Tesis de Maestría. FACULTAD DE CIENCIAS ECONOMICAS Y ESTADISTICA ; UNIVERSIDAD NACIONAL DE ROSARIO.

Proyecciones Probabilísticas en Demografía.

Lucia Andreozzi.

Cita:

Lucia Andreozzi (2016). *Proyecciones Probabilísticas en Demografía* (Tesis de Maestría). FACULTAD DE CIENCIAS ECONOMICAS Y ESTADISTICA ; UNIVERSIDAD NACIONAL DE ROSARIO.

Dirección estable: https://www.aacademica.org/lucia.andreozzi/33

ARK: https://n2t.net/ark:/13683/preH/2EG

Esta obra está bajo una licencia de Creative Commons.
Para ver una copia de esta licencia, visite
https://creativecommons.org/licenses/by-nc-nd/4.0/deed.es.

Acta Académica es un proyecto académico sin fines de lucro enmarcado en la iniciativa de acceso abierto*. Acta Académica fue creado para facilitar a investigadores de todo el mundo el compartir su producción académica. Para crear un perfil gratuitamente o acceder a otros trabajos visite: https://www.aacademica.org*.

UNIVERSIDAD NACIONAL DE ROSARIO

FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

CARRERA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

Proyecciones Probabilísticas en Demografía

Autor: Lic. Lucía Andreozzi

Director: Mgs. María Teresa Blaconá

Asesor: Dr. Rob J. Hyndman

Año 2016

Tribunal Examinador

Mgs. Nora E. Arnesi

Mgs. Javier Bussi

Dr. Martín Saino

Dedicado a

Susana, mi mamá,

y su inagotable fortaleza.

Agradecimientos

Gracias a...

a mi Directora, Mgs. M.T. Blaconá, por su guía y paciencia,

al Dr. Rob Hyndman por su amabilidad constante y la generosidad con la que compartió sus conocimientos,

a la Dra. Guiomar Bay, y los Dres. Leandro González, Eduardo Torres y Bruno Ribotta por responder a mis consultas y correos,

a la Dra. Dora Celton por haberme brindado la oportunidad de conocer el campo de la demografía,

al Dr. Eduardo Santillán Marcus por su orientación matemática,

al Dr. Carlos Guevel por su atención a la hora brindarme información,

a mis colegas Leticia, Gabriela, Nora A., Nora V., Virginia y Cristina por sus correcciones y sugerencias,

a Ivana, Anita y Ceci, mis amigas del alma, por acompañarme en todas,

a mi familia y a mis amigas por su aliento y compañía

y especialmente a Sebastián por aparecer en mi vida en el *sprint final* de esta tesis.

Resumen

En este trabajo se aplican modelos para datos funcionales (MDF) a las cifras de mortalidad y fecundidad de Argentina para obtener la descomposición de ambas componentes demográficas, sus pronósticos probabilísticos y el del total de la población a partir de las componentes pronosticadas. La aplicación de estos modelos a datos de Argentina demuestra que cuando se cuenta con información para períodos no muy extensos de tiempo y una calidad aceptable de las fuentes de datos, se pueden emplear con éxito para realizar pronósticos probabilísticos. Se comparan resultados obtenidos por MDF y modelos ARIMA. En segundo lugar se realiza el análisis de la componente migratoria, el comportamiento es una combinación del verdadero saldo migratorio, los errores de registro en las muertes por edad y en los nacimientos según edad de la madre y los errores en las estimaciones y/o proyecciones de población. Se aplica a datos de mortalidad la metodología alternativa de pronósticos coherentes con el fin de evitar una divergencia entre ambos géneros. Finalmente, a partir de los pronósticos de fecundidad (basados en dos hipótesis de fecundidad) y de mortalidad (independientes y coherentes) se elaboran pronósticos estocásticos de la población para hombres y mujeres. En todos los casos evaluados las cifras estimadas se encuentran por encima de las oficiales. Los resultados aquí obtenidos presentan medidas adecuadas de bondad de ajuste, cifras pronosticadas coherentes con la teoría demográfica y principalmente garantizan la consistencia probabilística.

Palabras Clave: Mortalidad, fecundidad, migración, datos funcionales y pronósticos coherentes.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Cuando en el año 1996 las Naciones Unidas proyectaron la población mundial para el año 2050 en base a la variante media, la cifra obtenida resultó menor, en 466 millones de personas, a la proyección realizada para el mismo año en 1994. Este cambio en las cifras fue tomado como una evidencia de que el crecimiento en la población no era un "problema" tan grave como se pensaba. Lo más importante es que ambas proyecciones tenían intervalos de incertidumbre que se superponían, y si bien los productores de las proyecciones comprendían este punto, la mayoría de los consumidores de los datos, no. A pesar de hechos como éste, las tendencias del tamaño de la población, la estructura de edad, nacimientos, muertes y otras variables demográficas resultan de interés para una amplia gama de analistas, políticos, científicos y planificadores de la industria y del gobierno. Contar con información de calidad en estos aspectos constituye un insumo básico para la toma de decisiones en las más diversas áreas. Por ejemplo, las tendencias demográficas a nivel nacional y mundial constituyen información necesaria para proyectar las demandas futuras de agua, alimentos y energía, además de que permiten estimar el impacto ambiental debido al consumo de recursos naturales. Así mismo, las proyecciones subnacionales contribuyen, por mencionar alguna de sus múltiples utilidades, a la planificación de la inversión, dado que en base a ellas se establecen los sitios donde deben construirse escuelas o caminos. Puntualmente, las proyecciones se utilizan, en conjunto con otros indicadores, para determinar la asignación de fondos destinados

a infraestructura en general.

Es inevitable que las proyecciones contengan incertidumbre y si bien la situación demográfica actual es conocida, las tendencias en los nacimientos, muertes y migraciones están sujetas a cambios impredecibles. Los errores cometidos en la elaboración de proyecciones varían de acuerdo al horizonte de pronóstico, el nivel de desarrollo del país y su tamaño. Mientras que las tendencias generales en la mortalidad, la fecundidad y la migración pueden discernirse y proyectarse para el futuro con una "confianza" (en el uso no estadístico de la palabra) razonable, una sustancial incertidumbre se haya asociada a la tendencia particular de un país o región. Cuantificar esta incertidumbre es de gran utilidad para los usuarios de las proyecciones y dado que los escenarios alternativos que se elaboran como parte de una proyección demográfica tienen diferentes implicancias en el futuro, conocer una medida de la incertidumbre asociada a cada resultado permite decidir en base a mayor información. Actualmente la posible variación en los resultados se expresa proveyendo escenarios alternativos con variantes para la fecundidad (y con menos frecuencia en la mortalidad y la migración). De este modo los escenarios alto y bajo se utilizan para indicar un rango de posibles futuros, no obstante, ninguna probabilidad específica se asocia a estos resultados.

En 1998 el Consejo Nacional de Investigaciones de la Academia de Ciencias de los Estados Unidos conformó un panel de expertos para analizar el tema de las proyecciones en profundidad (Bongaarts y Bulatao, 2000). El panel concluyó que las proyecciones elaboradas para el año 2050 se basaban en supuestos razonables y proveían pronósticos plausibles para las tendencias demográficas en las próximas décadas. Esta conclusión no abarcaba las proyecciones realizadas para países específicos, dado que el panel solamente examinó la metodología general de las proyecciones a nivel mundial. Además entre las conclusiones se destacaba que las proyecciones elaboradas no otorgan la importancia necesaria a un aspecto relevante: la incertidumbre asociada a los pronósticos, de modo que el panel indicó que sería adecuado que las cifras de proyecciones contemplaran de

forma explícita esta característica y remarcaba la necesidad de desarrollar métodos que la cuantifiquen.

La fecundidad, la mortalidad y la migración constituyen las componentes de crecimiento de la población y en la actualidad organismos tales como las Naciones Unidas y otras instituciones referentes en el área son los encargados de determinar los niveles y probables patrones para cada una de ellas. Luego de establecer ciertos supuestos, los niveles y patrones se combinan con información relativa a la estructura existente por edad y sexo y mediante una serie de cálculos conocidos como el método de las componentes se obtiene la proyección de la población por edad y sexo. Por otro lado, los métodos probabilísticos para el pronóstico de la población están ganando rápidamente reconocimiento. Los mismos poseen su principal ventaja en la consistencia probabilística a través de todas las variables pronosticadas y sus indices derivados (Lee y Tuljapurkar, 1994). En cambio, el método tradicionalmente utilizado presenta una inconsistencia cuando se pretende contemplar la incertidumbre asociada a las proyecciones mediante la construcción de escenarios bajo-medio-alto, dado que resulta imposible construir escenarios que reflejen simultáneamente la incertidumbre en todas las variables de interés. Si el rango alto-bajo se diseña para contener las fluctuaciones a corto plazo en las tasas demográficas, luego será demasiado amplio para indicar la incertidumbre del tamaño de la población a largo plazo, dado que algunas fluctuaciones a corto plazo suelen cancelarse en el largo plazo. Aún cuando estas fluctuaciones se mantienen dentro de los límites de los escenarios bajo y alto, otros parámetros demográficos derivados (tales como la proporción de la población de una determinada edad o ciertas relaciones de dependencia) pueden verse afectados mucho más y pueden exceder los niveles definidos por estos escenarios. En síntesis, los escenarios alto y bajo no representan extremos consistentes a lo largo de todos los parámetros demográficos. La inconsistencia probabilística y la dependencia de los métodos tradicionales a fuertes supuestos elaborados por demógrafos y expertos, entre otros aspectos, propiciaron el desarrollo de métodos probabilísticos de pronóstico, que, cada vez con mayor aceptación, son implementados por las agencias de

estadísticas oficiales para producir sus pronósticos nacionales, como las de Holanda y Estados Unidos, por mencionar algunos ejemplos.

La irrupción más clara y concisa de los métodos probabilísticos de pronóstico en el área demográfica la concretó el método propuesto por Lee y Carter (Lee y Carter, 1992). El mismo posee actualmente numerosas variantes y extensiones. Originalmente los autores proponen una metodología que permite modelar y extrapolar las tendencias observadas en las tasas de mortalidad a largo plazo e implementan dicha metodología para pronosticar la mortalidad de los Estados Unidos hasta el año 2065. Lee y Miller (2001) y Booth *et al.* (2002) han propuesto modificaciones al método de Lee-Carter, como la elección del período de ajuste, el método para el ajuste del parámetro de nivel y la elección de las tasas base para el pronóstico. La propuesta de Lee y Miller (2001) es ampliamente utilizada y la variante propuesta por Booth *et al.* (2002) ha demostrado ser al menos tan precisa como la de Lee y Miller en el corto plazo, ver Booth *et al.* (2005) y Booth *et al.* (2006a). Varios desarrollos han incorporado una estructura de error heterocedástica Poisson como Brouhns *et al.* (2002) y Wilmoth (1993)). Otros autores han extendido la aplicabilidad del modelo de Lee-Carter a los factores de reducción de la mortalidad, ver Renshaw y Haberman (2003a) o examinado el uso de más de un término en el modelo, como en el caso de Booth *et al.* (2002) y Renshaw y Haberman (2003b). También existen dos extensiones recientes que incluyen suavizados semi-paramétricos en el modelo; Jong y Tickle (2006) combinan un suavizado por *splines* y una estimación por medio del filtro de Kalman para ajustar una versión generalizada del modelo de Lee y Carter. Hyndman y Ullah (2007) proponen utilizar el paradigma de los datos funcionales para modelar las componentes demográficas; el mismo considera que los logaritmos de las tasas de cada año son una función continua de las edades. Hyndman y Ullah (2007) suavizan la mortalidad a través de regresiones *spline* penalizadas para luego ajustar un modelo mediante una descomposición en componentes principales. Estos métodos son comparados por Booth *et al.* (2006b)

Los mayores avances en relación al desarrollo de metodologías específicas se presentaron en relación a la mortalidad, en cambio, la fecundidad y las migraciones han recibido menor atención por parte de los investigadores. En cuanto al pronóstico de la fecundidad se destaca, en primer lugar, que los métodos elaborados para dicha componente se encuentran menos desarrollados que los destinados al análisis de la mortalidad. La fecundidad presenta dificultades a la hora de ser pronosticada debido a los cambios estructurales en su comportamiento altamente asociados a cambios en las pautas culturales. Lee (1993) encontró necesario preespecificar el valor medio a largo plazo de la fecundidad total e imponer valores límites para reducir la amplitud del intervalo de predicción e implementó un método paralelo al del Lee y Carter. Por otro lado, un enfoque de componentes principales fue empleado por Bozik y Bell (1987), quienes utilizan en su análisis los primeros cuatro componentes principales y modelos ARIMA[1] multivariados. Un método similar emplean Hyndman y Ullah (2007) como parte de un análisis funcional, dicho método es el que se utiliza en el presente trabajo. Los autores emplean un mismo modelo base para analizar la mortalidad y la fecundidad. La metodología propuesta por Hyndman y Ullah (2007) difiere en varios aspectos importantes con respecto al modelo de Lee-Carter; en primer lugar esta enmarcado en el paradigma de datos funcionales (Ramsay y Silverman, 2005), por otro lado emplea suavizados semi-paramétricos con el fin de reducir la aleatoriedad inherente a los datos observados y en la etapa de descomposición de las tasas suavizadas permite utilizar componentes principales clásicos o robustos. Estos últimos tienen en cuenta dentro del análisis, los posibles outliers. En relación al modelado de outliers, Li y Chan (2005) proponen una versión robusta del modelo de Lee-Carter, pero esta no se ubica dentro del paradigma de los datos funcionales.

Los métodos para pronosticar la migración están aún menos desarrollados y son habitualmente mucho más simples (George y Perreault, 1992). La razón principal de esta falta de desarrollo radica en la carencia de series temporales confiables, detalla-

[1]Modelo autorregresivo integrado promedio móvil

das y que representen adecuadamente los flujos migratorios internacionales. Además, las tendencias observadas en esta componente son más volátiles, debido a los acontecimientos políticos y económicos y los cambios en la legislación. Un método que puede solucionar la falta de datos es estimar la migración neta como la diferencia entre el incremento del tamaño de la población y el crecimiento natural utilizando la ecuación compensadora, también llamada ecuación de crecimiento. Más aún, para pronósticos de población subnacionales que involucran efectos de la migración interna, este método es el único disponible ya que es frecuente que no se registren datos sobre migración a ese nivel de desagregación, ver Miller (2003); Miller y Lee (2004). Además, el uso de la migración neta no contempla un aspecto importante: la bondad de pronóstico depende de la desagregación de acuerdo a diferentes grupos o tipos de inmigrantes y emigrantes (Hilderink *et. al.*, 2002 y Rogers, 1990). Por otro lado, Beer (1997) encontró consistencia entre pronósticos de series de tiempo de inmigración total, emigracion y migración neta. También es posible obtener los pronósticos de migración neta desagregados por edad (Keilman y Pham, 2004), utilizando una versión reducida del modelo multi-exponencial desarrollado en Rogers y Castro (1981) y Rogers y Little (1994).

Durante los últimos años se han desarrollado múltiples enfoques para pronósticos probabilísticos (Booth *et al.*, 2006a) y dentro de esta categoría han cobrado una particular relevancia los métodos para datos funcionales (Ramsay y Silverman, 2005). Estos métodos, de reciente aparición, constituyen un nuevo marco para el análisis de series de tiempo, que ha sido adoptado, entre otras finalidades para realizar pronósticos para todas las componentes demográficas (Hyndman y Ullah, 2007). Un aporte interesante a este enfoque lo hacen Hyndman *et al.* (2013) al introducir la idea de pronósticos coherentes en el paradigma de datos funcionales. La idea principal de esta propuesta radica en que la diferencia entre los pronósticos de grupos de interés debe permanecer constante a través del tiempo, reproduciendo la relación presente en los datos observados. Los grupos de interés pueden ser subregiones geográficas o géneros, por mencionar algunas posibilidades.

Finalmente, si bien resulta interesante el análisis y pronóstico de cada componente por separado, el objetivo último suele ser integrar los resultados para obtener una proyección probabilística de la población, de este modo, los componentes pronosticados (mortalidad, fecundidad y migraciones) son utilizados para generar pronósticos probabilísticos de población a través de la renovación estocástica de la población usando el método de las componentes Preston *et al.* (2001). El pronóstico de población se obtiene analiticamente a partir de la matriz estocástica de Leslie (Alho y Spencer, 1985, Lee y Tuljapurkar, 1994 y Sykes, 1969) o de un modo más simple por simulación de Monte Carlo, generando una distribución de los posibles resultados. Para ambos enfoques es necesario especificar la media (o mediana), una estructura de variancia-covariancia y la forma de la distribución de cada componente demográfica.

Capítulo 2

Objetivos

Objetivo General

En los últimos años se ha producido un auge en el desarrollo de métodos probabilísticos de pronóstico en el área demográfica y los mismos, cada vez con mayor aceptación, son implementados por las agencias de estadísticas oficiales. En este contexto el presente trabajo constituye un desafío para el caso de Argentina, dado que su principal objetivo es presentar una propuesta metodológica demográfico-estadística, que permite modelar y pronosticar las componentes demográficas, adaptada a datos del mencionado país. Asimismo, la metodología se basa en información disponible a nivel oficial, de modo que que permite que los modelos estimados sean replicados por cualquier usuario. Además, es importante destacar que, para el desarrollo de este trabajo, es necesario lograr una confluencia de dos ciencias: la estadística y la demografía, cada una con propios paradigmas y lenguajes, si bien el abordaje del objetivo se realiza desde la perspectiva estadística. En resumen, la principal finalidad de este trabajo es modelar y pronosticar las componentes demográficas a través de metodología de alto sustento teórico, enmarcada en un contexto de modelos estocásticos, ya que éstos se adaptan más adecuadamente al modelado de componentes que dependen del comportamiento humano.

Tareas para llevar adelante el objetivo

Aplicar modelos para datos funcionales a la mortalidad, la fecundidad y la migración internacional neta de Argentina, en el período 1980 a 2010, para obtener la descomposición de cada componente demográfica así como los pronósticos de las mismas y del total de la población. Para este punto es necesario organizar, depurar y compatibilizar los datos oficiales existentes, así como también adaptar la programación disponible al contexto de los datos.

Obtener pronósticos probabilísticos de la mortalidad, la fecundidad y las migraciones junto con sus intervalos de pronóstico con un 80 % de confianza.[1]

Comparar los pronósticos que se obtienen, con los métodos de pronósticos tradicionales y más comúnmente utilizados, los modelos ARIMA (Box y Jenkins, 1976).

Elaborar, en base a la información disponible, una estimación del saldo migratorio anual a través la ecuación de crecimiento, proporcionando cifras desagregadas a nivel edad simple.

Utilizar la metodología alternativa de pronósticos coherentes (Hyndman *et al.*, 2008) con el fin de evitar una divergencia entre ambos géneros y comparar los resultados obtenidos para cada género en forma independiente.

Combinar los pronósticos de la mortalidad, la fecundidad y las migraciones para obtener el pronóstico de la población por edades simples y género.

Implementar una simulación por Monte Carlo para obtener los intervalos de pronóstico probabilísticos de población por edad y sexo.

[1]Este intervalo propuesto, de menor amplitud que el tradicional 95 %, se adapta mejor a datos demográficos, sostiene el Dr. Rob Hyndman. En sí mismos, los datos demográficos son el resumen de un gran número de comportamientos humanos individuales y no el resultado de un experimento controlado que permite obtener resultados con una mayor precisión.

Comparar los pronósticos obtenidos con las proyecciones oficiales.

Capítulo 3

Pronósticos demográficos probabilísticos y datos funcionales

Las proyecciones de población son el producto demográfico más utilizado por los no demógrafos y sin embargo no es el principal objeto de estudio por parte de los científicos de población. Los usuarios de estos resultados podrían sorprenderse por la falta de una teoría rigurosa, o una base histórica para los escenarios que subyacen en las proyecciones más usadas. Si bien las tendencias generales de la fecundidad, la mortalidad y la migración pueden discernirse y proyectarse hacia el futuro con resultados razonables, existe una incertidumbre considerable vinculada a cada tendencia específica y a a un país o región particular. La cuantificación de esta incertidumbre es útil para los usuarios de las proyecciones porque permite tener en cuenta las diferentes implicancias de un pronóstico. En los pronósticos actuales, la incertidumbre se expresa típicamente proporcionando escenarios alternativos basados en la variación de la trayectoria para la fecundidad, pero rara vez, para la mortalidad y la migración. Escenarios altos y bajos se utilizan para abarcar una gama de posibles futuros. Sin embargo, ninguna medida de probabilidad específica se adjunta a esta gama, y su significado es, por lo tanto, ambiguo.

Un pronóstico que incluye distribuciones de probabilidad, permite incorporar la magnitud de la incertidumbre mediante una probabilidad proporcionando resultados más

informativos. Se pueden pensar los resultados demográficos futuros como variables aleatorias con una distribución de probabilidad. De este modo el pronóstico puntual es la media o la mediana de esta distribución, y, dada la distribución, los límites de un intervalo de confianza se pueden obtener fácilmente. Las distribuciones son condicionales a la información que se tiene en el momento de la previsión.

Por otro lado, el análisis de datos funcionales es aquella parte de la estadística que trabaja con muestras de funciones aleatorias. Ramsay *et al.* (1997), entre otros, introducen herramientas de análisis para este tipo de datos. Las medidas de tendencia central, de dispersión y de relación entre variables conocidas para muestras de variables aleatorias se pueden definir de manera análoga para muestras de datos funcionales. En realidad Ramsay y Silverman definen estas medidas de manera general para cualquier espacio que tenga definido un producto escalar, y luego deducen de ellas las definiciones para muestras de funciones.

En la presente tesis se emplea un enfoque de datos funcionales para generar pronósticos probabilísticos demográficos de la mortalidad, la fecundidad, la migración y la población. Este enfoque permite cuantificar la incertidumbre asociada a los pronósticos demográficos través del cálculo de intervalos de pronóstico, permitiendo además obtener una descomposición de cada componente demográfica, útil para describir su tendencia en el tiempo y su comportamiento particular para cada una de las edades. Su insumo básico son las tasas observadas, calculadas a partir de cifras de hechos vitales y población. Para continuar, se definen en la próxima sección el término tasa, las cantidades necesarias para construirla y el concepto de dato funcional con mayor profundidad.

3.1. Definiciones básicas

Tasas

Es importante destacar, que el conocimiento del número de casos de algún evento en particular, muertes o nacimientos, presentes en una población, tiene de por sí poca utilidad para epidemiólogos o demógrafos si no se relaciona dicha frecuencia con la población de la cual proceden los casos. Esta relación se establece generalmente a través de la construcción de una tasa.

El termino tasa se usa en muchos campos y su significado no es consistente en todos ellos. En demografía, las tasas se definen comúnmente como tasas de ocurrencia/exposición (Preston *et al.*, 2001). El numerador de este tipo de tasas contabiliza el número de ocurrencias de un evento de interés, mientras que el denominador combina dos factores: el número de personas en la población y la longitud del tiempo que enmarca el estudio. El denominador de una tasa de "ocurrencia/exposición", se conoce con el nombre de personas-tiempo en riesgo y es la suma del tiempo que cada persona permanece bajo observación y en riesgo de convertirse en un caso. El tiempo debe expresarse en unidades de medida apropiadas, por ejemplo, personas-año, en cuyo caso se dice que la tasa esta anualizada. Cuando los casos se obtienen a partir del registro de estadísticas vitales no se dispone de medidas directas de personas-tiempo en riesgo. Una estimación de dicho valor, para una población en un período dado, se puede calcular como el producto entre el número total de personas que componen la población en el punto medio del período de interés y la duración total del período. Este método proporciona estimaciones adecuadas siempre que la población permanezca estable durante ese período.

Las tasas se pueden calcular para una población entera o para subgrupos específicos de la misma. En el primer caso se las denomina tasas crudas o brutas y en el segundo tasas específicas por subgrupo de interés. Por ejemplo, en el estudio de la mortalidad resulta de interés el cálculo de tasas específicas por sexo y edad.

En las tasas específicas, el denominador de la tasa se define por separado para cada grupo. Por ejemplo, la tasa especifica de mortalidad según edad se define como el cociente entre el número de defunciones acaecidas en un grupo de edad específica de la población de un área geográfica dada durante un período determinado y el correspondiente valor de personas-tiempo en riesgo en ese grupo específico de edad del área geográfica y período bajo estudio. Si las defunciones se obtienen a partir de Registros de Estadísticas Vitales, el denominador se estima a partir de los datos censales.

Se definen a continuación los insumos básicos del modelo para datos funcionales (MDF) que se desarrolla en este trabajo:

$B_t(x)$ = Nacimientos en mujeres de edad x ocurridos durante el año calendario t;

$D_t(x)$ = Muertes de personas de edad x ocurridas en el año calendario t;

$P_t(x)$ = Población de edad x al primero de enero del año t;

$E_t(x)$ = Población de edad x expuesta al riesgo al 30 de junio del año t;

donde $x = 0, 1, 2, ..., p-1, p^+$ y $t = 1, ..., n$. Con p^+ se indica el grupo final abierto (generalmente "80 y más", "85 y más", etc).

La tasa de mortalidad específica por edad en el año calendario t es $m_t(x) = D_t(x)/E_t(x)$ y $f_t(x) = B_t(x)/E_t^M(x)$ es la tasa de fecundidad específica por edad para el año t (el supraíndice M se utiliza para indicar que se trata de la población de mujeres). Los nacimientos se dividen por sexo usando ρ, la razón de sexo al nacer.

La migración neta denotada con G_t se estima utilizando la ecuación de crecimiento:

$$G_t(x, x+1) = P_{t+1}(x+1) - P_t(x) + D_t(x, x+1) \quad con \quad x = 0, 1, 2, ..., p-2 \tag{3.1}$$

$$G_t(p-1^+, p^+) = P_{t+1}(p^+) - P_t(p-1) + D_t(p-1, p^+) \tag{3.2}$$

$$G_t(B, 0) = P_{t+1}(0) - B_t + D_t(B, 0), \tag{3.3}$$

donde $D_t(x, x+1)$ se refiere a las muertes en el año calendario t de personas de edad x al principio del año t; $D_t(p-1, p^+)$ indica las muertes en el año calendario t de personas de edad $p-1$ y mayores al comienzo del año t y $D_t(B, 0)$ son las muertes en el año calendario t de los nacidos durante el año t, de modo similar se definen para la migración neta, $G_t(x, x+1)$, $G_t(p-1, p^+)$ y $G_t(B, 0)$. Las muertes $D_t(x, x+1)$ se estiman utilizando un enfoque de tabla de vida estándar. Ver Preston *et al.* (2001) para más detalles.

Datos Funcionales

En una definición breve, Muller (2006) afirma: los datos funcionales son datos multivariados que poseen un ordenamiento en las dimensiones. Los datos funcionales están conformados por una muestra de curvas aleatorias independientes e idénticamente distribuidas y las mismas se suponen realizaciones de un proceso estocástico subyacente. Una gran diferencia entre los datos funcionales y los datos multivariados es que en el caso de los datos funcionales el orden y las relaciones de *vecindad* son relevantes y están claramente definidas. Este punto puede ilustrarse por el hecho de que uno puede reordenar componentes de un vector de datos multivariados y llegar exactamente al mismo análisis estadístico, que para el vector ordenado en la forma original. Para los datos funcionales la situación es completamente distinta.

La filosofía básica del análisis de datos funcionales radica en pensar a los datos observados como funciones en lugar de una secuencia de observaciones individuales. El término funcional, se refiere a la estructura intrínseca de los datos, más que a su forma explícita. En la práctica los datos observados se registran en forma discreta, como n

pares (t_j, y_j), donde y_j representa, de este modo, una instantánea de la función en el tiempo t_j afectada por el error de medida. Adicionalmente se supone, que la función subyacente es suave, es decir, un par de valores adyacentes y_j y y_{j+1} están necesariamente asociados hasta cierto punto y los mismos no difieren demasiado uno el otro. Por función suave, habitualmente, se entiende cuando la misma posee una o más derivadas. Si esta propiedad de suavidad no se aplicara, no habría ganancia al tratar los datos como funcionales en lugar de tratarlos simplemente como multivariados. Considerar una observación como funcional no implica que esta deba ser registrada para todo valor de t (de hecho esto no sería posible), en cambio esto significa que se supone la existencia de una función x que da origen a los datos.

Uno de los tipos de datos más adecuado para un análisis funcional son las series de tiempo no estacionarias, incluso con observaciones no equiespaciadas (Ramsay *et al.*, 1997). El enfoque de datos funcionales aplicado a datos temporales provee un método alternativo semi-paramétrico para el modelado de trayectorias individuales. La idea subyacente es ver a las trayectorias longitudinales individuales observadas como una muestra de funciones aleatorias que no están parametricamente especificadas. Las medidas observadas para un individuo corresponden entonces a valores de una trayectoria aleatoria, afectada por el error de medida. Un primer objetivo del enfoque de datos funcionales es reducir la dimensión de las trayectorias. Otro objetivo es predecir las trayectorias individuales a partir de las medidas realizadas para un sujeto, obteniendo información de la muestra completa de sujetos. La reducción de dimensión o regularización puede ser implementada a través de: Análisis de Componentes Principales Funcionales (ACPF), Yao *et al.* (2009). Otros métodos que han probado ser útiles en el Análisis de Datos Funcionales (ADF) incluyen Suavizado con *Splines*, Ke y Wang (2001), B-*splines*, Rice y Wu (2000) o P-*Splines*, Yao y Lee (2006).

En general se esta interesado en un conjunto de datos, y no en una función x únicamente. Una muestra proveniente de una función x_i consiste en n_i pares (t_{ij}, y_{ij}),

$j = 1, ..., n_i$. La suavidad, en el sentido de poseer un cierto número de derivadas, es una propiedad de la función x y puede no ser del todo obvio en el vector de observaciones $(y_1, y_2, ..., y_n)$ debido a la presencia de error observacional o al ruido proveniente del proceso de medición (pensando en el caso de una única función). Luego cada observación puede definirse como $y_j = x(t_j) + \varepsilon_j$ donde ε_j representa el error que contribuye a la rugosidad de los datos.

3.2. Pronósticos probabilísticos demográficos: el enfoque funcional

Con el fin de obtener pronósticos demográficos individualmente para cada componente demográfica, y luego para la población, primero se desarrollan modelos funcionales de series de tiempo sobre las cinco componentes demográficas específicas: $m_t^M(x)$, $m_t^V(x)$, (tasas de mortalidad por edad para mujeres y varones) $f_t(x)$ (tasas de fecundidad por edad de la madre), $G_t^M(x, x+1)$ y $G_t^V(x, x+1)$ (migración neta por edad de mujeres y varones). Para modelar cada cantidad se sigue el enfoque de Hyndman y Ullah (2007) y los resultados obtenidos a partir de los cinco modelos se emplean luego en la simulación de la población futura.

3.2.1. Modelo para datos funcionales definido para tasas

Se denota con $y_t^*(x)$ la cantidad a ser modelada, tasas de mortalidad, fecundidad o números de migración neta para la edad x en el año t. Si bien es posible plantear una transformación general de Box y Cox (1964) sobre y_t^*, que permite modelar una tasa cuya variación aumenta con el valor de y_t^*, es decir, cuando la variabilidad de las tasas es mayor a medida que las tasas son mayores, en la mayoría de las aplicaciones se implementa directamente la transformación logaritmo.

