Estadística Española • Volumen 58, número 189 / 2016, pp. 95-117

Consistencia en la desagregación de la población. El problema del ruido y el age heaping*

Silvia Bermúdez Parrado
Instituto de Estadística y Cartografía de Andalucía

Rafael Blanquero Bravo
Departamento de Estadística e Investigación Operativa. Universidad de Sevilla

Resumen

La población desagregada en edad simple es una herramienta básica para las oficinas estadísticas, pues es usada, por ejemplo, como denominador en el cálculo de indicadores. Sin embargo, las cifras de población para algunos ámbitos territoriales solo están disponibles en forma agrupada: normalmente, esta distribución de la población se publica en grupos quinquenales de edad más un intervalo superior abierto donde se acumula la población de mayor edad.

Un problema importante al que se suelen enfrentar las diferentes oficinas estadísticas, tanto de ámbito estatal como autonómico, es la desagregación de los datos de población en grupos de edad simple, permitiendo, cuando sea requerido, incluir conocimiento demográfico y mantener la consistencia de los resultados obtenidos con las agregaciones de población de los ámbitos territoriales superiores al desagregado o, incluso, la consistencia en la evolución de ésta a lo largo del tiempo.

En este trabajo se consideran técnicas de optimización para dar respuesta a este problema real, aunque muy poco estudiado en la literatura. Los procedimientos propuestos permiten también tratar un problema habitual en este tipo de fuentes estadísticas, como es la presencia de ruido de distinta naturaleza en los datos disponibles, y, en particular, el fenómeno conocido como *age heaping*.

Palabras clave: población, edad simple, desagregación, optimización matemática, age heaping.

*Clasificación AMS:*91D20, 62P25,65K05, 90C10.

* Este trabajo ha sido parcialmente financiado por los proyectos MTM2012-36163-C06-03 del Ministerio de Economía y Competitividad, P11-FQM-7603 y FQM-329 de la Junta de Andalucía.

Consistent degrouping of population data. The problem of noise and age heaping

Abstract

Official Statistics call for data by individual age, since a significant number of statistical operations, such as the calculation of demographic indicators, require the use of degrouped population figures. However, in some countries or regions population data are only available in a grouped form, usually as quinquennial age groups plus a large open-ended interval for elderly people.

A challenging problem faced by Official Statistics institutes is how to degroup data by individual age, allowing one, if needed, to include demographic knowledge or to be consistent with the heaped information.

In this paper Mathematical Optimization models are proposed to address this important, yet seldom studied problem. These models also consider a frequent issue in statistical sources: the presence of noise and errors, and, in particular, the phenomenon known as age heaping.

Keywords: population, individual age, degrouping, mathematical optimization, age heaping.

AMS classification: 91D20, 62P25,65K05, 90C10.

1. Introducción

La Optimización Matemática o Programación Matemática es una disciplina en la que, esencialmente, se trata de determinar el mejor elemento, con respecto a un criterio establecido previamente, entre un conjunto de posibilidades, permitiendo la toma de decisiones óptimas en contextos donde los recursos son limitados, (Williams, 2013; Winston et al., 2003). Entre los numerosos campos de aplicación de esta disciplina se encuentra la Estadística (Everitt, 2012; Rustagi, 1978), donde interviene de forma decisiva en el desarrollo de una gran variedad de técnicas. Sin embargo, cuando restringimos nuestra atención al campo concreto de la Estadística Oficial, la utilización de técnicas de Optimización Matemática tiene un carácter casi testimonial. Entre los ejemplos que se encuentran en la literatura, podemos citar la utilización de técnicas de optimización combinatoria para la protección de datos identificables individualmente en tablas estadísticas, (Salazar-González, 2008), o la determinación de parámetros en ajustes de curvas en Demografía (véase Bermúdez et al., 2012; Bermúdez, 2014) y las referencias allí contenidas). A pesar de la escasa interacción observada hasta ahora entre ambas disciplinas, tenemos la convicción de que un mayor acercamiento y colaboración entre profesionales de la Estadística Pública y expertos en el ámbito de la Investigación Operativa, puede contribuir a mejorar los procedimientos de resolución aplicados actualmente a determinados problemas presentes en este campo; como muestra de ello, presentamos aquí de forma sintética el trabajo que obtuvo el Premio en Estadística Oficial 2015 "Premios INE Eduardo García España", en el que abordamos mediante modelos de Optimización Matemática un problema clásico en Estadística Oficial, como

es la desagregación de datos poblacionales. Con posterioridad a la concesión del citado premio, el trabajo mencionado, que tiene su origen en (Bermúdez, 2012), ha sido aceptado para su publicación en la revista *Population Studies: A Journal of Demography*, (Bermúdez y Blanquero, 2015).

Los datos de población constituyen un elemento esencial en el ámbito demográfico, pese a lo cual no siempre se encuentran disponibles con el nivel de detalle deseable. Las distintas oficinas estadísticas nacionales e internacionales ofrecen datos para ámbitos territoriales que proceden de la subdivisión del ámbito nacional, pero son presentados con un nivel de desagregación menor que en aquél. Como ejemplo, la Oficina Estadística Europea, Eurostat, ofrece en su web, (Eurostat), datos de población desagregados por edad simple para ámbitos nacionales e inferiores, hasta las unidades territoriales estadísticas NUTS 2. Por debajo de este nivel, los datos aparecen agregados por grupos de edad, de mayor o menor amplitud, y sexo. En la mayoría de los casos, estos grupos de edad suelen ser quinquenales. Incluso para el ámbito nacional la información está disponible en grupos de mayor amplitud si la población objetivo es un subconjunto de la población total, como ocurre con la población de nacionalidad extranjera y otros grupos de interés.

El conocimiento de datos de población por edad simple es crucial en ciertas operaciones estadísticas, como el cálculo de indicadores demográficos. Por ejemplo, la estimación de las tasas específicas de fecundidad, como paso previo para la determinación del indicador coyuntural de fecundidad, precisa de datos poblacionales por edad simple. Estos datos no agrupados son también un elemento esencial en la elaboración de proyecciones, ya sean de base o derivadas, o en labores de planificación en áreas donde la edad juega un papel fundamental, como es el caso de la Educación. Parece, pues, de sumo interés el disponer de procedimientos que permitan obtener datos de población por edad simple a partir de datos agregados por grupos de edad.

La desagregación de una serie cronológica general, un problema íntimamente relacionado con el que aquí se aborda, ha sido ampliamente tratada en la literatura, (Lisman y Sandee, 1964; Cohen et al., 1971; Dentopn, 1971; Wei y Stram 1990; Guerrero, 2003; Pavía-Miralles, 2010; Proietti, 2011). En lo que respecta a datos demográficos, la desagregación de datos de mortalidad dados en grupos de edad ha recibido cierta atención, (Kostaki y Panousis, 2001), aunque el procedimiento allí propuesto no es aplicable a datos de población generales, al estar basado en la expansión de una tabla de mortalidad agrupada, haciendo uso de un modelo paramétrico para las probabilidades de muerte por año. Sin embargo, el problema de la desagregación de datos de población por edad simple apenas ha sido tratado en la literatura. Cuando los datos proceden de un único año, pueden aplicarse procedimientos sencillos, como el método de Sprague, (Siegel y Swanson, 2004), u otras técnicas de interpolación mediante funciones *spline*, (McNeil et al., 1977; Wegman yWright, 1983). Estas técnicas de propósito general carecen de una de las propiedades más importantes que todo procedimiento para desagregación de valores poblacionales debe poseer: los valores desagregados deben ser números enteros; esta carencia obliga a aplicar procedimientos de redondeo que, dependiendo del volumen de datos a tratar, pueden distorsionar el resultado final. Por otra parte, no es una tarea trivial el adaptar tales

procedimientos al caso en que el número de años a desagregar sea superior a uno, debido a la dificultad que supone asegurar la coherencia entre las poblaciones desagregadas de edades consecutivas en años también consecutivos.

