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

S. Bermúdez[(1)], R. Blanquero[(2)]

(1) Instituto de Estadística y Cartografía de Andalucía, Spain
silvia.bermudez.parrado@juntadeandalucia.es

(2) Facultad de Matemáticas, Universidad de Sevilla, Spain
rblanquero@us.es

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 ungrouped 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, several Mathematical Optimization models are proposed to address this important, yet seldom studied problem. The models also consider a frequent issue in statistical sources: the presence of noise and errors, and, in particular, age heaping.

Keywords: Degrouping population data, Integer Optimization, age-specific population, age heaping.

1 Introduction

Demography deals with human population and its dynamics; the study of the size, structure, and distribution of this population are encompassed in its task. This discipline has become a powerful tool for governments and business in carrying out effective planning and decision making for the future. Demography also helps to raise awareness of contemporary society and to evaluate the performance of political decisions.

Despite its importance, population data are not always published with the desired level of detail. Most National and International Statistical Offices provide figures for small geographical areas of countries, but at a low level of disaggregation. As an example, population data by single age can be found on the website of the Statistical Office of the European Union, Eurostat [8], but only for regions of sufficiently large size, namely those at the so-called NUTS 2 level or higher; however, for smaller regions, data appear grouped by age intervals. Furthermore, the information is only available in grouped form, even at the country level, if the target population is a subset of the overall population, as occurs with the foreign resident population and other groups of interest.

The knowledge of population figures by single age is crucial in some statistical operations, such as the calculation of demographic indicators. For instance, the estimation of the age-specific fertility rates, as a previous step to the assessment of total fertility rate, requires population data by single age. These ungrouped figures are also crucial in building projections of greater accuracy and for planning in fields where age plays a prominent role, such as in education.

The disaggregation of a general chronological series, a problem closely related to the one posed before, has been dealt with extensively, see [4, 5, 7, 11, 16, 20, 21, 26]. If we focus on demographic data, the expansion of an abridged series of mortality data given in age groups has received certain attention, see [5, 14] and references therein, while the disaggregation of series of population growth rates is addressed in [22]. Nevertheless, the problem of degrouping population figures by single age has barely been touched upon in the literature. When the data come from only one year, simple algorithms such as the Sprague method [23] or other spline interpolation techniques, e.g. [18, 25], can be used. These standard approaches lack one of the most desirable properties any procedure for disaggregating population should possess: the degrouped values at any age should be integer numbers, and these procedures fail to provide such numbers. On the other hand, it is far from trivial to adapt such procedures to the case in which the number of years under study is greater than one, due to the difficulty of assuring coherence between the populations of consecutive ages in

consecutive years. To the best of our knowledge, no research dealing with this problem in its full generality is available in the literature.

In this paper, this specific disaggregation problem is tackled; to this end, some mathematical optimization models are introduced that, starting from population data grouped into age intervals, allow to disaggregate them into single ages. In this manner, a singe value for each single age can be obtained in an optimal way in accordance with a certain criteria.

First we introduce the basic concepts of Mathematical Optimization, and we refer the reader to introductory texts such as [27, 28] for details. A mathematical optimization model, also known as a mathematical programming model, is a mathematical problem where, given a function $f : S \to \mathbb{R}$, with $S \subset \mathbb{R}^n$, we try to find a point $x^* \in S$ such that $f(x^*) \leq f(x)$ for all $x \in S$ (minimization) or $f(x^*) \geq f(x)$ for all $x \in S$ (maximization). These problems are usually represented in a compact form as follows,

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in S \end{array} \qquad \text{or} \qquad \begin{array}{ll} \text{maximize} & f(x) \\ \text{subject to} & x \in S \end{array}$$

The function f is called the *objective function* or simply *objective*, and the point x^*, the *optimal solution*. Since x is a vector in $\mathbb{R}^n$, it can be written as $x = (x_1, \ldots, x_n)$, where the components x_i are known as *decision variables*. The set S is the *feasible set* and is usually given by a set of constraints, i.e. expressions of the form $g(x) \leq 0$ or $h(x) = 0$. In that case, the optimization model can be written as follows (for minimization),

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_i(x) \leq 0 & i = 1, \ldots, m_1 \\ & h_j(x) = 0 & j = 1, \ldots, m_2 \end{array}$$

The difficulty of solving the optimization model depends on the type of functions involved. Problems where the objective and the constraints are linear functions, i.e. they have the form $c^T x = c_1 x_1 + c_2 x_2 + \ldots + c_n x_n$ with c_i constant values, are the easiest. Minimizing a convex quadratic function subject to linear constraints is also an easy-to-solve problem. A function is said to be quadratic when it can be written in the form $\frac{1}{2} x^T Q x + c^T x$, where Q is an $n \times n$ symmetric matrix, and where the function is convex when Q turns out to be positive semidefinite or, in other words, all its eigenvalues are non-negative. A quadratic function can easily be identified, since it is a summation of squared variables, products of different variables and single variables, all of which multiplied by

constant values. In certain optimization problems, like those considered in this paper, the variables are restricted to take on values from a discrete set, usually a subset of integer numbers, so that the model makes sense. The branch of optimization theory dealing with this kind of problem is known as Integer Optimization or Integer Programming, [29].

In this paper, we present mathematical optimization models for the problem of population degrouping, in which we seek mild transitions, longitudinally as well as transversally. To do that, for the different years and ages involved in the disaggregation process, we want to make the difference as small as possible between the number of individuals of ages i and $i+1$ in the year t and, at the same time, make the difference as small as possible between the number of individuals aged i in the year t and those aged $i+1$ in the year $t+1$. All these differences are aggregated into one single number, to be minimized. We consider here the most common aggregation method found in the literature, namely the L_2 model, in which the sum of squares of the differences is minimized. Other common aggregation methods, such as the L_1 and the L_∞ models, where the sum of the absolute values of the differences, and the highest absolute value of the differences are considered, respectively, were applied and compared with the L_2 models, and yielded worse results that those of the L_2 aggregation model and therefore these other methods have been excluded from this paper (see [1] for further details).

The remainder of this paper is structured as follows: in Section 2 we introduce some optimization models that allow population data grouped by age interval to be disaggregated. A brief description of the resolution techniques applied to solve these models, as well as some numerical results, can be found in Section 3, where the above-mentioned models are applied to the population figures of Andalusia, which is one of the largest autonomous regions of Spain. The paper finishes in Section 4 with a short summary of conclusions and with further extensions to the case in which data are affected by noise, whose main focus is on the age heaping phenomenon.

2 The models

We present several models aimed at the disaggregation of population figures in grouped form by age intervals. In Section 2.1 we begin by introducing a basic model that provides a distribution by age consistent with the intervals; the application of this model on empirical data shows that, generally speaking, the solutions provided are not fully satisfactory, mainly due to the large open-ended age interval. This steers us towards refining the basic model in Sections 2.2 and 2.3 by

using auxiliary information that is usually found at hand. Additionally, a model is introduced in Section 2.4 that deals with the disaggregation problem when geographic areas of different levels are considered within a hierarchical structure.

All of the models here proposed lead to quadratic problems in integer numbers, which can be faced by using a wide range of solvers, some of them freely available on the Internet at NEOS Server, [19]. As has been shown in [1], all of these models have a non-empty feasible set, which ensures that there exists at least one solution for the disaggregation problem. In fact, several feasible solutions can be obtained in a straightforward manner and they can be provided as initial solutions to the optimization software used to solve the problem.

2.1 Basic model

Given a geographic area s, we assume that, for each year t ($t = 1, \ldots, T$) the population is provided in G age intervals $E_j = \{L_j, L_j+1, \ldots, U_j-1, U_j\}$ ($j = 1, \ldots, G$) of variable length, such that $L_j = U_{j-1} + 1$. Let $P_{j,t}$ denote the population of the age group E_j in the calendar year t ($j = 1, \ldots, G$, $t = 1, \ldots, T$). In addition, let us denote by B_t the total births in the year $t-1$ in the area s. If the disaggregation process is being carried out over the foreign resident population, B_t stands for the total births to foreign-born women.

The decision variables of the disaggregation model are defined in a natural way in accordance with our aim. Hence, x_{it} will denote the population of the geographic area under consideration at the age i ($i = L_1, L_1 + 1, \ldots, U_G - 1, U_G$), for the year t ($t = 1, \ldots, T$). This yields a set of integer variables whose cardinality is given by

$$T \sum_{j=1}^{G} (U_j - L_j + 1) = T(U_G - L_1 + 1),$$

If only five-year age groups are considered, the previous expression leads to $5 \cdot G \cdot T$. If should be noted that our approach can hold the open-ended interval associated to the elderly population, provided that a reasonable upper bound U_G for that group is chosen.