Se supone el siguiente modelo para las observaciones transformadas $y_t(x)$:

$$y_t(x) = s_t(x) + \sigma_t(x)\varepsilon_{t,x} \tag{3.4}$$

$$s_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x), \tag{3.5}$$

donde $s_t(x)$ es una función suave subyacente de x, $\varepsilon_{t,x}$ son variables aleatorias, independientes e idénticamente distribuidas y la definición de $\sigma_t(x)$ permite a la variancia cambiar con la edad y con el tiempo. Esto significa que las observaciones transformadas son la suma de la cantidad a modelar, $s_t(x)$, una función suave de la edad y un error (Ecuación 3.4). La ecuación 3.5 describe la dinámica de $s_t(x)$ a través del tiempo, en esta ecuación, $\mu(x)$ es la media de $s_t(x)$ a través de los años, $\{\phi_k(x)\}$ es un conjunto de K funciones base ortogonales calculadas utilizando una descomposición en componentes principales funcionales de la matriz $[\hat{s}_t(x) - \hat{\mu}(x)]$ y $e_t(x)$ es el error del modelo (el cual se supone no correlacionado serialmente). La dinámica del proceso esta controlada por los coeficientes $\{\beta_{t,k}\}$, los cuales tienen un comportamiento independiente uno de otro (garantizado por la utilización del método de componentes principales). Existen tres fuentes de variación en el modelo: $\varepsilon_{t,x}$ representa la variación aleatoria con respecto a la distribución relevante, para los nacimientos, muertes y migrantes (Poisson o Normal); $e_t(x)$ representa el residuo que surge al modelar $s_t(x)$ utilizando un conjunto de funciones bases y además existe una aleatoriedad inherente al modelo de series de tiempo para cada $\{\beta_{t,k}\}$ que ejerce los cambios en la dinámica de la curva suave $\{s_t(x)\}$. Es importante destacar que es posible implementar este enfoque para edades simples y también para grupos quinquenales.

Este modelo fue propuesto inicialmente por Hyndman y Ullah (2007) para modelar tasas de mortalidad y fecundidad empleando una transformación logaritmo en lugar de plantear la transformación de Box-Cox. También ha sido utilizado por Erbas *et al.* (2007) para pronosticar tasas de mortalidad por cáncer de mama. Como señalan Hyndman y Ullah (2007), el modelo, es una generalización del conocido modelo de Lee y Carter

(1992) para pronosticar tasas de mortalidad. En el enfoque de Lee y Carter (1992), y_t^* representa a la tasa de mortalidad e $y_t(x)$ es el logaritmo de la mortalidad para el año t y la edad x, además, no incluye suavizados y por lo tanto $\sigma_t(x) = 0$ e $y_t(x) = s_t(x)$, finalmente $\mu(x)$ se estima como el promedio de $y_t(x)$ a través de los años. El número de componentes $K = 1$, y $\beta_{t,1}$ se obtienen a partir la primera componente principal de la matriz $[\hat{s}_t(x) - \hat{\mu}(x)]$ y los pronósticos se obtienen ajustando una serie de tiempo al coeficiente $\beta_{t,1}$; en la práctica el modelo que se obtiene resulta generalmente un paseo aleatorio con pendiente.

El modelo general, que se aplica a cada una de las componentes demográficas, se estima a través de la etapas que se enumeran a continuación:

1. Se estiman las funciones suaves $s_t(x)$ a través de regresión no paramétrica sobre $y_t(x)$ para cada año t, ecuación (3.5);

2. A $\hat{\mu}(x)$ se la define como la media de $\hat{s}_t(x)$ a través de los años;

3. Los coeficientes $\beta_{t,k}$ y las bases $\phi_k(x)$, con $k = 1, ..., K$ se estiman aplicando análisis de componentes principales funcionales sobre la matriz $[\hat{s}_t(x) - \hat{\mu}(x)]$;

4. Se ajusta un modelo de series de tiempo a $\hat{\beta}_{t,k}$ donde $k = 1, ..., K$. Para ello es posible utilizar un modelo ARIMA (Box y Jenkins, 1976) o modelos de espacio de estado de innovaciones (Hyndman *et al.*, 2008).

Aunque el valor de K debe ser especificado, Hyndman y Ullah (2007) sostienen que el método es insensible al valor elegido siempre y cuando sea lo suficientemente grande. Esto significa, que el costo al elegir K grande es pequeño (más allá del tiempo computacional), mientras que seleccionar un K pequeño puede producir menor exactitud en los pronósticos. Hyndman y Booth (2007) utilizan $K = 6$ para todos los componentes demográficos; esta cantidad parece ser mayor a la que cualquier componente demográfi-

ca requiere.

La variancia observacional, σ_t^2 depende de la naturaleza de los datos. Para las muertes la variancia observacional se estima a partir del logaritmo de las tasas ($y_t^* = m_t(x)$) suponiendo que las muertes se distribuyen Poisson (Brillinger, 1986) con parámetro medio $m_t(x)E_t(x)$ (donde $E_t(x)$ es la población de edad x expuesta al riesgo al 30 de junio del año t). Luego y_t^* tiene una variancia aproximada $E_t(x)^{-1}m_t(x)$ y la variancia de $y_t(x)$ (por aproximación de Taylor) resulta:

$$\hat{\sigma}_t^2 \approx [m_t(x)]^{-1} E_t(x)^{-1}. \tag{3.6}$$

Para los nacimientos se supone una distribución Poisson (Keilman *et al.*, 2002) con media $f_t(x)E_t^F(x)$, lo que implica,

$$\hat{\sigma}_t^2 \approx [f_t(x)]^{-1} E_t(x)^{-1}. \tag{3.7}$$

Para los datos de migraciones no se hacen supuestos distribucionales y se estima $\sigma_t^2(x)$ utilizando una regresión no paramétrica de $[y_t(x) - s_t(x)]^2$ sobre x.

El suavizado semi-paramétrico puede ser llevado a cabo utilizando alguno de los múltiples métodos de suavizado existentes, ver Ruppert *et al.* (1976) y Simonoff (1976). Puntualmente, (Hyndman y Ullah, 2007) sugieren utilizar regresión *spline* penalizada con restricciones para la mortalidad y la fecundidad, de modo que los pesos contemplen la heterogeneidad presente en este tipo de datos. Se impone además una restricción de monotonía para la mortalidad y una restricción de concavidad para la fecundidad. Más específicamente, para datos de mortalidad, se definen pesos iguales a la inversa de la variancia teórica (derivada del supuesto distribucional Poisson) y se utiliza una regresión *spline* penalizada, ver Wood (2003) y He y Hg (1999), para estimar las curvas $y_t(x)$ que representan a las tasas $m_t(x)$ luego de la transformación logarítmica. La restricción impuesta determina que las curvas sean monotonamente crecientes para $x > c$, es decir,

a partir de una edad determinada, permitiendo reducir el ruido en las curvas estimadas para edades avanzadas. La imposición resulta lógica dado que cuanto más anciana es una persona tiene más probabilidad de morir.

Para los datos de fecundidad se utilizan los pesos de modo análogo y se impone como restricción la concavidad de las curvas, respetando el perfil observado habitualmente en las curvas de fecundidad. El método que permite implementar esta restricción puede verse en He y Hg (1999). Para suavizar las tasas de fecundidad se utiliza una regresión cuantil semi-paramétrica por B-*Splines* con restricciones (He y Shi, 1996), la misma permite establecer restricciones a las funciones suavizadas tales como monotonía, convexidad, concavidad o límites.

Finalmente en el caso de las migraciones se utiliza simplemente un suavizado por regresión local, *loess*.

El éxito del modelo depende de que la superficie bivariada $\{s_t(x) - \mu_t(x)\}$ pueda ser aproximada por la suma de unos pocos productos de funciones univariadas del tiempo (t) y la edad (x). Hyndman y Ullah (2007) sugieren que el modelo es adecuado para producir pronósticos, pero no lo es tanto para estimar la variancia de pronóstico, por lo que proponen un ajuste para su cálculo.

3.2.1.1. Componentes principales funcionales

La finalidad de la aplicación de Componentes Principales Funcionales (CPF) en la tercera etapa del modelo propuesto para tasas demográficas es obtener un conjunto de K funciones ortonormales ϕ_m con $m = 1, 2..., K$, de modo tal que la "expansión" de cada curva en términos de las K bases funcionales se aproxime lo máximo posible a la curva $\hat{s}_t^*(x) = \hat{s}_t(x) - \hat{\mu}(x)$, donde $\hat{s}_t(x)$ es la función suavizada que estima a la función subyacente y desconocida $s_t(x)$ y $\hat{\mu}(x)$ es la media de las funciones suavizadas estimadas

a través de los años.

La expansión buscada tiene la forma

$$\hat{\hat{s}}_t^*(x) = \sum_{k=1}^{K} \beta_{t,k}\phi_k(x), \tag{3.8}$$

donde $\beta_{t,k}$ es el valor de

$$\int_x \hat{\hat{s}}_t^*(x)\phi_k(x)dx. \tag{3.9}$$

De este modo, los $\beta_{t,k}$ con $t = 1, ...n$ y $k = 1, ..., K$ son los *scores* obtenidos de la descomposición en componentes principales.

El criterio para el ajuste de cada curva se establece a partir de hacer mínimo el error cuadrático integrado

$$\| \hat{s}_t^* - \hat{\hat{s}}_t^* \|^2 = \int_x \left[\hat{s}_t^*(x) - \hat{\hat{s}}_t^*(x)\right]^2 dx \tag{3.10}$$

y en forma global, la suma de cuadrados del error de las componentes principales (conocido por su sigla en inglés PCASSE),

$$PCASSE = \sum_{t=1}^{n} \| \hat{s}_t^* - \hat{\hat{s}}_t^* \|^2 . \tag{3.11}$$

3.2.2. Pronósticos funcionales de las tasas

Se tienen datos hasta un tiempo $t = n$ y se desean estimar valores futuros de $y_t(x)$ para $t = n+1, ..., n+h$ y para $x = 0, 1, 2, ..., p-1, p^+$. Si con $\hat{\beta}_{n,k,h}$ se denota el pronóstico h pasos hacia adelante de $\beta_{n+h,k}$, y siendo $\hat{y}_{n,h}(x)$ el pronóstico h pasos hacia adelante de $y_{n+h}(x)$, $\hat{\hat{s}}_{n,h}(x)$ el pronóstico h pasos hacia adelante de $\hat{s}_{n+h}(x)$. Luego,

$$\hat{y}_{n,h}(x) = \hat{\hat{s}}_{n,h}(x) = \hat{\mu}(x) + \sum_{k=1}^{K} \hat{\beta}_{n,k,h}\hat{\phi}_k(x), \tag{3.12}$$

un pronóstico de y_t^* se encuentra a través de la transformación inversa a la realizada en el comienzo del análisis, en este caso la operación inversa del logaritmo.

Según Hyndman y Ullah (2007) se puede expresar la variancia del pronóstico como,

$$\hat{V}_h(x) = Var\left[\hat{\hat{s}}_{n+h}(x)|\mathcal{I}, \Phi\right] = \hat{\sigma}_\mu(x) + \sum_{k=1}^{K} u_{n+h,k}\hat{\phi}_k(x) + v(x), \tag{3.13}$$

donde $\mathcal{I} = \{y_t(x_i)\}$ denota todos los datos observados, $\hat{\sigma}_\mu^2(x)$ (la variancia del estimador suave $\hat{\mu}(x)$) se obtiene a través del método de suavizado empleado, $u_{n+h,k} = \hat{Var}(\hat{\beta}_{n+k,k}|\beta_{1,k}, ..., \beta_{n,k})$ se obtiene a partir del modelo de series de tiempo y $v(x)$ se estima promediando $\hat{e}_t^2(x)$ para cada x. El error de suavizado esta dado por el primer término, el error debido a la predicción de la dinámica esta dado por el segundo término, y el tercero es el error debido a la variación dinámica sin explicar. Si se tiene en cuenta el error observacional, se obtiene la siguiente expresión,

$$Var\left[y_{n+h}(x)|\mathcal{I}, \Phi\right] = V_h(x) + \sigma_t^2(x). \tag{3.14}$$

Se destaca que las correlaciones entre las edades se contemplan a través de las funciones suavizadas de la edad (x). Asimismo las correlaciones entre años se tienen en cuenta mediante el modelo de series planteado para los coeficientes $\hat{\beta}_{t,1}, ..., \hat{\beta}_{t,K}$.

Para valores bajos de h es posible chequear la validez de $V_h(x)$ calculando la variancia de pronostico empírica dentro de la muestra

$$W_h(x) = \frac{1}{n-h-m+1}\sum_{t=m}^{n-h}\left[\hat{s}_{t+h}(x) - \hat{\hat{s}}_{t,h}\right]^2, \tag{3.15}$$

donde m es el menor número de observaciones utilizadas para ajustar el modelo. En la práctica, pueden ocurrir considerables diferencias entre $W_h(x)$ y $V_h(x)$. En consecuencia se utiliza la siguiente expresión de la variancia ajustada:

$$Var\left[y_{n+h}(x)|\mathcal{I}, \Phi\right] = V_h(x)\, W_1(x)/\, V_1(x) + \sigma_t^2(x). \tag{3.16}$$

Este ajuste hace que la variancia de pronóstico un paso hacia adelante coincida con la variancia de pronóstico empírica un paso hacia adelante dentro de la muestra. Se supone que el mismo ajuste multiplicativo puede ser aplicado a mayores horizontes de pronóstico.

3.2.3. Pronósticos Coherentes

En demografía se denomina pronósticos "coherentes" a aquellos que para hombres y mujeres (u otros subgrupos) difieren siempre en la misma proporción través del tiempo, esencialmente, lo que la metodología impone es que la diferencia entre los pronósticos de ambos grupos sea estacionaria a través del tiempo. En el caso de la mortalidad, cuando se supone independencia, los pronósticos para subpoblaciones son generalmente divergentes, en el sentido de que la diferencia entre grupos se amplia a medida que aumenta el horizonte de pronóstico. El método de pronósticos coherentes pretende asegurar que los pronósticos para poblaciones relacionadas mantengan cierta relación estructural basada en información histórica y consideraciones teóricas. Por ejemplo, se ha observado que la mortalidad masculina es más alta que la mortalidad femenina para todas las edades y mientras la evidencia disponible sustenta tanto la hipótesis la biológica-genética como la social-cultural-ambiental-conductual, Kalben (2000) concluye que el factor determinante es el biológico. Esta diferencia en la mortalidad entre los sexos es de esperar que persista en el futuro y que además permanezca dentro de los límites observados.

El problema de obtener pronósticos coherentes ha sido considerado previamente por Lee (1993), Lee (2000) y Li y Chan (2005) en el contexto del modelo de Lee y Carter. Lee (2006) presenta un marco general para pronosticar la esperanza de vida como la suma de una tendencia común y la tasa de convergencia específica de cada subpoblación hacia esa tendencia común. Hyndman *et al.* (2013) proponen un método para el pronóstico coherente para las tasas de mortalidad de dos o más subpoblaciones, llamado también método de pronóstico funcional cociente-producto, el mismo modela y pronos-

tica la media geométrica de las tasas de las subpoblaciones y el cociente de las tasas de las subpoblaciones para producir pronósticos de las tasas que no divergen en el tiempo. Además, Hyndman *et al.* (2013) proponen una definición precisa de los pronósticos coherentes, sostienen que los mismos se obtienen cuando los pronósticos de la razón de tasas específicas por edad para dos subpoblaciones cualesquiera convergen a un conjunto de constantes apropiadas.

Inicialmente se plantea el problema de pronosticar las tasas especificas de mortalidad por edad para varones y mujeres, de todos modos, el método puede aplicarse a cualquier número de subpoblaciones e incluso a otras componentes demográficas, como por ejemplo, la migración.

Para aplicar el método del cociente-producto para varones y mujeres se definen las raíces cuadradas de los productos y cocientes entre las tasas suavizadas de cada sexo:

$$p_t(x) = \sqrt{s_{t,M}(x)s_{t,V}(x)}, \tag{3.17}$$

$$c_t(x) = \sqrt{s_{t,M}(x)/s_{t,V}(x)}. \tag{3.18}$$

Las cantidades 3.17 y 3.18 se modelan en lugar de las tasas de mortalidad especificas por edad observadas. La ventaja de este enfoque es que el cociente y el producto se comportaran de forma aproximadamente independiente uno de otro siempre que tengan variancias similares. En la escala logarítmica estas cantidades son sumas y diferencias aproximadamente no correlacionadas (si X e Y son dos variables aleatorias, luego $Corr(X-Y, X+Y) = Var(X) - Var(Y)$. Luego si X e Y tiene igual variancia, $X-Y$ y $X+Y$ son no correlacionadas). Si existen diferencias sustanciales en las variancias de las subpoblaciones luego el producto y el cociente no serán no correlacionados. Este hecho no produce sesgo en los pronósticos pero los hace menos eficientes. En cualquier caso, el método, continuaría siendo mejor que tratar las poblaciones de forma independiente

ya que aún de este modo se estarían imponiendo restricciones de coherencia. Es importante tener en cuenta que, el logaritmo de la raíz del producto representa el promedio entre el logaritmo de la mortalidad femenina y masculina, mientras que el logaritmo del la raíz del cociente es la mitad de la diferencia entre los logaritmos de hombres y mujeres.

Para modelar $p_t(x)$ y $c_t(x)$ Hyndman y Ullah (2007) utilizan modelos funcionales de series de tiempo:

$$log\left[p_t(x)\right] = \mu_p(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x), \tag{3.19}$$

$$log\left[c_t(x)\right] = \mu_r(x) + \sum_{\ell=1}^{L} \gamma_{t,\ell}\psi_\ell(x) + w_t(x), \tag{3.20}$$

donde las funciones $\{\phi_k(x)\}$ y $\{\psi_\ell(x)\}$ son las componentes principales obtenidas de la descomposición de $[log(p_t(x))]$ y $[log(c_t(x))]$ respectivamente, y $\beta_{t,k}$ y $\gamma_{t,\ell}$ son los correspondientes *scores* de las componentes principales. La función $\mu_p(x)$ es la media del conjunto de curvas $\{log(p_t(x))\}$ y $\mu_c(x)$ es la media de $\{log(c_t(x))\}$. Los términos de error, dados por $e_t(x)$ y $w_t(x)$, tienen media cero y son no correlacionados.

Los modelos se estiman utilizando el algoritmo de componentes principales ponderados de Hyndman y Shang (2009), el cual otorga más peso a datos recientes, y por ello evita el problema de que las funciones $\{\phi_k(x)\}$ y $\{\gamma_\ell(x)\}$ cambien a través del tiempo, (Lee y Miller, 2001).

Los coeficientes $\{\beta_{t,1}, ..., \beta_{t,K}\}$ y $\{\gamma_{t,1}, ..., \gamma_{t,L}\}$ se pronostican utilizando modelos de series de tiempo y para asegurar la coherencia se requiere que los coeficientes $\{\gamma_\ell(x)\}$ se comporten como un proceso estacionario. Los pronósticos de los coeficientes son multiplicados luego por las funciones base, dando como resultado curvas pronosticadas $p_t(x)$ y $c_t(x)$ para un futuro t. Si $p_{n+h|n}(x)$ y $c_{n+h|n}(x)$ son pronósticos h-pasos hacia delante de las funciones cociente y producto respectivamente, luego los pronósticos de las

tasas para cada sexo se obtienen utilizando $y^{M}_{n+h|n}(x) = p_{n+h|n}(x)c_{n+h|n}(x)$ y $y^{V}_{n+h|n}(x) = p_{n+h|n}(x)/c_{n+h|n}(x)$.

Los pronósticos de cada componente demográfica se obtienen a partir de los pronósticos de los coeficientes $\beta_{t,1}, ..., \beta_{t,K}$ y $\gamma_{t,1}, ..., \gamma_{t,L}$, realizados en forma independiente unos de otros. No hay necesidad de considerar modelos vectoriales dado que los coeficientes $\beta_{t,k}$ son no correlacionados por construcción (Hyndman y Ullah, 2007). Lo mismo sucede con los $\gamma_{t,k}$, que son aproximadamente no correlacionados entre si debido al uso de productos y cocientes.

Los coeficientes del modelo para la raíz del producto, $\{\beta_{t,1}, ..., \beta_{t,K}\}$ se pronostican utilizando modelos ARIMA[1]. Al ajustar el modelo ARIMA, se utiliza el algoritmo automático de selección desarrollado por Hyndman *et al.* (2008) para seleccionar el orden apropiado tanto para la parte promedio móvil como para la parte autorregresiva.

Los coeficientes del modelo para la raíz del cociente $\{\gamma_{t,1}, ..., \gamma_{t,L}\}$ se pronostican utilizando modelos estacionarios $ARMA(p, q)$[2] (Box *et al.*, 2008) o procesos $ARFIMA(p, d, q)$[3] (Granger y Joyeux (1980), Hosking (1981)), donde el orden de integración es fraccionario, también conocidos como procesos de memoria larga. La estacionariedad requerida asegura que los pronósticos sean coherentes. Hyndman *et al.* (2013) encontraron que los modelos ARFIMA son útiles para pronosticar los coeficientes de la función cociente, dado que suponen un comportamiento de memoria más larga con respecto a la que plantean los modelos ARMA. Al implementar modelos $ARFIMA(p, d, q)$ se estima el parámetro de diferenciación fraccionaria d utilizando el método de Haslett y Raftery (1989) y para seleccionar los ordenes p y q se emplea el algoritmo de Hyndman *et al.* (2008), luego se usa estimación máximo verosímil (Haslett y Raftery, 1989) y las ecuaciones de pronóstico desarrolladas por Peiris y Perera (1988). En los modelos ARFIMA

[1]Modelo autorregresivo integrado promedio móvil

[2]Modelo autorregresivo promedio móvil, con p y q se indica la cantidad de parámetros del polinomio autorregresivo y promedio móvil, respectivamente

[3]Modelo autorregresivo fraccionalmente integrado promedio móvil, donde d indica el orden de la diferenciación

estacionarios $-0{,}5 < d < 0{,}5$ y para el caso particular $d = 0$ el modelo $ARFIMA(p, d, q)$ se transforma en un modelo $ARMA(p, q)$.

Sea $\hat{\beta}_{n+h|n,k}$ el pronóstico h-pasos hacia adelante de $\beta_{n+h,k}$ y sea $\hat{\gamma}_{n+h|n,\ell,j}$ el pronóstico h-pasos hacia adelante de $\gamma_{n+h,\ell,j}$. Luego el pronóstico de la tasa h-pasos hacia adelante esta dado por

$$log\left[y^*_{n+h,j}(x)\right] = \hat{\mu}_j(x) + \sum_{k=1}^{K} \hat{\beta}_{n+h|n,k}\phi_k(x) + \sum_{\ell=1}^{L} \hat{\gamma}_{n+h|n,\ell,j}\psi_{\ell,j}(x). \tag{3.21}$$

Dado que todos los términos son no correlacionados es posible sumar las variancias, luego

$$\begin{aligned} Var\left\{log\left[y^*_{n+h,j}(x)\right] \mid \mathcal{I}_n\right\} = \hat{\sigma}^2_{\mu_j}(x) + \sum_{k=1}^{K} u_{n+h|n,k}\phi^2_k(x) \\ + \sum_{\ell=1}^{L} v_{n+h|n,\ell,j}\psi^2_{\ell,j}(x) + s_e(x) + s_{w,j}(x) + \sigma^2_{n+h,j}(x) \end{aligned} \tag{3.22}$$

donde $\mathcal{I}_n$ denota los datos observados hasta el momento n más las funciones base $\{\phi_k(x)\}$ y $\{\gamma_\ell(x)\}$, $u_{n+h|n,k} = Var(\beta_{n+h,k} \mid \beta_{1,k}, ..., \beta_{n,k}$ y $v_{n+h|n,k} = Var(\gamma_{n+h,k} \mid \gamma_{1,k}, ..., \gamma_{n,k})$ se pueden obtener a través de los modelos de series de tiempo, $\hat{\sigma}^2_{\mu_j}(x)$ (las variancias de las medias suavizadas) se puede obtener a partir del método e suavizado empleado, $s_e(x)$ se estima promediando $\hat{e}^2_t(x)$ para cada x y $s_{w,j}(x)$ se estima promediando $\hat{w}^2_{t,j}(x)$ para cada x. La variancia observacional $\sigma^2_{t,j}(x)$ es estable de año a año y por ello se estima tomando la variancia observacional media de los datos históricos. Resulta fácil construir un intervalo de pronóstico bajo el supuesto de que los errores están normalmente distribuidos.

3.2.4. Simulaciones estocásticas de población

Se simulan poblaciones a través de la ecuación compensadora adaptada para permitir un error aleatorio observacional. La población inicial es la observada al primero de enero en el año $n+1$. El algoritmo completo se describe paso a paso a continuación.

Para elaborar los pronósticos de población es necesario simular un gran número de trayectorias de cada componente demográfica; $m_t^M(x)$, $m_t^V(x)$, $f_t(x)$, $G_t^M(x, x+1)$ y $G_t^V(x, x+1)$. En forma secuencial y por simulación de Monte Carlo se obtienen N conjuntos de valores futuros para las tasas de mortalidad, fecundidad y en caso que sea posible para los saldos migratorios. A partir de las componentes generadas, empleando el algoritmo que se detalla a continuación es posible generar N poblaciones simuladas.

A continuación se retoman los conceptos presentados en la sección 3.1 y se agregan los necesarios para construir el algoritmo de simulación de la población:
$r_t(x)$: relación de sobrevivencia de la tabla de vida para $x = 0, 1, 2, ..., p-2$,
$r_t(B)$: relación de sobrevivencia de la tabla de vida entre el nacimiento y la edad cero,
$r_(p-1^+)$: relación de sobrevivencia de la tabla de vida desde la edad $p-1^+$ a p^+,
$R_t(x)$: Población ajustada por migración.

Inicialmente y con el fin de tener en cuenta los nacimientos y las muertes de los migrantes (emigrantes e inmigrantes) se ajusta la población por migración neta

$$R_t(x) = P_t(x) + \frac{G_t(x, x+1)}{2} \quad x = 0, ..., p-2, \tag{3.23}$$

de esta manera se supone que los migrantes permanecen la mitad del tiempo expuestos al riesgo. Luego, la migración neta de una cohorte al comienzo del año $n+h$ se calcula utilizando la media suavizada $\tilde{s}_{n+h}$ junto con el remuestreo de $y_t(x) - \hat{s}(x)$.
La mitad de la migración anual neta simulada $G_t(x, x+1)$ se suma a la población de edad x al primero de enero del año t y la otra mitad se suma a la población de edad

x+1 al final del año t. Se denota con R a la población ajustada por la primera mitad de la migración.

Para la migración neta de niños nacidos durante el año t, la mitad de $G_t(B,0)$ se suma a los nacimientos en el año y la otra la mitad se agrega a la población de edad cero al final del mismo año. Para las edades avanzadas $G_t(p-1,p^+)$ se divide equitativamente entre migrantes de edad $p-1$ y p^+ al primero del año t,

$$R_t(p-1) = P_t(p-1) + G_t(p-1,p^+)/4, \tag{3.24}$$

$$R_t(p^+) = P_t(p^+) + G_t(p-1,p^+)/4. \tag{3.25}$$

En el caso de las tasas de mortalidad para cada sexo se toma la trayectoria simulada de $\{m_t(x)\}$, para $t = n+1,...,n+h$, y con el fin de obtener un número aleatorio de muertes se utiliza la población a mitad de año, pero como a su vez esta depende del número de muertes del año anterior la circularidad se soluciona utilizando una tabla de vida basada en valores simulados de $\{m_t(x)\}$ y $R_t(x)$ para obtener un número esperado de muertes para la cohorte (se indica el valor esperado con la barra horizontal sobre la cantidad),

$$\bar{D}_t(x,x+1) = (1-r_t(x))R_t(x), \tag{3.26}$$

$$\bar{D}_t(p-1^+,p^+) = (1-r_t(p-1^+))\left[R_t(p-1) + R_t(p^+)\right] \quad x = 0,...,p-2. \tag{3.27}$$

En base a estas muertes se estima R_{t+1} y por lo tanto la población a mitad de año.

$$\bar{R}_{t+1} = R_t(x) - \bar{D}_t(x,x+1), \tag{3.28}$$

$$\bar{R}_{p^+} = R_t(p-1) + R_t(p^+) - \bar{D}_t(p-1^+,p^+) \quad x = 0,...,p-2. \tag{3.29}$$

En base a la población a mitad de año y $\{m_t(x)\}$ se define la distribución Poisson para las muertes $D_t(x)$,

$$\bar{E}_t(x) = \left[R_t(x) + \bar{R}_{t+1}(x)\right]/2 \quad x = 0, ..., p-1, p^+, \tag{3.30}$$

$$\bar{D}_t = m_t(x)\bar{E}_t(x) \quad x = 0, ..., p-1, p^+, \tag{3.31}$$

luego,

$$D_t(x) \sim Poisson(\bar{D}_t) \quad x = 0, ..., p-1, p^+. \tag{3.32}$$

Las muertes del año t se promedian para cada par adyacente de edades, de esta forma se obtienen las muertes de cada cohorte y luego la población ajustada R_{t+1},

$$D_t(x, x+1) = \left[D_t(x) + D_t(x+1)\right]/2 \quad x = 0, ..., p-2, \tag{3.33}$$

$$D_t(p-1, p^+) = \left[D_t(p-1)/2 + D_t(p^+)\right], \tag{3.34}$$

y

$$R_{t+1}(x+1) = R_t(x) - D_t(x, x+1) \quad x = 0, ..., p-2, \tag{3.35}$$

$$R_{t+1}(p^+) = R_t(p-1) + R_t(p^+) - D_t(p-1, p^+). \tag{3.36}$$

Por otro lado, el número de nacimientos se obtiene utilizando la tasa de fecundidad simulada $f_t(x)$ y la población ajustada por la mitad de la migración neta. Para los nacimientos se supone una distribución Poisson,

$$B_t(x) \sim Poisson(f_t(x)\left[R_t^F(x) + R_{t+1}^F(x)\right]/2) \quad x = 10, ..., 54. \tag{3.37}$$

Una muestra aleatoria de esta distribución determina $B_t(x)$ y estos valores se suman para obtener el valor simulado de B_t. Los nacimientos de varones (y por lo tanto de mujeres) se asignan a través de una binomial con parámetro ρ (relación de masculinidad al nacimiento)

$$B_t = \sum_{x=10}^{54} B_t(x), \tag{3.38}$$

$$B_t^M \sim Bi(B_t, \frac{\rho}{\rho+1}), \quad B_t^F = B_t - B_t^M, \tag{3.39}$$

$$R_t(B) = B_t + G_t(B,0)/2. \tag{3.40}$$

La mortalidad se aplica a los nacimientos a partir de la relación de sobrevivencia de la tabla de vida,

$$\bar{D}_t(B,0) = (1 - r_t(B))R_t(B), \tag{3.41}$$

se ajusta la población por las muertes obtenidas

$$\bar{R}_{t+1}(0) = R_t(B) - \bar{D}_t(B,0), \tag{3.42}$$

luego se obtiene la población estimada a mitad de año como

$$\bar{E}(0) = \left[R_t(0) + \bar{R}_{t+1}(0)\right]/2, \tag{3.43}$$

y siendo $\bar{D}(0) = m_t(0)\bar{E}_t(0)$ se define la distribución Poisson de la que provienen las muertes

$$D_t(0) \sim Poisson(\bar{D}_t(0)). \tag{3.44}$$

El número de muertes generado aleatoriamente $D_t(0)$ se divide en muertes de nacimientos ocurridos en el año t, $D_t(B,0)$ y muertes a la edad 0 ocurridas a nacimientos del año $t-1$, utilizando un factor de separación estimado como el cociente de las muertes esperadas entre el nacimiento y la edad 0 y las muertes esperadas a la edad 0 en el año t,

$$f_0 = \bar{D}_t(B,0)/\bar{D}_t(0), \tag{3.45}$$

$$D_t(B,0) = f_0 D_t(0). \tag{3.46}$$

Este factor de separación se utiliza también para obtener las muertes de la cohorte de la edad 0 a la edad 1 como el promedio ponderado por los factores de separación (f_0 y 0.5) de $D_t(0)$ y $D_t(1)$.

$$D_t(0,1) = (1 - f_0)D_t(0) + D_t(1)/2. \tag{3.47}$$

Luego, utilizando los nacimientos simulados, las muertes y los migrantes, se genera la población para el próximo año y se repite para los años $t = n + 1, ..., n + h$.

$$R_{t+1}(0) = R_t(B) - D_t(B,0), \tag{3.48}$$

$$R_{t+1}(1) = R_t(0) - D_t(0,1), \tag{3.49}$$

$$P_{t+1}(0) = R_{t+1}(0) + G_t(B,0)/2, \tag{3.50}$$

$$P_{t+1}(x+1) = R_{t+1}(x+1) + G_t(x,x+1)/2 \quad x = 0, ...p-1, p^+. \tag{3.51}$$

El procedimiento completo se repite para cada trayectoria de población. Se obtiene de este modo un gran número de trayectorias muestrales de poblaciones por edad y sexo y eventos vitales que se pueden utilizar para estimar, con incertidumbre controlada, cualquier variable demográfica que se pueda deducir de estos datos, tales como las esperanza de vida, las tasas totales de fecundidad y las relaciones de dependencia.

