A municipal database from the 2011 Spanish census
Francisco J. Goerlich
University of Valencia, Valencia, Spain and
Instituto Valenciano de Investigaciones Economicas, Valencia, Spain

Abstract

Purpose – The paper aims to describe the process to obtain a complete municipal database from the 2011 Spanish Census information. By complete, the authors mean variables for the full sample of the 8,116 municipalities as of the census reference date. In addition, the database should be consistent with the public census information released by the National Statistical Institute: microdata and customized tables.

Design/methodology/approach – The authors use mainly small area demographic and synthetic estimators that are reconciled using biproportional adjustment (iterative proportional fitting), when needed.

Findings – As a result, the authors obtain a complete and consistent municipal database composing 55 variables related to socio-demographic characteristics of persons.

Originality/value – The provision of a complete and consistent municipal database, available for download, which is absent in the original 2011 Spanish Census.

Keywords Municipalities, 2011 Spanish Census, Small area

Paper type Technical paper

1. Introduction
The 2011 Census marked a significant methodological turning point in the Spanish census tradition. It moved away from the classic census methodology, based on exhaustive fieldwork, toward a mixed system in which the population count and its most basic demographic characteristics are taken from administrative records –the Padrón, or Municipal Register– and the remaining population characteristics come from a large-scale survey of around 10 per cent of the population [Instituto Nacional de Estadística (INE), 2011].

© Francisco J. Goerlich. Published in Applied Economic Analysis. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

The author wishes to thank Jorge Luis Vega Valle, Carmen Tejeiro Breijo, Antonio Argüeso Jimenez and Ignacio Duque Rodriguez de Arellano from the Spanish Statistical Institute (INE) for their generous support in resolving innumerable methodological questions related to the census information, and is also grateful for feedback from members of the technical staff at the Instituto Valenciano de Investigaciones Economicas (Ivie), especially Irene Zaera and Carlos Albert, whose comments contributed to the iteration process in developing the disaggregation algorithms mentioned in the paper. The author is grateful for support from the FBBVA-Ivie research program, and from project ECO2015-70632-R. An extended version of this work (in Spanish) is available as a Working Paper, Goerlich (2016), at http://dx.medra.org/10.12842/MUNICIPIOS_CENSO_2011
Although this methodological change does not necessarily imply a loss in the quality of the resulting information (Goerlich et al., 2015), a number of caveats must be mentioned, not only in light of the final information published by the INE through its various census breakdowns –Persons, Households, Dwellings and Buildings– but also in relation to the territorial areas referred to –National, Autonomous Communities, Provinces, Municipalities or Census Sections.

The 2011 Census provides scant information for smaller territorial areas, including municipalities, which are the basic administrative unit in the division of the territory, and for which censuses offer the only opportunity to gather homogenous and comparable data that goes beyond purely demographic information.

This study describes and applies a simple method to obtain estimations for the large majority of census variables, and for the full set of 8,116 municipalities included in the 2011 Census. These estimations are consistent with the published census information. The frame of reference for obtaining variables at the municipal level is the census microdata, which is the source of information for all the non-demographic population characteristics. The final aim is to create a complete and consistent municipal database for a wide set of variables.

The paper is structured as follows. In Section 2, the basic elements of the census methodology are described. This stage is necessary to understand the process followed to disaggregate the information at the municipal level, which is described in Section 3. The resulting database and how it can be accessed are presented in Section 4. The paper ends with some brief conclusions in Section 5.

2. Structure of the information in the 2011 Census

2.1 Information on persons and households

The information on persons and their characteristics for the 2011 Census is based on two fundamental sources: the Municipal Register, for purely demographic information; and a large survey, in principle designed to be representative at the municipal level, for all other population characteristics (Instituto Nacional de Estadística [INE], 2011). These two pillars provide different information and an understanding of how they interrelate is needed to understand the process followed to create the database.