Departing from the parameters $P_{j,t}$ and B_t ($j = 1, \ldots, G$, $t = 1, \ldots, T$) and the decision variables $x_{i,t}$ ($i = L_1, \ldots, U_G$, $t = 1, \ldots, T$), we consider a first degrouping model that is based on the minimization of the sum of the squares of the differences previously described. The following convex quadratic problem with linear constraints and integer numbers, solvable by existing Mathematical

Optimization techniques [2, 3, 10, 17, 24], is proposed.

$$(L_2BAS) \quad \begin{array}{ll} \min & \displaystyle\sum_{t=1}^{T}\sum_{i=L_1+1}^{U_G}(x_{i,t}-x_{i-1,t})^2+\sum_{t=1}^{T}(x_{L_1,t}-B_t)^2+\sum_{t=2}^{T}\sum_{i=L_1+1}^{U_G}(x_{i,t}-x_{i-1,t-1})^2 \\ \text{s.t.} & \\ & \displaystyle\sum_{i=L_j}^{U_j} x_{i,t}=P_{j,t} \qquad j=1,\ldots,G \qquad t=1,\ldots,T \\ & x_{i,t}\in\mathbb{Z}^+ \qquad i=L_1,\ldots,U_G \qquad t=1,\ldots,T \end{array}$$

In the previous model, $\mathbb{Z}^+$ denotes the set of non-negative integer numbers, $\mathbb{Z}^+=\{z\in\mathbb{Z}\ :\ z\geq 0\}$.

Three different blocks can be identified within the objective function of (L_2BAS), whose goal is described below:

- $\sum_{t=1}^{T}\sum_{i=L_1+1}^{U_G}(x_{i,t}-x_{i-1,t})^2$: its aim is to achieve smooth transitions from one individual age to the following within the same year t.
- $\sum_{t=1}^{T}(x_{L_1,t}-B_t)^2$: for each year t, this strives to approximate the initial population of age 0 to the number of births occurring the previous year.
- $\sum_{t=2}^{T}\sum_{i=L_1+1}^{U_G}(x_{i,t}-x_{i-1,t-1})^2$: this is aimed at ensuring a suitable time evolution for each generational cohort, by seeking close values for the population of age i at year t and the population of age $i-1$ at the previous year $t-1$.

Apart from the integrality constraints over the decision variables, our model only contains a block of constraints which allows us to accomplish an exact correspondence between the original abridged data and the disaggregated values by individual age attained as a solution of (L_2BAS).

In order to illustrate how the model (L_2BAS) is constructed, a short example is shown below.

Example 1 The population data of Malta are considered and, for the sake of simplicity, only 3 years (2008, 2009, and 2010) and 3 age intervals of the same length (0-4, 5-9, and 10-14) are taken into account. Hence, $T=3$ and $G=3$.

The age intervals are $E_1=\{0,1,2,3,4\}$, $E_2=\{5,6,7,8,9\}$ and $E_3=\{10,11,12,13\}$, and the aggregated population values $P_{i,j}$ (taken from the Eurostat database) are

$$\begin{array}{lll} P_{1,1}=19810 & P_{1,2}=20097 & P_{1,3}=20372 \\ P_{2,1}=21374 & P_{2,2}=20730 & P_{2,3}=20216 \\ P_{3,1}=25404 & P_{3,2}=24731 & P_{3,3}=23939 \end{array}$$

where the first subindex corresponds to the age interval, and the second subindex, to the year. The parameters $B_t, t = 1, 2, 3$ are given by the total number of births in Malta in 2007, 2008, and 2009, respectively, namely $B_1 = 3765$, $B_2 = 4013$, $B_3 = 4029$. We have a set of 45 decision variables (15 single ages multiplied by 3 years), $x_{0,1}, \ldots, x_{14,1}, x_{0,2}, \ldots, x_{14,2}, x_{0,3}, \ldots, x_{14,3}$.

With all these elements at hand, the model (L_2BAS) can be written in extended form as follows:

$$
\begin{aligned}
\min\ & \Big[(x_{1,1} - x_{0,1})^2 + \ldots + (x_{14,1} - x_{13,1})^2 + (x_{1,2} - x_{0,2})^2 + \ldots + (x_{14,2} - x_{13,2})^2 + \\
& + (x_{1,3} - x_{0,3})^2 + \ldots + (x_{14,3} - x_{13,3})^2\Big] + \\
& + \Big[(x_{0,1} - B_1)^2 + (x_{0,2} - B_2)^2 + (x_{0,3} - B_3)^2\Big] + \\
& + \Big[(x_{1,2} - x_{0,1})^2 + \ldots + (x_{14,2} - x_{13,1})^2 + (x_{1,3} - x_{0,2})^2 + \ldots + (x_{14,3} - x_{13,2})^2\Big]
\end{aligned}
$$

$$
\begin{aligned}
\text{s.t.}\ & x_{0,1} + x_{1,1} + x_{2,1} + x_{3,1} + x_{4,1} = 19810 \\
& x_{0,2} + x_{1,2} + x_{2,2} + x_{3,2} + x_{4,2} = 20097 \\
& x_{0,3} + x_{1,3} + x_{2,3} + x_{3,3} + x_{4,3} = 20372 \\
& x_{5,1} + x_{6,1} + x_{7,1} + x_{8,1} + x_{9,1} = 21374 \\
& x_{5,2} + x_{6,2} + x_{7,2} + x_{8,2} + x_{9,2} = 20730 \\
& x_{5,3} + x_{6,3} + x_{7,3} + x_{8,3} + x_{9,3} = 20216 \\
& x_{10,1} + x_{11,1} + x_{12,1} + x_{13,1} + x_{14,1} = 25404 \\
& x_{10,2} + x_{11,2} + x_{12,2} + x_{13,2} + x_{14,2} = 24731 \\
& x_{10,3} + x_{11,3} + x_{12,3} + x_{13,3} + x_{14,3} = 23939 \\
& x_{i,t} \in \mathbb{Z}^+ \qquad i = 0, \ldots, 14 \qquad t = 1, 2, 3
\end{aligned}
$$

For the sake of completeness, the value of the decision variables obtained after solving the model, and the actual disaggregated population figures, taken from the Eurostat database, are shown in Table 1.

Age	Observed			Adjusted		
	2008	2009	2010	2008	2009	2010
0	3,867	4,150	4,158	3,911	4,052	4,092
1	3,889	3,902	4,148	3,964	4,022	4,096
2	4,233	3,917	3,911	3,978	4,023	4,082
3	3,852	4,257	3,901	3,976	4,012	4,067
4	3,969	3,871	4,254	3,981	3,988	4,035
5	3,917	3,982	3,865	4,015	3,970	3,968
6	4,085	3,931	3,980	4,100	4,007	3,951
7	4,328	4,097	3,933	4,234	4,095	3,988
8	4,374	4,344	4,094	4,410	4,235	4,080
9	4,670	4,376	4,344	4,615	4,423	4,229
10	4,808	4,695	4,371	4,835	4,644	4,447
11	4,941	4,812	4,695	4,998	4,836	4,652
12	5,091	4,965	4,807	5,115	4,986	4,829
13	5,140	5,107	4,966	5,200	5,093	4,965
14	5,424	5,152	5,100	5,256	5,172	5,046

Table 1: Actual and estimated population at each age from 0 to 14, (L_2BAS) model, Malta 2008-2010

2.2 Model with auxiliary information on the open-ended age interval

According to our experiments, [1], the previous model usually yields a suitable fit in all the age groups, with the exception of the open-ended age interval E_G, where strong discrepancies between the results provided by the model and the actual data can be found. This is mainly due to the large length of the open interval and the greater freedom that the last individual age possesses, since it is not bounded by later values in the same or in the following year. Under these conditions, the solver seldom finds a convenient solution among all the optimal solutions of the model. The solution discovered by the solver, despite being optimal according to the minimization criterion, is usually far from acceptable due to the major fitting errors observed at the extreme years and ages considered in the open interval.

In order to solve this problem, we propose setting the requirement that the relative frequencies of the ages in the open-ended interval take values within certain specific intervals derived from the available figures. These could come from statistical operations where population data by individual age are available for the open-ended interval in the same or a neighboring year. Examples of these operations include the Population Census, any survey conducted on the subpopulation under study (such as the National Migration Survey) and population figures by individual age for a higher-level

population. Following this approach we define:

- $\underline{f}_i$: Lower bound for the relative frequency of the population aged i in the open-ended interval ($i = L_G, L_G + 1, \ldots, U_G$).
- $\overline{f}_i$: Upper bound for the relative frequency of the population aged i in the open-ended interval ($i = L_G, L_G + 1, \ldots, U_G$).

For each age belonging to the open-ended interval E_G, it will be required that its relative frequency within that interval falls between $\underline{f}$ and $\overline{f}$, i.e.,

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

or, equivalently

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

If $\underline{f}_i$ and $\overline{f}_i$ are obtained, for instance, from the available census-based information, these bounds could be calculated using the relative distribution of the interval E_G in the immediately prior and immediately subsequent census to the period of years being disaggregated. If only one of those census is available, the relative frequencies of E_G obtained from this census could be used as the central values of the interval $[\underline{f}_i, \overline{f}_i]$ and their ranges could be determined proportionally to the above-mentioned central values. Another possibility consists of using the disaggregated age data from a higher-level population, if they are available.