Capítulo 4

Los Datos

Los datos, según el criterio de producción, pueden clasificarse en primarios o secundarios. En este trabajo los datos empleados como insumo para los modelos estadísticos son secundarios, dado que la fuente de los mismos esta disponible para múltiples usos, es preexistente a la investigación y no es el mismo investigador quien genera las cifras.

Los modelos para datos funcionales se aplican a tasas de mortalidad y fecundidad, para ello los datos requeridos son las defunciones por edad simple y género, los nacimientos según edad de la madre y la población por edad simple y género. Es importante destacar que durante la estimación es necesario contar con las cantidades netas tanto de eventos como de población y por ello no es la tasa en sí el dato base, sino que el mismo se construye a partir de los registros de nacimientos, defunciones y las cifras de población por edad.

En relación a los datos provenientes del registro de hechos vitales, los mismos son suministrados por la Dirección de Estadísticas e Información de Salud (DEIS), a través de bases de datos que contienen el registro de los hechos individuales que permiten generar la matriz por edades para un período de tiempo de aproximadamente 30 años. En cambio en relación a las cifras de población se genera una dificultad, no existe una única población oficial por edades simples para los últimos treinta años. En este punto hay

que destacar que la finalidad de este trabajo es presentar una propuesta metodológica adaptada a los datos de Argentina, basándose en la información disponible a nivel oficial, permitiendo que los modelos aquí estimados sean replicados por cualquier usuario interesado a partir de datos oficiales.

El Ministerio de Salud en sus Informes de Estadísticas Vitales publica en forma regular (aproximadamente anual) la población que el INDEC le proporciona y que constituye el denominador de las tasas que el informe contiene. Pero la construcción de estos denominadores es variable a través de los años, y en muchos casos no se informa el modo en el que han sido obtenidas o si se trata de proyecciones o estimaciones. En este punto no se pretende constituir una crítica sino un análisis de las cifras disponibles, su procedencia, su metodología de obtención y su coherencia a través de los años. Además surge otra dificultad, la constitución de los grupos etarios fluctúa a través de los años haciendo necesario implementar algún tipo de coeficiente para obtener la desagregación buscada, más aún, en ningún caso se informan números para edades simples.

A continuación se realiza un breve repaso sobre las cifras de población suministradas por el INDEC al DEIS, comenzando a partir del año 1980. El anuario correspondiente a 1980 y 1981 (conjuntamente) se publicó en 1984, las cifras de población del año 1980 son suministradas por el INDEC y provienen del Censo Nacional de Población y Vivienda de 1980, mientras que para el año 1981 la fuente indicada es el documento Estimaciones y Proyecciones 1950-2025 y estimaciones inéditas. Dicho documento, contiene proyecciones para el período 1980-2025, que son actualizaciones de las proyecciones realizadas a partir del Censo de 1970, en base a las cifras de población obtenidas en el Censo de 1980. Los grupos etarios disponibles son 1 a 4 años, 5 a 14 años, 15 a 49 y 50 años y más. En el año 1986 se publica el anuario que presenta las estadísticas vitales relativas al año 1982, en él se observa un cambio en el grupo abierto final, se cuenta con los grupos de 50 a 64 años y el grupo final es 65 años y más. La fuente declarada en esta publicación coincide con la del año 1982. También en el mismo año se publica el

anuario correspondiente a 1983, con iguales características en relación a grupos etarios y fuente, de igual modo siguen las publicaciones correspondientes a 1984, 1885 y 1986, publicados en 1988 los dos primeros y en 1989 el último mencionado. En el año 1987 la fuente relativa a las cifras de población por edad y género es el INDEC (basados en estimaciones inéditas) y continúan las mismas características para las publicaciones correspondientes a los años 1988, 1989 y 1990, difundidos en 1991 los dos primeros y en 1992 el último. En el año 1993 se publica el anuario correspondiente a la información del año 1991, en dicha publicación las cifras de población se basan en el Censo Nacional de Población y Vivienda de 1991. En el año 1994 se publica el anuario del año 1992, es importante destacar en este punto que la información no tiene regularidad en la publicación y carece del nivel de oportunidad que se observa en la actualidad (la oportunidad es característica relativa a las estadísticas que se define como la disponibilidad de los datos en el momento en que son requeridos). En el anuario de 1992 los grupos etarios continúan presentándose con la configuración del volumen anterior configuración y en la fuente consta "Los datos de población fueron estimados por el Instituto Nacional de Estadísticas y Censos (INDEC), en base al Censo Nacional de Población y Vivienda de 1991." A partir del anuario de 1993 se normaliza la periodicidad de las publicaciones, los anuarios están disponibles el año posterior al de los datos que contienen, además se incorpora un anexo con un cuadro exclusivo para informar la población por provincias (con anterioridad se presentaba junto a las tasas de mortalidad). Por otro lado, se presenta una nueva configuración de los grupos etarios de 0 a 4 años, 5 a 9 etc., es decir grupos quinquenales hasta el grupo abierto final de 80 y más. En este caso la fuente indicada para datos poblacionales es "Dirección de Análisis Demográfico, INDEC. Estimaciones inéditas provisionales", continuando del mismo modo hasta el año 1997. En 1998 la fuente es "Dirección de Estadísticas Poblacionales, INDEC" y prosiguen los anuarios con la misma fuente hasta 2002, recién en 2003 la misma nota se amplía "Los datos de población utilizados en el cálculo de las tasas fueron proporcionados por la Dirección de Estadísticas Poblacionales del Instituto Nacional de Estadística y Censos. Los mismos corresponden a la revisión de las proyecciones de población en base a re-

sultados definitivos del Censo 2001. En consecuencia, algunas jurisdicciones presentan fluctuaciones en las tasas que se deben fundamentalmente al cambio del denominador, así como a posibles variaciones en el volumen de los hechos vitales". Dicha fuente persiste en las publicaciones siguientes hasta el año 2006 cuando la fuente citada pasa a ser "Análisis Demográfico Nº31. INDEC." En este documento publicado por el INDEC se presentan las proyecciones provinciales por sexo y grupos de edad para el período 2001 a 2015, estas se basan en información de los censos 1991 y 2001 y defunciones y nacimientos correspondientes al período intercensal. A partir del año 2007 y hasta el 2010 la fuente informada es "Dirección de Estadísticas Poblacionales. Análisis Demográfico de datos inéditos", de modo que no es posible contar con la metodología empleada para la estimación o proyección informada. Luego para los años 2011 y 2012 se informa como fuente "INDEC (2005). Proyecciones provinciales de población por sexo y grupos de edad 2001-2015. Buenos Aires, Serie Análisis Demográfico Nº31 e información inédita."

En resumen la población informada para el período 1998 a 2012, esta constituida por información censal, estimaciones y proyecciones variando en su configuración etaria reiteradas veces durante el período mencionado.

En relación a la información proporcionada para el mismo período por el Centro Latinoamericano y Caribeño de Demografía (CELADE), los datos presentados en el Boletín Demográfico Nº66 de Julio de 2000 se refieren a las estimaciones y proyecciones para América Latina en conjunto y los países que la conforman, entre ellos Argentina, para el período 1950-2050 (si bien el boletín presenta los datos 1995-2005, la base está disponible para todo el período) y por edades simples. Luego, el boletín Nº71, del año 2003, actualiza estos datos presentando resultados que divergen de los anteriores a partir del año 2025, además consta una fuente "Estimaciones y proyecciones de población: Total del país, 1950-2050 (versión revisada) INDEC/CELADE", documento en el que se detalla que la información más reciente empleada en el proceso es el Censo de 1991 y estadísticas vitales hasta el año 1994. La siguiente publicación del organismo

relativa a proyecciones y estimaciones, el Boletín Demográfico Nº73 del año 2004, no presenta cambios con respecto al boletín anterior en relación a la estimaciones realizadas para la Argentina. Información más actualizada proporcionada por CELADE corresponde al año 2007, sin embargo las cifras se presentan para años múltiplos de 5 y grupos quinquenales. Esta publicación incorpora en sus proyecciones y estimaciones información correspondiente al Censo 2001 y de estadísticas vitales hasta 2004. Finalmente, en el año 2012 se presentan las proyecciones y estimaciones para el período 1980-2030, para grupos quinquenales cada diez años desde 1980 a 2000, y cada cinco años hasta 2030. En esta última publicación las cifras declaran las mismas fuentes que la publicación anterior y agregan la Encuesta Permanente de Hogares (EPH) 1990, 2000 y 2004 e Investigación de la Migración internacional en América Latína (IMILA). Las cifras presentan pequeñas divergencias con respecto a las cifras del boletín anterior, producto de su actualización en base a información más reciente.

Finalmente no existe una estimación oficial de la población en base a los cuatro últimos Censos, menos aún por edades simples. La idea del presente trabajo, en relación a las cifras de población, es utilizar una única fuente oficial que no requiera ajustes y sea coherente (preferentemente elaborada para todos los años mediante el mismo procedimiento o metodología), por ello, se decide utilizar la población por edades simples presentada por CELADE y estimada en forma conjunta con el INDEC. Dichos datos están disponibles para el periodo 1980-2012. Es importante destacar que, lamentablemente, las publicaciones más actualizadas no cuentan con estimaciones para todos los años requeridos y con desagregación por edades simples.

Capítulo 5

Resultados

A continuación se presentan los resultados de la aplicación a datos demográficos de Argentina del período (1980-2012), pero antes resulta interesante situar al país dentro del proceso de transición demográfica. Como sostienen Pantelides y Rofman (1983) el proceso de transición argentino sigue un modelo no ortodoxo, dado que la mortalidad no descendió antes que la fecundidad como sucedió en los países europeos y por otro lado las tasas de natalidad y mortalidad siguieron una trayectoria bastante paralela. En rasgos generales la mortalidad disminuye fuertemente entre 1895 y 1930, la esperanza de vida se duplica entre 1880 y 1960 (33 a 66 años), la mortalidad infantil disminuye hasta 1955, luego fluctúa y aumenta en 1970 y finalmente la fecundidad disminuye fuertemente entre 1895-1914 (7 a 5,3 hijos por mujer), más lento hasta 1970 (2,9 hijos por mujer) y aumenta hacia 1980. En las secciones siguientes se modela el comportamiento de las componentes demográficas en el período 1980 a 2012 y luego se pronostica el mismo para un período de 10 años.

5.1. Mortalidad

La información empleada para construir las tasas específicas por edad simple y sexo, para el período 1980 a 2012, proviene de la Dirección de Estadística e Información de Salud, fuente del número neto de defunciones y del Centro Latinoamericano y Caribeño de

Demografía (CELADE) que a través de sus publicaciones periódicas proporciona cifras de población con la desagregación adecuada para el presente tipo de análisis. Un análisis más detallado, relativo a las cifras de población y su proveniencia, se desarrolló en el capitulo 4.

Para la representación gráfica de los datos se emplea la escala arco iris propuesta y desarrollada por Rob Hyndman, si bien en la mayoría de sus trabajos no presenta una leyenda a modo de orientación, en el presente trabajo se incluyen tres años, el primero, el último y uno intermedio, a fines de ordenar la escala visualmente y permitir una mayor comprensión de las Figuras.

Para visualizar la estructura de los datos sobre las que se va a aplicar el modelo se presentan gráficamente los logaritmos de las tasas observadas para el total de la población (Figura 5.1a), en ella, las curvas para cada año muestran la forma típica del patrón de la mortalidad, alta del inicio de la vida, luego un descenso hasta antes de los 10 años seguido de un aumento hasta su pico alrededor de los 20 años, fenómeno presente principalmente en varones (ver Figura 5.1b) y en mucho menor medida en las mujeres (ver Figura 5.1c). Este valor alto esta vinculado, en ésta y en la mayoría de las poblaciones, a accidentes de tránsito, muertes relacionadas al consumo de drogas y muertes violentas en general. Según la Organización Mundial de la Salud (OMS) la adolescencia es el período de la vida en el cual el individuo adquiere su madurez reproductiva, transita los patrones psicológicos de la niñez a la adultez y esteblace su independencia socio-económica, comprendiendo varones y mujeres cuyas edades están entre los 10 y 24 años. Serfaty *et al.* (2007) sostiene que en Argentina este grupo asciende a un 27 % de la población y que en los últimos tiempos se ha profundizado el conocimiento que se tiene acerca de ellos; si bien no se enferman clínicamente con frecuencia, son más vulnerables a las causas de mortalidad vinculadas a la violencia: los accidentes, el suicidio y el homicidio. Luego de este pico las tasas presentan un leve descenso para volver a subir, de forma sostenida, hasta las edades avanzadas, en este punto se observa un

pico final que no se debe a la naturaleza de los datos observados sino a la agrupación forzada en mayores de 80 años, grupo que contiene edades, las cuales sería relevante que fueran analizadas en forma simple, ya que sobre este grupo recae un gran interés debido al fenómeno del envejecimiento poblacional. Sin embargo, como argumento a favor de la utilización de este intervalo, es importante tener en cuenta que el grupo etario de 80 años y más presenta serias inconsistencias en la declaración de su edad durante los procedimientos censales (Del Popolo, 2000). En su trabajo para CELADE, la autora sostiene que en base al análisis de las cifras surgen dos posturas: que los datos reflejan la realidad y la mortalidad en ancianos de América Latina sería menor a la esperada de acuerdo a su situación demográfica comparada con la de los países desarrollados y una segunda hipótesis que sostiene que existen errores en los datos básicos, específicamente una tendencia a declarar en forma exagerada la edad, y que este comportamiento se acentúa con el avance de ella. Si esta última hipótesis es válida, la desagregación del grupo de 80 años y más puede distorsionar el nivel de la mortalidad, mientras que trabajar con un grupo abierto final evitaría el problema.

Un aspecto que se destaca visualmente es la caída marcada en los niveles de mortalidad a través del tiempo, que se da en todas las edades, excepto para el pico de los 20 años, franja en la que parece observarse una caída más leve o fluctuante. La caída en general, en los niveles de la mortalidad se atribuye principalmente a las mejoras en la medicina cuando se evalúan dinámicas de largo plazo (períodos de estudio de 100 años o más), por lo cuál el descenso que se observa en el presente período de tiempo puede obedecer en parte a estas causas y a otras más específicas del proceso histórico particular de la Argentina.

Los logaritmos de las tasas observadas se convierten en datos funcionales al estimar curvas suavizadas a partir de las mismas (Figuras 5.2a, 5.2b y 5.2c). Para ello se aplica una regresión *spline* penalizada, dado que dicha metodología permite incorporar una restricción de monotonía: se supone que la función suavizada es monótona para

alguna edad $x > c$. En esta aplicación se especifica $c = 65$ años, en primer lugar, porque es la edad usada por Hyndman y Ullah (2007) y en segundo lugar, porque es la edad jubilatoria para hombres que marca la legislación Argentina. Esta restricción permite disminuir la alta variabilidad en las curvas estimadas para edades avanzadas y resulta razonable suponer, en este contexto, que para edades avanzadas, a mayor edad, mayor la probabilidad de muerte. Para más detalle acerca de la metodología empleada para el suavizado se puede consultar el Apéndice 6.

Los datos suavizados se descomponen mediante el uso de análisis de componentes principales funcionales, dicha metodología se presentó en la Subsección 3.2.1.1. En relación al número de bases funcionales se establece $K = 6$, por lo que se obtienen seis bases y sus correspondientes coeficientes asociados (es importante destacar que gráficamente se presentan siempre los primeros tres pares base-coeficiente por motivos de practicidad e interpretabilidad). Hyndman sostiene que es difícil hallar una explicación para los coeficientes y bases funcionales más allá del segundo o tercer par base-coeficiente. Las Figuras 5.3, 5.4 y 5.5 representan los resultados para el total de la población, los varones y las mujeres, respectivamente. La Figura 5.3 presenta la media, las bases, los coeficientes funcionales estimados y el pronóstico de los coeficientes con sus intervalos de pronóstico del 80 % de confianza para el total, mientras que las Figuras 5.4 y 5.5 contienen los mismos resultados para varones y mujeres respectivamente. Estas Figuras deben interpretarse de la siguiente manera: el primer recuadro de la fila superior representa el comportamiento promedio de la mortalidad a través de las edades. A partir de la segunda columna, la fila superior debe interpretarse en forma conjunta con el correspondiente recuadro de la fila inferior, es decir, cada base debe interpretarse en correspondencia con su coeficiente asociado. Por ejemplo, en el caso de la mortalidad total, el coeficiente muestra un decrecimiento de la mortalidad a través del tiempo, pero para interpretarlo correctamente este comportamiento debe ponerse en correspondencia

[1]Con el fin de orientar la escala de colores empleada (arco iris), en los gráficos subsiguientes se incluyen como referencia los colores relativos al primer año, el año intermedio y el último del período contemplado

con su base asociada (recuadro de la fila superior), ya que la misma indica en qué edades se manifiesta este descenso y con qué intensidad. En este caso la base funcional indica que el decrecimiento se da especialmente en los primeros años de vida (mortalidad infantil y primeros años)y en menor medida para los mayores de 40 años. Sin embargo ese decaimiento no es tan notorio para personas de alrededor de 20 años, este hecho se refleja en la base ya que la misma adquiere un valor cercano a cero para la mencionada edad. Como se dijo anteriormente, generalmente, no es posible hallar una explicación lógica con un correlato coherente con los hechos demográficos, más allá del segundo o tercer par de coeficientes y bases. Por otro lado, la media estimada ($\hat{\mu}$) representa el perfil promedio de la mortalidad a lo largo de la vida (obtenido como el promedio de las funciones a través de los años)

En primer lugar se analizan los resultados obtenidos para el total de la población, principalmente se destaca que en el comportamiento de la mortalidad total se solapan las tendencias de ambos géneros. Las seis primeras bases explican el 91.6 %, 3.8 %, 1.6 %, 0.7 %, 0.6 % y 0.4 % de la variación en los datos, dejando un 1.3 % sin explicar. La primera base presenta una estructura similar a la de los varones, pero más suave, mientras que la segunda base presenta un comportamiento similar a la misma base en mujeres. Por último la tercera base tiene un comportamiento similar a la tercera base estimada para los varones. En otras palabras, el comportamiento de la mortalidad total está dominado en la primer y tercer base por el comportamiento que presenta la mortalidad en los varones, mientras que la segunda base estaría dominada por la pauta que sigue la mortalidad de las mujeres.

Resulta más interesante analizar los resultados para cada género, en el caso de los varones, las seis primeras bases explican el 90.3 %, 5.0 %, 1.4 %, 0.8 %, 0.7 % y 0.5 % de la variación presente en los datos, dejando un 1.3 % sin explicar. El coeficiente correspondiente a la primer base representa una mortalidad que desciende a través del tiempo, con un pico en al año 1982 que podría atribuirse al conflicto bélico por la soberanía

de las Islas Malvinas en abril de ese año. La primera base funcional estimada indica de qué modo este descenso presente en el coeficiente se manifiesta para las distintas edades; en otras palabras indica que grupos etarios sufren en mayor o en menor medida el descenso a través del tiempo, es decir, el pico en las edades iniciales representa el descenso en la mortalidad infantil, seguido de valores altos para la niñez y las edades entre 40 y 50 años, que presentan también descensos, pero con menor magnitud. El descenso para la franja etaria 40-50 está vinculado a la caída en la mortalidad a nivel mundial atribuible en su mayoría a la merma en la muertes por afecciones cardíacas u otras patologías debido al avance de la medicina. El coeficiente asociado a la segunda base presenta períodos de subas moderadas alrededor de 1980 y 2010 y una suba marcada a mediados de la década de los 90, bajas menores se presentan en los restantes períodos de tiempo. La base que corresponde a este coeficiente indica que este comportamiento se manifiesta principalmente en la franja de entre 18 y 30 años aproximadamente. Por último, el tercer par base-coeficiente podría estar asociado a un diferencial en la mortalidad de varones de 10 a 20 años con respecto al resto de las edades y, su base asociada indicaría que la mortalidad para dicho grupo en relación a las demás edades fue baja en 1980, luego presentó una suba con fluctuaciones hasta alcanzar su valor máximo alrededor del año 1990, para luego comenzar a descender y alcanzar valores mínimos (relativos o locales) entre 2005 y 2010. En síntesis, para los hombres, las edades que dominan el cambio en la mortalidad resultan, la franja de edades cercanas a cero, el grupo de 40 a 50 años, el grupo de 18 a 30 años y el grupo formado por las edades en torno a los 15 años. La forma en la que ejercen el cambio esta dada por las bases, un descenso para los primeros tres grupos y fluctuaciones a lo largo del período para los restantes.

Si se analizan los resultados relativos a las mujeres, las bases explican el 92.2 %, 2.2 %, 1.8 %, 0.9 %, 0.6 % y 0.5 % de la variación en los datos respectivamente, dejando un 1.8 % sin explicar. Se destaca, que la franja etaria de 0 a 5 años presenta el valor más alto observado para la primera base estimada, esto indica que se trata de las edades que determinan en mayor medida el decrecimiento que presenta el coeficiente asociado a la

primera base. Interpretado de otro modo, es el grupo que presentó el mayor descenso en la mortalidad durante el período 1980-2010. En segundo lugar se encuentran los grupos de mujeres de alrededor de 30 y 75 años, que también presentan un descenso en el período en estudio, pero, con una menor magnitud o intensidad. Luego, el coeficiente asociado a la primera base claramente representa el descenso en la mortalidad de mujeres sucedido entre 1980 y 2010. Además, se destaca un leve aumento en la fluctuación de dicho coeficiente en período 2005-2010 y si bien el fenómeno podría ser producto del azar, es posible plantear algún tipo de hipótesis en relación a sus causas. El segundo par base-coeficiente se referiría a un comportamiento diferencial entre las mujeres de 15 a 20 años en relación al resto de las edades. Para este par, el coeficiente presenta una gran fluctuación a través de los años. Evaluando base y coeficiente en forma conjunta se puede interpretar que la mortalidad de mujeres de 15 a 20 años es la que menos presenta fluctuación. Por último, la tercer base, que es generalmente a partir de la cual resulta más difícil hallar una interpretación, podría estar representando la diferencia en la mortalidad para edades previas versus edades posteriores a los veinte años y su base asociada indicaría un comportamiento fluctuante durante el período en estudio, con valores altos al inicio del periodo y sobre el 2005, y valores bajos a principios de los 90 y a finales del período. Sin embargo, es importante tener mayor cautela al tratar de definir el comportamiento que modela el tercer par coeficiente-base. En síntesis los grupos que dominan los cambios en la mortalidad de mujeres desde 1980 a 2010 son las niñas de 0 a 5 años, mujeres de alrededor de 30 años y las adultas mayores de alrededor de 75 años de edad, en segundo lugar las mujeres de alrededor de 20 años y por ultimo las levemente mayores y menores a los 20 años. La forma en que estos grupos dominan los cambios esta dada por su base asociada, descenso en mayor o menor medida para los tres primeros grupos mencionados y fluctuaciones a lo largo del tiempo para los restantes.

Para evaluar la bondad del ajuste del modelo propuesto se utiliza un gráfico de contorno, el mismo sirve para verificar la independencia de los residuos. Bajo independencia se deberían observar zonas pequeñas y mezcladas de colores blanco, rosa y celeste. En

los residuos observados se presentan grupos o bandas rosas y blancas, este último color indica que las residuos son cercanos a cero, en cambio un color rosado más intenso refleja subestimación. Para las edades cercanas a cero se detectan manchas de color celeste, que son indicio de una leve sobreestimación. En los tres casos analizados, total, varones y mujeres, los residuos son cercanos a cero y en los gráficos de los mismos se alternan valores positivos en su mayoría y negativos en edades cercanas al cero. Bajo el supuesto de independencia la alternancia debería ser más marcada y la presencia de colores más equilibrada. Ver Apéndice B, Figuras 1, 2 y 3.

Una vez estimado el modelo se pronostican los coeficientes $\hat{\beta}$ utilizando modelos ARIMA, los modelos estimados se presentan en el Apéndice B. La selección de los modelos es automática, la misma se lleva cabo mediante la función de R^2 basada en una variación del algoritmo de Hyndman y Khandakar (Hyndman y Khandakar, 2008), que combina tests de raíces unitarias, minimización del AIC corregido y estimación máximo verosimil para obtener el mejor modelo ARIMA. Si bien en múltiples aplicaciones se plantea el uso de modelos de espacio de estados como método de pronóstico, Hyndman sugiere que los modelos ARIMA se adaptan mejor a las características de este tipo de datos. Los pronósticos de los coeficientes pueden verse en las mismas Figuras que presentan la descomposición, como una línea que se extiende más allá del año 2012, rodeada del sombreado color gris que representa el intervalo de pronóstico del 80 % de confianza.

Los coeficientes pronosticados $\hat{\beta}$ permiten construir las tasas a través de la ecuación (3.12) y obtener las variancias de pronóstico ajustadas mediante la ecuación (3.16) a fin de contar con intervalos de confianza para las tasas. En la Figura 5.6 se presentan las tasas observadas en gris y los pronósticos para el período 2013-2023 en colores según la escala arco-iris. Se destaca un comportamiento diferente entre géneros principalmente alrededor de los 20 años. En los varones no se detecta para la mencionada edad el mismo

[2]R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

descenso que se observa para la mayoría de las edades, es decir, que para los varones de entre 15 y 22 años la tasa de mortalidad pronosticada se mantiene en el mismo nivel aún cuando por lo general se presenta un descenso de la mortalidad. Este resultado es coherente con el valle que se observa en la Figura 5.4 alrededor de los 20 años, el mismo se refería justamente a que el descenso general observado en la mortalidad no se daba particularmente para el mencionado grupo etario. En el caso de las mujeres el descenso más marcado de presenta entre los 15 y 45 años y el menor descenso entre los 10 y 15 años y alrededor de los 60 años. A modo de resumen se presentan los pronósticos e intervalos de pronóstico (IP) del 80 % de confianza para edades y años seleccionados para total, varones y mujeres, Cuadros 5.1, 5.2 y 5.3 respectivamente.

Cuadro 5.1: Pronósticos de las tasas de mortalidad, por cada mil personas, e intervalos de pronóstico del 80 % de confianza, para edades y años seleccionados. Total, Argentina

Año	2013			2023		
Edad	Pronóstico	Lim. Inf.	Lim. Sup.	Pronóstico	Lim. Inf.	Lim. Sup.
0	11,245	9,914	12,755	7,893	6,758	9,219
5	0,226	0,200	0,254	0,187	0,164	0,213
10	0,200	0,176	0,227	0,165	0,144	0,190
15	0,526	0,482	0,573	0,470	0,428	0,516
20	1,041	0,961	1,128	0,960	0,871	1,059
25	1,023	0,942	1,111	0,915	0,822	1,018
30	1,060	0,976	1,151	0,919	0,826	1,023
35	1,226	1,142	1,318	1,038	0,956	1,127
40	1,761	1,637	1,895	1,444	1,328	1,571
45	2,711	2,537	2,898	2,268	2,099	2,451
50	4,324	4,082	4,579	3,731	3,490	3,988
55	7,108	6,781	7,451	6,359	6,018	6,720
60	11,161	10,720	11,620	10,157	9,682	10,656
65	16,350	15,734	16,991	14,873	14,242	15,532
70	25,727	24,556	26,954	23,090	21,921	24,321
75	32,188	30,454	34,021	28,871	26,922	30,960
80 y más	124,325	119,083	129,798	117,425	111,777	123,359

Fuente: Elaboración propia en base a datos de la DEIS-INDEC.

Para contar con un resultado comparativo en relación a la bondad de pronóstico fuera de la muestra de los resultados obtenidos mediante Modelos para Datos Funcionales, se calculan los pronósticos mediante modelos ARIMA univariados para cada edad. Tanto para modelos ARIMA como el enfoque funcional se ajustan los modelos dejando los cuatro últimos valores (2009-2012) para evaluar la calidad de los mismos. Los pronósticos MDF evaluados a través del MAPE (sigla en inglés para Error Porcentual Absoluto Medio) son levemente superiores a los obtenidos por modelos ARIMA, especialmente

Cuadro 5.2: Pronósticos de las tasas de mortalidad, por cada mil personas, e intervalos de pronóstico del 80 % de confianza. Para dades y años seleccionados. Varones, Argentina

Año	2013			2023		
Edad	Pronóstico	Lim. Inf.	Lim. Sup.	Pronóstico	Lim. Inf.	Lim. Sup.
0	12,226	10,758	13,894	8,405	7,115	9,930
5	0,249	0,216	0,288	0,202	0,172	0,238
10	0,214	0,182	0,250	0,179	0,150	0,212
15	0,670	0,604	0,742	0,623	0,561	0,693
20	1,561	1,428	1,706	1,547	1,396	1,714
25	1,503	1,373	1,645	1,398	1,244	1,572
30	1,496	1,367	1,638	1,314	1,165	1,481
35	1,613	1,490	1,747	1,377	1,254	1,512
40	2,185	2,017	2,366	1,779	1,616	1,958
45	3,340	3,098	3,602	2,705	2,463	2,972
50	5,527	5,170	5,908	4,639	4,270	5,040
55	9,499	8,993	10,034	8,283	7,745	8,858
60	15,219	14,531	15,941	13,707	12,961	14,495
65	22,903	21,934	23,915	20,858	19,834	21,936
70	35,853	34,184	37,605	32,740	31,010	34,566
75	47,158	44,731	49,716	43,278	40,575	46,161
80 y más	146,279	139,991	152,850	140,511	134,111	147,216

Fuente: Elaboración propia en base a datos de la DEIS-INDEC.