En este trabajo tratamos este problema específico de desagregación, para lo cual se presentan diversos modelos de optimización matemática que, partiendo de datos de población agrupados en intervalos de edad, permiten la desagregación de los datos en edades simples. De este modo, se obtiene un valor para cada edad simple de forma óptima, de acuerdo con un determinado criterio. Estos valores obtenidos tienen además la propiedad de ser números enteros, lo que evita el tener que recurrir a tediosos procedimientos de redondeo.

El resto del artículo está estructurado de la siguiente forma: en la Sección 2 se presentan los distintos modelos de optimización que permiten la desagregación de la población. Una breve descripción de las técnicas aplicadas para resolver estos modelos, así como algunos resultados numéricos, se presentan en la Sección 3, donde se aplican los modelos descritos a datos de población de España y de todas las comunidades autónomas que la integran. El artículo finaliza en la Sección 4 con un resumen de conclusiones, acompañado de ciertas extensiones de la metodología propuesta al caso de datos faltantes y de datos afectados por ruido, con mención especial al fenómeno conocido como *age heaping*.

2. Los modelos

Se presenta a continuación un conjunto de modelos matemáticos que permitirán, partiendo de datos en números enteros de población, agrupados en intervalos de edad de longitud mayor que uno, obtener una desagregación en edades simples con números enteros. De este modo, pasaremos a disponer de un valor entero de población para cada edad, obtenido de forma óptima con respecto a los criterios que se especificarán a continuación.

Con carácter general, los modelos propuestos tratan de conseguir que los resultados del proceso de desagregación den lugar a transiciones suaves, tanto longitudinal como transversalmente. Así, para los distintos años y edades simples consideradas en el proceso, se tratará de hacer "tan pequeña como sea posible" la diferencia entre el número de efectivos de edades i e i + 1 en el año t y, al mismo tiempo, se perseguirá también el mismo objetivo para la diferencia entre el número de efectivos de edad i en año t y de edad i + 1 en el año t + 1. Adicionalmente, y con objeto de fijar unas condiciones de partida lo más exactas posible, se tratará de hacer mínima la diferencia entre la población de edad 0 en cada año considerado y el número de nacimientos ocurridos en el año anterior, información ésta que se encuentra disponible habitualmente.

Todas las diferencias descritas son agregadas en un único valor, que se pretende minimizar; esta agregación se realiza empleando el criterio L_2, de uso más frecuente en la literatura, en el que se hace mínima la suma de los cuadrados de las diferencias. Otros métodos de agregación usuales, tales como el L_1 y el L_∞, en los que se pretende minimizar la suma de los valores absolutos de las diferencias o la mayor de éstas en valor absoluto, respectivamente, fueron utilizados y comparados con el método L_2,

proporcionando peores resultados, por lo que no han sido considerados en este trabajo (para más detalles, véase Bermúdez 2014).

Se presentan, en primer lugar, los diferentes modelos que se proponen para la desagregación de la población y, posteriormente, los resultados experimentales obtenidos tras su aplicación sobre conjuntos de datos de población de distinto nivel territorial.

2.1 Modelo básico

Dado un ámbito territorial s, supondremos que, para cada año t, t = 1,...,T, se dispone de la población desagregada en G grupos de edad $E_j = \{L_j, L_j+1, ..., U_j-1, U_j\}$, j = 1,...,G de longitud variable, tales que $Lj = U_{j-1}+1$. Así, $P_{j,t}$ denotará tal población para el grupo de edad E_j en el año calendario t, j = 1,...,G, t = 1,...,T. Por otra parte, representaremos por B_t el número total de nacimientos ocurridos a lo largo del año t−1 en el ámbito s. Para el caso en el que estemos estudiando la desagregación de la población extranjera, se considerarán los nacidos de madre extranjera. La Tabla 1 ilustra la definición de las constantes del problema.

Tabla 1

Definición de las constantes $P_{j,t}$

Grupos de edad	Años				
	1	⋯	t	⋯	T
E_1					
E_2					
⋮					
⋮		⋱			
E_j			$P_{j,t}$		
⋮					
⋮				⋱	
E_{G-1}					
E_G					

Las variables de decisión del modelo de desagregación se definen de forma natural de acuerdo con nuestro propósito. Así, x_{it} denotará la población del ámbito territorial considerado a la edad i en años cumplidos, $i = L_1, L_1+1, ..., U_G-1, U_G$, para el año calendario t, $t = 1,...,T$, lo que da lugar a un conjunto de variables enteras cuyo cardinal viene dado por:

$$T\sum_{j=1}^{G}\left(U_j - L_j + 1\right) = T\left(U_g - L_1 + 1\right)$$

que puede expresarse como $5 \cdot G \cdot T$ en el caso usual de grupos de edad quinquenales.

Obsérvese que el tratamiento del intervalo abierto, que habitualmente se asocia al grupo de mayor edad, tiene cabida en el planteamiento aquí considerado, a condición de que se establezca un límite superior U_G razonable para dicho grupo.

A partir de los parámetros $P_{j,t}$ y B_t, $j = 1,...,G$, $t = 1,...,T$, y de las variables de decisión $x_{i,t}$, $i = L_1,...,U_G$, $t = 1,...,T$, se considera un primer modelo de desagregación que viene dado por:

$$mín \sum_{t=1}^{T} \sum_{i=L_1+1}^{U_G} \left(x_{i,t} - x_{i-1,t}\right)^2 + \sum_{t=1}^{T} \left(x_{L_1,t} - B_t\right)^2 + \sum_{t=2}^{T} \sum_{i=L_1+1}^{U_G} \left(x_{i,t} - x_{i-1,t-1}\right)^2$$

s.a.

(L_2BAS)

$$\sum_{i=L_j}^{U_j} x_{i,t} = P_{j,t} \quad j = 1,...,G \quad t = 1,...,T$$

$$x_{i,t} \in \mathbb{Z}^+ \qquad i = L_1,...,U_G \quad t = 1,...,T$$

Se trata de un problema cuadrático convexo con restricciones lineales (resoluble, por tanto, empleando técnicas estándar en Optimización Matemática) donde la función objetivo presenta tres bloques claramente diferenciados, cuyo cometido se indica a continuación:

$\sum_{t=1}^{T} \sum_{i=L_1+1}^{U_G} (x_{i,t} - x_{i-1,t})^2$: tiene como propósito conseguir que las transiciones de cada edad simple a la siguiente, dentro del mismo año *t*, sean suaves.

$\sum_{t=1}^{T} (x_{L_1,t} - B_t)^2$: para cada año *t* trata de aproximar la población inicial (edad 0 años) al número de nacimientos del año anterior.

$\sum_{t=2}^{T} \sum_{i=L_1+1}^{U_G} (x_{i,t} - x_{i-1,t-1})^2$: trata de garantizar una adecuada evolución temporal de cada cohorte, buscando para ello una cierta concordancia entre la población de edad *i* en el año *t* y la de edad $i - 1$ en el año precedente.

