisis inter-censal (1993 y 2005)". En: *Revista Ciencias Biomédicas* 6.1, págs. 29-37.

Andreozzi, L. (2012). "Estimación y pronósticos de la mortalidad de Argentina utilizando el modelo de Lee-Carter." En: *Revista de la Sociedad Argentina de Estadística* 10.1, págs. 21-43.

Andreozzi, L. y M.T. Blaconá (2011). "Estimación y pronóstico de las tasas de mortalidad y la esperanza de vida en la República Argentina." En: Anales de las Decimosextas Jornadas Investigaciones en la Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario, Argentina.

Ayuso, M. (2007). *Estadística actuarial vida*. Vol. 51. Edicions Universitat Barcelona.

Banco-Mundial (2018). *Esperanza de vida al nacer - América Latina y el Caribe*. Ed. por Grupo Banco Mundial. URL: `https://datos.bancomundial.org/indicador/SP.DYN.LE00.IN?end=2018&locations=ZJ&most_recent_year_desc=true&start=1990`.

Barbieri, M. y R. Depledge (2013). "Mortality in France by département". En: *Population* 68.3, págs. 375-417.

Belliard, M. e I. Williams (2013). "Proyección estocástica de la mortalidad. Una aplicación de Lee-Carter en la Argentina." En: *Revista Latinoamericana de Población* 7.13, págs. 129-148.

Benneyan, J.C., R.C. Lloyd y P.E. Plsek (2003). "Statistical process control as a tool for research and healthcare improvement". En: *Quality & Safety in Health Care* 12.6, págs. 458-464.

Bertranou, E. (2008). *Tendencias demográficas y protección social en América Latina y el Caribe*. 82. CEPAL, Serie Población y desarrollo.

Bezdek, J.C., R. Ehrlich y W. Full (1984). "FCM: The fuzzy c-means clustering algorithm". En: *Computers & geosciences* 10.2-3, págs. 191-203.

Blaconá, M.T. y L. Andreozzi (2014). "Análisis de la mortalidad por edad y sexo mediante modelos para datos funcionales." En: *Estadística* 66.186-187, págs. 65-89.

Blum, A., A. Kalai y J. Langford (1999). "Beating the hold-out: Bounds for k-fold and progressive cross-validation". En: *International Conference on Computational Learning Theory*. ACM, págs. 203-208.

Booth, H., J. Maindonald y L. Smith (2002). "Applying Lee-Carter under conditions of variable mortality decline". En: *Population studies* 56.3, págs. 325-336.

Booth, H. y L. Tickle (2008). "Mortality modelling and forecasting: A review of methods". En: *Annals of actuarial science* 3.1-2, págs. 3-43.

Booth, H. y col. (2006). "Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions". En: *Demographic research* 15, págs. 289-310.

Breiman, L. (2001). "Random forests". En: *Machine learning* 45.1, págs. 5-32.

Briceño-León, R., A. Villaveces y A. Concha-Eastman (2008). "Understanding the uneven distribution of the incidence of homicide in Latin America". En: *International journal of epidemiology* 37.4, págs. 751-757.

Cairns, A.JG., D. Blake y K. Dowd (2006). "A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration." En: *Journal of Risk and Insurance* 73.4, págs. 687-718.

Cairns, A.JG. y col. (2009). "A quantitative comparison of stochastic mortality models using data from England and Wales and the United States." En: *North American Actuarial Journal* 13.1, págs. 1-35.

Cairns, A.JG. y col. (2011). "Mortality density forecasts: An analysis of six stochastic mortality models." En: *Insurance: Mathematics and Economics* 48.3, págs. 355-367.

Callot, L., N. Haldrup y M. Kallestrup-Lamb (2016). "Deterministic and stochastic trends in the Lee–Carter mortality model". En: *Applied Economics Letters* 23.7, págs. 486-493.

Canudas-Romo, V. (2008). "The modal age at death and the shifting mortality hypothesis." En: *Demographic Research* 19.30, págs. 1179-1204.

Carfora, M.F., L. Cutillo y A. Orlando (2017). "A quantitative comparison of stochastic mortality models on Italian population data". En: *Computational Statistics & Data Analysis* 112, págs. 198-214.