Cuadro 5.3: Pronósticos de las tasas de mortalidad, por cada mil personas, e intervalos de pronóstico del 80 % de confianza, para edades y años seleccionados. Mujeres, Argentina

Año	2013			2023		
Edad	Pronóstico	Lim. Inf.	Lim. Sup.	Pronóstico	Lim. Inf.	Lim. Sup.
0	10,440	9,154	11,908	7,489	6,191	9,059
5	0,208	0,177	0,243	0,173	0,145	0,206
10	0,174	0,146	0,206	0,156	0,129	0,188
15	0,365	0,319	0,417	0,308	0,265	0,359
20	0,507	0,447	0,574	0,398	0,344	0,461
25	0,521	0,461	0,589	0,417	0,360	0,482
30	0,629	0,562	0,705	0,510	0,445	0,584
35	0,866	0,781	0,960	0,695	0,613	0,788
40	1,340	1,222	1,470	1,106	0,987	1,239
45	2,093	1,939	2,259	1,840	1,677	2,018
50	3,180	2,988	3,384	2,893	2,687	3,114
55	4,930	4,654	5,223	4,605	4,305	4,927
60	7,552	7,166	7,960	6,995	6,573	7,444
65	10,932	10,424	11,465	9,828	9,277	10,412
70	17,725	16,753	18,752	15,733	14,663	16,880
75	22,663	21,121	24,319	19,905	18,040	21,963
80 y más	113,468	108,284	118,899	107,790	101,699	114,245

Fuente: Elaboración propia en base a datos de la DEIS-INDEC.

para el total y los varones en los dos últimos años evaluados (2011 y 2012), resultando similares o algo inferiores en el caso de las mujeres en años iniciales (2009 y 2010), Cuadro 5.4. Por lo general, los pronósticos demográficos y particularmente los referidos a mortalidad, son requeridos a largo plazo. La comprobación empírica de la calidad de los pronósticos con horizontes considerablemente extensos no es posible en esta aplicación debido a que se cuenta con información para un período de tiempo relativamente corto, pero, es de esperar que los pronósticos obtenidos por MDF se comporten mejor a largo plazo dado que la metodología reduce la variabilidad mediante el análisis funcional, esto se ve reflejado en la amplitud de los intervalos de pronóstico, siendo estos, en promedio, un 47,9 % más estrechos que los obtenidos mediante modelos ARIMA, con una variación entre el 24,7 y 79,9 %. Los intervalos MDF además de tener una menor amplitud (Ver Cuadro 5.5), en la mayoría de las categorías (sexo-edad), cubren el valor observado de la tasa.

Cuadro 5.4: Error medio absoluto porcentual (MAPE) de los pronósticos ARIMA y MDF de la mortalidad (Argentina, 2009-2012)

	MAPE											
	Total				Varones				Mujeres			
ARIMA	2009	2010	2011	2012	2009	2010	2011	2012	2009	2010	2011	2012
	4.3	5.62	6.07	9.05	5.37	7.77	7.84	10.47	7.48	7.07	7.87	9.49
		2009-2010	2009-2011	2009-2012		2009-2010	2009-2011	2009-2012		2009-2010	2009-2011	2009-2012
		4.96	5.33	8.12		6.57	6.99	7.86		7.28	7.47	7.98
	Total				Varones				Mujeres			
MDF	2009	2010	2011	2012	2009	2010	2011	2012	2009	2010	2011	2012
	5.41	5.70	5.73	8.12	6.01	7.46	6.56	7.89	8.79	7.56	8.83	10.61
		2009-2010	2009-2011	2009-2012		2009-2010	2009-2011	2009-2012		2009-2010	2009-2011	2009-2012
		5.55	5.61	6.24		6.73	6.67	6.98		8.17	8.39	8.94

Fuente: Elaboración propia en base a datos de la DEIS-INDEC.

Cuadro 5.5: Amplitud promedio de los intervalos de pronóstico ARIMA y MDF de la mortalidad (Argentina, 2009-2012).

	Amplitud Promedio IP											
	Total				Varones				Mujeres			
	2009	2010	2011	2012	2009	2010	2011	2012	2009	2010	2011	2012
ARIMA	0.0023	0.0031	0.0037	0.0042	0.0032	0.0040	0.0046	0.0051	0.0019	0.0025	0.0030	0.0033
MDF	0.0016	0.0017	0.0018	0.0008	0.0011	0.0023	0.0024	0.0024	0.0015	0.0015	0.0016	0.0016

Fuente: Elaboración propia en base a datos de la DEIS-INDEC.

5.2. Fecundidad

Se aplica el enfoque de datos funcionales para el análisis de las tasas de fecundidad según edad de la madre en Argentina durante el período 1980 a 2011. Las mismas se construyen en base a datos de nacidos vivos según edad de la madre proporcionados por la Dirección de Estadística e Información de Salud mientras que las cifras de población son suministradas por CELADE. Las tasas se definen como el total de nacidos vivos, durante un año calendario, de acuerdo a la edad de la madre por cada 1000 mujeres residentes en la población de la misma edad al 30 de junio del mismo año. Los datos corresponden a mujeres desde los 10 a los 54 años de edad.

Los datos observados (agrupados en intervalos con fines gráficos) se muestran como series temporales, mediante una línea para cada grupo etario, en la Figura 5.7. Las series representan la cantidad de hijos nacidos vivos por cada 1000 mujeres a través de los años en el período 1980-2011. Se distinguen tres series con valores cercanos a cero, las mismas correspondes a los intervalos de edad 10 a 14, 45 a 49 y 50 a 54, grupos para los que se espera una menor cantidad de nacimientos en relación a los restantes, si bien también es probable que algunos de estos hechos vitales provengan de errores en la etapa de registro. En segundo lugar se destaca, con una fecundidad levemente mayor a la de los grupos antes mencionados, el grupo de mujeres de 40 a 44 años de edad y con una fecundidad media se encuentran los grupos de 15 a 19 y 35 a 39 años. Estos últimos presentan un leve descenso al inicio del período en estudio pero se estabilizan en torno a un nivel constante para la segunda mitad del mismo. Por último, con la fecundidad más alta, se observan tres grupos diferenciados de los demás; 20 a 24, 25 a 29 y 30 a 34 años. La tendencia de estos tres grupos presenta hacia el año 1980 una tasa de 175 hijos cada mil mujeres y luego desciende a través del periodo en estudio llegando en al año 2011 a 100 hijos cada mil mujeres.

Otra forma de visualizar la información es elaborando un gráfico para las curvas co-

rrespondientes a cada año calendario a través de las edades, el mismo permite observar la estructura de la fecundidad según la edad de la madre y su cambio en los últimos 30 años (Figura 5.8). La Figura presenta un conjunto de curvas con valores máximos entre los 20 y los 30 años y que se va amesetando debido al descenso en la fecundidad a través de los años. Es importante destacar que dado que la serie analizada comienza en al año 1980 no es posible contemplar el importante cambio en esta componente demográfica causado por la aparición de la píldora contraceptiva. A pesar de ello, es posible que el descenso ininterrumpido que se observa en el presente período sea en gran parte producto de este profundo cambio cultural. Además se detecta un leve aumento en las tasas para mujeres de alrededor de los 40 años y presenta un corrimiento de las curvas hacia edades más avanzadas, este comportamiento refleja el retraso en la edad de la maternidad y el cambio en el rol de la mujer.

Estos últimos datos se convierten en datos funcionales estimando una curva suave para las tasas a través del tiempo. Las curvas suavizadas se estiman utilizando B-*splines* con restricción de concavidad mientras que σ_t^2 se estima a partir de $[f_t(x)]^{-1} E_t(x)^{-1}$. Como se mencionó Hyndman y Ullah (2007) imputan, previo al suavizado, los datos para las edades más extremas (le asigna 5 nacimientos por mil mujeres a edades menores de 15 y mayores de 49), sin embargo en la presente aplicación se decide respetar los datos oficiales. De este modo los resultados posiblemente reflejen no sólo tendencias demográficas sino también la tendencia en errores o subregistro, permitiendo un análisis de este aspecto.

El ajuste del modelo para datos funcionales aplicado a la fecundidad se muestra en las Figuras 5.10 y 5.11. Si bien bases y coeficientes de ambas Figuras son idénticos, las mismas se diferencian en los pronósticos de los coeficientes beta, es decir la etapa 4 del modelo, dado que, en este punto, se plantean dos hipótesis relativas a la fecundidad y su comportamiento en el futuro (se incluyen dos supuestos: una primera hipótesis que sostiene que la fecundidad continuará descendiendo y una segunda hipótesis que plantea que la fecundidad ya ha alcanzado su máximo descenso, por ello se seleccionan

dos modelos: uno con pendiente y otro sin pendiente). Las seis primeras bases estimadas explican el 92.0 %, 5.9 %, 0.9 %, 0.5 %, 0.3 % y 0.1 % de la variación en los datos respectivamente, dejando un 0.3 % sin explicar. Sin embargo resulta complejo encontrar un comportamiento lógico más allá de la segunda componente principal.

En las Figuras 5.10 y 5.11 se presentan las bases estimadas $\hat{\phi}_k(x)$ y sus coeficientes asociados $\hat{\beta}_{t,k}$. En ambos casos las Figuras deben interpretarse de la siguiente manera: el primer recuadro de la fila superior representa el comportamiento promedio de la fecundidad a través de las edades. A partir de la segunda columna, la fila superior debe interpretarse en forma conjunta con el correspondiente recuadro de la fila inferior, es decir, cada base debe interpretarse en correspondencia con su coeficiente asociado. Por ejemplo, el primer coeficiente presenta un decrecimiento en la fecundidad a través del tiempo que, al ponerlo en correspondencia con su base asociada (fila superior), indica que el comportamiento de ese decrecimiento se da particularmente en mujeres de aproximadamente 10 años de edad, para las mayores de 40 años y en forma más leve en las mujeres de alrededor de 25 años. Como el coeficiente marca un descenso sostenido es posible pensar que lo que principalmente se observa es un descenso en los registros con error (reflejados en la base mediante valores altos en los extremos) mientras que el pico que se destaca en el centro es realmente el descenso en las tasas de fecundidad que se da en las edades intermedias, el grupo de mujeres de entre 20 y 30 años. Además en los coeficientes de la primera base se destaca un valor alto en el año 1983, este valor posiblemente no se deba a errores de registro sino a un cambio social asociado con el retorno de un régimen de gobierno democrático en la República Argentina luego de varios años de dictadura militar.

En relación al segundo par coeficiente-base, la misma puede pensarse como la diferencia entre la fecundidad adolescente versus edades avanzadas. La base presenta valores bajos para edades próximas a los 10 años, valores relativamente altos para mujeres mayores de 40 años y valores cercanos a cero para las edades restantes, es por ello, que se la

puede asociar a un comportamiento diferente entre las edades que presentan los valores bajos y altos. Para dicha base el comportamiento a través del tiempo está determinado por el coeficiente asociado, el cuál presenta valores altos para los años 1982 y 1983, un descenso hacia el año 2000 seguido de un aumento hasta 2010, año en el que se observa una caída abrupta. Sin embargo es importante remarcar que poner en correspondencia las bases y coeficientes con comportamientos demográficos derivados de la teoría resulta particularmente complejo en el caso de la fecundidad, más allá de la primera o segunda base.

Finalmente, el tercer par base-coeficiente podría vincularse con el traslado de la edad de la maternidad hacia edades más avanzadas, de acuerdo a lo que indica la base la cuál presenta valores altos alrededor de los 40 años y el coeficiente que presenta una alta fluctuación con valores mas bajos al inicio del período y más altos hacia el final.

En resumen, la descomposición en bases y coeficientes permite identificar el patrón de descenso y las edades que lo determinan, el diferencial en la fecundidad adolescente versus edades avanzadas y el corrimiento de la maternidad a edades alrededor de los 40 años y su estabilización en torno a la misma.

La bondad del ajuste del modelo se evalúa a través de un gráfico de contorno, el cual permite verificar la independencia de los residuos. Si se cumple el supuesto, el gráfico debe mostrar zonas pequeñas y mezcladas de colores blanco, rosa y celeste. A diferencia de lo que se observa en los gráficos de residuos obtenidos para la mortalidad, el comportamiento de los residuos para la fecundidad presenta un mayor agrupamiento, es decir una menor alternancia de colores que la esperada bajo independencia. En las edades más jóvenes y más avanzadas se presentan valores altos de los residuos al inicio del período y entre el 2000 y 2010, (es más notorio y constante el error en edades avanzadas, alrededor de los 50, con subestimación antes de 1995 y sobreestimación después de este año) pero es importante tener en cuenta que estos grupos etarios son los que gene-

ralmente presentan una mayor cantidad de errores de registro. Ver Apéndice B, Figura 4.

Una vez que se estima el modelo para la fecundidad y se evalúa la bondad de su ajuste, se pronostican los coeficientes $\hat{\beta}_k$ mediante modelos ARIMA. La selección automática de modelos se realiza utilizando el algoritmo de Hyndman y Khandakar (Hyndman y Khandakar, 2008) y en esta etapa se incluyen dos supuestos: una primera hipótesis que sostiene que la fecundidad continuará descendiendo y una segunda hipótesis que plantea que la fecundidad ya ha alcanzado su máximo descenso, por ello se seleccionan dos modelos: uno con pendiente y otro sin pendiente, en correspondencia con los supuestos planteados. Los modelos estimados para cada coeficiente pueden verse en el Apéndice B. A partir de la combinación entre los coeficientes pronosticados y las bases estimadas por el modelo, se generan los pronósticos de las curvas de fecundidad.

Una vez estimado el modelo para datos funcionales, los pronósticos de los coeficientes $\hat{\beta}$ obtenidos bajo las dos hipótesis permiten construir los pronósticos de las tasas a través de la ecuación (3.12) y obtener las variancias de pronóstico ajustadas mediante la fórmula (3.13) a fin de contar con intervalos de confianza para las tasas. En la Figura 5.6 se representan las tasas observadas mediante lineas de color gris y los pronósticos para el período 2012-2022 en colores según la escala arco-iris. Los pronósticos de las tasas de fecundidad se ven claramente dominados por el supuesto establecido en relación a su tendencia, en la Figura 5.12a donde no se restringe la estimación a un modelo sin pendiente, el descenso presente en las tasas observadas continúa y se produce un amesetamiento desde los 20 a 35 años, e incluso se detecta una leve caída en torno a los 25 años. En la Figura 5.12 se presenta también una meseta, pero en este caso las tasas en la franja etaria de los 20 a 35 años de edad tienden a ubicarse en torno a un valor promedio a medida que aumenta el horizonte de pronóstico (h).

Además de la representación gráfica de las curvas pronosticadas, se presentan las tasas de fecundidad por edad de la madre (si bien se obtienen los resultados por edades simples desde los 10 a los 54 años, se presentan edades seleccionadas) junto con sus

intervalos de pronóstico del 80 % para los años 2012 y 2022 (primer y último año para el que se calcularon los pronósticos), Cuadros 5.6 y 5.7. Los Cuadros dan cuenta de la cantidad de resultados que provee la metodología empleada y el nivel de desagregación de la información.

Cuadro 5.6: Pronósticos de las tasas de fecundidad, cada mil mujeres, e Intervalos de Pronóstico del 80 % de confianza, para edades y años seleccionados. Hipótesis de fecundidad estable.

Año	2012			2022		
Edad	Pronóstico	Lim. Inf.	Lim. Sup.	Pronóstico	Lim. Inf.	Lim. Sup.
10	0,01	0,00	0,03	0,01	0,00	0,04
15	20,70	18,81	22,79	21,99	19,73	24,50
20	108,55	100,91	116,77	112,74	100,32	126,70
25	115,04	103,34	128,07	118,15	96,83	144,16
30	110,06	102,09	118,65	112,74	98,88	128,54
35	75,08	69,91	80,62	77,75	70,58	85,64
40	30,96	28,14	34,06	32,15	28,45	36,33
45	3,29	2,67	4,06	3,36	2,29	4,93
50	0,22	0,14	0,34	0,22	0,09	0,51
54	0,02	0,01	0,05	0,02	0,01	0,09

Fuente: Elaboración propia en base a datos de la DEIS-INDEC.

Cuadro 5.7: Pronósticos de las tasas de fecundidad, cada mil mujeres, e Intervalos de Pronóstico del 80 % de confianza, para edades y años seleccionados. Hipótesis de fecundidad decreciente.

Año	2012			2022		
Edad	Pronóstico	Lim. Inf.	Lim. Sup.	Pronóstico	Lim. Inf.	Lim. Sup.
10	0,01	0,00	0,02	0,00	0,00	0,01
15	21,07	19,13	23,20	21,90	19,60	24,46
20	108,20	100,67	116,31	103,84	94,34	114,29
25	109,76	98,82	121,91	99,93	85,82	116,36
30	106,24	98,66	114,40	101,29	91,20	112,50
35	74,44	69,35	79,90	73,33	67,40	79,78
40	30,61	27,83	33,66	30,09	26,96	33,59
45	3,14	2,56	3,86	2,45	1,84	3,26
50	0,19	0,13	0,30	0,10	0,05	0,19
54	0,02	0,01	0,04	0,01	0,00	0,02

Fuente: Elaboración propia en base a datos de la DEIS-INDEC.

Se comparan los pronósticos obtenidos mediante MDF versus una metodología de pronóstico clásica y de amplio uso como lo son los modelos ARIMA de Box y Jenkins (Box y Jenkins, 1976). Con este fin se calculan los pronósticos mediante modelos ARIMA univariados sobre las series de tasas observadas para cada edad. Tanto para modelos ARIMA como para el enfoque funcional se ajustan los modelos dejando los cuatro últi-

mos valores (2008-2011) fuera de la muestra para poder medir la bondad de pronóstico fuera de la muestra. Los pronósticos MDF evaluados a través del MAPE son claramente superiores a los obtenidos por modelos ARIMA, en ambos casos, es decir tanto para la hipótesis que contempla la pendiente que marca el descenso de la fecundidad como para la hipótesis que no supone pendiente y por ello considera que la fecundidad ha alcanzado un nivel estable (Cuadros 5.8 y 5.9).

Cuadro 5.8: Error medio absoluto porcentual (MAPE) de los pronósticos ARIMA y FDM de la fecundidad (Argentina, 2008-2011)

	MAPE							
	Sin Pendiente				Con Pendiente			
ARIMA	2008	2009	2010	2011	2008	2009	2010	2011
	27.81	29.10	28.42	42.95	35.75	38.76	40.81	43.25
		2008-2009	2008-2010	2008-2011		2009-2010	2009-2011	2009-2012
		28.46	28.45	32.07		37.26	38.44	39.64
MDF	2008	2009	2010	2011	2008	2009	2010	2011
	16.42	14.23	9.55	12.37	26.35	29.22	32.00	34.27
		2008-2009	2008-2010	2008-2011		2008-2009	2008-200	2008-2011
		15.32	13.40	13.14		27.79	29.19	30.46

Fuente: Elaboración propia en base a datos de la DEIS.

Cuadro 5.9: Amplitud promedio de los intervalos de pronóstico ARIMA y FDM de la fecundidad (Argentina, 2008-2011)

	Amplitud Promedio IP							
	Sin Pendiente				Con Pendiente			
	2008	2009	2010	2011	2008	2009	2010	2011
ARIMA	15.46	18.89	21.14	23.69	15.46	18.89	21.14	23.69
MDF	14.30	15.41	16.92	17.94	13.90	14.60	15.77	16.41

Fuente: Elaboración propia en base a datos de la DEIS.

Al igual que se observó para la mortalidad, los pronósticos de la fecundidad obtenidos por MDF tienen un mejor comportamiento a largo plazo dado que presentan una reducción en la variabilidad mediante el análisis funcional (más leve que para datos de mortalidad), esto se ve reflejado en la amplitud de los intervalos de pronóstico, siendo estos, en promedio, un 17,5 % más estrechos que los obtenidos mediante modelos ARIMA, con una variación entre el 7,49 y 30,73 %. Al evaluar los resultados de la comparación es notorio el descenso en el error medido a través de los MAPE que presentan los pronósticos MDF frente a los ARIMA, siendo más leve la disminución con respecto a la que se observa en la amplitud de los intervalos, Cuadro (5.9). Asimismo si se realiza

un análisis comparativo entre las hipótesis de fecundidad, la hipótesis que supone que la componente demográfica ha alcanzado un nivel estable presenta medidas de error menores en todos los años y períodos contemplados, sin embargo en relación a la amplitud de los intervalos de pronóstico los resultados son similares.

5.3. Migración

Los datos de migración en base a los que se estima el modelo para datos funcionales, a diferencia de aquellos relativos a la mortalidad y fecundidad, se obtienen a partir de la ecuación de crecimiento y datos de población, nacimientos y muertes. Es decir, los mismos no provienen de una fuente oficial en forma directa. Si bien los datos existentes, en materia de migración, para Argentina son escasos, se emplea la metodología propuesta a fin de ilustrar como se aplica a datos concretos, no esperando obtener resultados confiables.

El cálculo se realiza mediante el conjunto de ecuaciones 3.1, 3.2 y 3.3. Por otro lado, si bien teóricamente la ecuación se debe basar en la población correspondiente al 1 de enero, Hyndman y Ullah (2007) utilizan la población al 30 de junio y sostienen que la variación entre los valores correspondientes a ambas fechas no afecta los resultados que se obtienen. En el presente trabajo se aplica la función con una modificación para que la misma admita como entrada las poblaciones al 1 de enero y no al 30 de junio (dado que no existen datos oficiales para la fecha requerida es necesario realizar la correspondiente extrapolación para obtener las cifras al 1 de enero) sin embargo no se observa ninguna modificación sustancial con respecto a los resultados que se obtienen en base a la población al 30 de junio. Ambas estimaciones se alejan claramente de lo esperado y no coinciden con el comportamiento de los migrantes que se ha calculado para la Argentina a través de diversos estudios.

Las estimaciones de la migración neta para el período 1980-2011 se muestran en

las Figuras 5.13a y 5.13b. El comportamiento observado en los valores estimados es una combinación del verdadero comportamiento del saldo migratorio, los errores de registro en las muertes por edad y nacimientos según edad de la madre y los errores en las estimaciones y/o proyecciones de población. En primer lugar las cifras relativas a población tienen un origen que se analizó en el Capítulo 4, estas son estimaciones y/o proyecciones por lo que en su proceso de obtención pueden acarrear error, como sucede con cualquier estimación. Asimismo, los nacimientos también están afectados por otra fuente de error: el subregistro; en Argentina por ejemplo, para el período 1980-1985 el subregistro ascendía al 2.1 % de los nacimientos, 3.8 % para el período 1985-1990 y 5.5 % para 1995-2000 (Bay y Orellana, 2007). Este último porcentaje en valores netos representa aproximadamente entre 14.000 y 35.000 nacimientos sin registrar. El subregistro impacta en forma directa en las estimaciones de la migración neta de las edades iniciales, en ellas se observa un alto grado de variación (de forma sinusoidal), que surge de la compensación obligada que debe ejercer el algoritmo empleado al calcular migraciones y establecer una coherencia entre nacimientos y población en las edades iniciales. Una posible hipótesis de este fenómeno es que la diferencia sustancial entre las cifras mencionadas, nacimientos y población en las primeras edades, se debe en gran parte al subregistro presente en esta franja etaria. Otro aspecto destacado es el leve aumento de la variabilidad que se observa para las edades avanzadas, probablemente producto del sesgo en los datos y de la agrupación forzada en un grupo abierto final de 80 años y más, que como ya se destaco merecería una mayor desagregación.

Si bien los errores relativos a población y a defunciones afectan las estimaciones de la mortalidad, los errores relativos a las cifras de población y a los nacimientos afectan las estimaciones de la fecundidad, son todas estas fuentes de error en conjunto las que afectan a las estimaciones de migración. Además para evaluar este fenómeno se consultan demógrafos y expertos de manera informal con el fin de elaborar hipótesis acerca de la causa de haber obtenido estos resultados no convencionales y que no parecen representar el comportamiento de la migración neta. Los expertos, además de coincidir en que

la migración estimada con este procedimiento arrastra errores cometidos en todas las componentes demográficas empleadas durante el proceso de estimación, sostienen que el hecho de utilizar la población (estimada y/o proyectada) por edades simples puede ser un factor importante, debido a que el alto nivel de desagregación puede introducir aún más error. Es en parte por este motivo que en la actualidad la mayoría de los organismos de estadísticas oficiales generan estimaciones y proyecciones para grupos quinquenales.

En síntesis la complejidad que presenta la estimación de la componente migratoria está vinculada principalmente al error presente en las diversas fuentes de datos, entre ellas las cifras de población cuyo origen se analiza en el Capítulo 4. Este complejo fenómeno motiva a generar nuevas líneas de investigación, como por ejemplo, la posibilidad de trabajar a futuro con datos con menor desagregación (grupos quinquenales) que a su vez estén generados por interpolación entre los años censales y que no provengan de estimaciones y/o proyecciones demográficas (ya que las mismas se basan en múltiples supuestos). Sería interesante evaluar si los cambios introducidos modifican los resultados y si permiten generar cifras de migración que reflejen el verdadero comportamiento y no el agregado de errores acumulados como sucede en la presente aplicación.

5.4. Pronósticos Coherentes

Hyndman *et al.* (2013) proponen una metodología alternativa para realizar los pronósticos por género: los pronósticos coherentes. Los mismos tienen como finalidad evitar una divergencia entre los resultados para ambos subgrupos, de este modo los pronósticos para los varones y mujeres difieren siempre en la misma proporción través del tiempo y más aún en los pronósticos se mantiene la misma diferencia que se presenta en los datos observados.

Si bien sería posible aplicar esta metodología a datos de mortalidad y migración, esta

se aplica solamente a datos de mortalidad, debido a los resultados obtenidos expuestos en la sección 5.3 realtivos a la componente migratoria.

Se construyen las funciones $p_t(x)$ (Ecuación 3.17) y $c_t(x)$ (Ecuación 3.18) de las tasas de mortalidad de varones y mujeres. La función $p_t(x)$ es la media geométrica de las tasas suavizadas de hombres y mujeres y para la mortalidad esta medida declina a través del tiempo aunque se detectan leves ascensos en edades menores de 5 años y entre los 15 y 30 años, ver Figura 5.14. La función $c_t(x)$ es la raíz cuadrada del cociente entre las tasas suavizadas de cada sexo, en este caso se observa una alta variabilidad en las edades entre los 20 y 40 años ver Figura 5.15. Además entre los 40 y 70 años, aproximadamente, dicha función presenta un descenso a través del tiempo. Luego, bajo el paradigma de datos funcionales, se aplica el modelo dado por las ecuaciones 3.19 y 3.20.

En las Figuras 5.16 y 5.17 se presentan las funciones base, con los coeficientes correspondientes y sus pronósticos. Es importante destacar que el logaritmo de la raíz cuadrada del producto representa el promedio entre el logaritmo de la mortalidad femenina y masculina, mientras que el logaritmo de la raíz cuadrada del cociente es la mitad de la diferencia entre el logaritmo de la mortalidad de hombres y mujeres, es decir, que los resultados de bases y coeficientes se refieren a un comportamiento promedio de la mortalidad para el producto, y a una diferencia entre la mortalidad de ambos géneros para el cociente. En el modelo de $p_t(x)$, la primera componente principal muestra un descenso a través del tiempo, que según el comportamiento de la base correspondiente, resulta más rápido en la infancia y más lenta en las edades cercanas a los 20 años y edades avanzadas. Este término retiene el 92.9 % de la variabilidad. El porcentaje de variación debida a las seis primeras bases es 92.9 %, 2.8 %, 1.5 %, 0.8 %, 0.6 % y 0.4 %, respectivamente. En relación al modelo estimado para $c_t(x)$ se observa una tendencia creciente para la primer componente, que se estabiliza hacia el año 2000, y, de acuerdo a la base correspondiente se refiere a edades entre 20 y 40 años. Este par base-coeficiente esta asociado a la diferencia en la mortalidad entre géneros y estaría indicando que en

edades entre 20 y 40 años y mayores de 70 años, se incrementa la diferencia entre la mortalidad de hombres y mujeres. Esta componente representa el 78.8 % de la variabilidad, y de la segunda a la sexta base explican el 8.4 %, 3.2 %, 2.6 %, 1.5 % y 1.2 % respectivamente. En ambos modelos resulta complejo hallar un correlato con comportamientos demográficos teóricos. Finalmente, no se observan efectos de cohorte ni patrones de edad en los residuos (Apéndice 6: 5 y 6).

Cuadro 5.11: de confianza,

La última etapa del modelo consiste en el pronóstico de los coeficientes $\beta_{n,k}$ y $\gamma_{n,k}$, los mismos se realizan por modelos ARIMA en el caso de $p_t(x)$, y por modelos ARFIMA en el caso de $p_t(x)$. En el Apéndice 6 se pueden ver las estimaciones de los parámetros de los modelos de series de tiempo para el producto y el cociente.

A partir de los pronósticos de los coeficientes es posible calcular los pronósticos de las tasas para ambos géneros, los Cuadros 5.10 y 5.11 presentan los resultados para el año 2013 y 2023 y para edades seleccionadas, los pronósticos puntuales están acompañados además de sus intervalos de pronóstico del 80 % de confianza.