Al margen de las restricciones de integridad sobre las variables de decisión ($x_{i,t} \in \mathbb{Z}^+$) , el único bloque de restricciones del problema permite garantizar la concordancia entre los datos originales agregados por grupos de edad y los correspondientes valores desagregados obtenidos como solución de (L_2BAS).

2.2 Modelos con información auxiliar sobre el intervalo superior abierto

De acuerdo con la experimentación realizada, (Bermúdez 2014), el modelo anterior permite obtener ajustes adecuados en todos los grupos de edad, con la excepción del grupo de edad abierto final EG, en el que se producen discrepancias con los valores observados. A ello contribuye fundamentalmente la mayor libertad que posee el último dato de este grupo abierto, al no estar ligado a valores posteriores del mismo año o del siguiente.

Como solución al problema que acaba de ser descrito, se propone considerar las frecuencias relativas de las distintas edades que componen el intervalo abierto, e imponer que éstas tomen valores dentro de ciertos intervalos determinados a partir de la información disponible. Esta información puede proceder de resultados de Operaciones Estadísticas en las que dispongamos de datos de población desagregados por edad simple para el grupo superior abierto y para ese año en concreto o para un año próximo. Algunas de estas fuentes pueden ser: Censo de Población, encuesta específica realizada para la subpoblación bajo estudio (Encuesta Nacional de Inmigrantes, ENI, en el caso de población inmigrante en España, por ejemplo), información de población desagregada por edad simple para un ámbito de mayor nivel que el analizado (datos de un territorio de nivel NUTS 2 para el estudio de un ámbito NUTS 3, por ejemplo), etc. Así, se definen:

$\underline{f_i}$: límite inferior para la frecuencia relativa correspondiente a la edad i del intervalo abierto $i = L_G, L_G + 1, \ldots, U_G$.

$\overline{f}_i$: límite superior para la frecuencia relativa correspondiente a la edad i del intervalo abierto, $i = L_G, L_G + 1, \ldots, U_G$.

Para cada edad componente del intervalo E_G se exigirá que su frecuencia relativa dentro de dicho intervalo se encuentre comprendida entre $\underline{f}$ y $\overline{f}$, es decir:

$$\underline{f}_i \leq \frac{x_{i,t}}{P_{G,t}} \leq \overline{f}_i \qquad i = L_G, L_G + 1, \ldots, U_G \quad t = 1, \ldots, T$$

o, equivalentemente

$$\underline{f_i} P_{G,t} \leq x_{i,t} \leq \overline{f}_i P_{G,t} \qquad i = L_G, L_G + 1, \ldots, U_G \qquad t = 1, \ldots, T$$

Si la determinación de los límites $\underline{f}_i$ y $\overline{f}_i$ se realiza, por ejemplo, a partir de la información censal disponible, estos límites podrían obtenerse de la distribución relativa del intervalo E_G en los censos inmediatamente anterior y posterior al periodo de años en que se realiza la desagregación. Si sólo uno de estos censos está disponible, las frecuencias relativas del intervalo E_G obtenidas a partir de éste pueden usarse como valores centrales de los intervalos $[\underline{f_i}, \overline{f}_i]$ determinándose su amplitud de forma proporcional a dicho valor central. Otra posibilidad consiste en incorporar la información de un ámbito superior para el que dispongamos de los datos o de una encuesta que se realice de forma periódica.

Tras incorporar al modelo (L_2BAS) el nuevo grupo de restricciones que acaba de ser descrito, se obtiene el siguiente modelo de desagregación:

$$\min \sum_{t=1}^{T} \sum_{i=L_1+1}^{U_G} (x_{i,t} - x_{i-1,t})^2 + \sum_{t=1}^{T} (x_{L_1,t} - B_t)^2 + \sum_{t=2}^{T} \sum_{i=L_1+1}^{U_G} (x_{i,t} - x_{i-1,t-1})^2$$

s.a.

(L_2INT_1)

$$\begin{aligned}
&\sum_{i=L_j}^{U_j} x_{i,t} = P_{j,t} && j = 1, \ldots, G && t = 1, \ldots, T \\
&x_{i,t} \leq \overline{f}_i P_{G,t} && i = L_G, L_{G+1}, \ldots, U_G && t = 1, \ldots, T \\
&x_{i,t} \geq \underline{f}_i P_{G,t} && i = L_G, L_{G+1}, \ldots, U_G && t = 1, \ldots, T \\
&x_{i,t} \in \mathbb{Z}^+ && i = L_1, \ldots, U_G && t = 1, \ldots, T
\end{aligned}$$

En el supuesto de que la información auxiliar introducida en el modelo anterior no esté disponible, el problema se puede abordar de forma alternativa asumiendo que el volumen de población decrece a lo largo del intervalo superior abierto, dando lugar al modelo (L_2INT_2), cuya formulación puede consultarse en (Bermúdez y Blanquero, 2015). Esta hipótesis en la evolución de la población de mayor edad es, en general, bastante realista, con la excepción de aquellos ámbitos geográficos que puedan verse influenciados por una elevada inmigración en ese intervalo de edad, situación poco probable en edades tan avanzadas.

2.3 Modelos con información de contorno

Usualmente se dispone de datos de población por edad simple en años próximos a los que intervienen en el proceso de desagregación, como sucede con la información censal. Incluso, en ocasiones es frecuente contar con datos desagregados para el mismo año que los datos agregados de que se dispone, aunque procedentes de distinta Fuente Estadística. En tales circunstancias, es posible incorporar esta información al modelo para guiar el ajuste en los años inicial y/o final, mejorando los resultados proporcionados por el modelo básico. De esta forma, los nuevos elementos que se incorporan a dicho modelo serían:

- I_i: población para la edad simple i, $i = L_1,...,U_G$, en el año de referencia inicial.
- F_i: población para la edad simple i, $i = L_1,...,U_G$, en el año de referencia final.

Partiendo del modelo (L_2BAS), será preciso modificar la función objetivo de éstos para hacer que las variables de decisión correspondientes a los años inicial y final tomen valores próximos a I_i y F_i, respectivamente.

De acuerdo con lo anterior, tras incorporar la información de los años de referencia inicial y final al modelo (L_2BAS), éste quedará:

$$\min \sum_{t=1}^{T} \sum_{i=L_1+1}^{U_G} (x_{i,t} - x_{i-1,t})^2 + \sum_{t=1}^{T} (x_{L_1,t} - B_{t-1})^2 + \sum_{t=2}^{T} \sum_{i=L_1+1}^{U_G} (x_{i,t} - x_{i-1,t-1})^2 + \sum_{i=L_1}^{U_G} (x_{i,t} - I_i)^2 + \sum_{i=L_1}^{U_G} (x_{i,T} - F_i)^2$$

s.a.