First, one might naturally ask why the basic demographic characteristics of the population are taken from Municipal Register, even though they are not exactly the same as those from the Municipal Register for the census reference date, 1 November 2011. This is due to the very nature of the Municipal Register as a legally regulated administrative registry, which means that any alteration to it must have a legal basis; in other words, alterations cannot be made in the statistical adjustments. When the continuous Municipal Register was introduced in 1998, the population figures from the Municipal Register were disassociated from the census population figures, such that the population of the 2001 Census does not coincide with the population figures derived from the Municipal Register. It is well known that the way the Municipal Register is managed leads to an over estimation of the population, essentially associated with the register of foreign people, although problems have also been found in the upper and lower age distributions (Goerlich, 2007, 2012).

As a result, to find out the “population figure” of Spain and its territories, the Municipal Register – as the best statistical estimation of the resident population – had to be adjusted to give a more accurate reflection of the real situation. The INE therefore used the Municipal Register to build a pre-census file (PCF) that was adjusted as necessary to the increases and decreases in the natural population movement, and in which each registry entry had a count factor equal to 1 if the person could be proven to reside in Spain by crossing with other administrative records such as Social Security data, or was unknown if no conclusive proof
was available that the person was a resident in Spain. These registry entries were known as “doubtful”. Of the PCF registry entries, 97.2 per cent had a count factor equal to 1.

At the same time a large sampling survey was conducted with two objectives:
1) to determine the count factor in the PCF doubtful registry entries; and
2) to estimate the population characteristics.

The reference population for the sample is the population residing in main dwellings. The population living in institutional residences was therefore excluded from the sample and treated in a separate statistical operation: *Encuesta de Colectivos del Censo de Población y Viviendas 2011* (Instituto Nacional de Estadística [INE], 2013a).

2.2 The fit between the sample and the information from the pre-census file
The PCF and the sample are independent operations that must be reconciled. This reconciliation process, carried out by the INE, is based on two actions that are not wholly independent.

First, to determine the count factor of the doubtful entries both sets of information were partitioned into classes based on observable characteristics – age, nationality and place of residence – and a nominal crossing was made between the sample gathered from the fieldwork and the PCF, so the registry entries could be linked and those appearing in the PCF as doubtful could be identified if they were actually gathered in the sample.

From this identification, using the principle of analogy at the class level the count factors were estimated for the doubtful registry entries. The detailed procedure is described in Instituto Nacional de Estadística (INE) (2012) and Goerlich et al. (2015, chapter 1). What interests us here is that following this operation, each PCF entry has an assigned count factor. We therefore have a final weighted census file which determines the census population figure and its basic demographic characteristics. The resident population deriving from the census through this procedure was 46,815,916.

Second, the sample must be calibrated to the population to ensure consistency between the two in various dimensions referring to both population characteristics and territorial areas. However, the reference population of the survey is not derived from the final weighted census file, but the population in main family dwellings and excludes the population living in institutional residences. This population cannot be identified from the PCF.

The population living in institutional residences was estimated by the *Encuesta de Colectivos* as 444,101. However, not all the population living in institutional accommodation is officially registered as living there. According to this survey, only 241,187 people living in institutional establishments were officially registered as living there, whereas the remaining 202,914 were registered as living in main family dwellings, and are counted in the family dwellings for the effects of the sample, which is where they are officially registered. As a result, the population residing in main family dwellings is: 46,815,916 – 241,187 = 46,574,729 persons. That is, the elevation factors of the survey must include this population.

The calibration process uses the standard INE method: CALMAR (Deville and Särndal, 1992; Deville et al., 1993), is carried out at the municipal level, and is a function of the municipality size [Instituto Nacional de Estadística (INE), 2014].

Having two reference population groups – the resident population and the population living in main dwellings – significantly complicates the process of disaggregating the microdata to create the municipal database, since the disaggregated variables must be adjusted to population marginals that cannot be taken directly from the PCF. The PCF provides information for the total resident population, whereas the municipal database,
constructed from the microdata, must be adjusted to the population living in main dwellings.

For this reason, we first had to estimate the population living in main dwellings at the municipal level by sex and in two age groups: under the age of 16, and aged 16 and over. The methods used for this purpose are described in Section 3.