5.5. Pronósticos estocásticos de población

A partir de los pronósticos de fecundidad (basados en las dos hipótesis de fecundidad) y de mortalidad (independientes y coherentes) se elaboran pronósticos estocásticos de la población para hombres y mujeres. En relación a la componente migratoria, dado que el análisis de esta componente arrojó resultados atípicos, se supone que las cantidades de migrantes y emigrantes son iguales generando así un saldo migratorio nulo. Teniendo en cuenta que los datos de fecundidad y mortalidad se modelan hasta 2011 y 2012, respectivamente, la simulación se basa en datos de las dos componentes demográficas para el período 1980-2011 y se toma como población base la del año 2012.
En base a N=1000 trayectorias simuladas para cada componente y a través del procedimiento detallado paso a paso en la sección 3.2.4 para generar los pronósticos de

ra 5.37: Pirámides de población año 2020 (Argentina) - Pronósticos Independientes

población y sus intervalos de confianza, se obtienen los resultados presentados en el

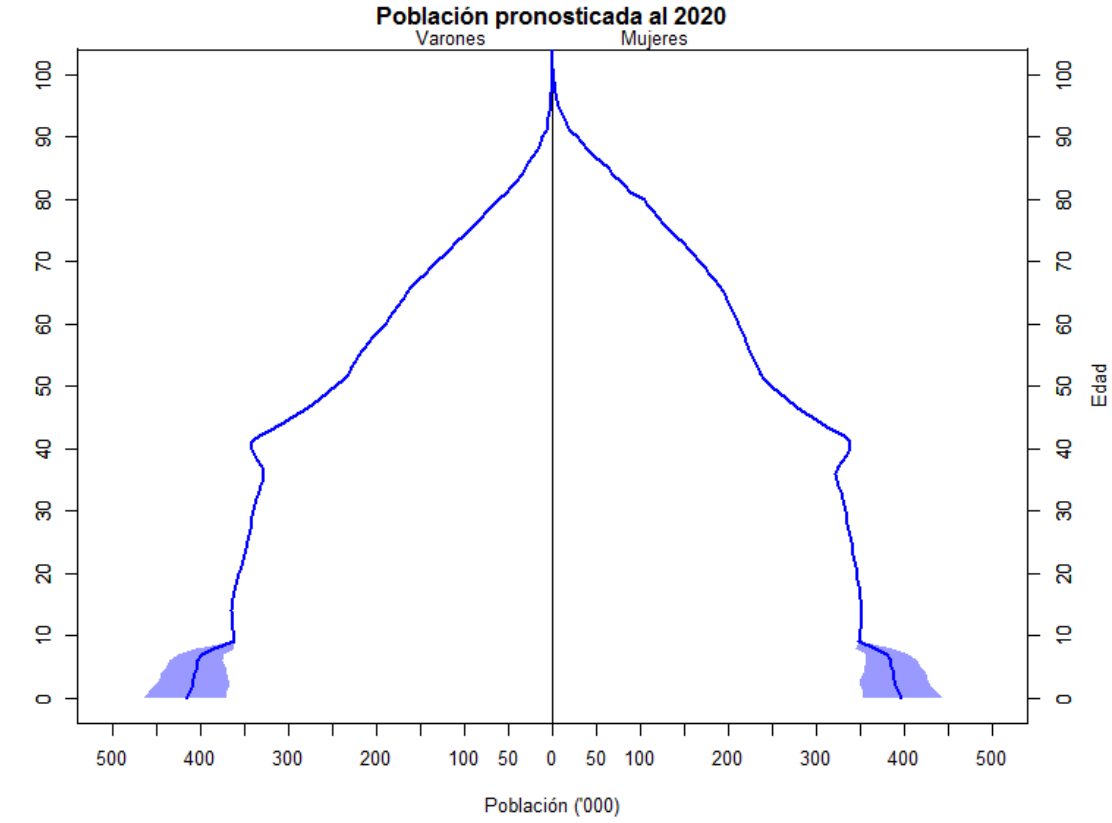


Figura 5.38: sin pendiente

(Hyndman *et al.*, 2008), donde se presenta el pronóstico de población de Australia para

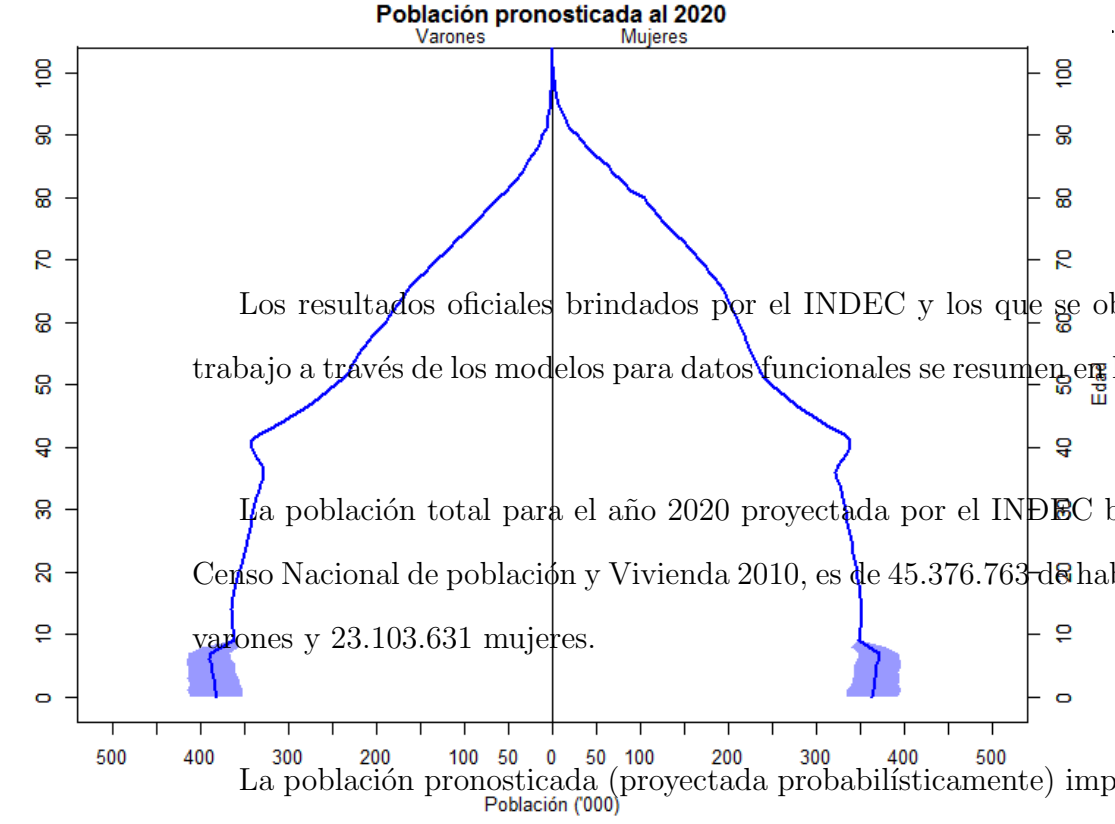


Figura 5.39: con pendiente

Los resultados oficiales brindados por el INDEC y los que se obtienen en presente trabajo a través de los modelos para datos funcionales se resumen en las siguientes cifras:

La población total para el año 2020 proyectada por el INDEC basada en datos del Censo Nacional de población y Vivienda 2010, es de 45.376.763 de habitantes, 22.273.132 varones y 23.103.631 mujeres.

La población pronosticada (proyectada probabilísticamente) imponiendo coherencia entre géneros es de 46.479.523 habitantes (46.016.405, 46.983.910)[3] , 22.876.092 varones

(22.641.598, 23.135.384) y 23.603.430 mujeres (23.376.855, 23.850.370), para el supuesto de fecundidad estable y 46.128.313 habitantes (45.813.878, 46.468.696), 22.696.071 varones (22.533.200, 22.866.642) y 23.432.242 mujeres (23.279.489, 23.596.665) para el supuesto de fecundidad descendente.

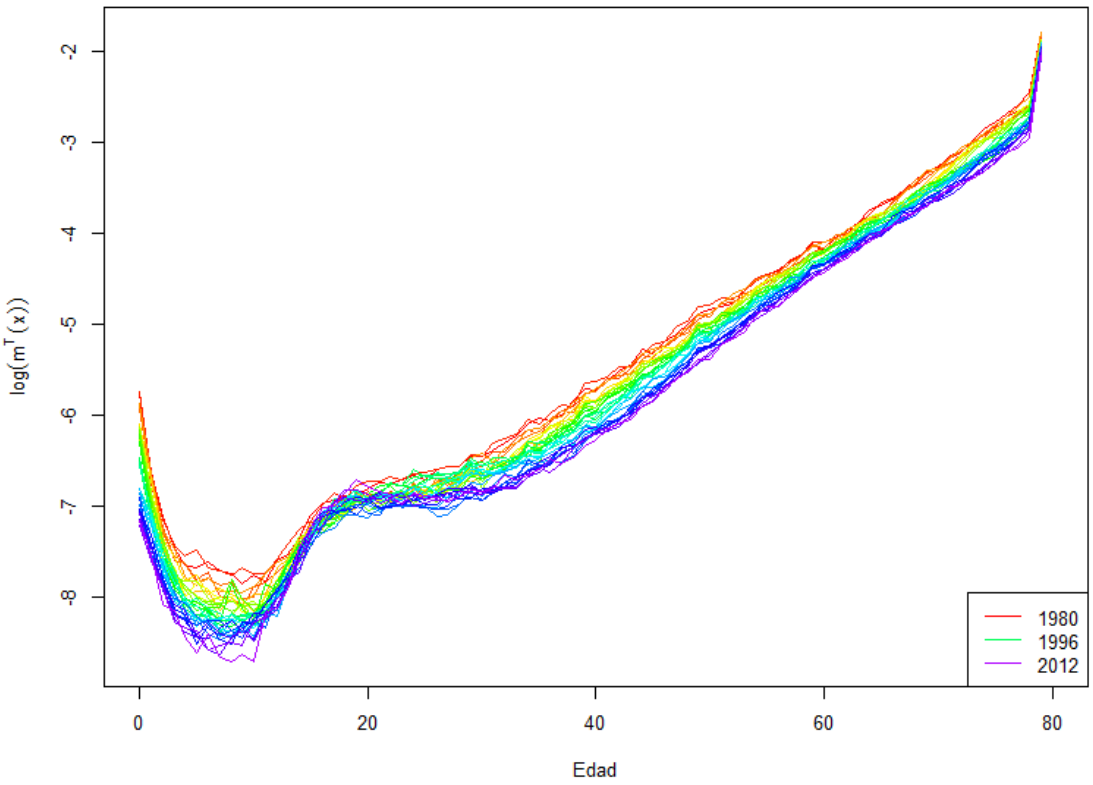
La población pronosticada con pronósticos independientes entre géneros es de 46.113.514 habitantes (45.633.580, 46.619.389), 22.690.469 varones (22.436.905, 22.951.933) y 23.423.045 mujeres (23.188.880, 23.668.318) para el supuesto de fecundidad estable y 46.104.694 habitantes (46.099.118, 46.586.127), 22.685.361 varones (22.445.377, 22.932.227) y 23.419.333 mujeres (23.192.940, 23.658.456) para el supuesto de fecundidad descendente.

A fin de hacer un resumen de las cifras que facilite la comparación se calculan las diferencias, para el total, varones y mujeres de las cuatro cifras estimadas versus las oficiales, Cuadro 5.13. En la comparación, en todos los casos las cifras estimadas se encuentran por encima de las oficiales con una diferencia promedio del 2.6 %. Se destaca que la mayor diferencia con la proyección oficial se observa para varones, consistentemente, a lo largo de las cuatro estimaciones. Las cifras que se obtienen a través de pronósticos coherentes con hipótesis de fecundidad estable presentan los porcentajes más altos de diferencia tanto para el total como para ambos géneros; 2.43 %, 2.71 % y -2.16 %. Mientras que las cifras calculadas en forma independiente para el total y cada género, correspondientes a la hipótesis de fecundidad descendente, son las que presentan la menor diferencia con la proyección oficial, 1.58 %, 1.82 % y 1.35 %.
Es importante remarcar que esta comparación no es indicador de bondad de ajuste, ya que todas las cifras que se comparan son pronósticos, en cambio da cuenta de la divergencia en las metodologías y supuestos que subyacen a los números obtenidos. La similitud entre las cifras oficiales y las obtenidas por medio de pronósticos independientes e hipótesis de fecundidad descendente radica, posiblemente, en que en ambos procesos no se impone coherencia entre las cifras de varones y mujeres, además en ambos

[3]Entre paréntesis se presenta el intervalo de pronóstico del 80 % de confianza

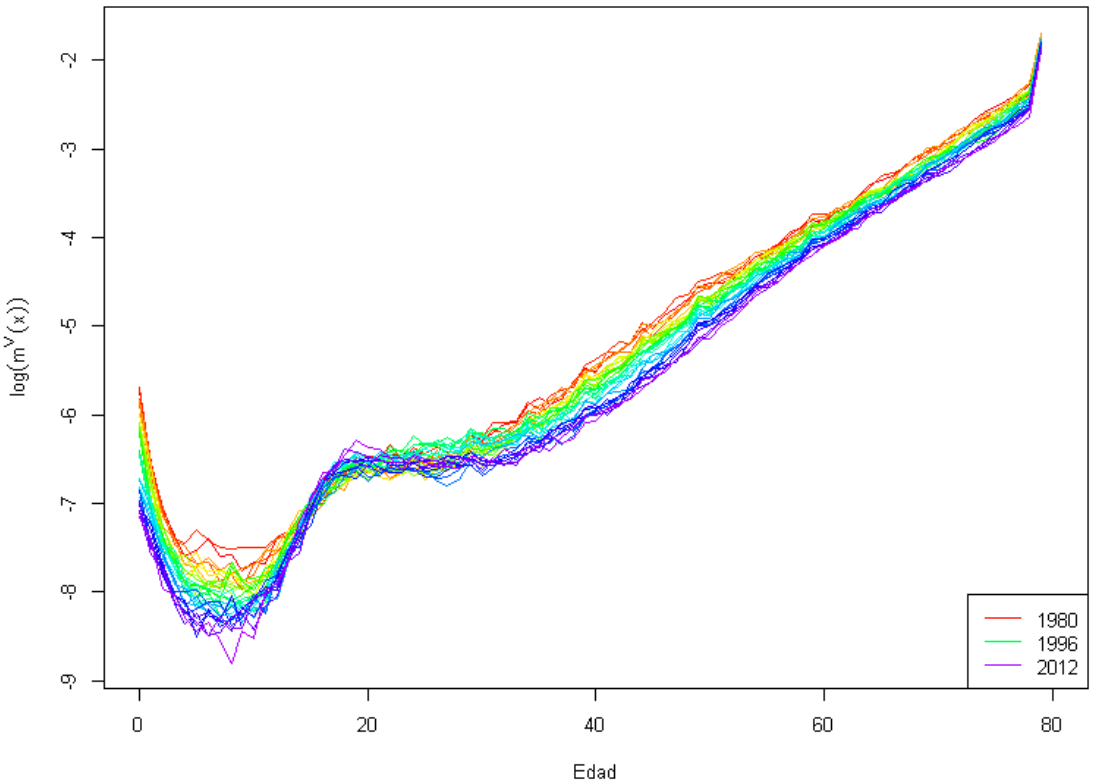
casos se supone una fecundidad descendente. El INDEC en su proyección oficial, por medio de una función logística y en base a la tendencia histórica que presenta la Tasa Global de Fecundidad (TGF) desde al año 1981, genera tasas proyectadas hasta el año 2040 (Cuadro 5.14)(INDEC-CELADE, 2013). Estas proyecciones conforman el supuesto de la fecundidad que subyace en la población obtenida. Si bien el documento técnico que detalla la metodología empleada para la elaboración de las proyecciones destaca que esta componente presenta un descenso mucho más atenuado que el observado en décadas pasadas y en base a ello recomienda tener cautela.

Figura 5.1: Logaritmo de las tasas de mortalidad observadas (Argentina, 1980-2012) [1]



.5

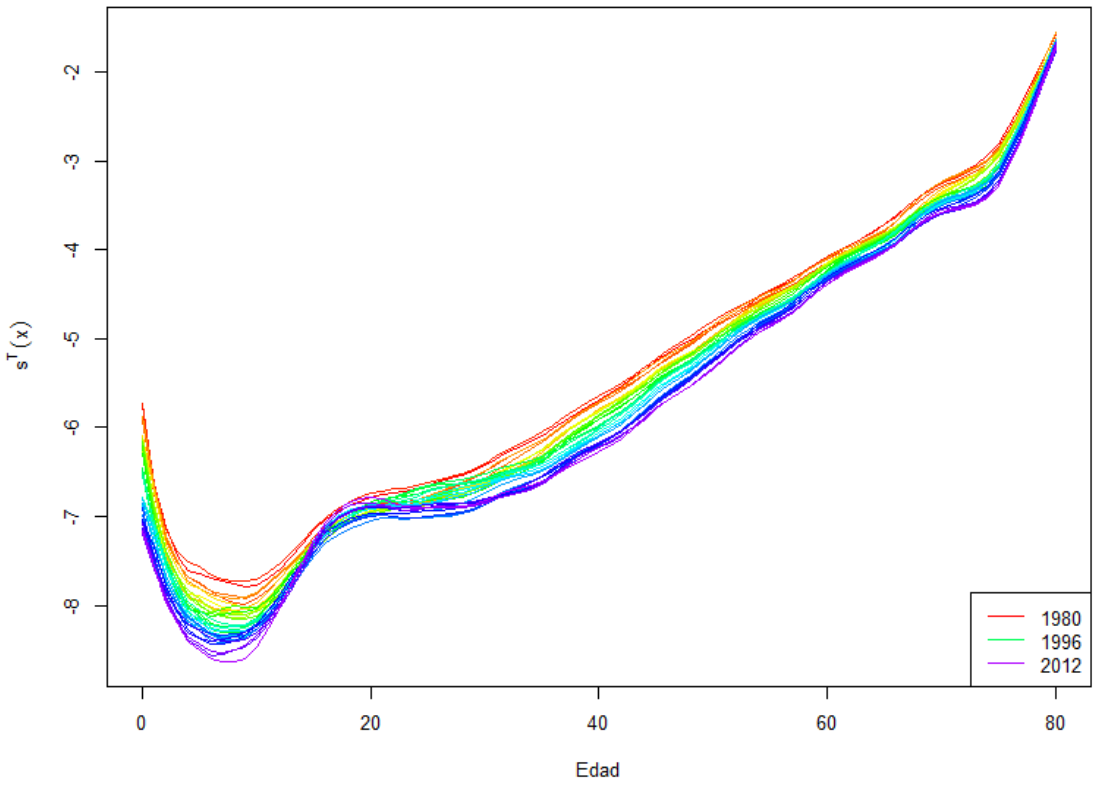
Figura 5.2: Total



.5

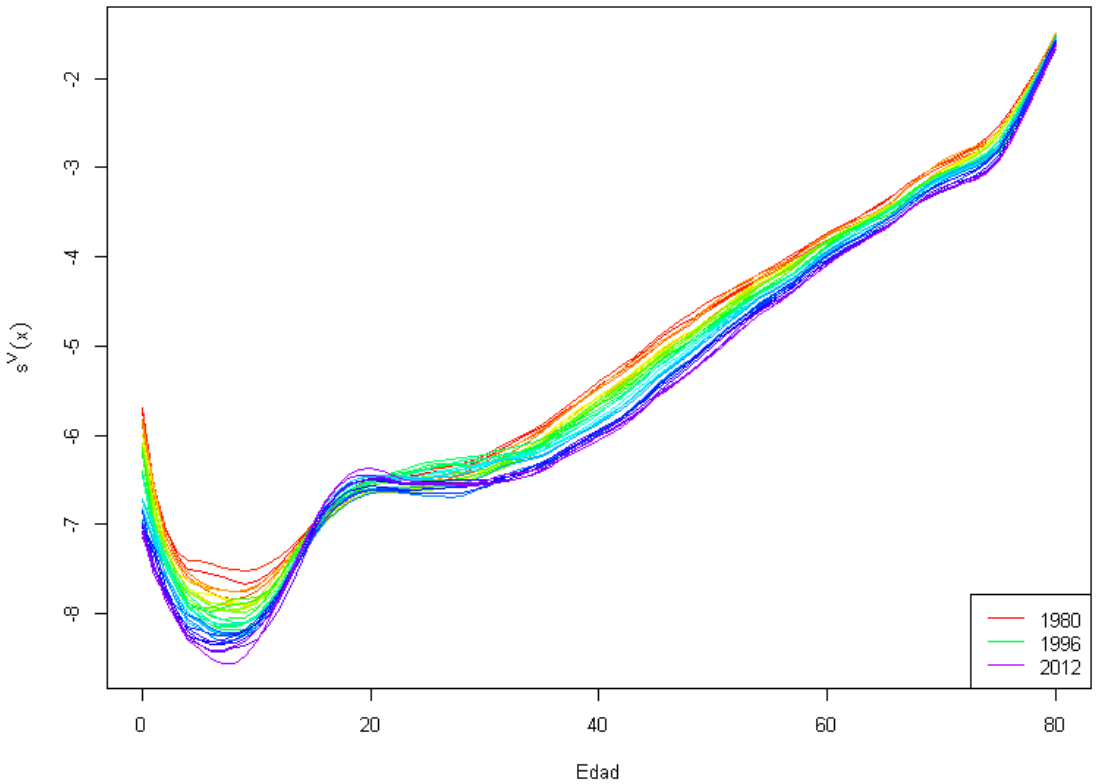
Figura 5.3: Varones

Figura 5.5: Suavizado del logaritmo de las tasas de mortalidad (Argentina, 1980-2012)



.5

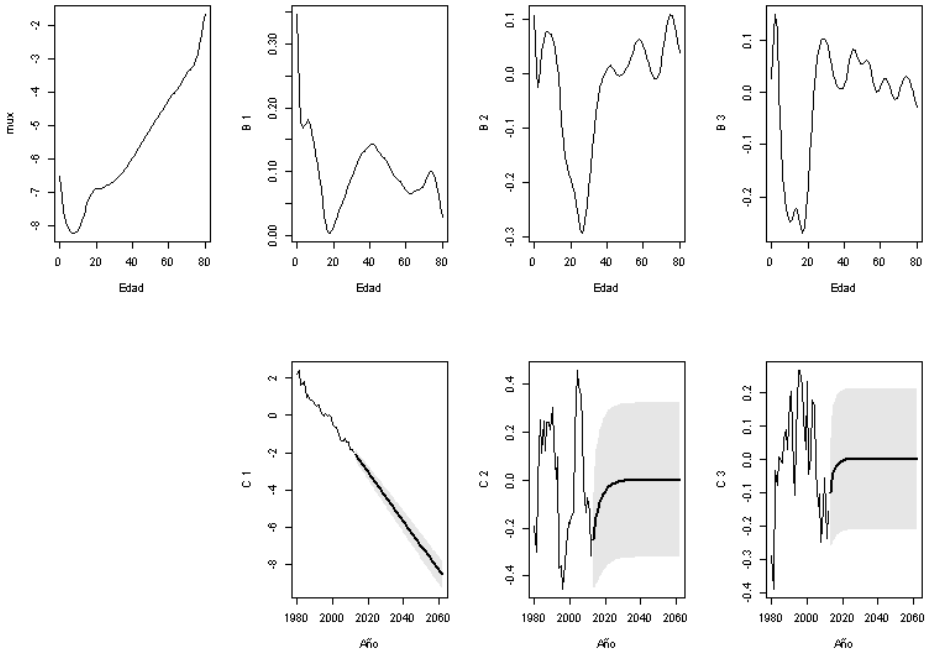
Figura 5.6: Total



.5

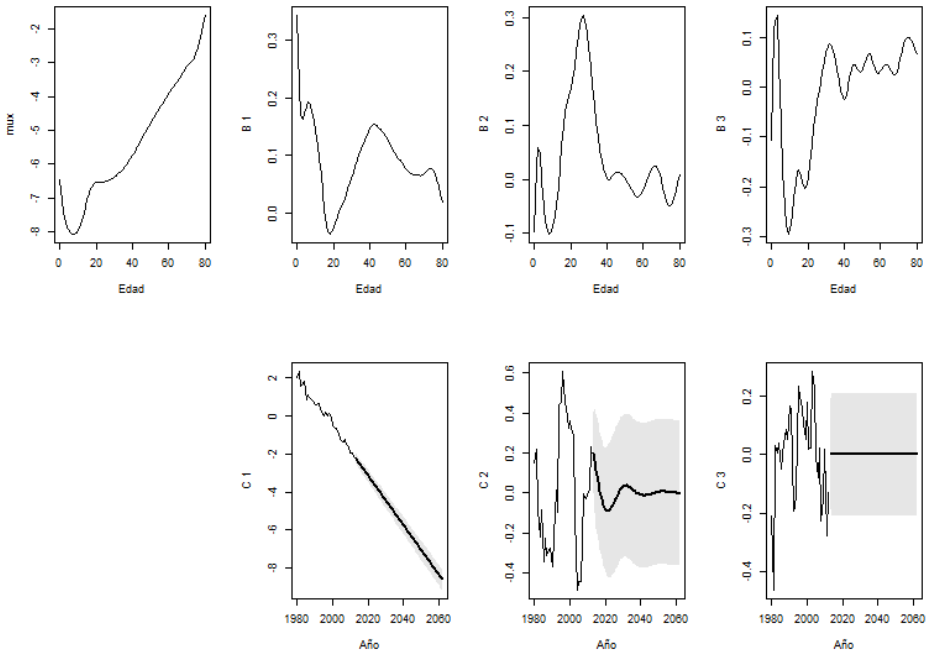
Figura 5.7: Varones

Figura 5.9: Funciones base, coeficientes asociados, pronósticos e intervalos de pronóstico del 80 % de confianza. Modelo de datos funcionales para la mortalidad Total.



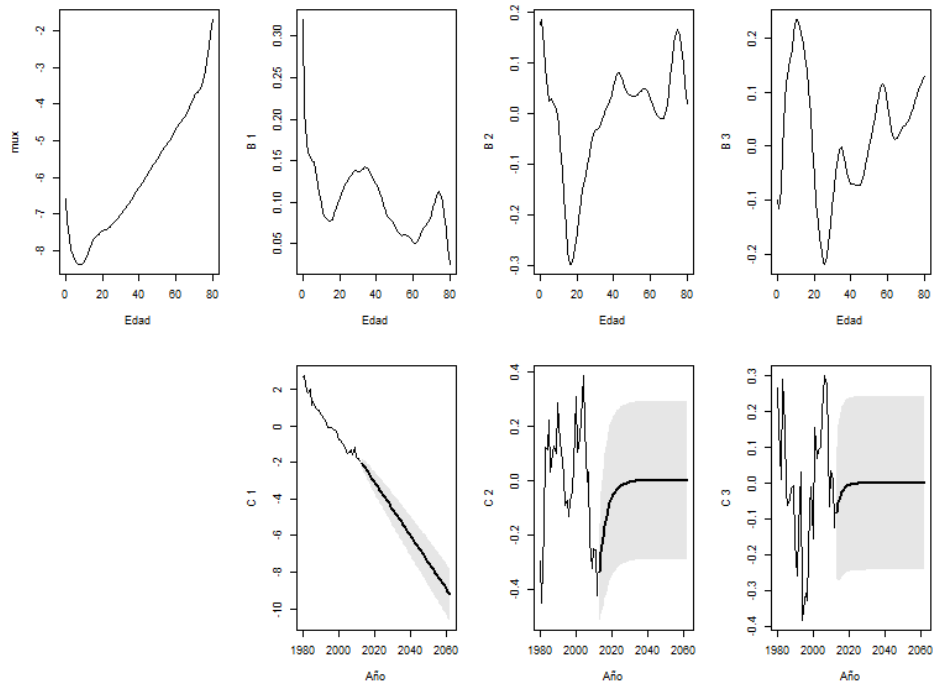
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.10: Funciones base, coeficientes asociados, pronósticos e intervalos de pronóstico del 80 % de confianza. Modelo de datos funcionales para la mortalidad de Varones.



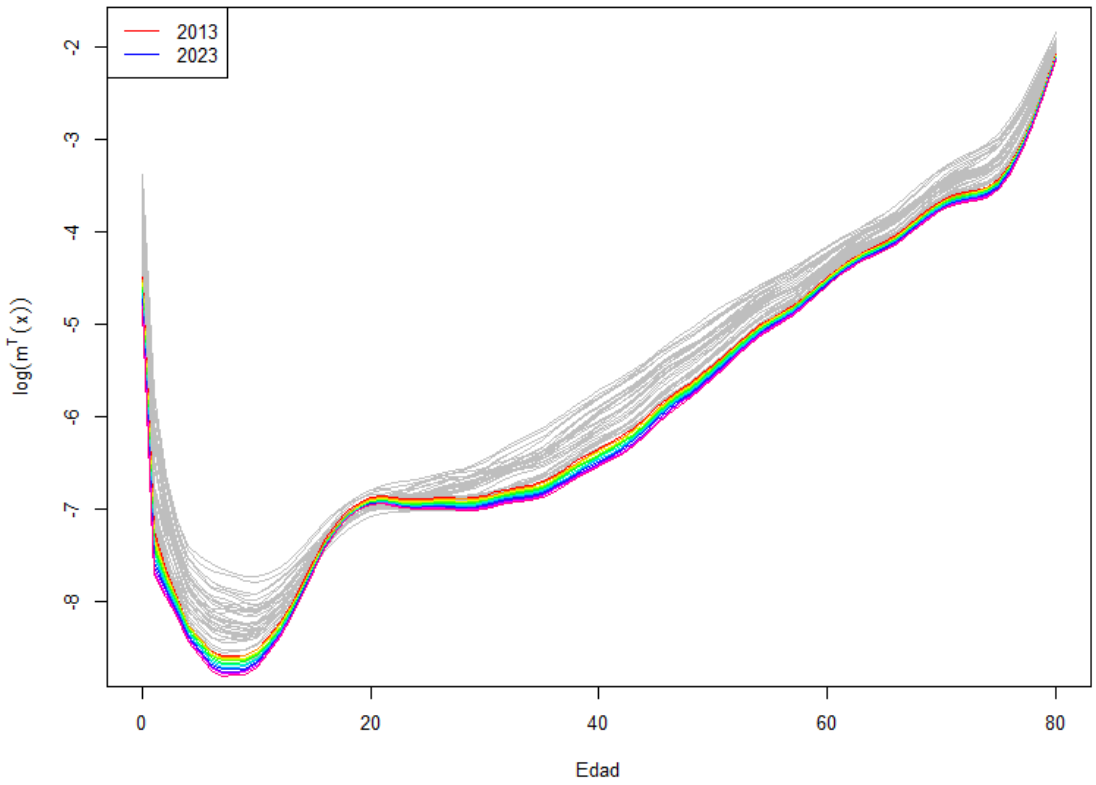
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.11: Funciones base, coeficientes asociados, pronósticos e intervalos de pronóstico del 80 % de confianza. Modelo de datos funcionales para la mortalidad de Mujeres.



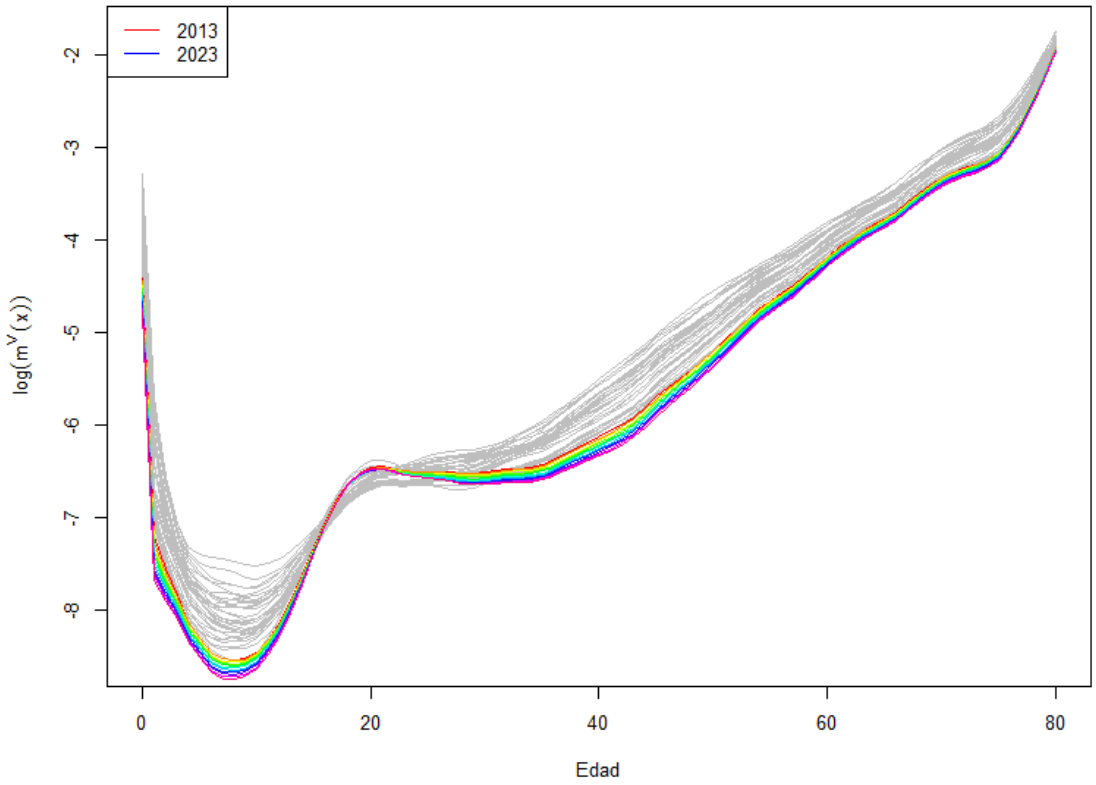
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.12: Tasas de mortalidad pronosticadas (Argentina, 2013-2023)



.5

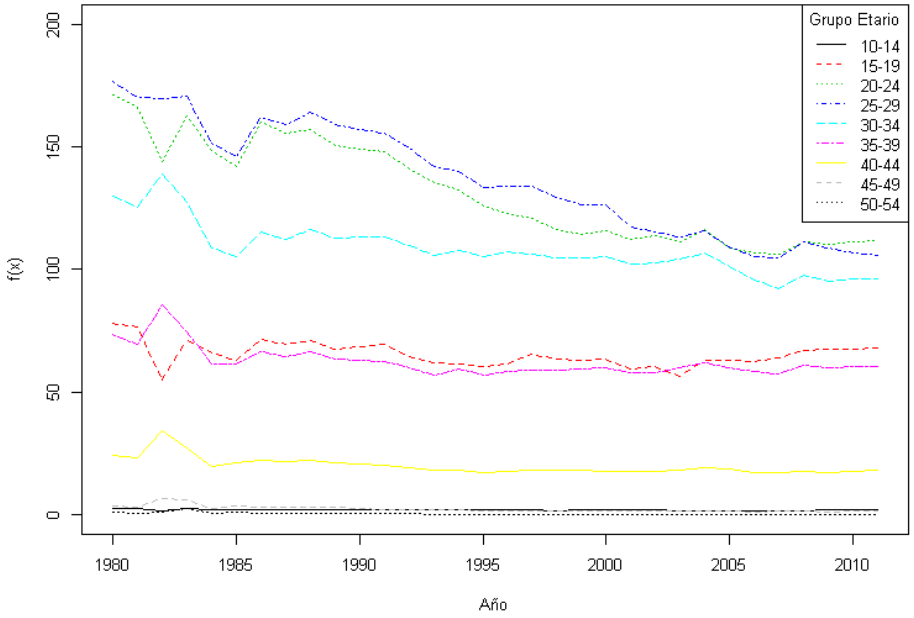
Figura 5.13: Total



.5

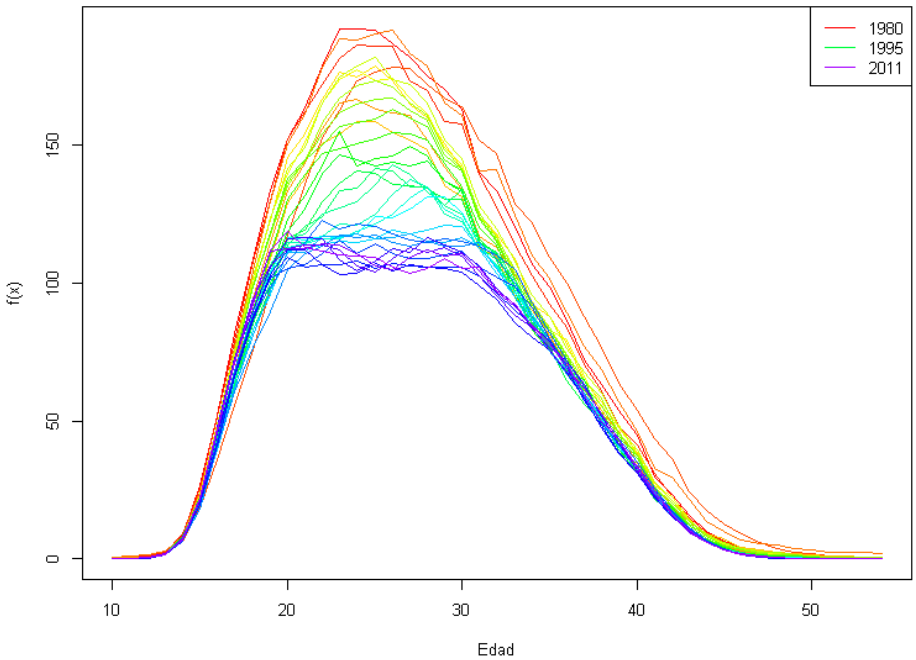
Figura 5.14: Varones

Figura 5.16: Series temporales de las tasas de fecundidad por edad por cada mil mujeres (Argentina, 1980-2011).