(L_2CON)

$$\sum_{i=L_j}^{U_j} x_{i,t} = P_{j,t} \quad j = 1,...,G \quad t = 1,...,T$$

$$x_{i,t} \in \mathbb{Z}^+ \qquad i = L_1,...,U_G \quad t = 1,...,T$$

2.4 Modelos con información de ámbito superior

Otra variante de interés de los modelos presentados con anterioridad se obtiene si se dispone de la población desagregada en edades simples correspondiente a un nivel superior al de los ámbitos para los cuales se tienen los datos agregados, de tal modo que la unión de estos subámbitos da lugar al ámbito superior. Ésta es una situación frecuente en la práctica y así, por ejemplo, el Instituto Nacional de Estadística publicaba datos de población provinciales en grupos quinquenales, mientras que la información poblacional de las comunidades autónomas se facilitaba desagregada en edades simples. En la actualidad, las cifras de población aparecen agregadas, por ejemplo, cuando se clasifican atendiendo a ciertas variables de cruce, como pueden ser la nacionalidad o el lugar de nacimiento.

Para el desarrollo de este modelo supondremos que se dispone, para cada ámbito territorial s, $s = 1,2,...,S-1,S$, y cada año calendario t, $t = 1,...,T$, de la población desagregada en G grupos de edad de longitud variable $E_j = \{L_j, L_j+1,...,U_j-1, U_j\}$ $j = 1,...,G$. Así, $P_{j,s,t}$ representará tal población para el grupo de edad E_j, en el ámbito territorial s y en el año calendario t, $j = 1,...,G$, $s = 1,...,S$, $t = 1,...,T$.

Supondremos también conocida la población por edad simple del ámbito de nivel superior, que se representará por $Q_{i,t}$, donde el subíndice i corresponde a la edad y el subíndice t al año calendario, $i = L_1,...,U_G$, $t = 1,...,T$. Por la propia definición de $P_{j,s,t}$ y $Q_{i,t}$, es obvio que debe verificarse la siguiente relación:

$$\sum_{s=1}^{S} P_{j,s,t} = \sum_{i=L_j}^{U_j} Q_{i,t} \qquad j = 1,\dots,G \quad t = 1,\dots,T \qquad [1]$$

Por último, representaremos por $B_{s,t}$ el número de nacimientos ocurridos a lo largo del año $t-1$ en el ámbito territorial s, $s = 1,...,S$, $t = 1,...,T$.

Las variables de decisión del modelo se definen en la forma habitual, aunque teniendo en cuenta la existencia de los S subámbitos. De este modo, la variable de decisión $x_{i,s,t}$

representará la población de edad i, $i = L_1,...,U_G$, en el subámbito s, $s = 1,...,S$, para el año calendario t, $t = 1,...,T$.

El primer modelo que se considera en el contexto que ahora nos ocupa, trata de hacer mínima la suma de los cuadrados de las diferencias correspondientes a las características a que debe responder el modelo, y que han sido recogidas con anterioridad:

- Transición suave entre edades consecutivas dentro de cada ámbito y año.
- Aproximación de la población inicial de cada ámbito al número de nacimientos del año precedente.
- Transición suave entre edades consecutivas de años consecutivos dentro de cada ámbito.

Esto da lugar al siguiente problema cuadrático convexo con restricciones lineales:

$$mín \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{i=L_1+1}^{U_G}\left(x_{i,s,t} - x_{i-1.s,t}\right)^2 + \sum_{s=1}^{S}\sum_{t=1}^{T}\left(x_{L_1,s,t} - B_{s,t}\right)^2 + \sum_{s=1}^{S}\sum_{t=2}^{T}\sum_{i=L_{1+1}}^{U_G}\left(x_{i,s,t} - x_{i-1,s,t-1}\right)^2$$

(L_2SUP) s.a.

$$\sum_{i=L_j}^{U_j} x_{i,s,t} = P_{j,s,t} \qquad j = 1,\dots,G \qquad s = 1,\dots,S \qquad t = 1,\dots,T$$

$$\sum_{s=1}^{S} x_{i,s,t} = Q_{i,t} \qquad i = L_1,\dots,L_G \qquad t = 1,\dots,T$$

$$x_{i,s,t} \in \mathbb{Z}^+ \qquad i = L_1,\dots,U_G \qquad s = 1,\dots,S \qquad t = 1,\dots,T$$

El primer bloque de restricciones garantiza que, para cada intervalo de edad, año calendario y subámbito, la suma de la población desagregada por edad simple coincide con el total del intervalo de edad correspondiente; éstas son, esencialmente, las restricciones que aparecían en el modelo (L_2BAS).

El segundo bloque de restricciones tiene como objetivo garantizar la concordancia entre los valores poblacionales de los subámbitos y los del ámbito de nivel superior. Así, estas restricciones garantizan que, para cada edad simple y año calendario, la población del ámbito superior es igual a la suma de las poblaciones de los subámbitos que lo componen.

3. Resolución de los modelos de desagregación

Los modelos presentados en los apartados anteriores han sido resueltos utilizando el software de optimización IBM ILOG Cplex Optimization Studio en su versión 12.2. Cplex es el optimizador de referencia para la resolución de problemas en números enteros, tanto lineales como cuadráticos, si bien hubiera sido posible igualmente la utilización de otras herramientas con igual funcionalidad, tales como Gurobi, (FICO Xpress Optimization Suite) o Xpress, (Gurobi Optimization). Los diferentes modelos han sido formulados utilizando el lenguaje algebraico de modelado utilizado por Cplex Optimization Studio, denominado OPL y que, siguiendo la tendencia actual en este tipo de lenguajes, permite una perfecta separación entre el modelo y los datos a los que éste se aplica.

Con carácter general, la resolución de problemas en números enteros requiere tiempos de cómputo elevados; sin embargo, los tiempos de resolución para los modelos aquí presentados son muy pequeños, bastando, por lo general, algunos segundos para la obtención de la solución óptima.

Con el objetivo de comprobar la validez de los modelos propuestos, éstos han sido aplicados a datos de población correspondientes a España y todas las comunidades autónomas que la integran. En concreto, para testar los modelos (L_2BAS), (L_2INT_1), (L_2INT_2) y (L_2CON), se han utilizado los datos de la comunidad autónoma de Andalucía, mientras que en el caso del modelo jerárquico, (L_2SUP), se ha trabajado con los datos de todas las comunidades autónomas que conforman el nivel superior, España. Es decir, se tienen, como datos de partida del modelo, en el periodo 2002 − 2011, por un lado, las cifras de población agrupadas en intervalos quinquenales de edad de cada una de las comunidades autónomas y, por otro, la población desagregada en edad simple de España para el mismo periodo; como salida se obtienen, para cada una de las comunidades autónomas en el periodo de diez años considerado, las cifras de población desagregadas por edad simple consistentes con las del total nacional para cada edad y año, además de verificar el resto de condiciones impuestas al modelo, Sección 2.4. Otros experimentos pueden ser consultados en Bermúdez y Blanquero, 2015.

Los resultados obtenidos en los experimentos realizados, a los que nos referiremos como *datos estimados*, han sido comparados con los valores de población desagregados reales, a los que denominaremos *datos observados*. Para realizar estas comparaciones se ha utilizado una medida del error de ajuste cometido en el proceso de desagregación, en concreto, la raíz cuadrada del error cuadrático medio relativo, *RMSRE* (Root Mean Squared Relative Error), que, para un ámbito s y un año t, viene definida por:

$$RMSRE(s,t) = \sqrt{\frac{1}{U_G - L_1 + 1}\sum_{i=L_1}^{U_G}\left(\frac{O_{i,s,t} - x_{i,s,t}}{O_{i,s,t}}\right)^2} \qquad [2]$$

donde $O_{i,s,t}$ son los valores observados y $x_{i,s,t}$ los estimados ($i = L_1,...,U_G, s = 1,...,S, t = 1,...,T$).