2.3 Territorial structure in the microdata from the 2011 census
The microdata from the census only provide information at the municipal level for municipalities with more than 20,000 inhabitants. The remaining municipalities are grouped into four strata by size for each province as follows:

1. up to 2,000 inhabitants (Code 991);
2. between 2,001 and 5,000 inhabitants (Code 992);
3. between 5,001 and 10,000 inhabitants (Code 993); and
4. between 10,001 and 20,000 inhabitants (Code 994).

The distribution of the municipalities by province and strata are reported in Table I.

The 394 municipalities with more than 20,000 inhabitants can be perfectly identified in the microdata. In addition, the eight cases in which there is only one municipality per stratum can also be identified, together with the smallest municipality in Spain in demographic terms, Illán de Vacas in the province of Toledo, which has just one inhabitant. We can therefore directly identify 403 municipalities in the microdata; for the remaining 7,713 municipalities, we can only know the aggregated values of the stratum to which they belong. The database in this study obtains information on certain variables for these municipalities.

2.4 The customized tables system in the published census information
In addition to the microdata file, the published findings from the 2011 Census include a Customized Table query system in which users can select the variables they are interested in from within a geographical area and domain.

The Customized Tables system is constructed from the sample and the reference population is therefore those living in main dwellings and as such is consistent with the microdata. However, for various reasons the system is fairly limited for obtaining complete generalized information for all the municipalities. On one hand, it is subject to a series of confidentiality norms that restricts the information provided, and which in no case covers all the municipalities. On the other hand, to ensure statistical secrecy all data is rounded to the closest multiple of five.

The information in the Customized Tables is, however, of unquestionable value since, following some experimentation, their incorporation was shown to notably improve the municipal estimations using the procedure described below. The information available in the Customized Tables was therefore incorporated as the starting point for the disaggregation process.