After adding the new group of above-described constraints to (L_2BAS), the following model is obtained:

$$\begin{aligned}
& \min \quad \sum_{t=1}^{T} \sum_{i=L_1+1}^{U_G} (x_{i,t} - x_{i-1,t})^2 + \sum_{t=1}^{T} (x_{L_1,t} - B_t)^2 + \sum_{t=2}^{T} \sum_{i=L_1+1}^{U_G} (x_{i,t} - x_{i-1,t-1})^2 \\
& \text{s.t.} \\
(L_2INT_1) \qquad & \sum_{i=L_j}^{U_j} x_{i,t} = P_{j,t} \qquad j = 1, \ldots, G \qquad t = 1, \ldots, T \\
& x_{i,t} \leq \overline{f}_i P_{G,t} \qquad i = L_G, L_G + 1, \ldots, U_G \quad t = 1, \ldots, T \\
& x_{i,t} \geq \underline{f}_i P_{G,t} \qquad i = L_G, L_G + 1, \ldots, U_G \quad t = 1, \ldots, T \\
& x_{i,t} \in \mathbb{Z}^+ \qquad i = L_1, \ldots, U_G \qquad t = 1, \ldots, T
\end{aligned}$$

If the external information required remains unavailable, then the problem can be addressed in an alternative manner: by assuming that population decreases with the age in the open-ended interval, thereby yielding the following model.

$$
\begin{aligned}
& \min \quad \sum_{t=1}^{T} \sum_{i=L_1+1}^{U_G} (x_{i,t} - x_{i-1,t})^2 + \sum_{t=1}^{T} (x_{L_1,t} - B_t)^2 + \sum_{t=2}^{T} \sum_{i=L_1+1}^{U_G} (x_{i,t} - x_{i-1,t-1})^2 \\
& \text{s.t.} \\
(L_2INT_2) \qquad & \sum_{i=L_j}^{U_j} x_{i,t} = P_{j,t} \qquad j = 1, \ldots, G \qquad t = 1, \ldots, T \\
& x_{i,t} \geq x_{i-1,t} \qquad i = L_G, L_G + 1, \ldots, U_G \quad t = 1, \ldots, T \\
& x_{i,t} \in \mathbb{Z}^+ \qquad i = L_1, \ldots, U_G \qquad t = 1, \ldots, T
\end{aligned}
$$

Observe that such a monotonicity assumption is rather realistic, with the exception of countries deeply influenced by late-age immigration.

2.3 Model with contour information

When degrouping population data, one usually has auxiliary information at hand which may be of help to model cohorts, in addition to births. An interesting case arises when population figures by individual age are available for the years surrounding to those taking part in the disaggregation process, as can be the case of the population census, which is carried out every ten years in many countries.

Occasionally, one can even count on individual population figures for the extreme years of the period under consideration, although they are supplied by other statistical sources. Under those circumstances, this disaggregated information can be added to the model in order to lead the fitting process at the initial and/or final years, thereby improving the solutions provided by the basic model. With this aim in mind, we define:

- I_i: population of individual age i ($i = L_1, \ldots, U_G$) at the initial or a closely preceding year.
- F_i: population of individual age i ($i = L_1, \ldots, U_G$) at the final or a closely succeeding year.

Starting from the model (L_2BAS), its objective function must be modified so that the decision variables regarding the initial and final years take values near I_i and F_i, respectively. This yields the following optimization problem,

$$
(L_2CON)\quad
\begin{aligned}
\min\ & \sum_{t=1}^{T}\sum_{i=L_1+1}^{U_G}(x_{i,t}-x_{i-1,t})^2+\sum_{t=1}^{T}(x_{L_1,t}-B_{t-1})^2+\sum_{t=2}^{T}\sum_{i=L_1+1}^{U_G}(x_{i,t}-x_{i-1,t-1})^2+\\
& +\sum_{i=L_1}^{U_G}(x_{i,1}-I_i)^2+\sum_{i=L_1}^{U_G}(x_{i,T}-F_i)^2\\
\text{s.t.}\ & \\
& \sum_{i=L_j}^{U_j}x_{i,t}=P_{j,t} \qquad j=1,\ldots,G \qquad t=1,\ldots,T\\
& x_{i,t}\in\mathbb{Z}^+ \qquad i=L_1,\ldots,U_G \qquad t=1,\ldots,T
\end{aligned}
$$

2.4 Model with higher-level information

Another variant of interest of the models introduced previously arises when the disaggregated population of a geographic area of higher level than those involved in the degrouping process is considered, in such a way that the union of these areas gives the former area. This is a common situation in Official Statistics and, for instance, the National Statistics Institute (INE) of Spain published provincial population figures in quinquennial age groups, whereas the regional figures were provided in a disaggregated way in terms of a single year of age. In January 2002, the dissemination of the so-called *Population Now-Cast* began, thereby providing unabridged population figures at provincial level. However, the provincial population figures of the period 1970-2001 remain aggregated by age group on the website of the INE, [13].

In order to develop this model, it is assumed that, for each geographic area s ($s=1,2,\ldots,S-1,S$), and each calendar year t ($t=1,\ldots,T$), G age population groups of variable length $E_j=\{L_j,L_j+1,\ldots,U_j-1,U_j\}$ ($j=1,\ldots,G$) are available. Let us denote by $P_{j,s,t}$ the population of the age group E_j, in the area s, for the calendar year t ($j=1,\ldots,G$, $s=1,\ldots,S$, $t=1,\ldots,T$).

We also assume that the population by single age in the higher-level area is known; this will be denoted by $Q_{i,t}$, where the subindex i refers to the age and the subindex t stands for the calendar year ($i=L_1,\ldots,U_G$, $t=1,\ldots,T$). From the definition of $P_{j,s,t}$ y $Q_{i,t}$, it follows that the following relation must be fulfiled:

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

Finally, $B_{s,t}$ stands for the number of births registered over the year $t-1$ in the territorial area s ($s = 1, \ldots, S$, $t = 1, \ldots, T$).

The decision variables of this model are defined in the usual manner, but S subareas are now taken into account. In this way, the decision variable $x_{i,s,t}$ will denote the population of age i ($i = L_1, \ldots, U_G$) in the territorial area s ($s = 1, \ldots, S$) for the calendar year t ($t = 1, \ldots, T$).

As in the previous cases, the first model taken into consideration in this context strives to minimize the sum of the squares of the differences related to the main goals of the disaggregation problem under study, namely:

- Mild transitions from one individual age to the following within each year and territorial area.
- Approximation of the initial population of age 0 to the number of births occurring the previous year in each area.
- Mild transitions between consecutive ages of consecutive years in each territorial area.

Using the previous set of variables and bearing in mind the above-mentioned aims, the disaggregation model can be formulated as the following convex quadratic problem with linear constraints:

$$
\begin{aligned}
& \min \quad \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{i=L_1+1}^{U_G}(x_{i,s,t}-x_{i-1,s,t})^2+\sum_{s=1}^{S}\sum_{t=1}^{T}(x_{L_1,s,t}-B_{s,t})^2+ \\
& \qquad +\sum_{s=1}^{S}\sum_{t=2}^{T}\sum_{i=L_1+1}^{U_G}(x_{i,s,t}-x_{i-1,s,t-1})^2 \\
(L_2HIG) \qquad & \text{s.t.} \\
& \qquad \sum_{i=L_j}^{U_j} x_{i,s,t} = P_{j,s,t} \quad j = 1, \ldots, G \qquad s = 1, \ldots, S \quad t = 1, \ldots, T \\
& \qquad \sum_{s=1}^{S} x_{i,s,t} = Q_{i,t} \qquad i = L_1, \ldots, L_G \qquad\qquad\qquad t = 1, \ldots, T \\
& \qquad x_{i,s,t} \in \mathbb{Z}^+ \qquad\quad i = L_1, \ldots, U_G \quad s = 1, \ldots, S \quad t = 1, \ldots, T
\end{aligned}
$$

The first set of constraints allows us to ensure, for each age interval, calendar year, and territorial area, that the sum of the disaggregated population figures is equal to the total of the related age interval; these are essentially the constraints appearing in the model (L_2BAS).

The second set of constraints aims to ensure the agreement between the population figures in the territorial areas and those in the higher-level area. Thus, these constraints guarantee that, for each individual age and calendar year, the population figures of the higher level area agree with the sum of the corresponding values in the subareas of which it is comprised.

3 Solving the disaggregation models

The models introduced in the previous section have been solved using the solver Cplex v12.2, which is a component of the IBM ILOG Cplex Optimization Studio, [6]. Cplex is a well-known software product aimed at the resolution of linear and quadratic integer problems. Other software tools offering the same functionality, such as Gurobi, [12], and Xpress, [9], could equally have been applied.