En la experimentación realizada, los datos de partida se encontraban desagregados en intervalos de edad quinquenal, con la excepción del grupo superior abierto de 85 y más

años. La serie temporal considerada abarca un periodo de 10 años en la mayoría de las pruebas realizadas. Los datos utilizados en esta experimentación se encuentran disponibles en los bancos de datos libres que Eurostat, (Eurostat), y el Instituto Nacional de Estadística, (INE), ofrecen en sus respectivos sitios webs. Los resultados, en términos de la raíz cuadrada del error cuadrático medio relativo, [2], se muestran en las Tablas 2 y 3. Por cuestión de brevedad, la Tabla 3 sólo contiene los resultados de un número reducido de comunidades autónomas, aunque en el proceso de desagregación han intervenido la totalidad de ellas.

Tabla 2

Raíz cuadrada del error cuadrático medio relativo según el modelo de desagregación, Andalucía

MODELOS	RMSRE												Media	Desviación típica
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010		
Modelo básico														
L_2BAS	1,66	0,66	0,52	0,35	0,22	0,10	0,09	0,14	0,20	0,24			0,42	0,45
Modelo con información del intervalo superior abierto														
L_2INT_1	0,11	0,09	0,08	0,09	0,09	0,10	0,12	0,11	0,11	0,12			0,10	0,01
L_2INT_2	1,49	0,61	0,50	0,35	0,24	0,11	0,09	0,13	0,20	0,24			0,40	0,40
Modelo con información de contorno														
L_2CON				0,27	0,18	0,08	0,10	0,13	0,17	0,17	0,18	0,20	0,17	0,05

Tabla 3

Raíz cuadrada del error cuadrático medio relativo para el modelo (L_2SUP) según comunidad autónoma

Modelo con información del ámbito superior	RMSRE										Media	Desviación típica
	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011		
L_2SUP												
Aragón	0,22	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,17	0,02
Canarias	0,25	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,17	0,03
Castilla-León	0,09	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,02
Cataluña	0,42	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,19	0,08
Madrid	0,08	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,02
País Vasco	0,19	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,17	0,01

Las figuras que siguen a continuación muestran gráficamente los resultados obtenidos al desagregar la población agrupada en grupos quinquenales, tras la aplicación de los distintos modelos de desagregación descritos en la Sección 2. A modo de ejemplo, solo

se presentan los resultados para un año concreto, el 2004, en las comunidades de Andalucía, Aragón, Canarias, Castilla y León, Cataluña, Madrid y País Vasco. Un análisis más minucioso, así como otros ejemplos, también basados en datos reales, pueden encontrarse en Bermúdez, 2014.