3. Methodology: from microdata to municipalities
The previous section describes the census information structure with regard to small areas – municipalities. The next question is how to combine all this information so we obtain estimations for all municipalities for a large set of variables. Whatever method is followed it must comply with a basic condition: the estimations must be consistent with the microdata. The reference population is therefore the population living in main dwellings.
<table>
<thead>
<tr>
<th>Province</th>
<th>Up to 2,000 Inhab.</th>
<th>2,001 to 5,000 Inhab.</th>
<th>5,001 to 10,000 Inhab.</th>
<th>10,001 to 20,000 Inhab.</th>
<th>Over 20,000 Inhab.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 Alava</td>
<td>42</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>02 Albacete</td>
<td>62</td>
<td>17</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>87</td>
</tr>
<tr>
<td>03 Alacant/Alicante</td>
<td>66</td>
<td>18</td>
<td>20</td>
<td>13</td>
<td>24</td>
<td>141</td>
</tr>
<tr>
<td>04 Almeria</td>
<td>62</td>
<td>19</td>
<td>9</td>
<td>6</td>
<td>6</td>
<td>102</td>
</tr>
<tr>
<td>05 Avila</td>
<td>233</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>248</td>
<td></td>
</tr>
<tr>
<td>06 Badajoz</td>
<td>97</td>
<td>41</td>
<td>17</td>
<td>4</td>
<td>5</td>
<td>164</td>
</tr>
<tr>
<td>07 Illes Balears</td>
<td>14</td>
<td>13</td>
<td>17</td>
<td>11</td>
<td>12</td>
<td>67</td>
</tr>
<tr>
<td>08 Barcelona</td>
<td>121</td>
<td>58</td>
<td>51</td>
<td>37</td>
<td>44</td>
<td>311</td>
</tr>
<tr>
<td>09 Burgos</td>
<td>360</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>371</td>
</tr>
<tr>
<td>10 Cáceres</td>
<td>188</td>
<td>21</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>221</td>
</tr>
<tr>
<td>11 Cádiz</td>
<td>6</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>15</td>
<td>44</td>
</tr>
<tr>
<td>12 Castellón/Castelló</td>
<td>104</td>
<td>11</td>
<td>9</td>
<td>3</td>
<td>8</td>
<td>135</td>
</tr>
<tr>
<td>13 Ciudad Real</td>
<td>62</td>
<td>16</td>
<td>11</td>
<td>8</td>
<td>5</td>
<td>102</td>
</tr>
<tr>
<td>14 Córdoba</td>
<td>23</td>
<td>24</td>
<td>14</td>
<td>6</td>
<td>8</td>
<td>75</td>
</tr>
<tr>
<td>15 A Coruña</td>
<td>12</td>
<td>29</td>
<td>31</td>
<td>11</td>
<td>11</td>
<td>94</td>
</tr>
<tr>
<td>16 Cuenca</td>
<td>222</td>
<td>9</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>238</td>
</tr>
<tr>
<td>17 Girona</td>
<td>159</td>
<td>29</td>
<td>14</td>
<td>11</td>
<td>8</td>
<td>221</td>
</tr>
<tr>
<td>18 Granada</td>
<td>95</td>
<td>34</td>
<td>18</td>
<td>14</td>
<td>7</td>
<td>168</td>
</tr>
<tr>
<td>19 Guadalajara</td>
<td>267</td>
<td>13</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>288</td>
</tr>
<tr>
<td>20 Guipúzcoa</td>
<td>45</td>
<td>10</td>
<td>13</td>
<td>14</td>
<td>6</td>
<td>88</td>
</tr>
<tr>
<td>21 Huelva</td>
<td>35</td>
<td>24</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>79</td>
</tr>
<tr>
<td>22 Huesca</td>
<td>189</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>202</td>
</tr>
<tr>
<td>23 Jaén</td>
<td>33</td>
<td>36</td>
<td>13</td>
<td>9</td>
<td>6</td>
<td>97</td>
</tr>
<tr>
<td>24 León</td>
<td>178</td>
<td>21</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>211</td>
</tr>
<tr>
<td>25 Lleida</td>
<td>193</td>
<td>23</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>231</td>
</tr>
<tr>
<td>26 La Rioja</td>
<td>153</td>
<td>12</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>174</td>
</tr>
<tr>
<td>27 Lugo</td>
<td>24</td>
<td>30</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>67</td>
</tr>
<tr>
<td>28 Madrid</td>
<td>69</td>
<td>31</td>
<td>31</td>
<td>15</td>
<td>33</td>
<td>179</td>
</tr>
<tr>
<td>29 Málaga</td>
<td>44</td>
<td>29</td>
<td>9</td>
<td>3</td>
<td>16</td>
<td>101</td>
</tr>
<tr>
<td>30 Murcia</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>13</td>
<td>17</td>
<td>45</td>
</tr>
<tr>
<td>31 Navarra</td>
<td>213</td>
<td>37</td>
<td>12</td>
<td>7</td>
<td>3</td>
<td>272</td>
</tr>
<tr>
<td>32 Ourense</td>
<td>61</td>
<td>21</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>92</td>
</tr>
<tr>
<td>33 Asturias</td>
<td>36</td>
<td>11</td>
<td>10</td>
<td>14</td>
<td>7</td>
<td>78</td>
</tr>
<tr>
<td>34 Palencia</td>
<td>180</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>191</td>
</tr>
<tr>
<td>35 Palmas de Gran Canaria (Las)</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>9</td>
<td>13</td>
<td>34</td>
</tr>
<tr>
<td>36 Pontevedra</td>
<td>4</td>
<td>21</td>
<td>12</td>
<td>16</td>
<td>9</td>
<td>62</td>
</tr>
<tr>
<td>37 Salamanca</td>
<td>349</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>362</td>
</tr>
<tr>
<td>38 Santa Cruz de Tenerife</td>
<td>6</td>
<td>16</td>
<td>12</td>
<td>8</td>
<td>12</td>
<td>54</td>
</tr>
<tr>
<td>39 Cantabria</td>
<td>55</td>
<td>27</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>102</td>
</tr>
<tr>
<td>40 Segovia</td>
<td>198</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>209</td>
<td></td>
</tr>
<tr>
<td>41 Sevilla</td>
<td>14</td>
<td>25</td>
<td>30</td>
<td>19</td>
<td>17</td>
<td>105</td>
</tr>
<tr>
<td>42 Soria</td>
<td>175</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>183</td>
</tr>
<tr>
<td>43 Tarragona</td>
<td>122</td>
<td>32</td>
<td>14</td>
<td>6</td>
<td>10</td>
<td>184</td>
</tr>
<tr>
<td>44 Teruel</td>
<td>225</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>236</td>
</tr>
<tr>
<td>45 Toledo</td>
<td>112</td>
<td>63</td>
<td>15</td>
<td>11</td>
<td>3</td>
<td>204</td>
</tr>
<tr>
<td>46 Valencia/València</td>
<td>132</td>
<td>55</td>
<td>28</td>
<td>20</td>
<td>31</td>
<td>266</td>
</tr>
<tr>
<td>47 Valladolid</td>
<td>201</td>
<td>13</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>225</td>
</tr>
<tr>
<td>48 Vizcaya</td>
<td>60</td>
<td>19</td>
<td>13</td>
<td>9</td>
<td>11</td>
<td>112</td>
</tr>
<tr>
<td>49 Zamora</td>
<td>244</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>248</td>
</tr>
</tbody>
</table>