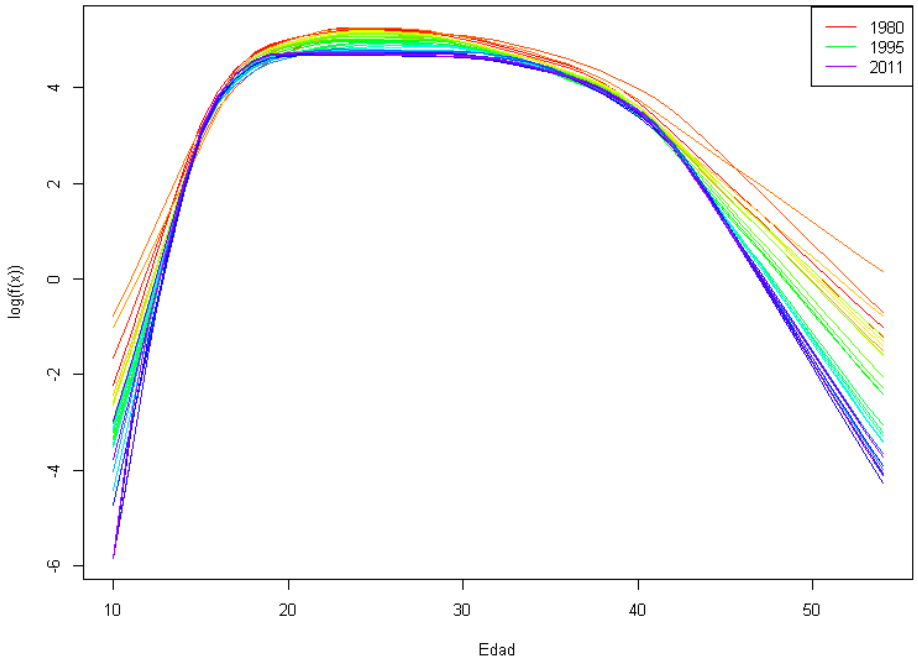
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.17: Tasas de fecundidad por edad por cada mil mujeres (Argentina, 1980-2011).



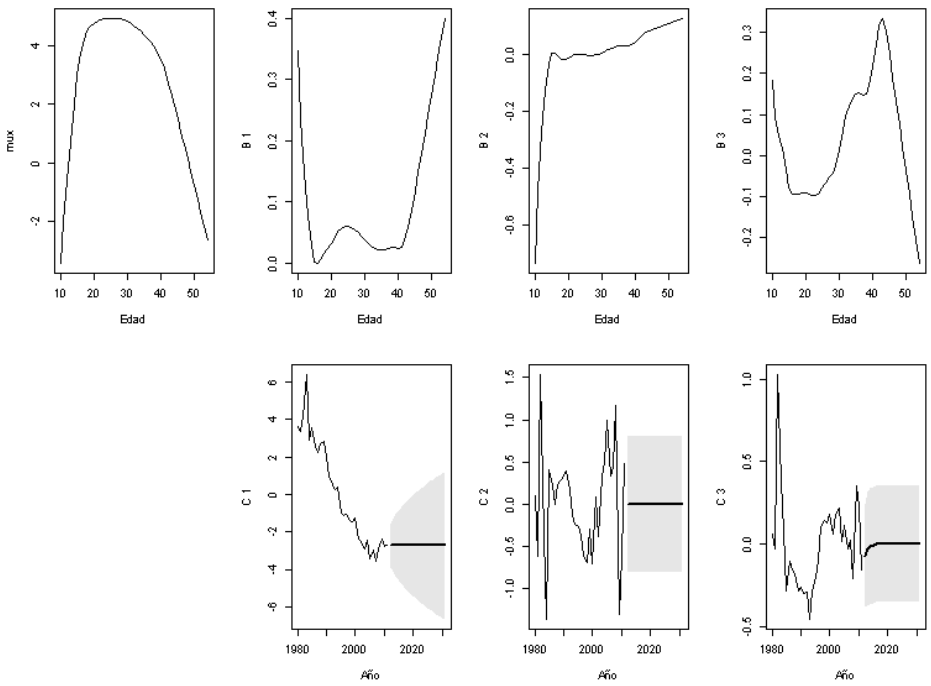
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.18: Suavizado del logaritmo de las tasas de fecundidad por edad (Argentina, 1980-2011)



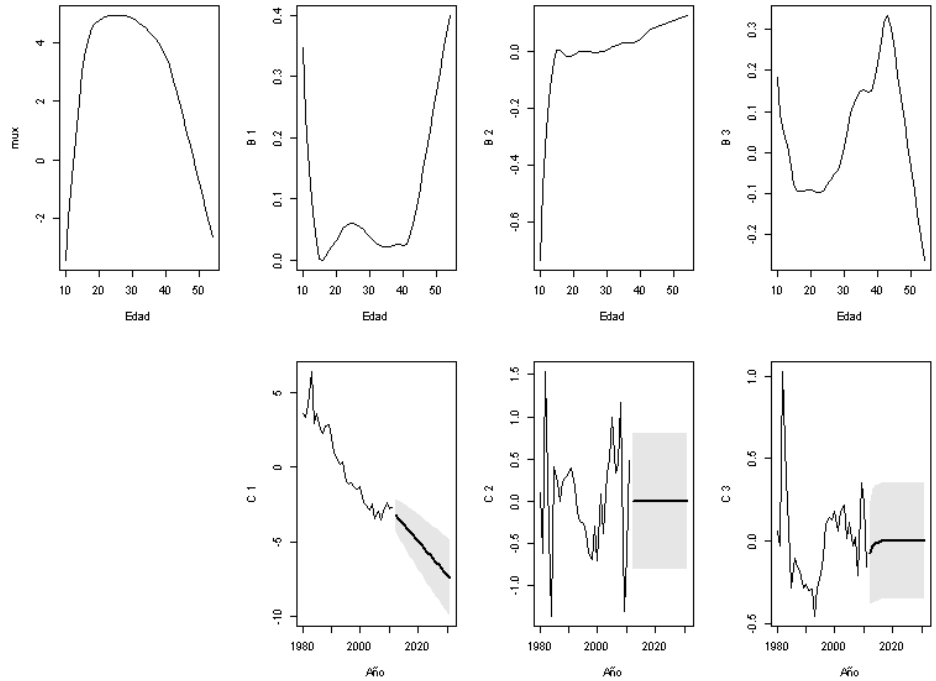
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.19: Funciones base, coeficientes asociados, pronósticos e intervalos del pronóstico del 80 % de confianza. Modelo para datos funcionales de fecundidad - sin pendiente



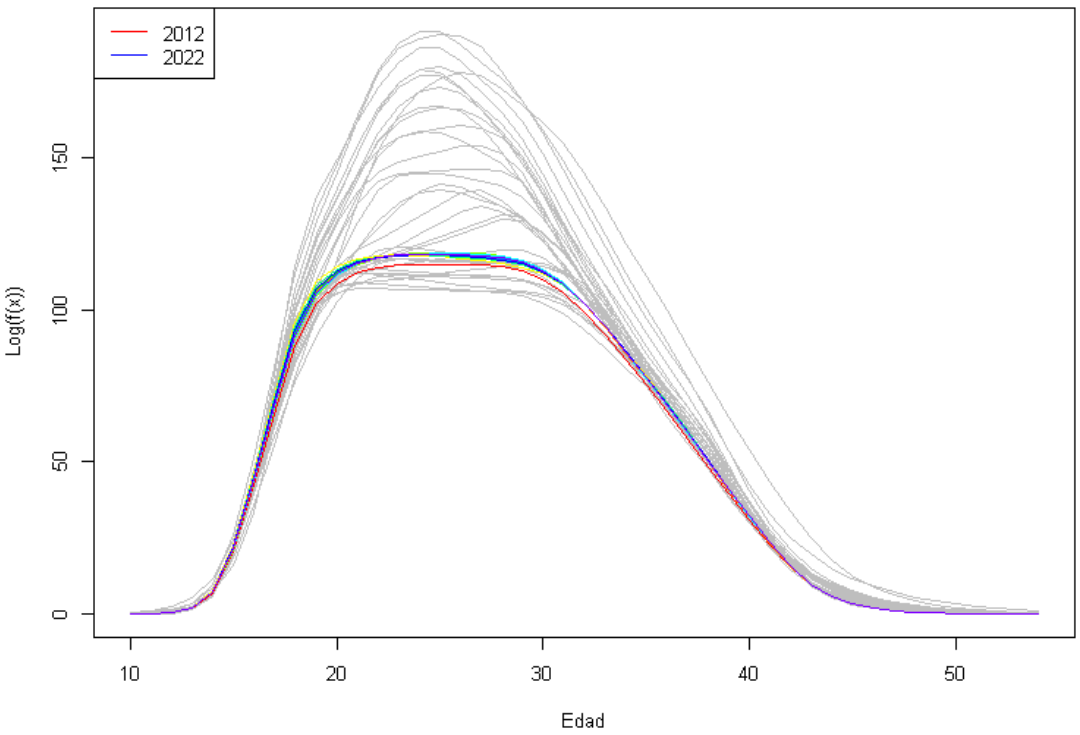
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.20: Funciones base, coeficientes asociados, pronósticos e intervalos de pronóstico del 80 % de confianza. Modelo para datos funcionales de fecundidad - con pendiente



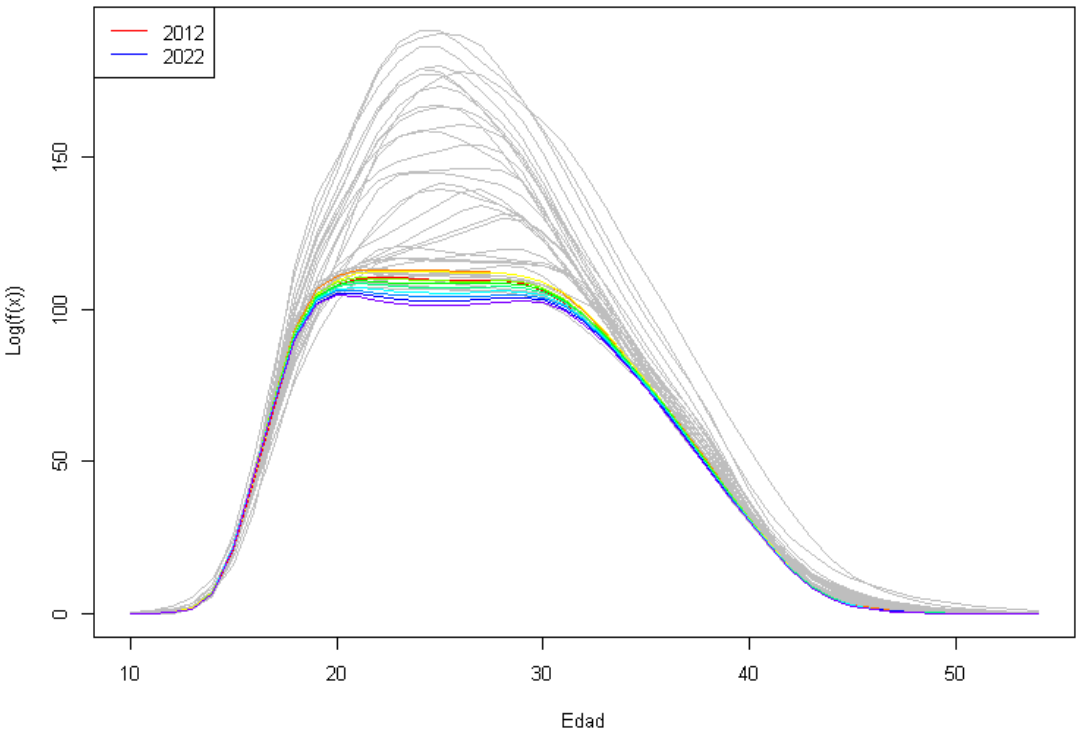
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.21: Curvas de fecundidad pronosticadas (Argentina, 2012-2022)



.5

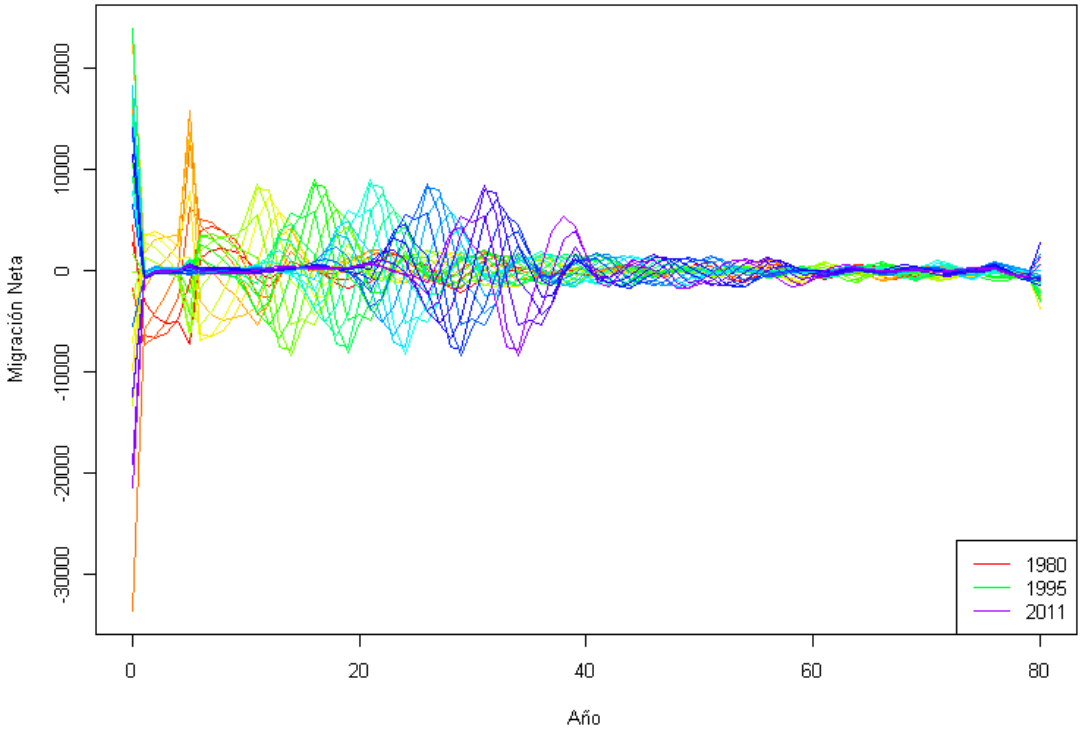
Figura 5.22: sin pendiente



.5

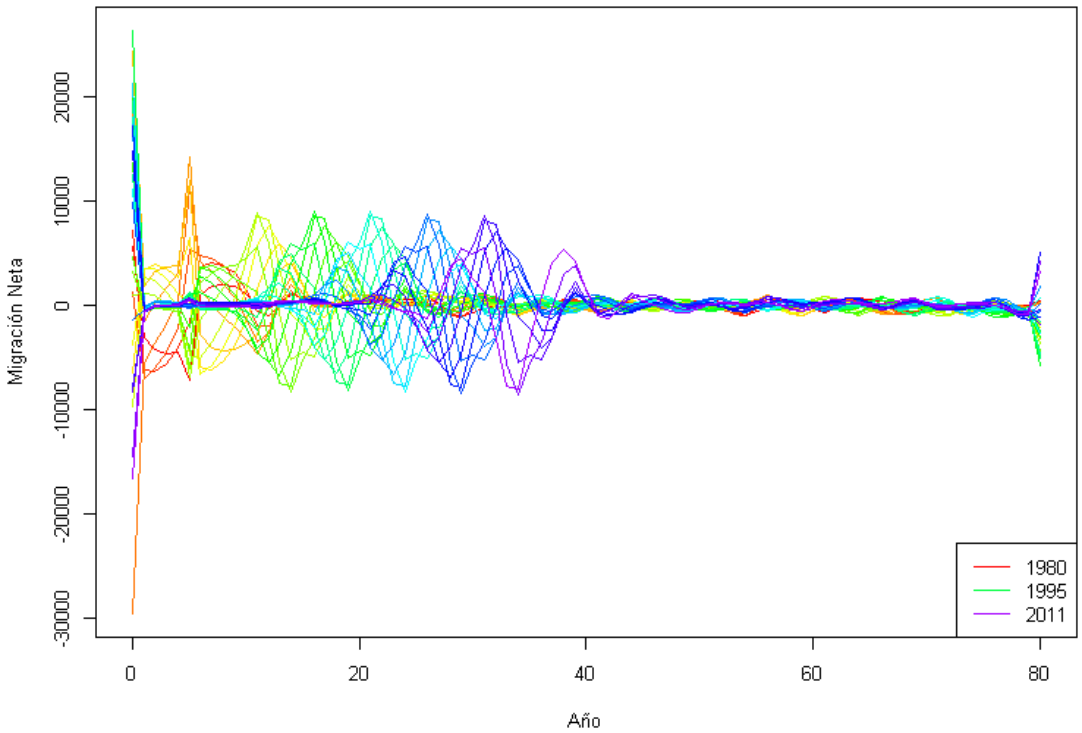
Figura 5.23: con pendiente

Figura 5.24: Migración neta estimada (Argentina, 1980-2011)



.5

Figura 5.25: Varones

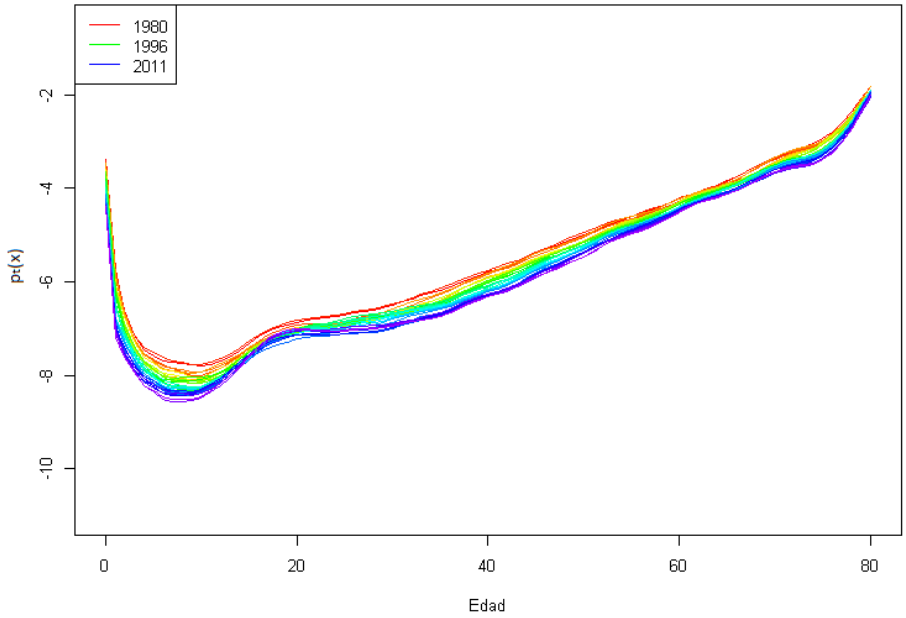


.5

Figura 5.26: Mujeres

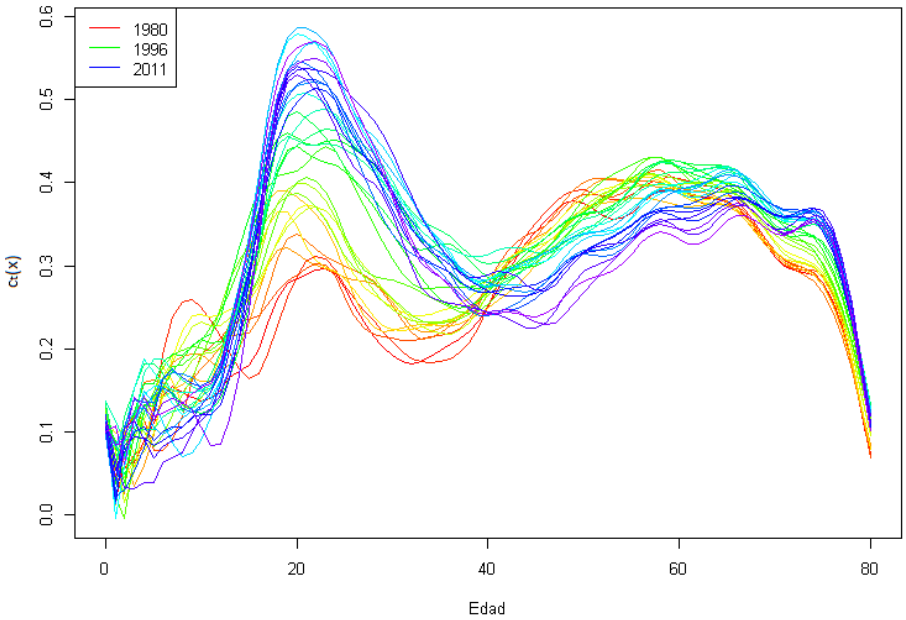
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.27: Función $p_t(x)$ - Mortalidad (Argentina, 1980-2012)



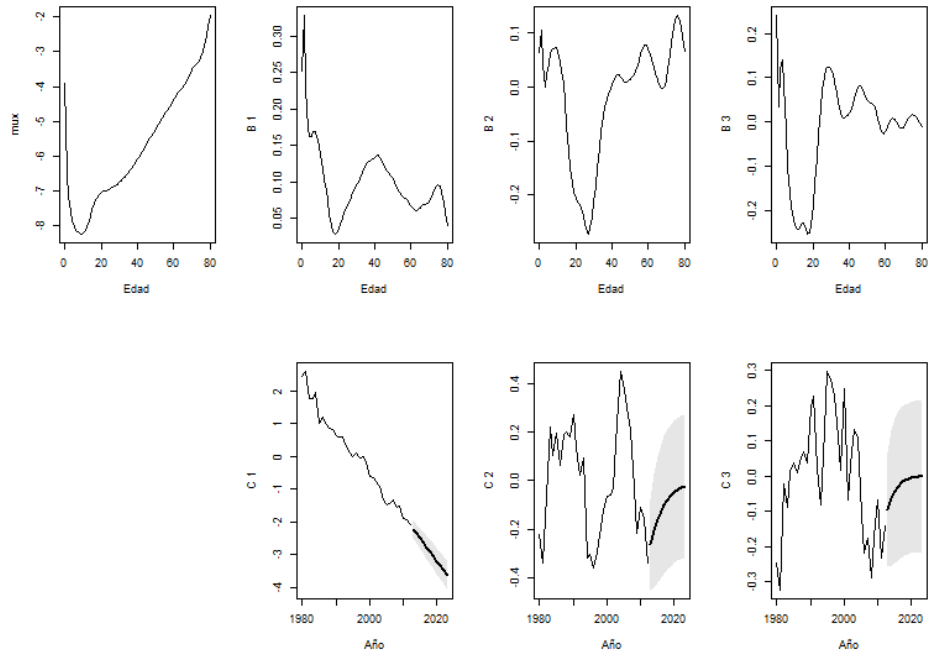
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.28: Función $c_t(x)$ - Mortalidad (Argentina 1980-2012)



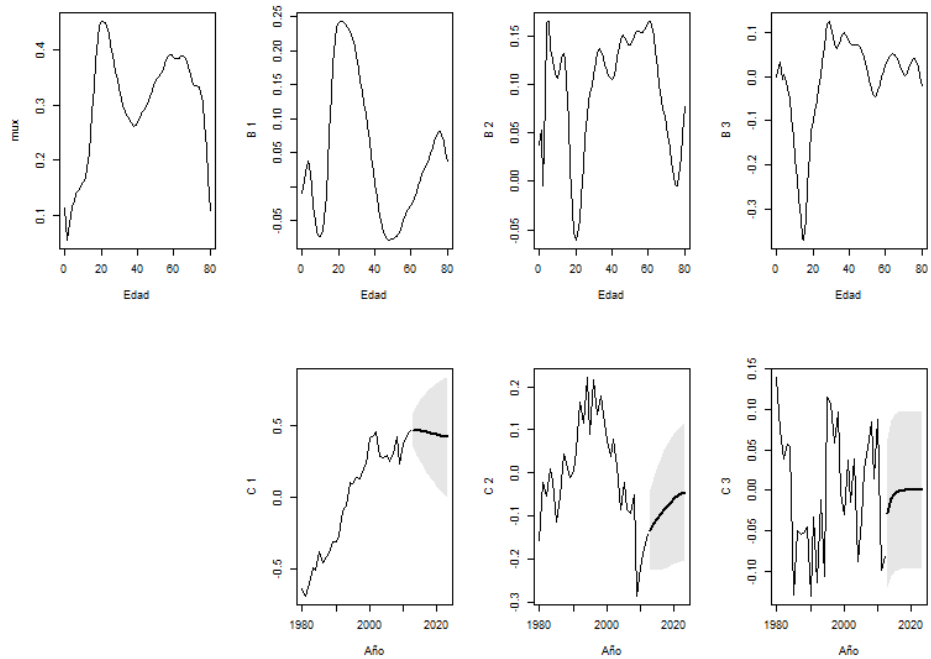
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.29: Funciones base, coeficientes asociados, pronósticos e intervalos de pronóstico del 80 % de confianza. Modelo para datos funcionales de la Mortalidad - $p_t(x)$ (Argentina, 1980-2012)



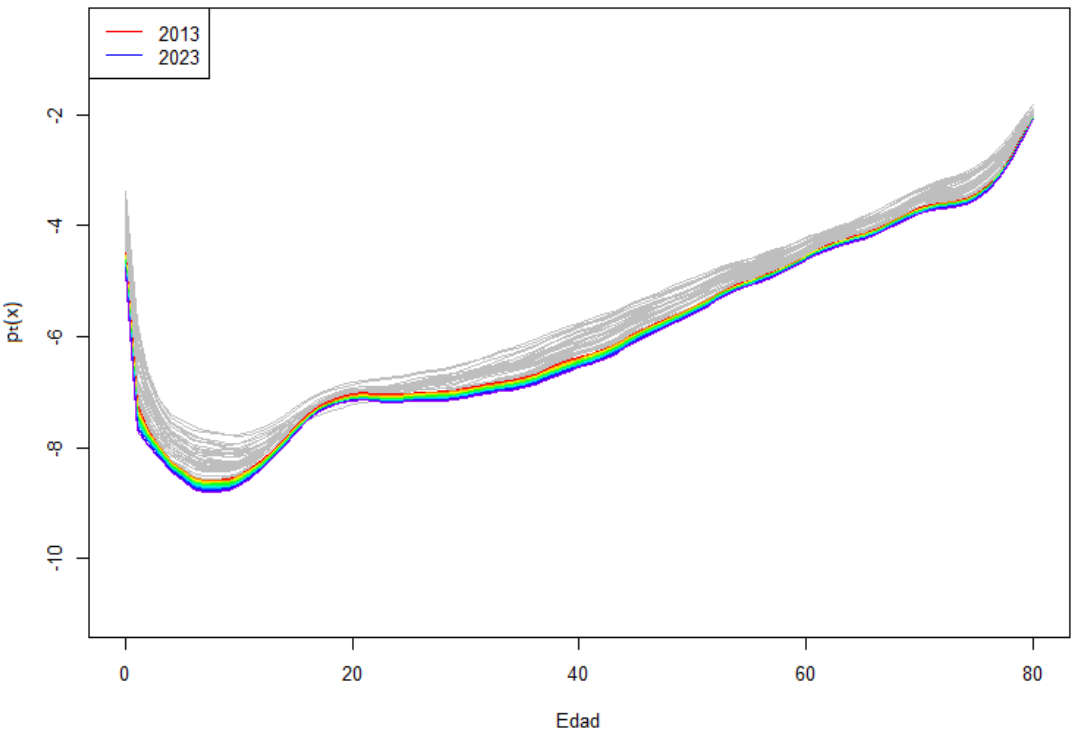
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.30: Funciones base, coeficientes asociados, pronósticos e intervalos de pronóstico del 80 % de confianza. Modelo para datos funcionales de la Mortalidad - $c_t(x)$ (Argentina, 1980-2012)



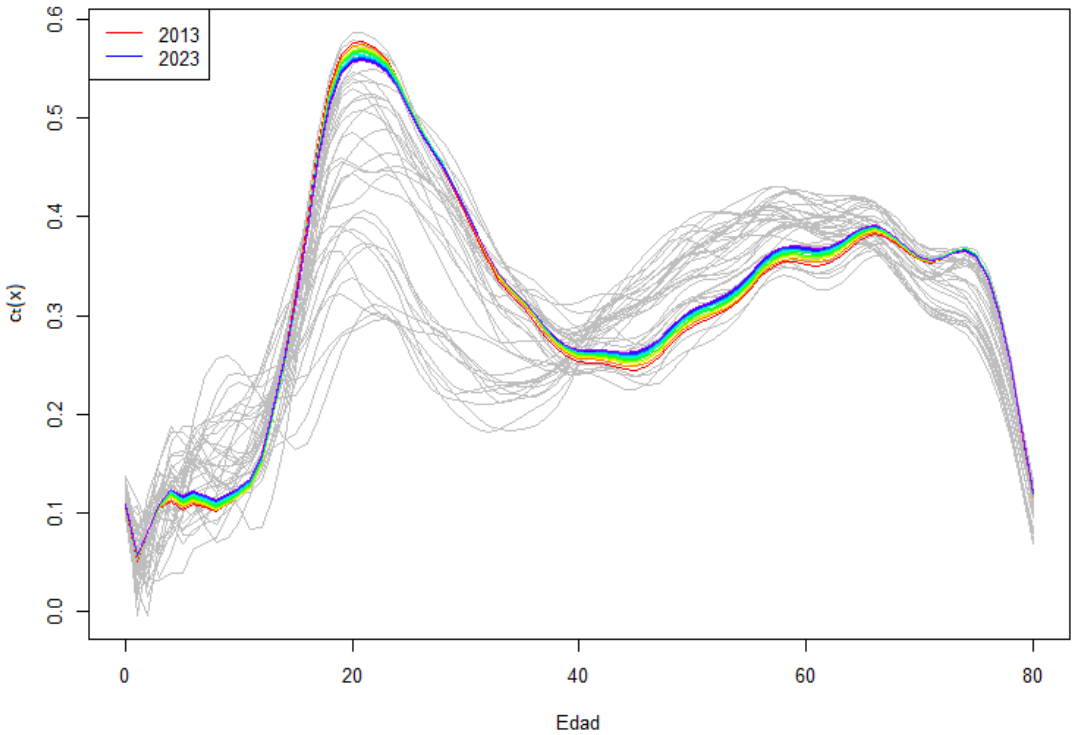
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5.31: Pronósticos del producto y cociente de Mortalidad (Argentina, 2013-2023)



.5

Figura 5.32: Producto



.5

Figura 5.33: Cociente

Fuente: Elaboración propia en base a datos de la DEIS.