Figura 1

Población desagregada en edad simple, modelo (L_2BAS), Andalucía, 2004

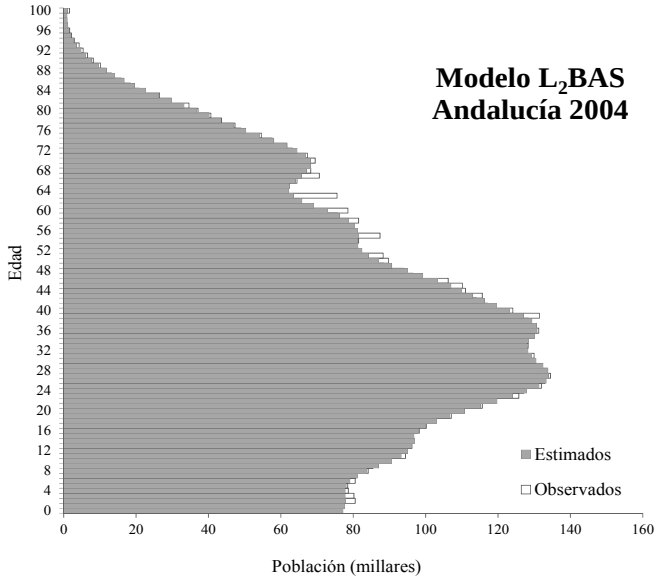


Figura 2

Población desagregada en edad simple, modelo (L_2INT_1), Andalucía, 2004

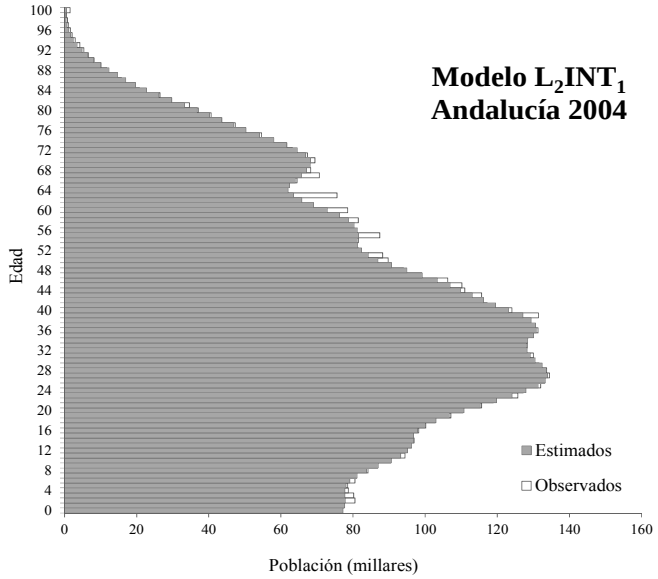


Figura 3

Población desagregada en edad simple, modelo (L_2CON), Andalucía, 2004

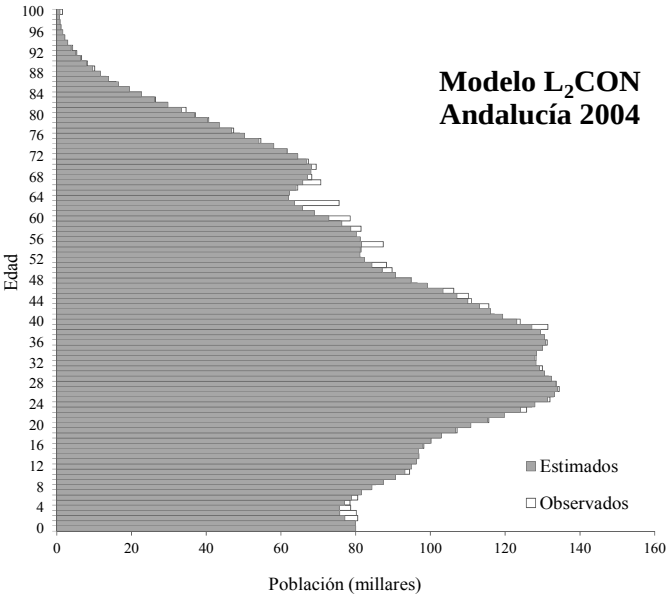


Figura 4

Población desagregada en edad simple, modelo (*L₂SUP*), Aragón, 2004

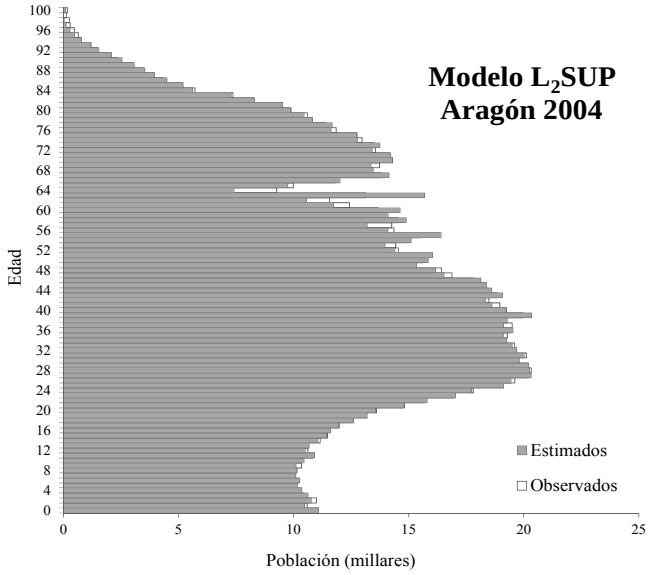


Figura 5

Población desagregada en edad simple, modelo (*L₂SUP*), Canarias, 2004

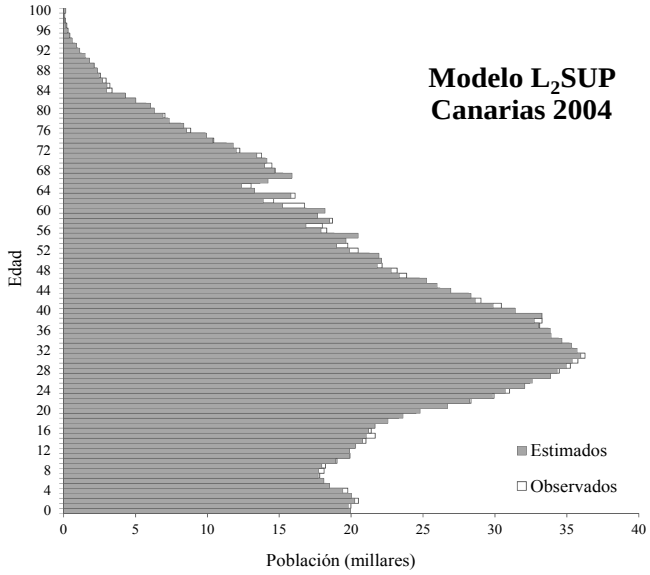


Figura 6:

Población desagregada en edad simple, modelo (*L_2SUP*), Castilla y León, 2004

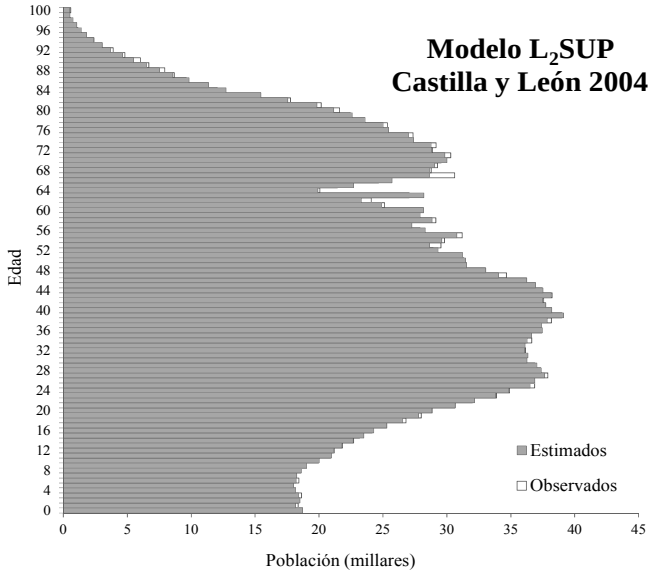


Figura 7

Población desagregada en edad simple, modelo (*L_2SUP*), Cataluña, 2004

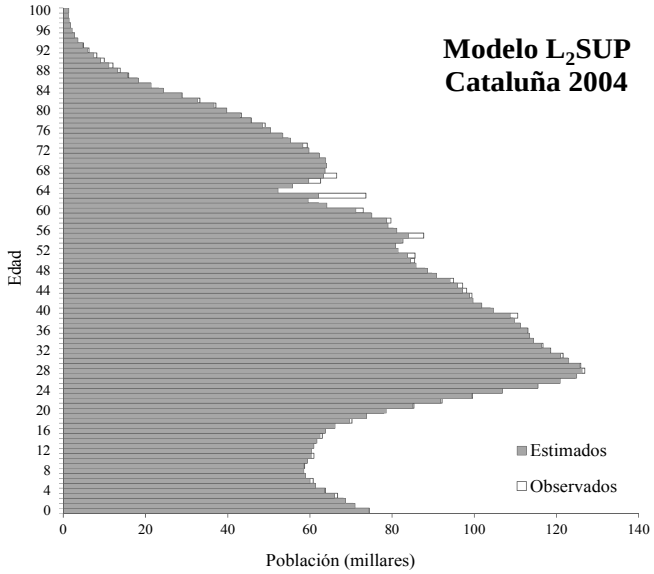


Figura 8

Población desagregada en edad simple, modelo (*L_2SUP*), Comunidad de Madrid, 2004

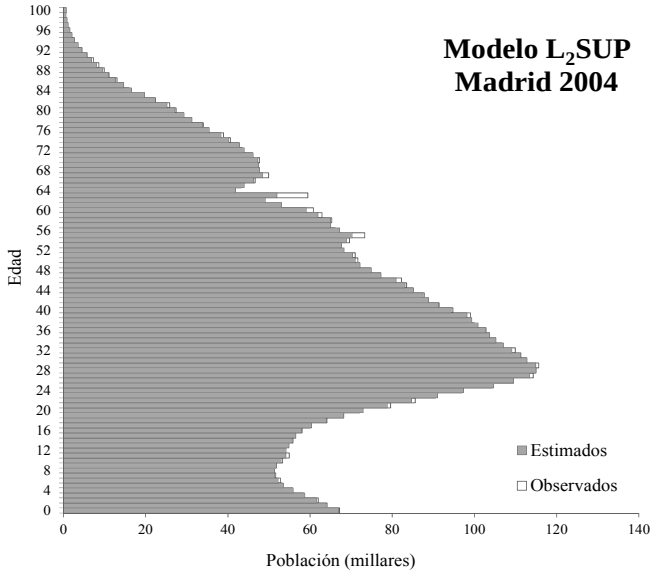
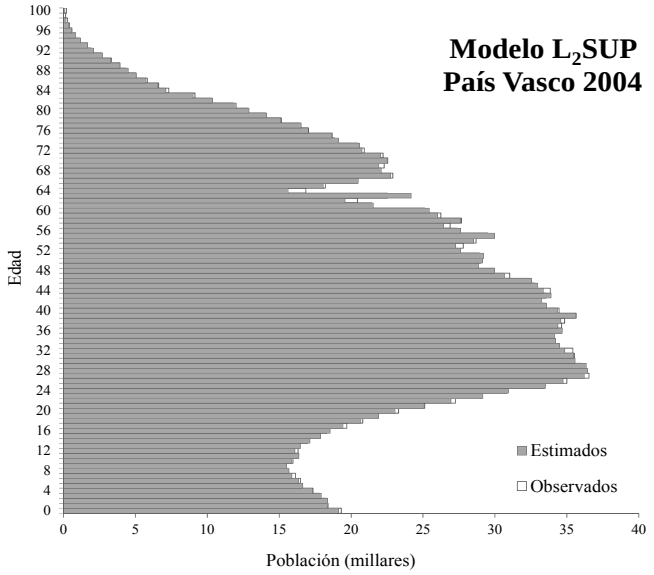


Figura 9

Población desagregada en edad simple, modelo (*L_2SUP*), País Vasco, 2004



Los modelos propuestos en este artículo proporcionan soluciones coherentes y consistentes, no sólo en sentido transversal, como se puede comprobar en las figuras anteriores, sino también en sentido longitudinal. Los resultados presentados en Bermúdez y Blanquero, 2015 muestran que nuestro enfoque proporciona resultados precisos también cuando se realiza un análisis longitudinal de los datos empíricos, garantizando la coherencia de las distintas generaciones presentes en los datos desagregados.