Table I.
Geography by municipality size in 2011 census microdata (continued)
Consistency with the microdata implies that: (i) for each municipality, values disaggregated by categories of one variable must coincide with the value for the same variable at the municipal level, and (ii) for each stratum of the microdata, the sum of the values disaggregated at the municipal level must coincide with the values for that stratum. The information (i) must be found externally, and the information (ii) comes from the microdata. In addition, the estimations for the 403 municipalities that can be identified in the microdata are taken directly from that source and are used to validate the method.

3.1 Disaggregation of the population living in main dwellings

As noted above, the reference population for creating the municipal database is the population living in main dwellings. In some cases, the corresponding group is the total of the population living in main dwellings (PRVP), but in other cases the group is limited to the classification by sex or age groups –below the age of sixteen, and sixteen years and over– and occasionally it is necessary to cross these variables or previous estimations of the microdata classification variables. These are the groups that act as marginals to which the estimations must be adjusted.

For this reason, the first stage was to disaggregate the PRVP according to the above-mentioned criteria. The procedure followed was very simple. For the 5,608 municipalities that do not have a population registered as living in institutional accommodation this information is available in the PCF and is taken from there. These municipalities are not estimated and form part of the validation set. For the rest we distinguish between two cases:

1. municipalities with more than 20,000 inhabitants; and
2. municipalities with up to 20,000 inhabitants.

The first group is also identified in the microdata and information for this group was taken directly from there. For the second group, following an initial estimation, an iterative proportional fitting (ipf – Deming and Stephan, 1940; Stephan, 1942) procedure was applied at the stratum level, more commonly known in economics as the RAS method (Bacharach, 1965)[1].

3.2 Disaggregation of variables of persons in the microdata

We start from the following general frame. Let us consider a categorical variable, $X$, for a municipality $m$, which takes $J$ possible values. For example, the variable "Relation to economic activity", $\text{RELA}$, takes 6 possible values, and is not applied when the person is below the age of 16 years. Therefore, when the population is restricted to the population aged 16 and over, in this example $J = 6$. 

<table>
<thead>
<tr>
<th>Province</th>
<th>Up to 2,000 inhab.</th>
<th>2,001 to 5,000 inhab.</th>
<th>5,001 to 10,000 inhab.</th>
<th>10,001 to 20,000 inhab.</th>
<th>Over 20,000 inhab.</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zaragoza</td>
<td>256</td>
<td>22</td>
<td>9</td>
<td>4</td>
<td>2</td>
<td>293</td>
</tr>
<tr>
<td>Ceuta</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Melilla</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Spain</td>
<td>5,808</td>
<td>1,000</td>
<td>553</td>
<td>361</td>
<td>394</td>
<td>8,116</td>
</tr>
</tbody>
</table>

Source: Instituto Nacional de Estadística (INE) (2013a)
Given that each person in the municipality estimated must belong to one of the possible $J$ categories, the population of that municipality, $N^m$, can be written as $N^m = \sum_{j=1}^{J} X^m_j$, where the superscript $m$ indicates the corresponding municipality. The values of $X^m_j$ are unknown for each $j$ and $m$, and are the variables we are trying to estimate. We know the population of the municipality, $N^m$, from the final weighted census file, and also $X_j$ for the stratum to which the municipality belongs, $X_j = \sum_{m \in S} X^m_j$ where $S$ represents the stratum, taken from the microdata. In other words, seen in table format we know the marginal distributions, but not the whole distribution.