All the models have been formulated using OPL, the algebraic modeling language used by Cplex Optimization Studio. The syntax of this kind of language is similar to the mathematical notation used to formulate the optimization problem and allows a perfect separation between the model and the data.

The models considered in this paper involve integer programming problems, which require large computational times as a rule; however, the run times for our disaggregation models are very short and a few seconds are usually sufficient to find an optimal solution.

In order to evaluate and compare the performance of the above-mentioned models, empirical population data from the region of Andalusia (Spain) covering several years were used. These were taken from the so-called Municipal Register of Inhabitants, which is an administrative register where every inhabitant of Spain has to be accounted for. We began by aggregating the population figures by single year of age as found in the municipal register in five-year age groups plus a large open-ended interval for the ages 85 and over. The models previously described were then applied to these abridged data, thereby yielding disaggregated population figures that will be referred to as *estimated data*. The empirical values from the municipal register, referred on what follows as *observed data*, were compared to the estimated ones in order to test the plausibility of our disaggregation scheme. This comparison was carried out by using the Root Mean Squared Relative Error ($RMSRE$) as accuracy measure, which, for an area s and a year t, is defined as:

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

where $O_{i,s,t}$ are the observed values and $x_{i,s,t}$ the estimated values ($i = L_1, \ldots, U_G, s = 1, \ldots, S, t = 1, \ldots, T$).

A period of ten years, ranging from 1999 to 2008 inclusive, is considered in all the experiments, except for the models with contour information, (L_2CON), where the period 2002-2010 is considered in order to make use of the census carried out in Spain in 2001. The population data used in this computational experience were collected from the website of the National Statistics Institute of Spain, [13]. The accuracy results, in terms of Root Mean Squared Relative Error (2), are shown in Tables 2 – 3. The best-performing model seems to be (L_2INT_1). The model (L_2INT_2) provides worse results than the previous model, with a slight improvement with respect to the model (L_2BAS). The difference in the accuracy of (L_2INT_1) and (L_2INT_2) can be caused by the stronger assumptions used in the former model.

MODELS	RMSRE												Mean	Standard deviation
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010		
Basic model														
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
Model with open-ended age interval information														
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
Model with contour information														
L_2CON				0.27	0.18	0.08	0.10	0.13	0.17	0.17	0.18	0.20	0.17	0.05

Table 2: Root Mean Squared Relative Error by disaggregation model, Andalusia

Model with higher-level information	RMSRE										Mean	Standard deviation
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008		
L_2HIG												
Almería	0.21	0.17	0.17	0.10	0.06	0.13	0.22	0.28	0.41	0.60	0.24	0.15
Cádiz	0.20	0.20	0.16	0.16	0.16	0.12	0.09	0.11	0.08	0.09	0.14	0.04
Córdoba	0.22	0.21	0.14	0.14	0.10	0.10	0.09	0.06	0.06	0.05	0.12	0.06
Granada	0.15	0.13	0.11	0.09	0.09	0.07	0.04	0.03	0.06	0.10	0.09	0.03
Huelva	0.21	0.14	0.17	0.15	0.12	0.09	0.17	0.28	0.52	0.69	0.25	0.19
Jaén	0.20	0.19	0.11	0.05	0.06	0.05	0.06	0.07	0.13	0.15	0.11	0.06
Málaga	0.14	0.13	0.09	0.09	0.06	0.03	0.04	0.06	0.13	0.15	0.09	0.04
Seville	0.47	0.28	0.19	0.09	0.06	0.05	0.09	0.11	0.15	0.20	0.17	0.12

Table 3: Root Mean Squared Relative Error for models (L_2HIG) by province

The result obtained from the disaggregation process are now shown graphically. Figures 1-4 provide

illustrations of the performance of the various models applied to the observed data of Andalusia. For the sake of concision, the year 2004 in those figures is dealt with exclusively.

Regarding the model (L_2HIG) and for the same reason, only results for the Andalusian provinces of Granada and Seville are depicted graphically, (Figures 5-6).

A more thorough analysis and further results can be found in [1].

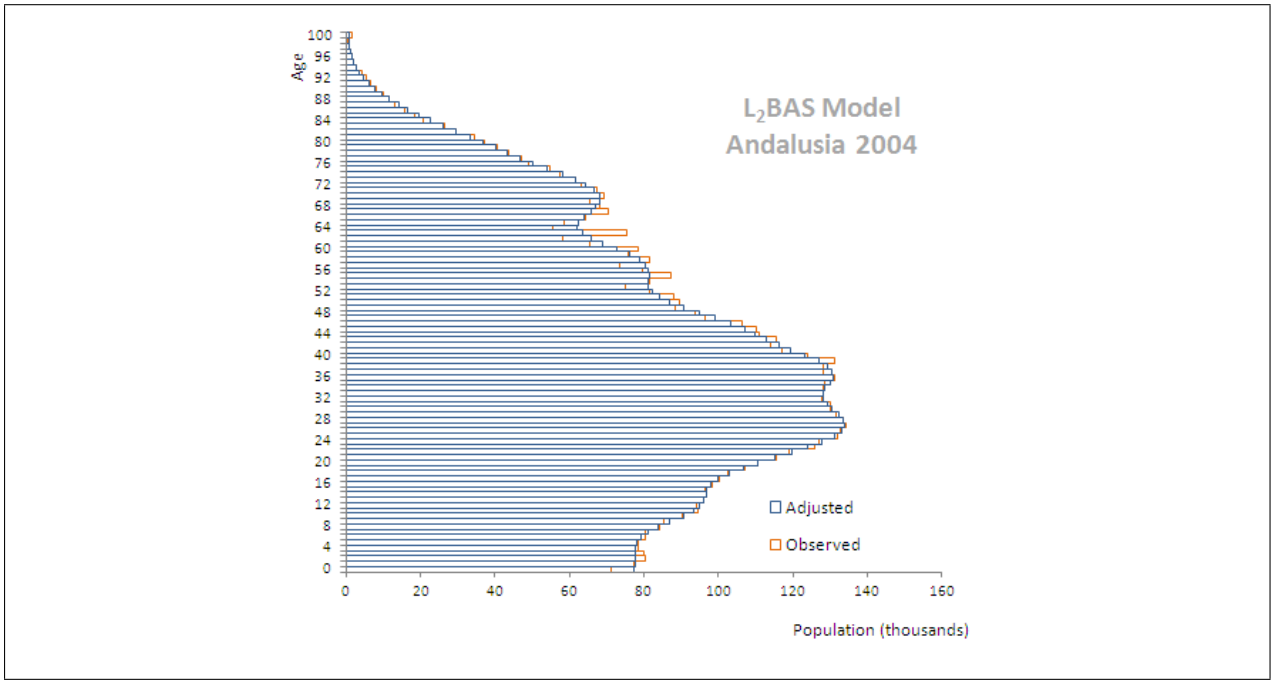

Figure 1: Age-disaggregated population, (L_2BAS) model, Andalusia, 2004.

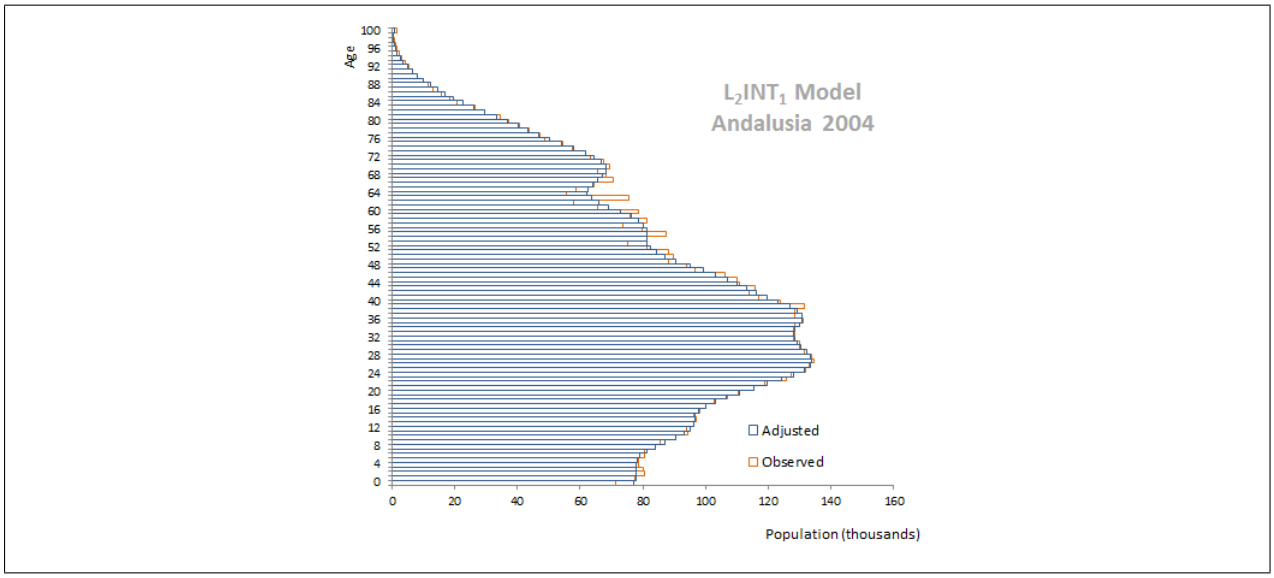

Figure 2: Age-disaggregated population, (L_2INT_1) model, Andalusia, 2004.