Carmona-Fonseca, J. (2005). "Cambios demográficos y epidemiológicos en Colombia durante el siglo XX". En: *Biomédica* 25, págs. 464-480.

Carracedo, P. y col. (2018). "Detecting spatio-temporal mortality clusters of European countries by sex and age". En: *International journal for equity in health* 17.1, págs. 1-19.

CELADE (1996). *Impacto de las Tendencias Demográficas sobre los Sectores Sociales en América Latina: contribución al diseño de políticas y programas*. Inf. téc. 45. CEPAL/CELADE/BID,Santiago de Chile.

CEPAL (2014). *La nueva era demográfica en América Latina y el Caribe. La hora de la igualdad según el reloj poblacional*. URL: `https://repositorio.cepal.org/handle/11362/37252`.

Chamberlin, W.H. y col. (1993). "Monitoring intensive care unit performance using statistical quality control charts". En: *International journal of clinical monitoring and computing* 10.3, págs. 155-161.

Champ, C.W. y L.A. Jones-Farmer (2007). "Properties of multivariate control charts with estimated parameters". En: *Sequential Analysis* 26.2, págs. 153-169.

Christiansen, M.C., E. Spodarev, V. Unseld y col. (2015). "Differences in European mortality rates: A geometric approach on the age-period plane". En: *Astin Bulletin* 45.3, págs. 477-502.

Coelho, E. y L.C. Nunes (2011). "Forecasting mortality in the event of a structural change". En: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174.3, págs. 713-736.

Collett, D. (2003). *Modelling binary data*. 2ª ed. Chapman & Hall/CRC, FL, USA.

Comisión Económica para América Latina y el Caribe (CEPAL) (2017). *Observatorio Demográfico (LC/PUB.2017/20-P)*.

Cristancho, C.A. (2017). "Niveles, tendencias y determinantes de la mortalidad reciente en Colombia". Tesis doct. Universitat Autónoma de Barcelona.

Currie, I.D., M. Durban y P.H.C. Eilers (2006). "Generalized linear array models with applications to multidimensional smoothing." En: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.2, págs. 259-280.

DANE (2007a). *Conciliación censal 1985-2005. Colombia. Estimación de la Mortalidad 1985-2005*. Inf. téc. DANE.

— (2007b). *División Político Administrativa*. Ed. por Colombia DANE. URL: https://www.dane.gov.co/index.php/servicios-al-ciudadano/72-espanol/clasificaciones/geografica/488-division-polistico-administrativa.

— (2018). *Censo nacional de Población y vivienda 2018, Colombia*. Ed. por Colombia DANE. URL: https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018/cuantos-somos.

— (2021). *Boletín técnico. Producto Interno Bruto por departamento*. DANE, Colombian.

Díaz, G. y A. Debón (2016). "Tendencias y comportamiento de la mortalidad en Colombia entre 1973 y 2005." En: *Estadística Española* 58.191, págs. 277-300.

Debón, A., F. Martínez-Ruiz y F. Montes (2012). "Temporal evolution of mortality indicators: application to Spanish data". En: *North American Actuarial Journal* 16.3, págs. 364-377.

Debón, A., F. Montes y F. Puig (2008). "Modelling and Forecasting mortality in Spain". En: *European Journal of Operation Research* 189.3, págs. 624-637.

Debón, A., F. Montes y R. Sala (2009). *Tablas de mortalidad dinámicas. Una aplicación a la hipoteca inversa en España*. Fundación ICO. Publicaciones de la Universitat de València,València.

Debón, A., F. Martínez-Ruiz y F. Montes (2010). "A geostatistical approach for dynamic life tables: The effect of mortality on remaining lifetime and annuities". En: *Insurance: Mathematics and Economics* 47.3, págs. 327-336.

Debón, A. y col. (2017). "Characterization of between-group inequality of longevity in European Union countries". En: *Insurance: Mathematics and Economics* 75, págs. 151-165.