4. Conclusiones y extensiones

En este artículo se han propuesto diversos modelos de Optimización Matemática para desagregar en edad simple cifras de población agrupadas en intervalos de longitud mayor que uno. Estos modelos cumplen una serie de requisitos que deberían ser de obligado cumplimiento para cualquier metodología de desagregación que involucre magnitudes no fraccionarias:

- La solución se presenta en números enteros. Esto evita el hacer redondeos posteriores en las cifras de población.
- Para un t fijo, se obtiene coherencia en el corte transversal de los datos.
- Para una edad fijada, las soluciones muestran coherencia generacional, respetando las peculiaridades de cada una de las generaciones.

Partiendo de un modelo básico, que incorpora los requerimientos mínimos exigidos en el proceso de desagregación, éste es enriquecido haciendo uso de información adicional disponible, con objeto de proporcionar soluciones más adecuadas desde un punto de vista demográfico. Los diferentes modelos propuestos dan lugar a problemas de optimización en números enteros, que pueden ser resueltos en un ordenador personal empleando Cplex u otro optimizador de características similares, en tiempos inferiores a un minuto.

La validación de los modelos considerados ha sido realizada utilizando las cifras de población de España y sus comunidades autónomas en diferentes años. Tanto desde un punto de vista gráfico como numérico, se comprueba que estos modelos proporcionan buenos ajustes a los valores poblacionales observados. De acuerdo con los resultados numéricos obtenidos, de entre los modelos en que interviene un único ámbito territorial, el que proporciona los mejores ajustes es el modelo (L_2INT_1).

Otra de las ventajas que avalan la potencia de los modelos descritos es su gran versatilidad, de forma que pueden ser fácilmente adaptados a hipótesis más generales que las enunciadas hasta el momento. Por ejemplo, hemos asumido que las cifras de población de partida son exactas, aunque la realidad demuestra que esto no siempre es así. Debido a la naturaleza de los datos, registros administrativos o encuestas en su mayoría, estos pueden estar afectados por errores que distorsionen el proceso de desagregación. Para evitar esto, los distintos bloques de restricciones del modelo donde se exige la igualdad de $P_{j,t}$ con los valores desagregados correspondientes se transforman en *restricciones blandas*; con ello ya no se exige el cumplimiento estricto

de las citadas restricciones, si bien su violación queda penalizada en la función objetivo; luego se impone que, para cada año t, la población total estimada sea igual a la población total de partida. Un desarrollo más detallado se encuentra en Bermúdez y Blanquero, 2015.

Otra de las situaciones que pueden ser abordadas es la falta de información para algún año del periodo temporal considerado. Hasta ahora, hemos asumido que disponemos de las cifras de población para cada año. Desafortunadamente, esto no ocurre siempre así, especialmente en países en desarrollo.

La falta de datos para algunos años se puede solucionar mediante la aplicación de técnicas de uso común (por ejemplo, interpolación de los datos disponibles, ya sea por medio de funciones spline o siguiendo algún modelo de distribución que se considere adecuado a la vista de los datos). Sin embargo, los modelos propuestos en este artículo pueden abordar esta problemática directamente, sin necesidad de recurrir a técnicas para el tratamiento de datos faltantes. En un principio se podría pensar que las variables de decisión relacionadas con intervalos de edad con valores desconocidos pudieran tomar valores arbitrarios; sin embargo, la función objetivo obliga a estas variables a tomar valores próximos a las variables correspondientes a los intervalos más cercanos, tanto longitudinal como transversalmente. Es decir, si $P_{j,t}$ es desconocida, las variables de decisión $x_{i,t}$ relacionadas deben tomar valores cercanos a aquellos correspondientes a los intervalos E_{j-1} y E_{j+1} en el año t, y también a los correspondientes a E_{j-1} en el año $t-1$ y a E_{j+1} en el año $t+1$. En [4] se proporcionan los detalles de la adaptación de la metodología a esta problemática y se presenta un ejemplo de aplicación, donde se pone de manifiesto la robustez de los modelos propuestos frente a la presencia de datos faltantes.

La metodología propuesta en este artículo también es válida para abordar el fenómeno denominado *age heaping*, que se presenta cuando la persona (especialmente las de mayor edad o bajo nivel educativo) no proporciona su edad exacta en la encuesta o registro administrativo, sino que redondea su año de nacimiento al múltiplo más cercano de un valor determinado (usualmente 5 o 10), Figura 10. Este tipo de irregularidades pueden detectarse con ayuda de determinados índices, como son los de Whipple, Myers, Bachi o Zelnik, (Siegel y Swanson, 2004; Spoorenberg, 2008).

Se pueden aplicar los modelos aquí presentados para corregir los datos de población afectados por el fenómeno que acaba de ser descrito. Para ello, las cifras originales de población desagregadas se agrupan en intervalos de edad centrados en múltiplos de cinco o diez, según el tipo de apilamiento observado en los datos. Así, en el caso de 5 años, los grupos de edad serían 13-17, 18-22, 23-27, y así sucesivamente (las primeras edades se suelen mantener desagrupadas al no presentar este tipo de alteración). Teniendo en cuenta que en nuestra metodología los valores extremos de los intervalos, L_j y U_j, son de libre elección, podemos hacer uso de ella para desagregar los datos, reduciendo así el impacto del apilamiento de edades. Para ilustrar esta aplicación de los modelos propuestos, la Figura 11 muestra los datos de población originales (con ruido) de la población de España en 1930, claramente afectados por el fenómeno del age heaping, y los corregidos tras aplicar el procedimiento descrito.

Figura 10

Población observada, España 1930.