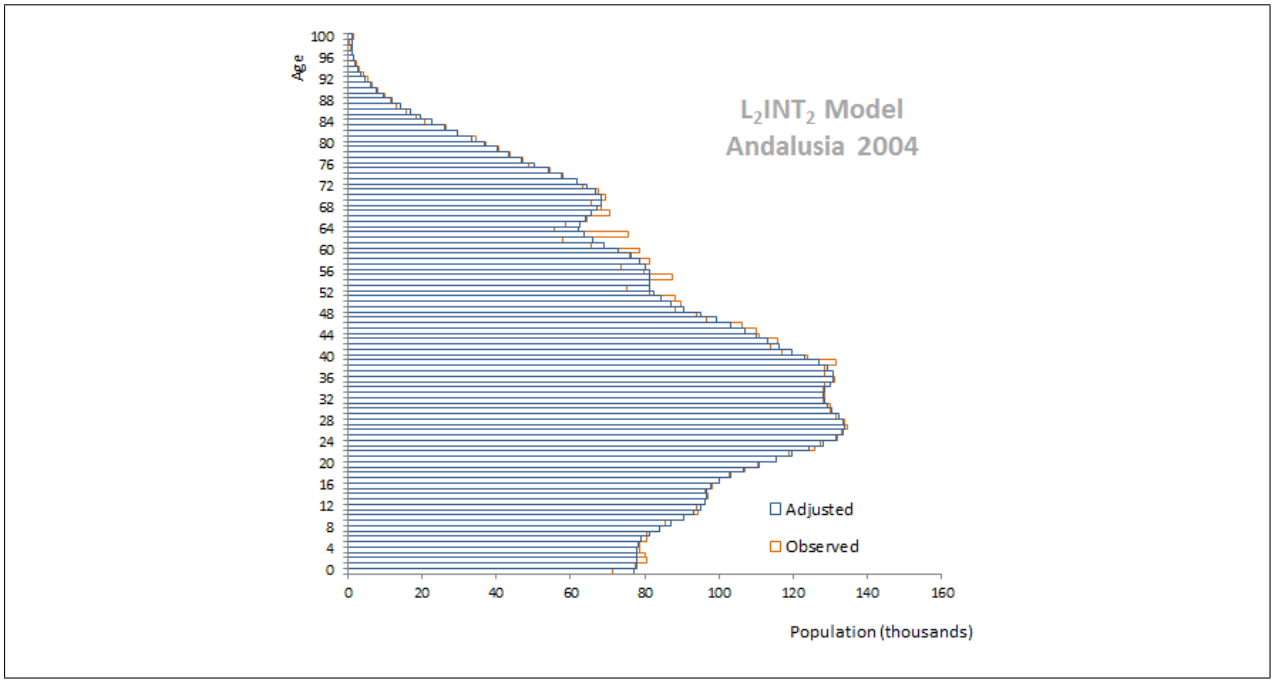

Figure 3: Age-disaggregated population, (L_2INT_2) model, Andalusia, 2004.

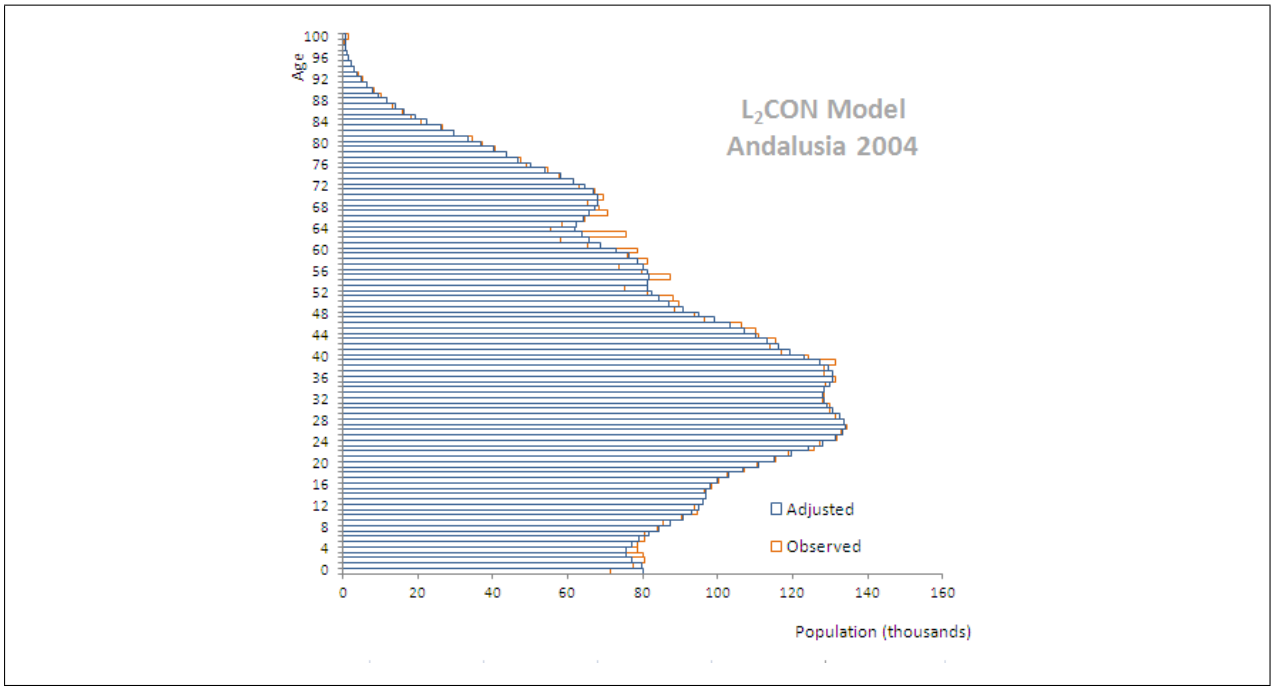

Figure 4: Age-disaggregated population, (L_2CON) model, Andalusia, 2004.

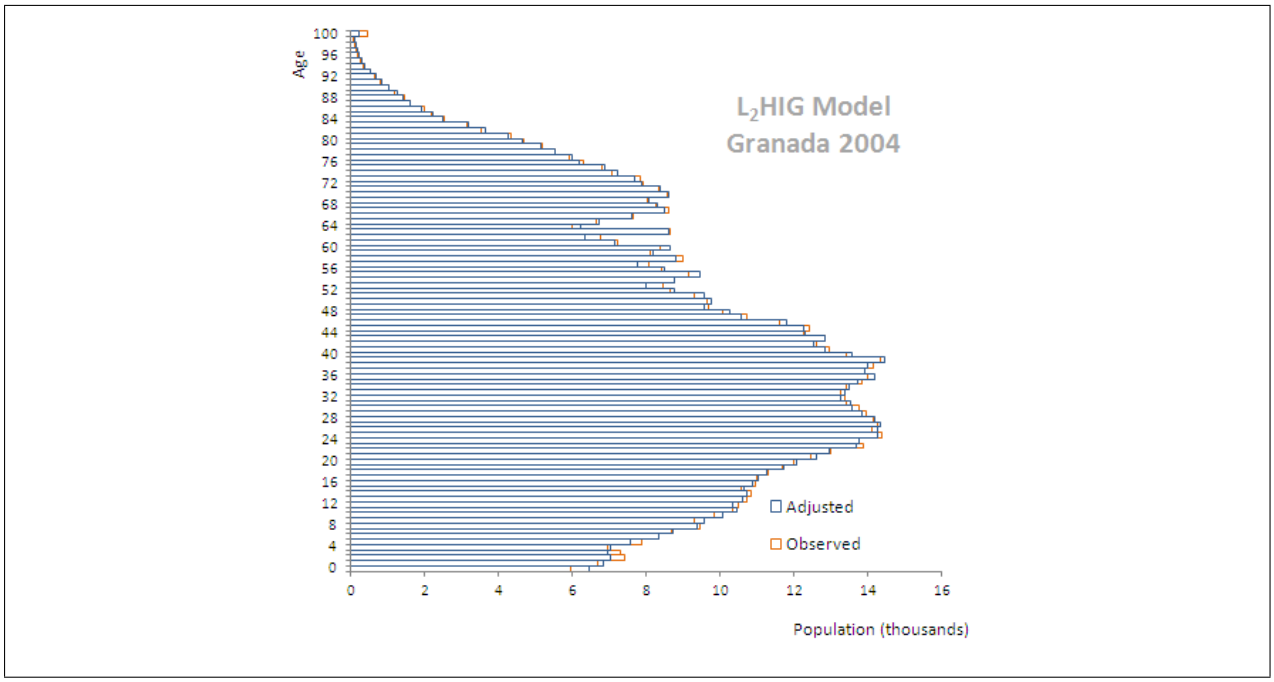

Figure 5: Age-disaggregated population, (L_2HIG) model, Granada, 2004.