Delwarde, A. y M. Denuit (2003). "Importance de la période d'observation et des âges considérés dans la projection de la mortalité selon la méthode de Lee-Carter". En: *Belgian Actuarial Bulletin* 3.1, págs. 1-21.

Díaz, G., A. Debón y V. Giner-Bosch (2018). "Mortality forecasting in Colombia from abridged life tables by sex". En: *Genus, Journal of Population Sciences* 74.15.

Díaz-Rojo, G., A. Debón y J. Mosquera (2020). "Multivariate Control Chart and Lee-Carter Models to Study Mortality Changes". En: *Mathematics* 8.2093.

ENDS, 2015 (2015). *Encuesta Nacional de Demografía y Salud - ENDS 2015*. Ministerio de Salud y Protección Social y Profamilia, Colombia.

Fahad, A. y col. (2014). "A survey of clustering algorithms for big data: Taxonomy and empirical analysis". En: *IEEE transactions on emerging topics in computing* 2.3, págs. 267-279.

Fernández-Delgado, M. y col. (2014). "Do we need hundreds of classifiers to solve real world classification problems?" En: *The journal of machine learning research* 15.1, págs. 3133-3181.

Fritzell, J. y col. (2013). "Cross-temporal and cross-national poverty and mortality rates among developed countries". En: *Journal of environmental and public health* 2013.

Fukuyama, Y. y M. Sugeno (1989). "A new method of choosing the number of clusters for the fuzzy c-mean method". En: *Proc. 5th Fuzzy Syst. Symp., (in Japanese)*, págs. 247-250.

García-Guerrero, V.M. y M.O. Mellado (2012). "Proyección estocástica de la mortalidad mexicana por medio del método de Lee-Carter." En: *Estudios Demográficos y Urbanos* 27.2, págs. 409-448.

Garfield, R. y C.P. Llanten (2004). "The public health context of violence in Colombia". En: *Revista Panamericana de Salud Pública* 16.4, págs. 266-271.

Gaviria, A. (2000). "Increasing returns and the evolution of violent crime: the case of Colombia". En: *Journal of development economics* 61.1, págs. 1-25.

Giordani, P. y H.A. Kiers (2007). "Principal component analysis with boundary constraints". En: *Journal of Chemometrics: A Journal of the Chemometrics Society* 21.12, págs. 547-556.

Guibert, Q. y col. (2020). "Bridging the Li-Carter's gap: a locally coherent mortality forecast approach". En.

Hennig, C. y col. (2016). *Handbook of cluster analysis*. CRC Press.

Herrera, A. Y. (2019). *Análisis de Situación de Salud (ASIS). Colombia, 2019*. Inf. téc. Ministerio de Salud y Protección Social, Bogotá, Colombia.

Holford, T.R. (2006). "Approaches to fitting age-period-cohort models with unequal intervals". En: *Statistics in Medicine* 25.6, págs. 977-993.

Horiuchi, S. (1999). "Epidemiological transitions in human history". En: *Health and mortality: Issues of global concern*, págs. 54-71.

Hotelling, H. (1947). "Techniques of statistical analysis". En: ed. por C. Eisenhart, M.W. Hastay y W.A. Wallis. New York: McGraw-Hill. Cap. Multivariate quality control illustrated by the testing of sample bombsights, págs. 113-184.

Hunt, A. y A.M. Villegas (2015). "Robustness and convergence in the Lee-Carter model with cohort effects." En: *Insurance: Mathematics and Economics* 64, págs. 186-202.

Husson, F. y col. (2020). *FactoMineR*. R package version 2.4.

Hyndman, R.J. (2016). *Forecast:Forecasting functions for time series and linear models*. R package version 7.3.

Hyndman, R.J. y M.S. Ullah (2007). "Robust forecasting of mortality and fertility rates: a functional data approach." En: *Computational Statistics & Data Analysis* 51.10, págs. 4942-4956.

Imam, N. y col. (2019). "Statistical Process Control Charts for Monitoring Staphylococcus aureus Bloodstream Infections in Australian Health Care Facilities". En: *Quality Management in Healthcare* 28.1, págs. 39-44.

INE (2009). *Tablas de mortalidad. Metodología*. Inf. téc. Instituto Nacional de Estadística, España.