A mechanical application using an iterative proportional fitting process based on an initial uniform distribution yields very poor results, indicating that the key is to incorporate auxiliary information into the estimation of this joint distribution, in other words, to look for a reasonable initial estimation for each municipality that serves as an initial value in the iterative fitting process.

For the municipalities for which information is available in the Customized Tables system, this initial value can be taken from that source. Because this information is not available for the remaining municipalities we must find a reasonable alternative estimation. Let us suppose that we have another partition of the municipality’s population into $K$ exhaustive and mutually exclusive classes. We can also now write the population of the municipality as $N^m = \sum_{k=1}^{K} N^m_k$, where $N^m_k$ is now known from the information in the final weighted census file.

Let us now consider the problem of estimating $X^m_j$. By definition:

$$X^m_j = \sum_{k=1}^{K} X^m_{k,j} = \sum_{k=1}^{K} N^m_k \frac{X^m_{k,j}}{N^m_k}$$

(1)

The estimator proposed for these municipalities estimates the rates that appear in (1), $\frac{X^m_{k,j}}{N^m_k}$, from the stratum to which the municipality belongs, $S$, with the information available in the microdata, and applies these rates to the partition of the population considered at the municipal level. That is:

$$\hat{X}^m_j = \sum_{k=1}^{K} N^m_k \frac{X^S_{k,j}}{N^S_k}$$

(2)

where $N^S_k = \sum_{m \in S} N^m_k$ and $X^S_{k,j} = \sum_{m \in S} X^m_{k,j}$. Consequently, (2) substitutes the real rates in (1), $\frac{X^m_{k,j}}{N^m_k}$, $\forall k$, with estimated rates at the level of the stratum to which the municipality belongs, $\frac{X^S_{k,j}}{N^S_k}$, $\forall k$, and applies these rates to all the municipalities in that stratum.

The method for obtaining $\hat{X}^m_j$ from (2) is simple and falls within the so-called traditional demographic methods in the context of small area estimations (Rao, 2003, chapter 3), or synthetic estimators (Rao, 2003, chapter 4.2) and can be implemented in a generalized and automatic way for several different census microdata variables when the Customized Tables system provides no information for the municipality in question.

An estimator is known as synthetic if a reliable direct estimator for a large area covering several small areas is used to obtain an indirect estimator for these small areas, under the assumption that the small areas have the same characteristics as the large area. Clearly (2) falls within this definition, where the implicit assumption is that all the municipalities in
stratum $S$ present the same rates, $X_k^S$, $\forall k$, and the municipalities of this stratum are only differentiated by their demographic structure. This method is also known as the propensity method (Bell et al., 1995), and is applied by the Instituto Nacional de Estadística (INE) (2013b) in a range of contexts.

An alternative way of looking at (2) is:

$$\hat{X}_m = \sum_{k=1}^{K} \frac{N_k^m}{N_k^S} X_{k,j}^S$$

which highlights the way the value of $X_j$ at the stratum level for each element in the partition, $X_{k,j}^S$, is rescaled by the proportion that the population of the municipality represents in the stratum, $\frac{N_k^m}{N_k^S}$.

Colom et al. (2015) provide an explanation of the method within the framework of traditional sampling superpopulation models when it is not possible to identify the registers of the specific units within a broader domain. This is the case of the microdata structure in the 2011 Census. These authors show how in this context, (2) is an unbiased although inefficient estimator. Nonetheless, the estimated standard errors are very small and of a similar magnitude to that provided by the INE in many of its sample surveys. In addition, this procedure yields practically identical results to those obtained by modeling the variable to be disaggregated using discrete choice models.