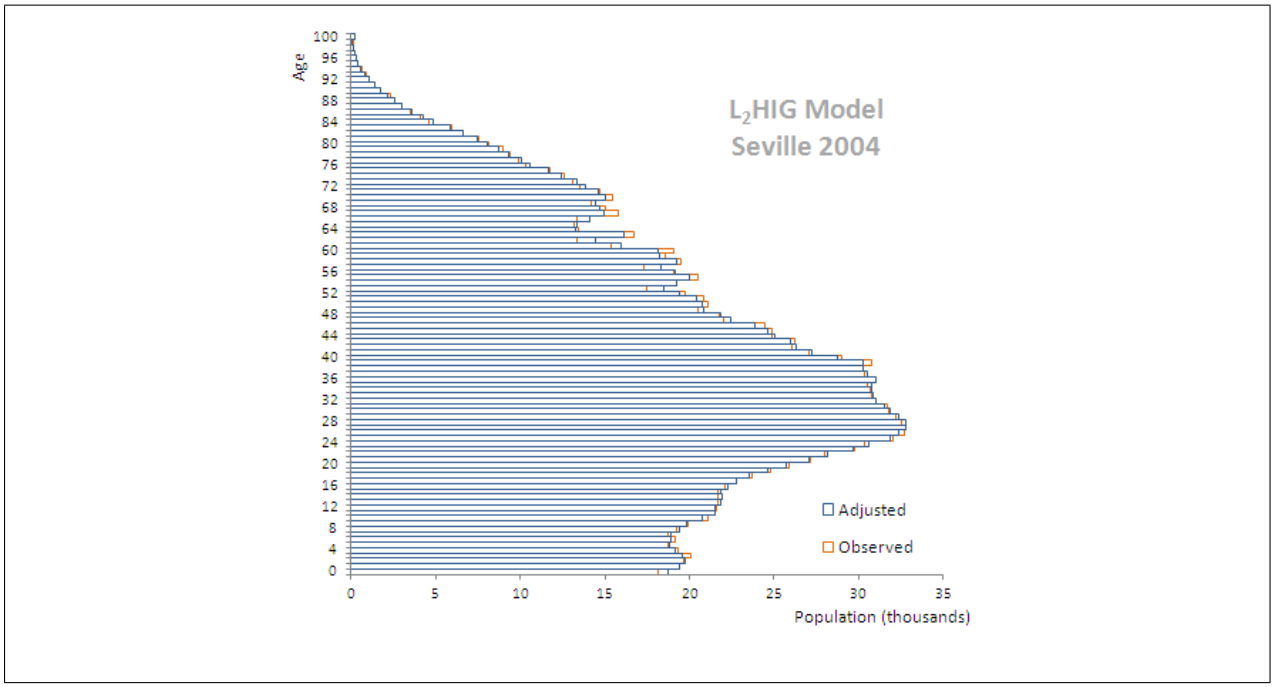

Figure 6: Age-disaggregated population, (L_2HIG) model, Seville, 2004.

For the sake of completeness, other territorial areas have also been considered in the disaggregation process. As an illustration, Figure 7 presents the results obtained after applying the model (L_2CON) to the empirical data of Ireland obtained from the website of Eurostat.

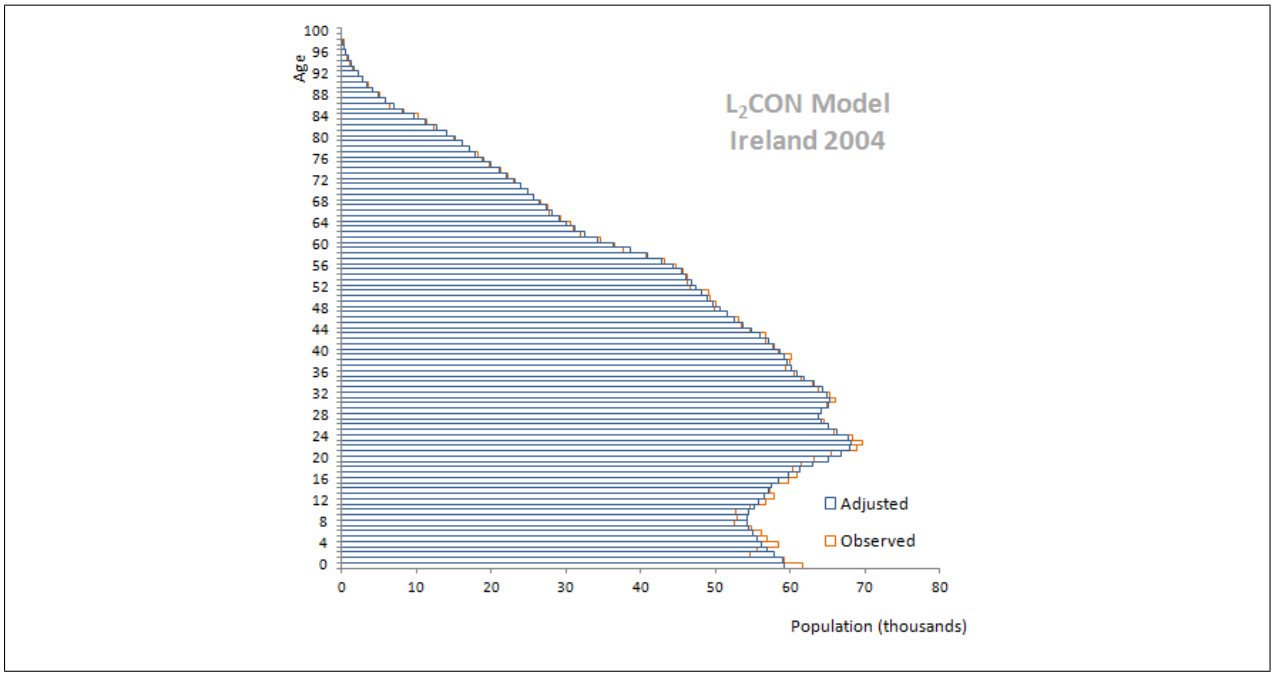

Figure 7: Age-disaggregated population, (L_2CON) model, Irlanda, 2004.

The models proposed in this article strive to achieve mild transitions across ages, not only transversally but also longitudinally. The figures that have been presented previously show that our approach provides accurate results when a transversal analysis of empirical and observad data is performed. Now we focus on the longitudinal analysis. Figure 8 depicts the evolution of three generational cohorts throughout the period 1999-2008; it presents empirical data as well as adjusted ones provided by the (L_2INT) model. Table 4 shows the result of calculating the RMSRE over the three cohorts in the period 1999-2008.

Two main conclusions emerge from the longitudinal analysis: firstly, the evolution of each cohort is smooth, which fulfils one of our aims, and secondly, small differences are found between observed and adjusted figures. Similar conclusions can also be reached if we consider the other disaggregation models. Hence, we conclude that both transversal and longitudinal adjustments are suitable.