— (2020). *Índicadores Demográficos Básicos. Metodología*. Inf. téc. Instituto Nacional de Estadística, España.

Janitza, S., E. Celik y A.L. Boulesteix (2018). "A computationally fast variable importance test for random forests for high-dimensional data". En: *Advances in Data Analysis and Classification* 12.4, págs. 885-915.

Johnson, R. y D. Wichern (2014). *Applied Mltivariate Statistical Analysis*. Sixth. Prentice hall Upper Saddle River, NJ.

Johnson, R.A., D.W. Wichern y col. (2002). *Applied multivariate statistical analysis*. Vol. 5. 8. Prentice hall Upper Saddle River, NJ.

Jones, J.H. y B. Ferguson (2006). "The marriage squeeze in Colombia, 1973–2005: The role of excess male death". En: *Social Biology* 53.3-4, págs. 140-151.

Kendall, M.G. (1975). *Multivariate Analysis*. Londres: Griffin.

Kennes, T. (2017). "The Convergence and Robustness of Cohort Extensions of Mortality Models". En: *MaRBLe* 1, págs. 36-53.

Kleinow, T. (2015). "A common age effect model for the mortality of multiple populations". En: *Insurance: Mathematics and Economics* 63, págs. 147-152.

Kuhn, M. (2021). *caret: Classification and Regression Training*. R package version 6.0-88.

LAHMD, ed. (2015). *Latin American Human Mortality Database*. URL: `http://www.lamortalidad.org`.

Lee, R. y R. Rofman (1994). "Modelación y Proyección de la Mortalidad en Chile". En: *Notas de Poblacion* 6.59, págs. 183-213.

Lee, R.D. y L. Carter (1992). "Modelling and Forecasting U.S. Mortality". En: *Journal of the American Statistical Association* 87, págs. 659-671.

Lee, W.C. (1997). "Characterizing exposure-disease association in human populations using the Lorenz curve and Gini index". En: *Statistics in Medicine* 16.7, págs. 729-739.

Levitt, S. y M. Rubio (2000). *Understanding crime in Colombia and what can be done about it*. Inf. téc. 20. FEDESARROLLO.

Liaw, A. y M. Wiener (2002). "Classification and Regression by randomForest". En: *R News* 2.3, págs. 18-22.

Linares, G. (1990). *Análisis de datos*. Ministerio de Educación Superior, La Habana: Universidad de La Habana, Facultad de Matemática y Cibernética.

Llorca, J., M.D. Prieto y M. Delgado-Rodriguez (2000). "Medición de las desigualdades en la edad de muerte: cálculo del índice de Gini a partir de

las tablas de mortalidad". En: *Revista Española de Salud Pública* 74.1, págs. 5-12.

Llorca, J. y col. (1998). "Age differential mortality in Spain, 1900-1991." En: *Journal of Epidemiology & Community Health* 52, págs. 259-261.

Lora, E. (2008). *Técnicas de medición económica. Metodología y aplicaciones en Colombia.* Fourth. Bogotá D.C.: Alfaomega Colombiana S.A.

Marshall, T. y M.A. Mohammed (2007). "Case-mix and the use of control charts in monitoring mortality rates after coronary artery bypass". En: *BMC health services research* 7.1, pág. 63.

Mason, R.L., N.D. Tracy y J.C. Young (1995). "Decomposition of T^2 for multivariate control chart interpretation". En: *Journal of Quality Technology* 27, págs. 99-108.

Mason, R.L. y J.C. Young (2002). *Multivariate Statistical Process Control with Industrial Applications.* ASA-SIAM.

Meyer, D. y col. (2021). *e1071.* R package version 1.7-6.

Millossovich, P., A.M. Villegas y V.K. Kaishev (2018). "StMoMo: An R Package for Stochastic Mortality Modelling". En: *Journal of Statistical Software* 84.3.

MinSalud (2013). *Envejecimiento Demográfico. Colombia 1951-2020. Dinámica demográfica y estrucuras poblacionales.* Inf. téc. Ministerio de Salud y Protección Social, Colombia.