Once we have $\hat{X}_m$ for the $J$ categories of the variable, and for all the municipalities in the stratum, either from the procedure described above or from the information provided by the Customized Tables system, these initial estimations are adjusted to the total known marginals, $N_m^m$ and $X_m$, by means of an iterative bi-proportional fitting process (Deming and Stephan, 1940; Stephan, 1942). The estimation is therefore carried out at the stratum level and yields a final estimator $\hat{X}_m$.

We use as a partition the municipal population by sex and simple ages up to 100 years and above since this partition is available from the final weighted census file, which generates a total of 202 cells, 101 for each sex, and therefore $K = 202$ in (2).

The application of (2) rests on the assumption that the municipality for which we perform the estimation has the same characteristics as the stratum to which it belongs, and that the differences between the municipalities in this stratum reside in their demographic structure. This implies that the closer the variable in question is related to the demography, and the more homogenous the municipalities within the stratum, the lower the estimation errors will be.

Because the method we describe above can be applied to municipalities that are clearly identified in the microdata, these data constitute the validation set against which to measure the aggregate estimation error. It should be noted, however, that these are mostly municipalities with more than 20,000 inhabitants, which undoubtedly means it is a biased validation set.

For these municipalities, $X_m^m$ is known, so we can calculate the absolute error (AE): $|\hat{X}_m - X_m^m|$. From this discrepancy we calculate standard error means, the mean of the absolute relative errors (MARE), as a percentage:

$$MARE = \frac{100}{M \times J} \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{|\hat{X}_m^m - X_m^m|}{X_m^m}$$

and an overall error mean, as the total absolute relative error (TARE), as a percentage:
ranging between 0 and 1, since the sum of the $AE$, $\sum_{m=1}^{M} \sum_{j=1}^{J} |X_{j}^{m} - \bar{X}_{j}^{m}|$, ranges between 0 when no error is made, $X_{j}^{m} = \bar{X}_{j}^{m}$, $\forall m,j$, and twice the reference population, $N = \sum_{m=1}^{M} \sum_{j=1}^{J} X_{j}^{m}$, when the error is the maximum possible in each case, and can be interpreted as the percentage of the population erroneously distributed in the set [2]. An analysis of errors showed negligible errors for the validation municipalities in all cases.

4. Database: content and access

The procedure described above allowed us to disaggregate the 55 variables reported in Table II, together with the variables related to the population living in main dwellings according to certain classification criteria, and that are not generally available at the municipal level from the 2011 Census.

The advantages of this database derive from the availability of data for all municipalities without exception, unlike the information available from the census, yet at the same time it is wholly consistent with the published census information. It can therefore be used in research whose territorial scope is the municipality or certain arbitrary aggregations of municipalities such as, for example, districts or rural areas (Reig et al., 2016), and morphological (Goerlich and Cantarino, 2013) or functional urban areas (Goerlich et al., 2019).

The database is available in an Access file at this link (https://nuvol.uv.es/owncloud/index.php/s/aWLV2KzUbodR5bQ). It should be used in conjunction with the design of the census microdata register, and it is structured as follows. For each variable included in Table II, a table is provided in which the rows represent the municipalities, identified by a code, and include as many columns as there are values for the corresponding variable. The columns are named according to the following criterion: given the variable in question, the name of which appears in the last column of Table II, and the values it takes, each column is identified with the name of the variable to which its code is added. The final column indicates the marginal to which the variable in question is added.

For example, the variable “Relation to economic activity”, RELA, takes 6 possible values: 1 – Employed, 2 – Unemployed with previous work experience, 3 – Unemployed in search of first job, 4 – Person with permanent work disability, 5 – Retired, early retiree, pensioner or rentier and 6 – Other situation; and is defined for the population living in main dwellings aged 16 years or over, PRVP16M. Thus, the first column in the table “20_RELA” in the Access file has the code for the municipality, codmun, followed by 6 columns, RELA#, # = 1 to 6, and a final column, PRVP16M, such that RELA1 gives the number of people in employment in each municipality, and RELA5 the retired, early retirees, pensioners or rentiers.

A final table contains only the codes and names of the municipalities as they appear in the census.