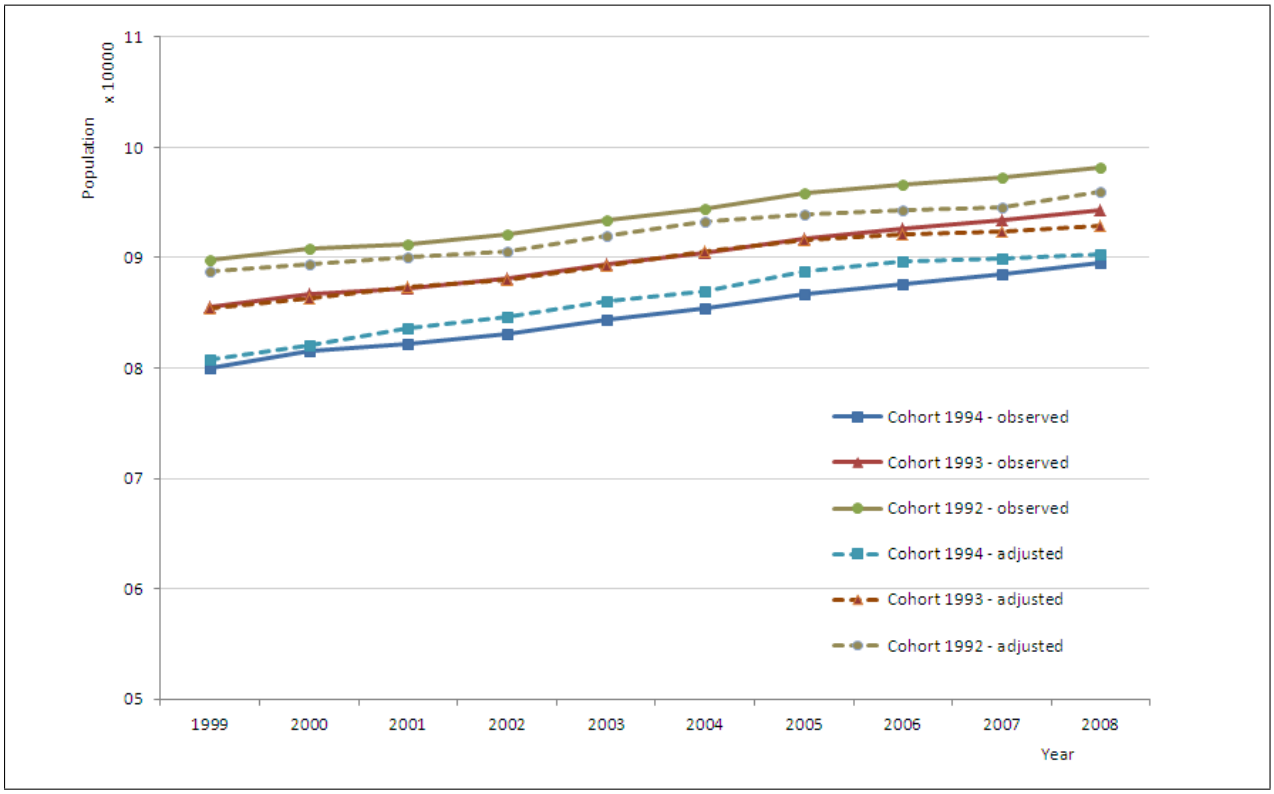

Figure 8: Age-disaggregated population by cohort, (L_2INT_1) model, Andalusia, 1999 – 2008.

MODELS	RMSRE			Mean	Standard deviation
	1994	1993	1992		
Basic model					
L_2BAS	0.02	0.01	0.02	0.01	0.01
Model with open-ended age interval information					
L_2INT$_1$	0.02	0.01	0.02	0.01	0.01
L_2INT$_2$	0.02	0.01	0.02	0.01	0.01
Model with contour information					
L_2CON	0.02	0.01	0.02	0.02	0.00

Table 4: Root Mean Squared Relative Error by disaggregation model and cohort, Andalusia, 1999 – 2008.

4 Concluding remarks and extensions

This article describes several optimization models for the disaggregation of population figures by single age from empirical data grouped in intervals. These models fulfil certain reasonable properties for any disaggregation methodology, namely:

- The disaggregated population values are integer numbers. Hence, no ad-hoc rounding is needed.
- For a given year, cross-sectional consistency is shown.
- For a given age, cross-cohort consistency is shown, thereby maintaining the idiosyncrasy of each cohort.

Starting from a basic model that fulfils the basic requirements demanded in the disaggregation process, this model is enriched so that, from a demographic point of view, solutions that are more suitable can be obtained. The proposed models lead to quadratic integer optimization problems, which can be solved in a few seconds on a personal computer using *Cplex* or other similar optimization software.

Model validation was carried out using population figures from the region of Andalusia (Spain) and its provinces. A graphical and numerical analysis of the results reveals that the proposed methodology provides suitable longitudinal and transverse fits to the empirical data.

The above-described models can be easily modified to deal with hypotheses of a more general nature than those ones considered previously in this paper. For instance, it has been implicitly assumed that the population figures are accurate. Nevertheless, these figures are usually based on censuses and surveys, both of which may contain errors that, when they are sufficiently large, could distort the disaggregation process. For the sake of simplicity, we only consider the model (L_2BAS), and a similar approach could be used for all the models described so far. Assuming that the values $P_{j,t}$ may be affected by errors, the set of constraints

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

might transfer the errors to the disaggregated population. In order to face this problem, the previous constraints are transformed into *soft constraints* as it is described below. Let

$$X_{j,t} = \sum_{i=L_j}^{U_j} x_{i,t}$$

be the estimated population of each age group. Note that $X_{j,t}$ are not actually decision variables, since they are perfectly defined by the actual variables $x_{i,t}$. For each year, it is then imposed that the total estimated population equals the total reference population,

$$\sum_{j=1}^{G} X_{j,t} = \sum_{j=1}^{G} P_{j,t} \qquad t = 1, \ldots, T.$$

The objective function must be modified so that the model seeks solutions close to the reference data. This is achieved by adding the following term,

$$\sum_{t=1}^{T} \sum_{j=1}^{G} (X_{j,t} - P_{j,t})^2$$

In this way, a more flexible model is obtained. Without disregarding the reference values, the model can, to a certain extent, separate the solutions from these values with the aim of obtaining a better fit according to the minimization criterion considered in the model.

According to the previous considerations, the model (L_2BAS) could be rewritten as follows.

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

s.t.

$$\sum_{j=1}^{G} X_{j,t} = \sum_{j=1}^{G} P_{j,t} \qquad t = 1, \ldots, T$$

$$X_{j,t} = \sum_{i=L_j}^{U_j} x_{i,t} \qquad j = 1, \ldots, G \qquad t = 1, \ldots, T$$

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

Throughout this article, it has been assumed that population figures are available every year, but this is not necessarily true, especially in developing countries. The lack of data in certain years may be overcome by applying commonly used techniques (e.g. interpolation of the available data, either by means of spline functions or by following a distributional model that is considered suitable in view of the data). However, the models proposed in this paper can deal directly with this situation, without resorting to missing data techniques. Indeed, in the model (L_2BAS) the indices t with

unknown B_t (number of births in the previous year) should be excluded from the middle summation; in addition, if a parameter $P_{j,t}$ is not available, the relevant constraint

$$\sum_{i=L_j}^{U_j} x_{i,t} = P_{j,t}$$

should be removed from the optimization problem.

At first glance, the decision variables $x_{i,t}$ related to age intervals with unknown $P_{j,t}$ might take arbitrary values. However, the objective function has to lead them to take values close to the variables corresponding to the nearest intervals, both longitudinally and transversely. More precisely, if $P_{j,t}$ is unknown, the relevant variables $x_{i,t}$ have to take values close to those corresponding to the intervals E_{j-1} and E_{j+1} in the calendar year t, and also close to those corresponding E_{j-1} in the year $t-1$ and E_{j+1} in the year $t+1$. To illustrate this, the data of Malta used in Example 1 is again considered, assuming now that the figures of 2009 (index $t = 2$) are unknown. Hence, the term $(x_{0,2} - B_2)^2$ is removed from the objective function, as well as the constraints involving $P_{1,2}$, $P_{2,2}$ and $P_{3,2}$, thereby yielding the following model.

$$\begin{aligned}
\min\ & \Big[(x_{1,1} - x_{0,1})^2 + \ldots + (x_{14,1} - x_{13,1})^2 + (x_{1,2} - x_{0,2})^2 + \ldots + (x_{14,2} - x_{13,2})^2 + \\
& + (x_{1,3} - x_{0,3})^2 + \ldots + (x_{14,3} - x_{13,3})^2\Big] + \\
& + \Big[(x_{0,1} - B_1)^2 + (x_{0,3} - B_3)^2\Big] + \\
& + \Big[(x_{1,2} - x_{0,1})^2 + \ldots + (x_{14,2} - x_{13,1})^2 + (x_{1,3} - x_{0,2})^2 + \ldots + (x_{14,3} - x_{13,2})^2\Big]
\end{aligned}$$

$$\begin{aligned}
\text{s.t.}\ & x_{0,1} + x_{1,1} + x_{2,1} + x_{3,1} + x_{4,1} = 19810 \\
& x_{0,3} + x_{1,3} + x_{2,3} + x_{3,3} + x_{4,3} = 20372 \\
& x_{5,1} + x_{6,1} + x_{7,1} + x_{8,1} + x_{9,1} = 21374 \\
& x_{5,3} + x_{6,3} + x_{7,3} + x_{8,3} + x_{9,3} = 20216 \\
& x_{10,1} + x_{11,1} + x_{12,1} + x_{13,1} + x_{14,1} = 25404 \\
& x_{10,3} + x_{11,3} + x_{12,3} + x_{13,3} + x_{14,3} = 23939 \\
& x_{i,t} \in \mathbb{Z}^+ \qquad i = 0, \ldots, 14 \qquad t = 1, 2, 3
\end{aligned}$$

As shown in Table 5, the fitting procedure provides accurate results, despite the missing data on 2009. In fact, the population estimates of that year by age group turn out to be $\hat{P}_{1,2} = 20054$, $\hat{P}_{2,2} = 20836$, and $\hat{P}_{3,2} = 24570$, which are close to the actual values, $P_{1,2} = 20097$, $P_{2,2} = 20730$, and $P_{3,2} = 24731$.