O'hare, C. e Y. Li (2017). "Models of mortality rates-analysing the residuals". En: *Applied Economics*, págs. 1-15.

Olivieri, A. (2001). "Uncertainly in mortality projections: an actuarial perspective". En: *Insurance: Mathematics and Economics* 29.4, 231?245.

Ornelas, A. (2015). "La mortalidad y la longevidad en la cuantificación del riesgo actuarial para la población de México." Tesis doct. Universitat de Barcelona.

Padilla, J.C., D.P. Rojas y R. Sáenz (2012). *Dengue en Colombia: epidemiología de la reemergencia a la hiperendemia*. Guías de Impresión Ltda.

Perz, S.G. (2004). *Population change. In Siegel JS and Swanson DA. (eds.) The Methods and Materials of Demography*. 2ª ed. Elsevier Academic Press, California, USA.

Postigo-Boix, M., R. Agüero y J.L. Melús-Moreno (2019). "An alternative procedure to obtain the mortality rate with non-linear functions: Application to the case of the Spanish population". En: *PloS ONE* 14.10, e0223789.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ramírez, J.C. y J.M. de Aguas (2017). *Escalafón de la competitividad de los departamentos de Colombia, 2017*. Cepal, Naciones Unidas.

— (2021). *Escalafón de la competitividad de los departamentos de Colombia, 2019*. 36. Cepal, Naciones Unidas. Cap. Estudios y Perspectivas.

Remund, A., C.G. Camarda, T. Riffe y col. (2017). *A cause-of-death decomposition of the young adult mortality hump*. Inf. téc. Max Planck Institute for Demographic Research, Rostock, Germany.

Renshaw, A.E. y S. Haberman (2003). "Lee-Carter mortality forecasting with age-specific enhancement". En: *Insurance: Mathematics and Economics* 33.2, págs. 255-272.

— (2006). "A cohort-based extension to the Lee–Carter model for mortality reduction factors". En: *Insurance: Mathematics and economics* 38.3, págs. 556-570.

— (2008). "On simulation-based approaches to risk measurement in mortality with specific reference to Poisson Lee–Carter modelling". En: *Insurance: Mathematics and Economics* 42.2, págs. 797-816.

Reyes, A.R. (2010). "Una aproximación al costo fiscal en pensiones como consecuencia del envejecimiento de la población en Colombia y el efecto de la sobre-mortalidad masculina". En.

Rezaee, M.R., B.P. Lelieveldt y J.H. Reiber (1998). "A new cluster validity index for the fuzzy c-mean". En: *Pattern recognition letters* 19.3, págs. 237-246.

Richards, S.J. (2008). "Detecting year-of-birth mortality patterns with limited data". En: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171.1, págs. 279-298.

Riva, S. Muriel de la, M. Cantalapiedra y F. Lopéz (2010). *Towards Advanced Methods for Computing Life Tables*. Inf. téc. Technical Report.

Rodríguez, A. y C. Vial (2020). *IDERE LATAM. Índice de desarrollo regional-Latinoamérica*. Instituto Chileno de Estudios Municipales (ICHEM) de la Universidad Autónoma de Chile y el Instituto de Economía (IECON) de la Facultad de Ciencias Económicas y de Administración de la Universidad de la República del Uruguay.

Rodríguez, J. (2007). "Desigualdades socioeconómicas entre departamentos y su asociación con indicadores de mortalidad en Colombia en 2000." En: *Rev Panam Salud Publica* 21.2/3, págs. 111-124.

Ryan, T (2011). *Statistical methods for quality improvement*. 3ª ed. John Wiley & Sons, New Jersey, USA.

Saad, P.M., T. Miller y C. Martínez (2009). "Impacto de los cambios demográficos en las demandas sectoriales en América Latina". En: *Revista Brasileira de Estudos de População* 26.2, págs. 239-261.

Salhi, Y. y S. Loisel (2017). "Basis risk modelling: a cointegration-based approach". En: *Statistics* 51.1, págs. 205-221.

Schnürch, S., T. Kleinow y R. Korn (2021). "Clustering-Based Extensions of the Common Age Effect Multi-Population Mortality Model". En: *Risks* 9.3, pág. 45.