5. Conclusions

This study describes the process followed to create a municipal database for a large set of variables based on the 2011 Census. This information is not available in a general form for all municipalities. The methods for creating the database are simple, although time-consuming, but have the advantage that they are compatible with the published census information, and allow the incorporation of external information derived from the INE’s Customized Tables system, which is essential to improve the accuracy of the estimations.
### Variables acting as marginals in the disaggregation process

1. Population living in main dwellings by age group
   - Under the age of 16: PRPVM16
   - 16 years old and above: PRVP16M

2. Population living in main dwellings by sex
   - Male: PRVPVAR
   - Female: PRVPMUJ

3. Population living in main dwellings by sex and age
   - Males under the age of 16: PRVPVARM16
   - Males aged 16 years old and above: PRVPVAR16M
   - Females under the age of 16: PRVPMUJM16
   - Females aged 16 years old and above: PRVPMUJ16M

### Microdata classification variables

4. Current municipality of residence and Previous municipality of residence: RES_ANTERIOR
5. Current municipality of residence and Municipality of residence 1 year ago: RES_UNANO
6. Current municipality of residence and Municipality of residence 10 years ago: RES_DANO
7. Spending more than 14 nights in second municipality: SEG_VIV
8. Having a dwelling in second municipality: SEG_DISP
9. Marital status: ECIVIL
10. Attending school: ESCOLAR
11. Level of completed studies (qualifications): GRADOS
12. Level of completed studies (details): ESREAL
13. Type of studies undertaken: TESTUD
14. Caring for a child under the age of 15: TAREA1
15. Caring for a person with health problems: TAREA2
16. Charitable work or social volunteering: TAREA3
17. Responsible for most of the domestic tasks in the home: TAREA4
18. Indicator of whether the woman has had children: HIJOS
19. Principal relation with economic activity (employed/unemployed): ACTIVO
20. Principal relation with economic activity (detail): RELA
21. Type of working day: JORNADA
22. Occupation code: OCUPACION
23. to 1 digit: CNO
24. to 2 digits: CNAE
25. Economic activity code to 2 digits: RAMA
26. Branch: LETRA
27. Letter: LETRA
28. Professional situation: SITU
29. Socioeconomic status: CSE
30. Students (ESCUR1): Yes/No: ESTUDIANTE
31. Current studies: Type of Studies
   - 01 – Compulsory secondary education (ESO), Adult secondary education: ESCUR01
   - 02 – Initial Professional Qualification Programs: ESCUR02
   - 03 – High school (bacalaureate): ESCUR03
   - 04 – Middle Grade Vocational Training, Plastic Arts and Design, and Sports Education or equivalent: ESCUR04
   - 05 – Official Language School Education: ESCUR05
   - 06 – Professional Music and Dance Education: ESCUR06
   - 07 – Higher Grade Vocational Training, Plastic Arts and Design, and Sports Education or equivalent: ESCUR07
   - 08 – University diploma, Technical architecture, Technical engineering or equivalent: ESCUR08

Table II. Variables of persons disaggregated by the methods described in the paper
The procedures used must overcome numerous small inconsistencies between the two main pillars of the 2011 Census—the final weighted census file and the survey—which provide all the population characteristics beyond simple demographic data. Apart from these small inconsistencies the estimations generated are wholly consistent at the municipal level and at the level of the strata to which the municipalities in the microdata belong. Although all the disaggregated variables in the database are at the individual person level, identical methods can be used for household variables. Similar methods could also be used for the dwellings and buildings variables.

Finally, a few words of caution. The results must be interpreted for what they are—estimations based on a census sample—with the aim of providing statistics for all municipalities, and they should be used with that caution in mind. The information derived from the Customized Tables system has been exploited to the full, but in some cases it is limited or partial and in no case is it available in a general sense for all municipalities.

Notes
1. There are two exceptions to the above rules due to the lack of consistency between the PCF and the calibration of the microdata. In both cases, to maintain consistency with the final database we prioritized the use of the microdata. The details of the process followed in these cases are described in Goerlich (2016).
2. That is, assigned to a cell to which it does not correspond.
References


**Further reading**


**Corresponding author**

Francisco J. Goerlich can be contacted at: Francisco.J.Goerlich@uv.es