Capítulo 6

Conclusiones

En primer lugar se aplican modelos para datos funcionales a las cifras de mortalidad y fecundidad de Argentina (período 1980 a 2012 para la mortalidad y 1980 a 2011 para la fecundidad) para obtener la descomposición de ambas componentes demográficas, sus pronósticos probabilísticos y el del total de la población a partir de las componentes pronosticadas. La aplicación a datos de Argentina demuestra que los modelos para datos funcionales se pueden emplear con éxito para pronosticar tasas de mortalidad y fecundidad cuando se cuenta con información para períodos no muy extensos de tiempo y una calidad aceptable de las fuentes de datos. Es importante remarcar, que este enfoque de modelado es altamente adaptable; el mismo modelo básico se aplica a las tres componentes permitiendo transformaciones y suavizados específicos para cada una de ellas.

La aplicación del modelo funcional a datos de mortalidad se realiza primero para cada género de forma independiente. En relación al análisis de datos de los varones, las edades que dominan el cambio en la mortalidad resultan; la franja de edades cercanas a cero, el grupo de 40 a 50 años, el grupo de 18 a 30 años y los levemente menores y mayores de 15 años. La forma en la que ejercen el cambio esta dada por las bases, un descenso para los primeros tres grupos y fluctuaciones a lo largo del período para los restantes. En las mujeres, los grupos que dominan los cambios en la mortalidad de

mujeres desde 1980 a 2010 son las niñas de 0 a 5 años, mujeres de alrededor de 30 años y las adultas mayores de alrededor de 75 años de edad, en segundo lugar las mujeres de alrededor de 20 años y por ultimo las levemente mayores y menores a los 20 años. La forma en estos grupos dominan los cambios esta dada por su base asociada, descenso en mayor o menor medida para los tres primeros grupos mencionados y fluctuaciones a lo largo del tiempo para los restantes. Además, la aplicación del modelo funcional a datos de fecundidad, y por ende la descomposición en bases y coeficientes, permite identificar el patrón de descenso y las edades que lo determinan, la diferencia en la fecundidad adolescente versus edades avanzadas y el corrimiento de la maternidad a edades alrededor de los 40 años y su estabilización en torno a la misma.

En base a los modelos estimados se obtienen pronósticos probabilísticos de la mortalidad y la fecundidad junto con sus intervalos de pronóstico de un 80 % de confianza. Se evalúan resultados obtenidos por MDF y modelos ARIMA (Box y Jenkins, 1976) a fin de establecer una comparación del nuevo método versus una de las metodologías de pronóstico más clásicas. Si se evalúa la amplitud de los intervalos de pronóstico obtenidos, para las tasas de mortalidad, los intervalos, en promedio, son un 47,9 % más estrechos que los obtenidos mediante modelos ARIMA, con una variación entre el 24,7 y 79,9 %. Al igual que se observó para la mortalidad, los pronósticos de la fecundidad obtenidos por MDF tienen un mejor comportamiento a largo plazo, dado que presentan una reducción en la variabilidad mediante el análisis funcional (más leve que para datos de mortalidad), esto se ve reflejado en la amplitud de los intervalos de pronóstico, siendo estos, en promedio, un 17,5 % más estrechos que los obtenidos mediante modelos ARIMA, con una variación entre el 7,49 y 30,73 %.

Al evaluar los resultados de la comparación mediante el cálculo de los MAPE, es notorio el descenso en el error que presentan los pronósticos MDF frente a los ARIMA. Para la mortalidad esta mejora de los resultados se da principalmente en varones. Para la fecundidad también se observa un descenso marcado en el MAPE calculado para

pronósticos derivados del MDF con respecto a los obtenido a partir de modelos ARIMA. Asimismo al realizar un análisis comparativo entre las hipótesis de fecundidad, la hipótesis que supone que la componente demográfica ha alcanzado un nivel estable presenta medidas de error menores en todos los años y períodos contemplados, sin embargo en relación a la amplitud de los intervalos de pronóstico los resultados son similares.

En segundo lugar se realiza el análisis de la componente migratoria. El análisis de la mortalidad y la fecundidad no presenta dificultades, en cambio al analizar los datos de migración, que provienen de una estimación, se presentan problemas. Las estimaciones de la migración neta para el período 1980-2011 presentan un comportamiento que es una combinación del verdadero comportamiento del saldo migratorio, los errores de registro en las muertes por edad y nacimientos según edad de la madre y los errores en las estimaciones y/o proyecciones de población. Como ya se destacó, las cifras relativas a población tienen un origen que se analizó en el Capítulo 4, estas son estimaciones y/o proyecciones por lo que en su proceso de obtención pueden acarrear error, como sucede con cualquier estimación. Asimismo, los nacimientos también están afectados por otra fuente de error: el subregistro; en Argentina por ejemplo, para el período 1980-1985 el subregistro ascendía al 2.1 % de los nacimientos, 3.8 % para el período 1985-1990 y 5.5 % para 1995-2000 (Bay y Orellana, 2007). Este último porcentaje en valores netos representa aproximadamente entre 14.000 y 35.000 nacimientos sin registrar. El subregistro impacta en forma directa en las estimaciones de la migración neta de las edades iniciales, en ellas se observa un alto grado de variación (de forma sinusoidal), que surge de la compensación obligada que debe ejercer el algoritmo empleado al calcular migraciones y establecer una coherencia entre nacimientos y población en las edades iniciales. Una posible hipótesis de este fenómeno es que la diferencia sustancial entre las cifras mencionadas, nacimientos y población en las primeras edades, se debe en gran parte al subregistro presente en esta franja etaria. Otro aspecto destacado es el leve aumento de la variabilidad que se observa para las edades avanzadas, probablemente producto del sesgo en los datos y de la agrupación forzada en un grupo abierto final de 80 años y

más, que como ya se destaco merecería una mayor desagregación.

Luego se aplica la metodología alternativa de pronósticos coherentes propuesta por Hyndman *et al.* (2008) con el fin de evitar una divergencia entre ambos géneros. Si bien sería posible aplicar esta metodología a datos de mortalidad y migración, esta se aplica solamente a datos de mortalidad, debido a los problemas que presentan los resultados relativos a la componente migratoria. El modelo que impone coherencia entre géneros se basa en el producto y el cociente de las tasas de mortalidad de varones y mujeres. En el modelo para el producto, la primera componente principal muestra un descenso a través del tiempo, que según el comportamiento de la base correspondiente, resulta más rápido en la infancia y más lenta en las edades cercanas a los 20 años y edades avanzadas. En relación al modelo para el cociente se observa una tendencia creciente para la primera componente, que se estabiliza hacia el año 2000, y, de acuerdo a la base correspondiente se refiere a edades entre 20 y 40 años. Este par base-componente esta asociado a la diferencia en la mortalidad entre géneros, y estaría indicando que en edades entre 20 y 40 años y mayores de 70 años, mueren "menos" las mujeres. En ambos modelos resulta complejo hallar un correlato con comportamientos demográficos teóricos. Es importante remarcar que la metodología de pronósticos coherentes produce resultados que difieren, para valores pronosticados, en la misma forma que difieren las cifras observadas. En síntesis, mantiene la coherencia entre géneros presente en las tasas a la hora de generar los pronósticos, arrojando resultados más acordes con la realidad en comparación al pronóstico de los géneros en forma separada.

Finalmente, a partir de los pronósticos de fecundidad (basados en las dos hipótesis de fecundidad) y de mortalidad (independientes y coherentes) se elaboran pronósticos estocásticos de la población para hombres y mujeres. En relación a la componente migratoria, dado que el análisis de esta componente arrojó resultados atípicos, se supone que las cantidades de migrantes y emigrantes son iguales generando así un saldo migratorio nulo. Teniendo en cuenta que los datos de fecundidad y mortalidad están disponibles

hasta 2011 y 2012, respectivamente, la simulación se basa en datos de las tres componentes demográficas para el período 1980-2011 y se toma como población base la del año 2012. En base a N=1000 trayectorias simuladas para cada componente y a través del procedimiento detallado paso a paso en la sección 3.2.4 se generan los pronósticos de población y sus intervalos de confianza.

En todos los casos evaluados las cifras estimadas se encuentran por encima de las oficiales con una diferencias promedio del 2.6 %. Se destaca que la mayor diferencia con la proyección oficial se observa para varones, consistentemente, a lo largo de las cuatro estimaciones obtenidas. La cifras encontradas a través de pronósticos coherentes con hipótesis de fecundidad estable presentan las diferencias relativas más altas tanto para el total como para ambos géneros; 2.43 %, 2.71 % y 2,16 %. Mientras que las cifras obtenidas en forma independiente para el total y cada género correspondientes a la hipótesis de fecundidad descendente son las que presentan la menor diferencia con respecto a la proyección oficial, 1.58 %, 1.82 % y 1.35 %. Como ya se mencionó, resulta importante remarcar que esta comparación no es indicador de bondad de ajuste, ya que todas las cifras que se comparan son pronósticos, en cambio da cuenta de la divergencia en las metodologías y supuestos que subyacen en los números obtenidos. La similitud entre las cifras oficiales y las obtenidas por medio de pronósticos independientes e hipótesis de fecundidad descendente radica, posiblemente, en que en ambos procesos no se impone una coherencia entre las cifras de varones y mujeres, además en ambos casos se supone una fecundidad descendente; el INDEC en su proyección oficial, por medio de una función logística y en base a la tendencia histórica que presenta la Tasa Global de Fecundidad (TGF) desde al año 81, genera tasas proyectadas hasta el año 2040. Estas proyecciones conforman el supuesto de la fecundidad que subyace a la población obtenida. Si bien el documento técnico que detalla la metodología empleada para la elaboración de las proyecciones destaca que esta componente presenta un descenso mucho más atenuado que el observado en décadas pasadas y en base a ello recomienda tener cautela.

En síntesis, al realizar un pronóstico de población es importante tener en cuenta múltiples fuentes de variación. El método que se utiliza en esta tesis comprende las fuentes más importantes de variabilidad: la variabilidad observada a través del modelo Poisson que genera las muertes y nacimientos y la variación dinámica de las tasas a través de los modelos de series de tiempo. Este aspecto garantiza una representación más completa que la mayoría de los enfoques propuestos para este tipo de datos. En los pronósticos de población, las correlaciones entre los errores de pronóstico de las componentes demográficas afecta la amplitud de los intervalos de predicción, en este enfoque las correlaciones a través de las edades se tienen en cuenta mediante el uso de suavizados y la correlación temporal mediante el uso de modelos para series de tiempo. Estas correlaciones se incorporan naturalmente en las simulaciones de cada componente demográfica. De todos modos, dado que los métodos de series de tiempo que se utilizan en esta tesis son útiles para pronosticar a corto y a mediano plazo, la amplitud de los intervalos muestra que los métodos se vuelven menos informativos a lo largo del tiempo y recomiendan no utilizarlos con un horizonte mayor a los 20 años (Hyndman y Ullah, 2007).

Los resultados aquí obtenidos presentan medidas adecuadas de bondad de ajuste, cifras pronosticadas coherentes con la teoría demográfica y principalmente garantizan la consistencia probabilística. Las tasas y cifras de población pronosticadas están acompañadas de su intervalo de pronóstico, anexando de este modo una cuantificación de la incertidumbre asociada al resultado informado.

Si bien no se lleva adelante en el presente trabajo, es posible, a través de la utilización de métodos estándar de tablas de vida (Preston *et al.*, 2001), obtener pronósticos de la esperanza de vida a partir de los pronósticos de las tasas de mortalidad específicas por edad que genera el MDF. Para calcular los intervalos de pronóstico para la esperanza de vida se simula un gran número de tasas de mortalidad futuras, como se describe en Hyndman *et al.* (2008), y se obtienen las esperanzas de vida para cada una. Luego los

intervalos de pronósticos se construyen a partir de los percentiles de las esperanzas de vida simuladas. Por otro lado, a partir de los datos de fecundidad, es posible obtener la tasa global de fecundidad pronosticada. En futuros trabajos se desarrollará el cálculo de estas y otras medidas derivadas que surgen de la aplicación de los MDF. Por último sería importante investigar el desempeño del método propuesto en esta tesis en otros países de la región con características demográficas disimiles.

Bibliografía

Alho, J. y Spencer, B. (1985). Uncertain population forecasting. *Journal of the American Statistical Association*, 80:306–314.

Arriaga, E. (2003). *El Análisis de la Población con Microcomputadoras*. Doctorado en Demografía. Facultad de Ciencias Económicas, Universidad Nacional de Córdoba.

Bay, G. y Orellana, H. (2007). La calidad de las estadísticas vitales en la américa latina (versión preliminar para discusión). Presentado en Taller de expertos en el uso de estadísticas vitales: alcances y limitaciones. Santiago, Chile 13-14 de diciembre 2007.

Beer, J. D. (1997). The effect of uncertainty of migration on national population forecasts: the case of the Netherlands. *Journal of Official Statistics*, 13:227–243.

Beers, H. (1945). Six term formulas for routine actuarial interpolation. *The Record of the American Institute of Actuaries*, 34(68).

Bongaarts, J. y Bulatao, R. A. (2000). *Beyond Six Billion: Forecasting the World's Population. Panel on Population Projections, Committee on Population, National Research Council*. NATIONAL ACADEMY PRESS.

Booth, H., Hyndman, R., Tickle, L., y de Jong, P. (2006a). Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research*, 15(9):289–310.

Booth, H., Hyndman, R. J., Tickle, L., y de Jong, P. (2006b). Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions.

Booth, H., Maindonald, J., y Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, 56(3):325–336.

Booth, H., Tickle, L., y Smith, L. (2005). Evaluation of the variants of the Lee-Carter method of forecasting mortality: A multi-country comparison. *New Zealand Population Review*, 31(1):13–34.

Box, G. y Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.

Box, G. y Jenkins, G. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Box, G., Jenkins, G., y Reinsel, G. (2008). *Time Series Analysis: Forecasting and Control*. Wiley, Hoboken, New Jersey.

Bozik, J. y Bell, W. (1987). Forecasting age-specific fertility using principal components.

Brillinger, D. (1986). The natural variability of vital rates and associated statistics. *Biometrics*, (42):693–734.

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

CELADE (2000). Boletín demográfico n66. américa latina: Población por año calendario y edad simple. 1995-2005.

CELADE (2003). Boletín demográfico n71. américa latina: Estimaciones y proyecciones de población. 1995-2005.

CELADE (2004). Boletín demográfico n73. américa latina: Estimaciones y proyecciones de población. 1950-2050.

CELADE (2005). Cambios en la estructura poblacional: Una pirámide que exige nuevas miradas. *Temas de Población y Desarrollo*, (1).

de Boor, C. (1978). *A Practical Guide to Splines*. Springer, New York.

de Estadísticas de Salud del Ministerio de Salud y Acción Social de la Nación, D. (1986). Estadisticas vitales - información básica años 1982.

de Población, C. D. (2007). Observatorio demográfico. proyeccion de población. 1950-2050.

DEIS (1984). Estadisticas vitales - información básica años 1980 - 1981.

DEIS (1987). Estadisticas vitales - información básica años 1983.

DEIS (1988). Estadisticas vitales - información básica años 1984 - 1985.

DEIS (1989). Estadisticas vitales - información básica años 1986.

DEIS (1990). Estadisticas vitales - información básica años 1987.

DEIS (1991a). Estadisticas vitales - información básica años 1988.

DEIS (1991b). Estadisticas vitales - información básica años 1989.

DEIS (1992). Estadisticas vitales - información básica años 1990.

DEIS (1993). Estadisticas vitales - información básica años 1991.

DEIS (1994a). Estadisticas vitales - información básica años 1982.

DEIS (1994b). Estadisticas vitales - información básica años 1993.

DEIS (1995). Estadisticas vitales - información básica años 1994.

DEIS (1996). Estadisticas vitales - información básica años 1995.

DEIS (1997). Estadisticas vitales - información básica años 1996.

DEIS (1998). Estadisticas vitales - información básica años 1997.

DEIS (1999). Estadisticas vitales - información básica años 1998.

DEIS (2000). Estadisticas vitales - información básica años 1999.

DEIS (2001). Estadisticas vitales - información básica años 2000.

DEIS (2002). Estadisticas vitales - información básica años 2001.

DEIS (2003). Estadisticas vitales - información básica años 2002.

DEIS (2004). Estadisticas vitales - información básica años 2003.

DEIS (2005). Estadisticas vitales - información básica años 2004.

DEIS (2006). Estadisticas vitales - información básica años 2005.

DEIS (2007). Estadisticas vitales - información básica años 2006.

DEIS (2008). Estadisticas vitales - información básica años 2007.

DEIS (2009). Estadisticas vitales - información básica años 2008.

DEIS (2010). Estadisticas vitales - información básica años 2009.

DEIS (2011). Estadisticas vitales - información básica años 2010.

DEIS (2012). Estadisticas vitales - información básica años 2011.

DEIS (2013). Estadisticas vitales - información básica años 2012.

Del Popolo, F. (2000). Los problemas en la declaración de la edad de la población adulta mayor en los censos. *Serie población y desarrollo - CELADE*, 8.

Eilers, P. y Brian, M. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11:98–102.

Erbas, B., Hyndman, R., y Gertig, D. (2007). Forecasting age-specific breast cancer mortality using functional data models. *Statistics in Medicine*, 2(26):458–470.

George, M. y Perreault, J. (1992). Methods of external migration projections and forecasts. *National population forecasting in industrialized countries*, p. 87103.

Granger, C. W. J. y Joyeux, R. (1980). An introduction to long memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1(1):15–29.

Hardin, J., Hilbe, J., y Hilbe, J. (2007). *Generalized Linear Models and Extensions, Second Edition*. A Stata Press publication. Taylor & Francis.

Haslett, J. y Raftery, A. (1989). Space-time modelling with long-memory dependence: Assessing ireland's wind power resource (with discussion). *Journal of the Royal Statistical Society, series C - Applied Statistics*, 38:1–50.

Hastie, T. J. y Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

He, X. y Hg, P. (1999). Cobs: qualitatively constrained smoothing via linear programming. *Computational Statistics*, 14:315–337.

He, X. y Shi, P. (1996). Monotone b spline smoothing. *Journal of the American Statistical Asociation*, 93:643–650.

Hilderink, H., der Gaag, N. V., Wissen, L. V., Jennissen, R., Román, A., Salt, J., Clarke, J., y Pinkerton, C. (2002). Analysis and forecasting of international migration by major groups, part iii. *Working papers and studies*, 3.

Hosking, J. (1981). Fractional differencing. *Biometrika*, 68(1):165–176.

Hyndman, R. y Booth, H. (2007). Stochastic population forecasts using functional data models for mortality, fertility and migration.

Hyndman, R., Booth, H., y Yasmeen, F. (2013). Coherent Mortality Forecasting: The Product-Ratio Method With Functional Time Series Models. *Demography*, 50(1):261–283.

Hyndman, R. y Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for r. *Journal of Statistical Software*, 3(27).

Hyndman, R., Koehler, A., Ord, J., y Snyder, R. (2008). *Forecasting with Exponential Smoothing The State Space Approach*. Springerlink.

Hyndman, R. y Shang, H. (2009). Forecasting functional time series (with discussion). *Journal of the Korean Statistical Society*, 38(3):199–221.

Hyndman, R. y Shang, H. (2010). Rainbow plots, bagplots and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19:29–45.

Hyndman, R. y Ullah, M. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis*, 51:4942–4956.

INDEC (2005). Proyecciones provinciales de población por sexo y grupos de edad: 2001-2015.

INDEC-CELADE (1982). Argentina: Estimaciones y proyecciones de población: 1950-2025 (versión preliminar).

INDEC-CELADE (1995). Proyecciones provinciales de población por sexo y grupos de edad: 1950-2050.

INDEC-CELADE (2013). Estimaciones y proyecciones de población 2010-2040: total del país.

Jong, P. D. y Tickle, L. (2006). Extending Lee-Carter mortality forecasting. *Mathematical Population Studies*, 13(1):1–18.

Kalben, B. (2000). Why men die younger. *North American Actuarial Journal*, 4(4):83–111.

Ke, C. y Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications. *Journal of American Statistist Association*, 96:272–1281.

Keilman, N. y Pham, D. (2004). Empirical errors and predicted errors in fertility, mortality and migration forecasts in the european economic area. *Discussion Papers, Research Department of Statistics Norway*, 386.

Keilman, N., Pham, D., y Hetland, A. (2002). Why population forecasts should be probabilistic - illustrated by the case of Norway. *Demographic Research*, 6(15):409–454.

Koenker, R. (1994). *Confidence Intervals for Regression Quantiles*. Contributions to Statistics. Physica Verlag HD.

Lee, R. (1993). Modeling and forecasting the time series of us fertility: age distribution, range, and ultimate level. *International Journal of forecasting*, 9:187–202.

Lee, R. (2000). The Lee-Carter method for forecasting mortality, with various extensions and applications. *North American actuarial journal*, 4(1):80–91.

Lee, R. (2006). *Mortality Forecasts and Linear Life Expectancy Trends*, pp. 19–40. National Social Insurance Board, Stockholm.

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

Lee, R. y Miller, T. (2001). Evaluating the performance of the Lee-Carter Method for Forecasting Mortality. *Demography*, 38:537–549.

Lee, R. y Tuljapurkar, S. (1994). Stochastic population forecasts for the United States: Beyond high, medium, and low. *Journal of the American Statistical Association*, 89:1175–1189.

Li, S. y Chan, W. (2005). Outlier analysis and mortality forecasting: The United Kingdom and scandinavian countries. *Scandinavian Actuarial Journal*, 2005(3):187–211.

Miller, T. (2003). California s uncertain population future. technical appendix. *Special Report: the growth and aging of California s population: demographic and fiscal prjections, characteristics and service needs. center of Economics and Demography of aging. CEDA.*

Miller, T. y Lee, R. (2001). Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography*, 38(4):537–549.

Miller, T. y Lee, R. (2004). A probabilistic forecast of net migration to the united states. *Report I in Stochastic infinite horizon forecasts for social security and related studies*, 10917.

Muller, H. (2006). Functional Modeling of Longitudinal Data.

NU (2000). Replacement migration. is it a solution to declining and ageing populations?

Pantelides, E. A. y Rofman, A. (1983). La transición demográfica argentina: un modelo no ortodoxo. *Desarrollo Económico*, 22(88):511–534.

Peiris, M. y Perera, B. (1988). On prediction with fractionally differenced arima models. *Journal Of Time Series Analysis*, 9(3):215–220.

Preston, S., Heuveline, P., y Guillot, M. (2001). *Demography, Measuring and Modelling Population Processes*. Blackwell, Oxford.

Ramsay, J., Silverman, B., Liu, A., Belliveau, J., y Dale, J. (1997). Functional Data Analysis.

Ramsay, J. O. y Silverman, B. W. (2005). *Functional data analysis 2nd ed.* Springer, New York.

Renshaw, A. y Haberman, S. (2003a). Lee-Carter mortality forecasting: A parallel generalized linear modelling approach for england and wales mortality projections. *Applied Statistics*, 52(1):119–137.

Renshaw, A. y Haberman, S. (2003b). Lee-Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, 33(2):255–272.

Renshaw, A. y Haberman, S. (2003c). On the forecasting of mortality reduction factors. *Insurance: Mathematics and Economics*, 32(3):379–401.

R.H., S. y D.S., S. (2010). *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer.

Rice, J. y Wu, C. (2000). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 50:253–259.

Rogers, A. (1990). Requiem for the net migrant. *Geographical Analysis*, 22:283–300.

Rogers, A. y Castro, L. (1981). Model migration schedules. *Research Report, International Institute for Applied Systems Analysis*, pp. 81–30.

Rogers, A. y Little, J. S. (1994). Parameterizing age patterns of demographic rates with the multiexponential model schedule. *Mathematical Population Studies*, 4:175–195.

Ruppert, D., Wand, M., y Carrol, R. (1976). *Semiparametric regression.* Oxford University Press, New York.

Serfaty, E., Foglia, L., Masaútis, A., y Negri, G. (2007). Mortalidad por causas violentas en adolescentes y jóvenes de 10- 24 años. *Rev Vertex.*, 40:25–30.

Simonoff, J. (1976). *Smoothing methods in statistics.* Springer-Verlag, New York.

Sykes, Z. (1969). Some stochastic versions of the matrix model for population dynamics. *Journal of the American Statistical Association*, 44:111–130.

Wilmoth, J. (1993). Computational methods for fitting and extrapolating the Lee–Carter model of mortality change.technical report. department of demography. university of california, berkeley.

Wood, S. (1994). Monotonic smoothing splines fitted by cross validation. *SIAM J. Sci Comput*, 15(5):1126–1133.

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society*, Serie B 65(1):95–114.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R.* Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Yao, F. y Lee, T. (2006). Penalized spline models for functional principal component analysis. *Royal Statistical Society B*, 68:325.

Yao, F., Muller, H. G., y Wang, J. (2009). Functional linear regression analysis for longitudinal data. *Annals of Statistics*, 33:2873–2903.

Apéndice A

Técnicas empleadas para el suavizado: *Splines*

Los *splines* son regresiones por partes unidas por puntos llamados nodos. En su forma más simple, los *splines* son modelos de regresión que incluyen un conjunto de variables *dummy* como variables independientes, cuya finalidad es "forzar" al ajuste de la regresión a cambiar la dirección en ciertos puntos a lo largo del rango de las abscisas. Además es necesario fijar el grado polinómico de las funciones por tramo y la cantidad y ubicación de los nodos.

Los *splines* de regresión simple no son adecuados en la mayoría de los problemas aplicados. Es altamente restrictivo estimar únicamente funciones lineales entre nodos cuando el objetivo principal es estimar formas funcionales más curvilíneas. La solución es combinar funciones de regresión por tramos con regresión polinómica. Los polinomios por tramo poseen dos ventajas; admiten funciones lineales entre los nodos, y en segundo lugar aseguran que la primera derivada esté definida en los nodos, garantizando que la estimación *spline* no tenga esquinas "afiladas" o "agudas".

En los *splines* cúbicos la matriz del modelo tiende a estar altamente correlacionada, dado que cada columna es una versión transformada de la x, induciendo una considerable colinearidad. La colinearidad puede generar una matriz del modelo cercana a la no singularidad e imprecisión en el ajuste *spline*. Para solucionar este inconveniente puede representarse un *spline* cúbico (y en cualquier otra base polinómica) como una base

B-*spline*. Una base B-*spline* de k nodos puede definirse como sigue:

$$f(x) = \sum_{i=1}^{k} B_i^2(x)\beta_i, \tag{1}$$

donde las funciones base se definen como:

$$B_i^2(x) = \frac{x - c_i}{c_{i+2+1} - c_i} B_i^{2-1}(x) + \frac{c_{i+2+1} - x}{c_{i+2+1} - c_{i+1}} B_{i+1}^{2-1}(x) \tag{2}$$

y

$$B_i^{-1}(x) \begin{cases} 1 & \text{si } c_i \leq x \leq c_{i+1} \\ 0 & \text{en otro caso} \end{cases} \tag{3}$$

donde con c_i se indican los nodos.

La función base B-*spline* es en esencia un "re-escalamiento" de cada una de las funciones por tramo. La idea es similar al "re-escalamiento" de un conjunto de variables restándoles la media para reducir la colinealidad. El re-escalamiento en la base B-*spline* reduce la colinealidad en las funciones base de la matriz del modelo. El *spline* resultante es numéricamente más estable que el spline cúbico. Esto es particularmente cierto si se utilizan una gran cantidad de nodos y Mínimos Cuadrados Ordinarios en el ajuste. Ver de Boor (1978) y Eilers y Brian (1996).

Por otro lado existen los *splines* penalizados, una técnica de regresión no paramétrica que depende de los principios de la teoría estadística y cuya finalidad es minimizar la posibilidad de sobreajuste. Un modelo estadístico sobre-ajustado posee demasiados parámetros en relación a la cantidad de datos empleados para su estimación y esto provoca que la variación aleatoria en los datos sea vista como efectos sistemáticos. La solución esta en la reducción del número de parámetros, de acuerdo a algún criterio específico, el Criterio de Información de Akaike, por ejemplo, el cuál puede utilizarse para el caso de los *splines*. Otra solución es la estimación penalizada, de modo que se agrega una penalización al modelo por cada parámetro que se incluye en el mismo.

Dado que los modelos *spline* se estiman por medio de Mínimos Cuadrados, comparten propiedades con los modelos de regresión lineal. Esto implica que el *spline* estimado, $\hat{f}$, minimiza la suma de cuadrados entre y y la estimación no paramétrica, $f(x_i)$

$$SS(f) = \sum_{i=1}^{n} [y - f(x)]^2.$$

El punto radica en que estimar f que minimice la ecuación anterior puede implicar el uso de demasiados parámetros. Luego de minimiza $SS(f)$ pero sujeto a una restricción o término penalizado de acuerdo al número de parámetros utilizados. La penalización planteada para los modelos *spline* es

$$\lambda \int_{x_1}^{x_n} [f''(x)]^2 \, dx, \tag{4}$$

y la estimación *spline* resulta

$$SS(f, \lambda) = \sum_{i=1}^{n} [y - f(x)]^2 + \lambda \int_{x_1}^{x_n} [f''(x)]^2 \, dx. \tag{5}$$

Al término que se agrega a la fórmula básica se lo conoce como "penalización por rugosidad" y la misma consta de dos partes. En primer lugar, λ es llamada habitualmente parámetro de suavizado y la segunda es la integral del cuadrado de la derivada segunda de $f(x)$, esta inclusión resulta lógica dado que la derivada segunda mide la tasa de cambio en la pendiente para una función o curva. Un valor alto de la derivada segunda implica mayor curvatura y viceversa. Mediante el uso de la integral del cuadrado se tiene en cuenta una medida de curvatura a lo largo del rango completo de la estimación no paramétrica, cuando es grande, $f(x)$ es más rugosa, y cuando es más pequeña $f(x)$ es más suave.