Scrucca, L. (2017). *qcc: Quality Control Charting*.

Shewhart, W.A. (1927). "Quality control". En: *The Bell System Technical Journal* 6.4, págs. 722-735.

Shkolnikov, V.M., E.M. Andreev y A. Begun (2003). "Gini coefficient as a life table function: computation from discrete data, decomposition of differences and empirical examples". En: *Demographic Research* 8, págs. 305-358.

Siegel, J.S. y D.A. Swanson (2004). *The methods and materials of demography.* 2ª ed. Elsevier Academic Press, California, USA.

Singh, A. y col. (2017). "Trends in inequality in length of life in India: a decomposition analysis by age and causes of death". En: *Genus* 73.1, pág. 5.

Speiser, J.L. y col. (2019). "A comparison of random forest variable selection methods for classification prediction modeling". En: *Expert systems with applications* 134, págs. 93-101.

Tabeau, E. (2001). "A review of demographic forecasting models for mortality". En: *Forecasting Mortality in Developed Countries. European Studies of Population.* Vol. 9. Springer, Dordrecht, págs. 1-32.

Thacker, S.B. y col. (1995). "Public health surveillance for chronic conditions: a scientific basis for decisions". En: *Statistics in Medicine* 14.5-7, págs. 629-641.

Therneau, T. y B. Atkinson (2019). *rpart: Recursive Partitioning and Regression Trees.* R package version 4.1-15.

Tracy, N.D., J.C. Young y R.L. Mason (1992). "Multivariate Control Chart for individual observations". En: *Journal of Quality Technology* 24, págs. 88-95.

Turner, H. y D. Firth (2015). *Generalized nonlinear models in R: An overview of the gnm package.* R package version 1.0-8.

UNICEF (2016). *www.unicef.org, 1 de Diciembre de 2016.*

Urdinola, B.P. y N. Rojas-Perilla (2013). "Quality Control Charts as a Tool to Correct Adult Mortality Under-Registration." En.

Urdinola, B.P., F. Torres y J.A. Velasco (2015). *Latin American Human Mortality Database.*

Urdinola, B.P., F. Torres y J.A. Velasco (2017). "The Homicide Atlas in Colombia: Contagion and Under-Registration for Small Areas." En: *Cuadernos de Geografía - Revista Colombiana de Geografía* 26.1, págs. 101 -118. ISSN: 0121-215X.

Vetter, T.R. y D. Morrice (2019). "Statistical Process Control: No Hits, No Runs, No Errors?" En: *Anesthesia & Analgesia* 128.2, págs. 374-382.

Villegas, A.M., P. Millossovich y V.K. Kaishev (2018). "StMoMo: Stochastic Mortality Modeling in R". En: *Journal of Statistical Software* 84.3.

Wang, D. y P. Lu (2005). "Modelling and forecasting mortality distributions in England and Wales using the Lee–Carter model". En: *Journal of Applied Statistics* 32.9, págs. 873-885.

Ward, J.H. (1963). "Hierarchical grouping to optimize an objective function". En: *Journal of the American statistical association* 58.301, págs. 236-244.

Williamson, G.D. y G.W. Hudson (1999). "A monitoring system for detecting aberrations in public health surveillance reports". En: *Statistics in Medicine* 18.23, págs. 3283-3298.

Woodall, W.H. (2006). "The Use of Control Charts in Health-Care and Public-Health Surveillance". En: *Journal of Quality Technology* 38.2, págs. 89-104.

Xie, X. y G. Beni (1991). "A validity measure for fuzzy clustering". En: *IEEE Transactions on pattern analysis and machine intelligence* 13.8, págs. 841-847.

Yue, J. y col. (2017). "A new VLAD-based control chart for detecting surgical outcomes". En: *Statistics in Medicine* 36.28, págs. 4540-4547.

Yusuf, F. y col. (2014). *Methods of demographic analysis*. Springer.

Zarruk, A., A.M. Villegas y F. Ortiz (2011). *Tablas de Mortalidad. Evolución en el Sector Asegurador Colombiano*. Inf. téc. Fasecolda.