	Observed			Adjusted		
Age	2008	2009	2010	2008	2009	2010
0	3,867	4,150	4,158	3,909	4,058	4,100
1	3,889	3,902	4,148	3,959	4,014	4,103
2	4,233	3,917	3,911	3,973	4,008	4,082
3	3,852	4,257	3,901	3,977	3,996	4,061
4	3,969	3,871	4,254	3,992	3,978	4,026
5	3,917	3,982	3,865	4,023	3,987	3,953
6	4,085	3,931	3,980	4,108	4,030	3,946
7	4,328	4,097	3,933	4,238	4,121	3,991
8	4,374	4,344	4,094	4,406	4,259	4,087
9	4,670	4,376	4,344	4,599	4,439	4,239
10	4,808	4,695	4,371	4,826	4,626	4,463
11	4,941	4,812	4,695	4,989	4,808	4,658
12	5,091	4,965	4,807	5,109	4,952	4,826
13	5,140	5,107	4,966	5,199	5,056	4,957
14	5,424	5,152	5,100	5,281	5,128	5,035

Table 5: Actual and estimated population at each age from 0 to 14, (L_2BAS) model, Malta 2008-2010. The estimates were obtained without information of the year 2009.

The methodology proposed in this article is also useful to address the phenomenon known as *age heaping* or *digit preference*, that occurs because many people (particularly older people or people with a low level of education) tend not to give their exact age in surveys or censuses. Instead, they usually round their age up or down to the nearest number ending in 0 or 5. These irregularities can be detected by means of age heaping indices, such as Whipple's index, Myers' index, Bachi's index or Zelnik's index.

Our models can also be used to restore population data without age heaping. To this end, the original age-specific population counts are grouped by age intervals, and have multiples of five as the central age in each group. Hence, the age groups 13-17, 18-22 and so on (the first ages remain ungrouped) are considered. The fact that the extreme values L_j and U_j of the intervals can be freely chosen in our methodology, can be used to disaggregate the data, thereby reducing the impact of age heaping. These artificially-grouped data are then disaggregated by solving the optimization problem (L_2BAS) or its variants. By construction, the so-obtained disaggregated data have mild transitions, both transversally and longitudinally, and thus the age heaping is eliminated.

As an illustration, Figure 6 shows both original (noisy) and restored population data of Mexico in 2005, which are clearly affected by age heaping.

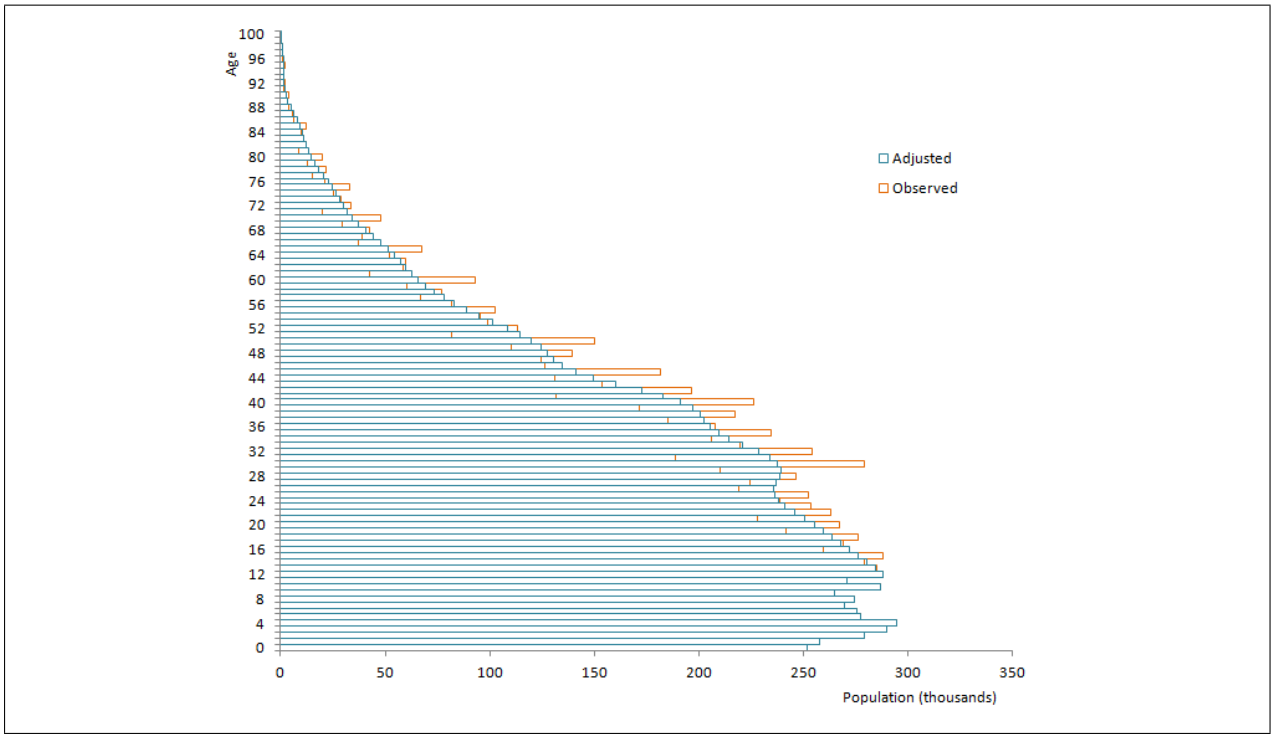

Table 6: Actual and estimated population, Mexico 2005.

References

[1] Bermúdez, S. (2014), *Avances Metodológicos en Demografía*, PhD Dissertation, University of Seville, Spain. Available at *http://fondosdigitales.us.es/tesis/tesis/2553/avances-metodologicos-en-demografia/*

[2] Boot, J.C.G. (1964), *Quadratic Programming: Algorithms, Anomalies, Applications*, North-Holland Publishing Company.

[3] 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.

[4] 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.

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

[6] *IBM ILOG CPLEX Optimizer.* *http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/*

[7] 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.

[8] *EUROSTAT, Statistical Office of the European Union.*
http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

[9] *FICO Xpress Optimization Suite.*
http://www.fico.com/en/Products/DMTools/Pages/FICO-Xpress-Optimization-Suite.aspx

[10] Frank M. and P. Wolfe. 1956. An Algorithm for Quadratic Programming, *Naval Research Logistics Quarterly* 3(1–2): 95–110.

[11] Guerrero, V.M. 2003. Monthly disaggregation of a Quarterly Time Series and Forecasts of Its Unobservable Monthly Values, *Journal of Official Statistics* 19(3): 215–235.

[12] *Gurobi Optimization.*
http://www.gurobi.com

[13] *INE, Instituto Nacional de Estadística (Spain).*
http://www.ine.es

[14] Kostaki, A. and V. Panousis. 2001. Expanding an abridge life table, *Demographic Research* 5: 1–22.

[15] Kostaki, A. and J. Lanke. 2000. Degrouping mortality data for the elderly, *Mathematical Population Studies* 7(4): 333–341.

[16] 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.

[17] Monteiro, R.D.C. and Adler, I., "Interior path following primal-dual algorithms. part II: Convex quadratic programming". *Mathematical Programming*, **44** (1989) 43–66.

[18] McNeil, D.R., T.J. Trussel, and J.C. Turner. 1977. Spline Interpolation of Demographic Data, *Demography* 14(2): 245–252.

[19] *NEOS, Server for Optimization.*
http://www.neos-server.org/neos/

[20] Pavía-Miralles, J.M. 2010. A Survey of Methods to Interpolate, Distribute and Extrapolate Time Series, *Journal of Service Science and Management* 4: 449–463.

[21] Proietti, T. 2011. Multivariate Temporal Disaggregation with Cross-sectional Constraints, *Journal of Applied Statistics* 38(7): 1455–1466.

[22] 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.

[23] Shryock, H.S., J.S. Siegel, and Associates. 1980. *The Methods and Material of Demography*, U.S. Governemnt Printing Office.

[24] Van De Panne, C. and Whinston, A. (1964), "The Simplex and the Dual Method for Quadratic Programming", *Operations Research Quarterly* **15**, 355–388.

[25] Wegman, J.E. and W.I. Wright. 1983. Splines in Statistics, *Journal of the American Statistical Association* 78(382): 351–365.

[26] 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.

[27] Williams, H.P. 2013. *Model building in mathematical programming*, John Wiley & Sons.

[28] Winston, W.L., M. Venkataramanan and J.B. Goldberg. 2003. *Introduction to mathematical programming*, Thomson/Brooks/Cole.

[29] Wolsey, L.A. 1998. *Integer Programming*, John Wiley & Sons.