El parámetro λ, no negativo, controla directamente la cantidad de peso dada a la derivada segunda como medida de la suavidad de f. Por lo tanto, λ establece una compensación entre la proximidad del ajuste a los datos y la penalización, con una función

análoga a la amplitud de la ventana en un suavizado por *loess*. A medida que λ decrece $\hat{f}(x)$ interpola los datos y se obtiene una estimación más "rugosa". A medida que $\lambda \to \infty$ la integral de la derivada segunda está restringida a cero y el resultado es un ajuste suavizado por mínimos cuadrados. De un modo más general, cuando λ aumenta se obtiene un ajuste más suave de los datos pero probablemente más sesgado, y a medida que el parámetro decrece el ajuste muestra menos sesgo pero un aumento en la variancia.

Mientras un valor pequeño de λ es cercano a hacer una interpolación de los datos, un valor grande devuelve un ajuste por mínimos cuadrados, valores intermedios de λ no tienen una efecto interpretable en la magnitud del suavizado aplicado a los datos. Para resolver este problema se puede transformar el parámetro para que sea una aproximación de los grados de libertad, lo que es lógico desde que el parámetro controla el número de parámetros locales utilizados en la estimación no paramétrica. Los grados de libertad de los *splines* de suavizado se obtienen de un modo similar a como se obtienen para modelos estimados por Mínimos Cuadrados Ordinarios o regresión local polinómica, pero la penalización agrega algunas complicaciones. En un modelo de regresión lineal el número de parámetros es igual a la $tr(\mathbf{H})$ donde $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Para un modelo *spline* estándar, el cual se basa en una matriz del modelo alterada y un ajuste por mínimos cuadrados, los grados de libertad también se obtiene a partir de $tr(\mathbf{H})$.

Si se escribe la ecuación 5 en forma matricial

$$SS(f, \lambda) = \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\beta'\mathbf{D}\beta, \tag{6}$$

donde

$$\mathbf{D} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times k} \\ \mathbf{0}_{k\times 2} & \mathbf{0}_{k\times k} \end{bmatrix}, \tag{7}$$

siendo k el número de nodos y la matriz *hat* de una regresión lineal estándar se altera para incluir el término de la penalización. Ruppert, Wand y Caroll (2003) derivaron la

matriz *hat* o matriz de suavizado para *splines* penalizados:

$$\mathbf{S}_\lambda = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda^{2p}\mathbf{D})^{-1}\mathbf{X}', \tag{8}$$

donde p es el orden del polinomio de la función base. En esta forma el término de penalización es un escalar λ^{2p} multiplicado por la matriz $\mathbf{D}$. La traza de $\mathbf{S}_\lambda$, tal como sucede en regresión lineal representa el número de grados de libertad del modelo *spline* y es aproximadamente equivalente al número de parámetros en el ajuste. Se fijan luego los valores de λ y de p y se construye una matriz del modelo con un conjunto de funciones base idénticas a las utilizadas en un *spline* estándar. Utilizando estos valores se forma la matriz $\mathbf{S}_\lambda$ y se la aplica al vector de observaciones para obtener los predichos

$$\hat{\mathbf{y}} = \mathbf{S}_\lambda \mathbf{y}. \tag{9}$$

Las funciones predichas que se obtienen son la estimación *spline* penalizada. La simplicidad de la matriz del *spline* de suavizado hace que la programación de este tipo de modelos sea levemente más compleja que la de un *spline* cúbico. El proceso es el mismo que el de un *spline* de regresión en el cuál la y se regresa sobre una matriz del modelo utilizando mínimos cuadrados. De todos modos, aunque es posible la estimación por mínimos cuadrados, este método es a menudo numéricamente inestable y es por ello que se emplean métodos tales como el de matriz ortogonal o descomposición de Cholesky o espectral.

Restricciones

Las técnicas más populares de suavizado tienen dificultades para incluir restricciones cualitativas tales como la monotonía, la convexidad o condiciones límite sobre las funciones ajustadas. (He y Hg, 1999) proponen entonces un suavizado B-*Spline* con restricciones (COBS), el mismo incorpora información previa y por ello a menudo mejora la eficiencia de los estimadores. Se define inicialmente la función cuantil $g_\tau(x)$ de Y dado

$X = x$ de modo que

$$P(Y \leq g_\tau(x) \mid X = x) = \tau. \tag{10}$$

Luego, la función mediana condicional ($\tau = 0{,}5$) provee una medida de tendencia central y puede ser usada para describir la relación general entre X e Y. Dado n pares de realizaciones $\{(x_i, y_i)\}_{i=1}^n$con $a = x_0 < x_1 < ... < x_n < x_{n+1} = b$, una función suave f y una función de verificación $\rho_\tau(u) = 2[\tau - I(u < 0)]u$ donde $I(\cdot)$ es una función indicadora y se define la "fidelidad" a los datos como

$$\text{“}fidelidad\text{''} = \sum_{i=1}^{n} \rho_\tau(y_i - f(x_i)). \tag{11}$$

Koenker (1994) presentan el suavizado *spline* por cuantiles τ-ésimo L_p, el cuál es la solución a

$$\min_g \text{ “fidelidad” } + \lambda \text{ “rugosidad } L_p\text{”}$$

resulta en el estimador semi-paramétrico de $f_\tau(x)$.

Para la definición de la metodología de interés los autores se centran en el caso especial que $\tau = 0{,}5$ y de este modo la "fidelidad" se mide a través de la norma L_1

$$\text{“fidelidad”} = \textstyle\sum_{i=1}^{n} \mid y_i - g(x_i) \mid.$$

Es sabido que cualquier suavizado por *spline* de orden m con nodos $x_1, ..., x_n$ tiene una representación B-*spline* equivalente en la misma secuencia de nodos. Una malla de nodos más general resulta $T = t_{i\,i=1}^{N+2m}$ con $t_1 = ... = t_m < t_{m+1} < ... < t_{N+m} < t_{N+m} < t_{N+m+1} = ... = t_{N+2m}$. Esta generalización esta relacionada con consideraciones relativas a la eficiencia computacional. Luego, la representación B-*spline* resulta

$$s(x) = \textstyle\sum_{j=1}^{N+m} a_j B_j(x),$$

donde N es el número de nodos internos, $B_j(x)$ son funciones base B-*Spline* normalizadas, a_j son los coeficientes de las funciones base y $S_{m,T}$ es el espacio de *splines*

polinómicas de orden m con malla T. Ver de Boor (1978) para más detalles.

En el caso de las B-*splines* lineales, la función objetivo puede escribirse como

$$\min_{\theta \in R^{N+m}} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{N+m} a_j B_j(x) \right| + \lambda \sum_{i=1}^{N} \left| \sum_{j=1}^{N+m} a_j B'_j(t_{i+m}) - \sum_{j=1}^{N+m} a_j B'_j(t_{i+m-1}) \right|, \quad (12)$$

donde $\theta = (a_1, ...a_{N+m})$. Y en forma compacta

$$\min_{\theta \in R^{N+m}} \sum_{i=1}^{N+m} |\tilde{y}_i - \tilde{x}_i \theta|, \quad (13)$$

donde

$$\tilde{y} = \begin{pmatrix} y \\ \mathbf{0} \end{pmatrix}, \quad (14)$$

es un vector de *pseudo respuesta* de dimensiones $(n + N) \times 1$,

$$\tilde{X} = \begin{pmatrix} \mathbf{B} \\ \lambda \mathbf{C} \end{pmatrix}, \quad (15)$$

es una matriz de *pseudo diseño* de dimensiones $(n + N) \times (n + m)$, con

$$\mathbf{B} = \begin{bmatrix} B_1(x_1) & \cdots & B_{N+m}(x_1) \\ \vdots & \cdots & \vdots \\ B_1(x_n) & \cdots & B_{N+m}(x_n) \end{bmatrix}, \quad (16)$$

y

$$\mathbf{C} = \begin{bmatrix} B'_1(t_{m+1}) - B'_1(t_m) & \cdots & B'_{N+m}(t_{m+1}) - B'_{N+m}(t_m) \\ \vdots & \cdots & \vdots \\ B'_1(t_{N+m}) - B'_1(t_{N+m-1}) & \cdots & B'_{N+m}(t_{N+m}) - B'_{N+m}(t_{N+m-1}) \end{bmatrix}, \quad (17)$$

y la curva ajustada, $\hat{m}_{\lambda,L_1} = \sum_{j=1}^{N+m} \hat{a}_j B_j(x)$ es un suavizado B-*spline* lineal.

Para imponer restricciones adicionales de índole cualitativa sobre las curvas ajustadas se agregan desigualdades o igualdades al ajuste.

Monotonía

Para el *spline* lineal se agrega

$$\mathbf{H}\theta \geq \mathbf{0}, \tag{18}$$

para funciones crecientes, y

$$\mathbf{H}\theta \leq \mathbf{0}, \tag{19}$$

para funciones decrecientes, donde

$$\mathbf{B} = \begin{bmatrix} B_1'(t_m) & \cdots & B_{N+m}'(t_m) \\ \vdots & \cdots & \vdots \\ B_1'(t_{N+m+1}) & \cdots & B_{N+m}(t_{N+m+1}) \end{bmatrix}, \tag{20}$$

Concavidad

Para la *spline* lineal se imponen N restricciones

$$\mathbf{C}\theta \geq \mathbf{0}, \tag{21}$$

mientras que para fijar concavidad se revierte la desigualdad.

Apéndice B

Modelos

Modelos ajustados para el pronóstico de los coeficientes $\hat{\beta}$ del modelo de datos funcionales.

Mortalidad

Mujeres

Coeficiente $\hat{\beta}_1$: ARIMA(1,1,0) con pendiente.

$\hat{\phi}_1$= -0.4809 (0.1543), $\hat{\mu}$= -0.146 (0.028) y $\hat{\sigma}^2$= 0.05404

AIC=2.79, AICc=3.64 y BIC=7.18.

Coeficiente $\hat{\beta}_2$: ARIMA(1,0,0).

$\hat{\phi}_1$= 0.7914 (0.1182) y $\hat{\sigma}^2$= 0.01929

AIC=-31.65, AICc=-31.25 y BIC=-28.66.

Coeficiente $\hat{\beta}_3$: ARIMA(1,0,0).

$\hat{\phi}_1$= 0.6562 (0.1319) y $\hat{\sigma}^2$= 0.02025

AIC=-30.47, AICc=-30.07 y BIC=-27.47.

Coeficiente $\hat{\beta}_4$: ARIMA(0,0,1).

$\hat{\theta}_1$= 0.4911 (0.1594) y $\hat{\sigma}^2$= 0.01472

AIC=-41.28, AICc=-40.88 y BIC=-38.29.

Coeficiente $\hat{\beta}_5$: ARIMA(0,0,0) con media cero.

$\hat{\sigma}^2$= 0.01174

AIC=-51.02, AICc=-50.89 y BIC=-49.52.

Coeficiente $\hat{\beta}_6$: ARIMA(1,0,0).

$\hat{\phi}_1$= 0.2800 (0.1651) y $\hat{\sigma}^2$= 0.008035

AIC=-61.46, AICc=-61.06 y BIC=-58.47.

Varones

Coeficiente $\hat{\beta}_1$: ARIMA(0,1,1) con pendiente.

$\hat{\theta}_1$= -0.7365 (0.1783), $\hat{\mu}$= -0.1299 (0.0110) y $\hat{\sigma}^2$= 0.04216

AIC=-4.4, AICc=-3.55 y BIC=-0.01.

Coeficiente $\hat{\beta}_2$: ARIMA(2,0,1).

$\hat{\phi}_1$= 1.7392 (0.0842), $\hat{\phi}_2$= -0.8321 (0.0808), $\hat{\theta}_1$= -0.9173 (0.1147) y $\hat{\sigma}^2$= 0.02468

AIC=-18.51, AICc=-17.09 y BIC=-12.53.

Coeficiente $\hat{\beta}_3$: ARIMA(0,0,1).

$\hat{\theta}_1$= 0.4732 (0.1661) y $\hat{\sigma}^2$= 0.02163

AIC=-28.61, AICc=-28.21 y BIC=-25.62.

Coeficiente $\hat{\beta}_4$: ARIMA(0,0,1).

$\hat{\theta}_1$= 0.5684 (0.2151) y $\hat{\sigma}^2$= 0.01195

AIC=-48.06, AICc=-47.66 y BIC=-45.06.

Coeficiente $\hat{\beta}_5$: ARIMA(1,0,0).

$\hat{\phi}_1$= 0.4096 (0.1600) y $\hat{\sigma}^2$= 0.009365

AIC=-63.72, AICc=-63.6 y BIC=-62.23.

Coeficiente $\hat{\beta}_6$: ARIMA(0,0,0) con media cero.

$\hat{\sigma}^2$= 0.00799

AIC=-63.72, AICc=-63.6 y BIC=-62.23.

Total

Coeficiente $\hat{\beta}_1$: ARIMA(0,1,1) con pendiente.

$\hat{\theta}_1$= -0.6185 (0.1671), $\hat{\mu}$= -0.1309 (0.0148) y $\hat{\sigma}^2$= 0.04202

AIC=-4.8, AICc=-3.95 y BIC=-0.41.

Coeficiente $\hat{\beta}_2$: ARIMA(1,0,0).

$\hat{\phi}_1$= 0.7862 (0.1049) y $\hat{\sigma}^2$= 0.02411

AIC=-24.32, AICc=-23.92 y BIC=-21.33.

Coeficiente $\hat{\beta}_3$: ARIMA(1,0,0).

$\hat{\phi}_1$= 0.6432 (0.1395) y $\hat{\sigma}^2$= 0.01583

AIC=-38.64, AICc=-38.24 y BIC=-35.64.

Coeficiente $\hat{\beta}_4$: ARIMA(1,0,0).

$\hat{\phi}_1$= 0.4973 (0.1670) y $\hat{\sigma}^2$= 0.009253

AIC=-56.6, AICc=-56.2 y BIC=-53.61.

Coeficiente $\hat{\beta}_5$: ARIMA(1,0,0).

$\hat{\phi}_1$= 0.3976 (0.1582) y $\hat{\sigma}^2$= 0.008097

AIC=-61.11, AICc=-60.71 y BIC=-58.12.

Coeficiente $\hat{\beta}_6$: ARIMA(0,0,0) con media cero.

$\hat{\sigma}^2$= 0.006416

AIC=-70.96, AICc=-70.83 y BIC=-69.47.

Fecundidad

Modelo 1

Coeficiente $\hat{\beta}_1$: ARIMA(1,1,0).

$\hat{\phi}_1$= -0.3160 (0.1668) y $\hat{\sigma}^2$= 0.7932

AIC=84.9, AICc=85.33 y BIC=87.77.

Coeficiente $\hat{\beta}_2$: ARIMA(0,0,0) con media cero.

$\hat{\sigma}^2$= 0.3941

AIC=63.02, AICc=-31.25 y BIC=-28.66.

Coeficiente $\hat{\beta}_3$: ARIMA(1,0,0).

$\hat{\phi}_1$= 0.4585 (0.1533) y $\hat{\sigma}^2$= 0.05986

AIC=4.95, AICc=5.36 y BIC=7.88.

Coeficiente $\hat{\beta}_4$: ARIMA(2,1,0).

$\hat{\phi}_1$= -0.9646 (0.1648), $\hat{\phi}_2$= -0.7024 (0.1604) y $\hat{\sigma}^2$= 0.02267

AIC=-21.66, AICc=-20.78 y BIC=17.36.

Coeficiente $\hat{\beta}_5$: ARIMA(0,0,1) con media cero.

$\hat{\theta}$=0.5166 (0.1382)y $\hat{\sigma}^2$= 0.02437

AIC=-23.74, AICc=-23.33 y BIC=-20.81.

Coeficiente $\hat{\beta}_6$: ARIMA(0,1,1).

$\hat{\theta}_1$= -0.7750 (0.1155) y $\hat{\sigma}^2$= 0.01237

AIC=-43.26, AICc=-42.83 y BIC=-40.39.

Modelo 2

Coeficiente $\hat{\beta}_1$: ARIMA(0,1,1) con pendiente.

$\hat{\theta}_1$= -0.5091 (0.1834), $\hat{\mu}$= -0.2239 (0.0783) y $\hat{\sigma}^2$= 0.7064

AIC=80.02, AICc=80.9 y BIC=84.32.

Coeficiente $\hat{\beta}_2$: ARIMA(0,0,0) con media cero.

$\hat{\sigma}^2$= 0.3941

AIC=63.02, AICc=63.15 y BIC=64.48.

Coeficiente $\hat{\beta}_3$: ARIMA(1,0,0).

$\hat{\phi}_1$= 0.4585 (0.1533) y $\hat{\sigma}^2$= 0.05986

AIC=4.95, AICc=5.36 y BIC=7.88.

Coeficiente $\hat{\beta}_4$: ARIMA(0,1,2).

$\hat{\theta}_1$= -1.1568 (0.1933), $\hat{\theta}_2$= 0.5671 (0.1666) y $\hat{\sigma}^2$= 0.02407

AIC=-20, AICc=-19.11 y BIC=-15.69.

Coeficiente $\hat{\beta}_5$: ARIMA(0,0,1) con media cero.

$\hat{\theta}_1$= 0.5166 (0.1382) y $\hat{\sigma}^2$= 0.02437

AIC=-23.74, AICc=-23.33 y BIC=-20.81.

Coeficiente $\hat{\beta}_6$: ARIMA(2,1,0).

$\hat{\phi}_1$= -0.7395 (0.1947), $\hat{\phi}_2$= -0.4112 (0.1910) y $\hat{\sigma}^2$= 0.01326

AIC=-39.34, AICc=-38.45 y BIC=-35.03.

Mortalidad - Coherentes

Producto

Coeficiente $\hat{\beta}_1$: ARIMA(0,1,1) con pendiente.

$\hat{\theta}_1$= -0.5996 (0.1717), $\hat{\mu}$= -0.1399 (0.0161) y $\hat{\sigma}^2$= 0.04624

AIC=-1.87, AICc=-1.01 y BIC=2.53.

Coeficiente $\hat{\beta}_2$: ARIMA(1,0,0) con media cero.

$\hat{\phi}_1$= 0.7833 (0.1094) y $\hat{\sigma}^2$= 0.02067

AIC=-29.41, AICc=-29.01 y BIC=-26.42.

Coeficiente $\hat{\beta}_3$: ARIMA(1,0,0) con media cero.

$\hat{\phi}_1$= 0.6654 (0.1321) y $\hat{\sigma}^2$= 0.01567

AIC=-38.91, AICc=-38.51 y BIC=-35.91.

Coeficiente $\hat{\beta}_4$: ARIMA(1,0,0) con media cero.

$\hat{\phi}_1$= 0.7328 (0.1198) y $\hat{\sigma}^2$= 0.006517

AIC=-67.68, AICc=-67.28 y BIC=-64.69.

Coeficiente $\hat{\beta}_5$: ARIMA(0,0,1) con media cero.

$\hat{\theta}_1$= 0.2929 (0.1721) y σ^2= 0.009752

AIC=-55.06, AICc=-54.66 y BIC=-52.07.

Coeficiente $\hat{\beta}_6$: ARIMA(0,0,0) con media cero.

$\hat{\sigma}^2$= 0.00716

AIC=-67.35, AICc=-67.22 y BIC=-65.85.

Cociente

Varones

Coeficiente $\hat{\beta}_1$: ARFIMA(1,d,0).

$\hat{\phi}_1$= 0.9794, $\hat{\sigma}^2$= 0.0892 y $\hat{d}$=0.0828

Coeficiente $\hat{\beta}_2$: ARFIMA(2,d,0).

$\hat{\phi}_1$= 0.4872, $\hat{\phi}_2$= 0.3538, $\hat{\sigma}^2$= 0.0728 y $\hat{d}$=0.0395

Coeficiente $\hat{\beta}_3$: ARFIMA(1,d,0).

$\hat{\phi}_1$= 0.3491, σ^2= 0.0734 y $\hat{d}$=0.00004

Coeficiente $\hat{\beta}_4$: ARFIMA(0,d,0).

σ^2= 0.0663, y $\hat{d}$=0.0001

Coeficiente $\hat{\beta}_5$: ARFIMA(0,d,0).

$\hat{\sigma}^2$= 0.0506 y $\hat{d}$=0.0001

Coeficiente $\hat{\beta}_6$: ARFIMA(0,d,0).

$\hat{\sigma}^2$= 0.0454 y $\hat{d}$=0.0001

Mujeres

Coeficiente $\hat{\beta}_1$: ARFIMA(1,d,0).

$\hat{\phi}_1$= 0.9814, $\hat{\sigma}^2$= 0.0908 y $\hat{d}$=0.0463

Coeficiente $\hat{\beta}_2$: ARFIMA(2,d,0).

$\hat{\phi}_1$= 0.492, $\hat{\phi}_1$= 0.3514, $\hat{\sigma}^2$= 0.0725 y $\hat{d}$=0.0339

Coeficiente $\hat{\beta}_3$: ARFIMA(1,d,0).

$\hat{\phi}_1$=0.3491, $\hat{\sigma}^2$= 0.0706 y $\hat{d}$=0.0001

Coeficiente $\hat{\beta}_4$: ARFIMA(0,d,0).

$\hat{\sigma}^2$= 0.0663 y $\hat{d}$=0.0001

Coeficiente $\hat{\beta}_5$: ARFIMA(0,d,0).

$\hat{\sigma}^2$= 0.0505 y $\hat{d}$=0.0001

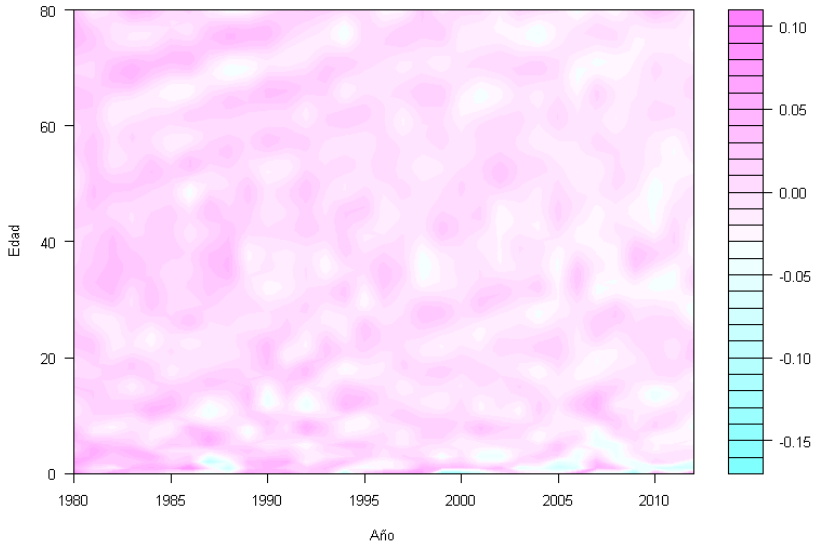
Coeficiente $\hat{\beta}_6$: ARFIMA(0,d,0).

σ^2= 0.0454 y $\hat{d}$=0.0001

Residuos del ajuste MDF

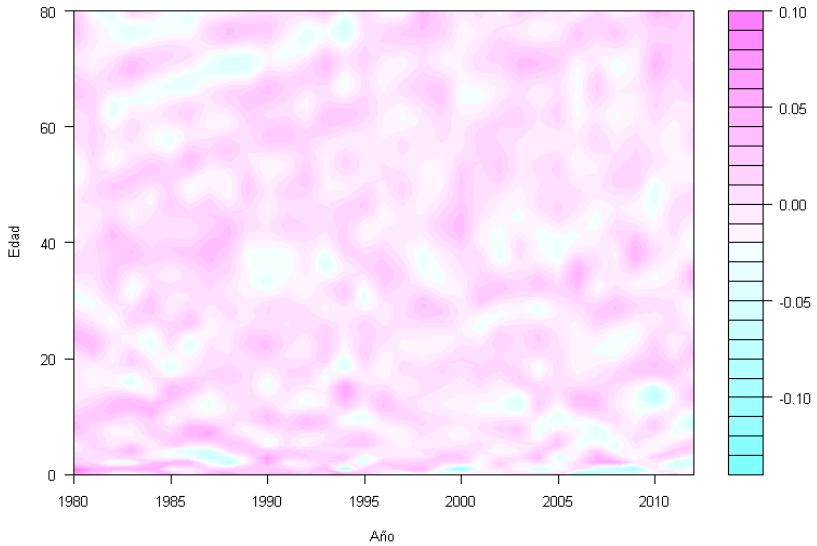
A continuación se presentan los gráficos de residuos para cada uno de los modelos estimados en la sección anterior.

Figura 1: Residuos del Modelo para Datos Funcionales (Mortalidad) - Total



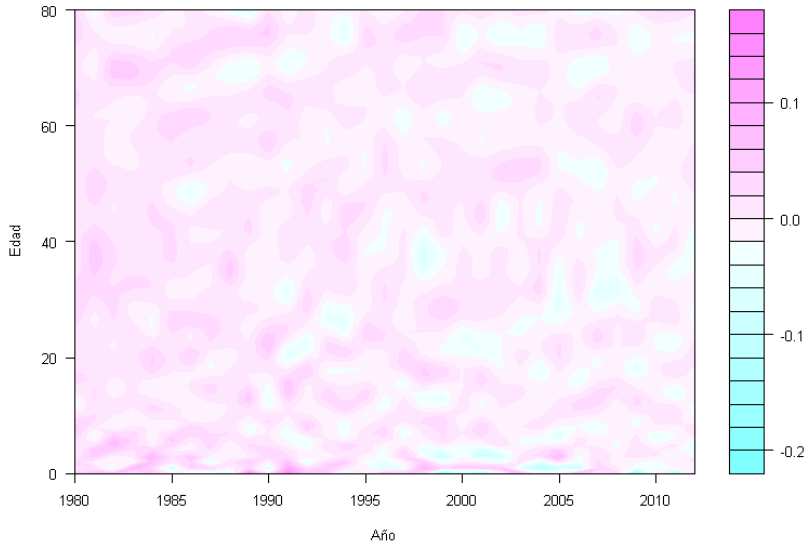
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 2: Residuos del Modelo para Datos Funcionales (Mortalidad) - Varones



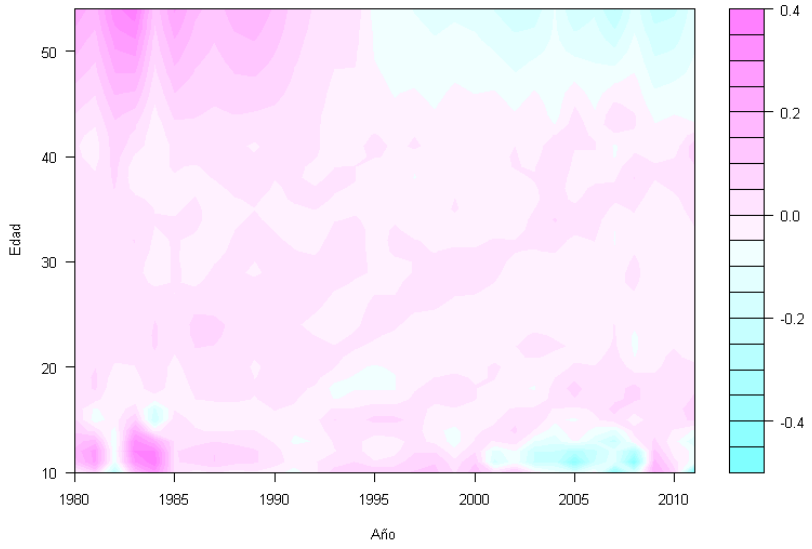
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 3: Residuos del Modelo para Datos Funcionales (Mortalidad) - Mujeres



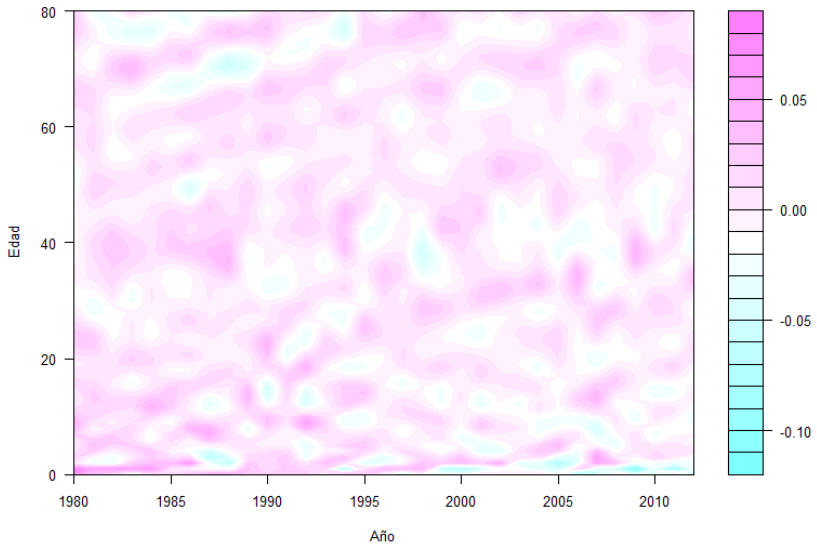
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 4: Residuos del Modelo para Datos Funcionales (Fecundidad).



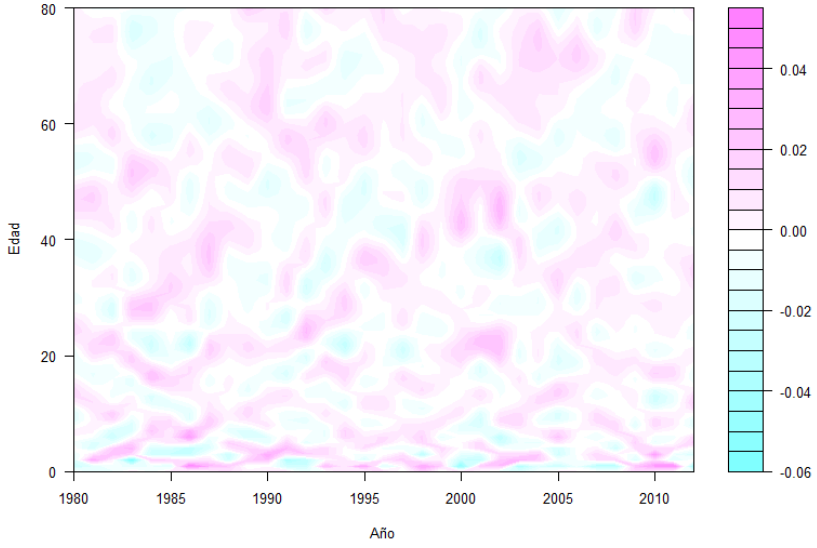
Fuente: Elaboración propia en base a datos de la DEIS.

Figura 5: Residuos del Modelo para Datos Funcionales Coherentes (Mortalidad - Función $p_t(x)$).



Fuente: Elaboración propia en base a datos de la DEIS.

Figura 6: Residuos del Modelo para Datos Funcionales Coherentes (Mortalidad - Función $c_t(x)$).



Fuente: Elaboración propia en base a datos de la DEIS.