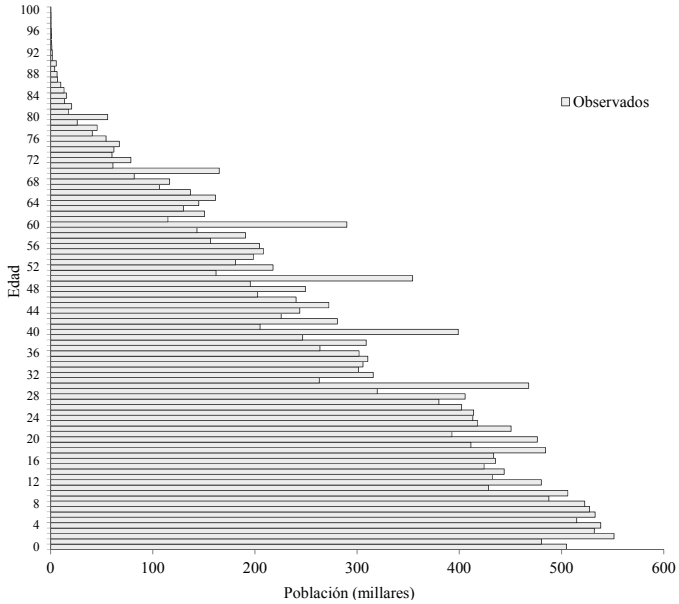
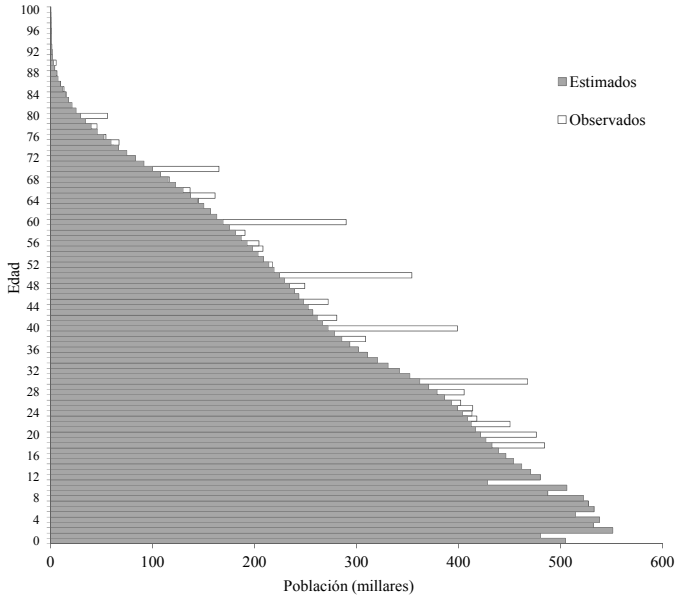


Figura 11

Población observada y estimada, España 1930.



Referencias

BERMÚDEZ, S., R. BLANQUERO, J.A. HERNÁNDEZ AND J. PLANELLES (2012). «A new parametric model for fitting fertility curves», *Population Studies: A Journal of Demography* 66(3): 297—310.

BERMÚDEZ, S. (2014). «Avances Metodológicos en Demografía», *Tesis Doctoral, Universidad de Sevilla*, España. Disponible en *http://fondosdigitales.us.es/tesis/tesis/2553/avances-metodologicos-en-demografia/*

BERMÚDEZ, S. (2014). «Incorporating demographic knowledge in parametric models for forecasting households size», *Estadística Española* 56(185): 381—399.

BERMÚDEZ, S. AND R. BLANQUERO (2016). «Optimization models for degrouping population data», *Population Studies: A Journal of Demography*. Por aparecer. Disponible en http://dx.doi.org/10.1080/00324728.2016.1158853

BOOT, J.C.G. (1964). «Quadratic Programming: Algorithms, Anomalies, Applications», *North-Holland Publishing Company*.

BUCHHEIM, C., A. CAPRARA, AND A. LODI (2012). «An effective branch-and-bound algorithm for convex quadratic integer programming», *Mathematical Programming* 135(1–2): 369–395.

COHEN, K.J., W. MULLER, AND M.W. PADBERG (1971). «Autoregressive Approaches to Disaggregation of Time Series Data», *Journal of the Royal Statistical Society. Series C* 20(2): 119–129.

CONGDON, K. (1993). «Statistical Graduation in Local Demographic Analysis and Projection», *Journal of the Royal Statistical Society. Series A* 156(2): 237–270.

IBM ILOG CPLEX OPTIMIZER.
http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/

DENTON, F.T. (1971). «Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization», *Journal of the American Statistical Association* 66(333): 99–102.

EUROSTAT, STATISTICAL OFFICE OF THE EUROPEAN UNION.
http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

EVERITT, B.S. (2012). «Introduction to Optimization Methods and their Application in Statistics», *Springer*.

FICO XPRESS OPTIMIZATION SUITE.
http://www.fico.com/en/Products/DMTools/Pages/FICO-XpressOptimization-Suite.aspx

FRANK M. AND P. WOLFE (1956). «An Algorithm for Quadratic Programming», *Naval Research Logistics Quarterly* 3(1–2): 95–110.

GUERRERO, V.M. (2003). «Monthly disaggregation of a Quarterly Time Series and Forescasts of Its Unobservable Monthly Values», *Journal of Official Statistics* 19(3): 215–235.

GUROBI OPTIMIZATION. .*http://www.gurobi.com*INE,

INSTITUTO NACIONAL DE ESTADÍSTICA (ESPAÑA). *http://www.ine.es*

KOSTAKI, A. AND V. PANOUSIS (2001). «Expanding an abridge life table», *Demographic Research* 5: 1–22.

KOSTAKI, A. AND J. LANKE (2000). «Degrouping mortality data for the elderly», *Mathematical Population Studies* 7(4): 333–341.

LISMAN, J.H.C. AND J. SANDEE (1964). «Derivation of Quarterly Figures from Annual Data», *Journal of the Royal Statistical Society. Series C* 13(2): 87–90.

MONTEIRO, R.D.C. AND ADLER, I. (1989). «Interior path following primal-dual algorithms. Part. II: Convex quadratic programming». *Mathematical Programming* 44(1): 43–66.

MCNEIL, D.R., T.J. TRUSSEL, AND J.C. TURNER (1977). «Spline Interpolation of Demographic Data», *Demography* 14(2): 245–252.

NEOS, SERVER FOR OPTIMIZATION. *http://www.neos-server.org/neos/*

PAVÍA-MIRALLES, J.M. (2010). «A Survey of Methods to Interpolate, Distribute and Extrapolate Time Series», *Journal of Service Science and Management* 3(4):449–463.

PROIETTI, T. (2011). «Multivariate Temporal Disaggregation with Cross-sectional Constraints», *Journal of Applied Statistics* 38(7): 1455–1466.

RUSTAGI, J.S. (1978). «Optimization in statistics: an overview», *Communication in Satistics - Simulation and Computation* 7(4): 303–307.

SALAZAR-GONZÁLEZ, J. J. (2008). «Statistical confidentiality: Optimization techniques to protect tables», *Computers & Operations Research* 35(5), 1638-1651.

SIEGEL, J.S. AND SWANSON D.A. (2004). «The Methods and Materials of Demography», *Elsevier Academic Press*.

SILVA, E., V.M. GUERRERO, AND D. PEÑA (2011). «Temporal disaggregation and restricted forecasting of multiple population time series», *Journal of Applied Statistics* 38(4): 799 – 815.

SPOORENBERG, T. (2008). «Quality of age reporting: extension and application of the modified Whipple's index», *Population* 62(4): 729–741.

VAN DE PANNE, C. AND WHINSTON, A. (1964), «The Simplex and the Dual Method for Quadratic Programming», *Operations Research Quarterly* 15(4): 355–388.

WEGMAN, J.E. AND W.I. WRIGHT (1983). «Splines in Statistics», *Journal of the American Statistical Association* 78(382): 351–365.

WEI, W.W.S. AND D.O. STRAM (1990). «Disaggregation of Time Series Models», *Journal of the Royal Statistical Society. Series B* 52(3): 453–467.

WILLIAMS, H.P. (2013). «Model building in mathematical programming», *John Wiley & Sons*.

WINSTON, W.L., M. VENKATARAMANAN AND J.B. GOLDBERG (2003). «Introduction to mathematical programming», *Thomson/Brooks/Cole*.

WOLSEY, L.A. (1998). «Integer Programming», *John Wiley & Sons*.