The continuous sample of working lives: improving its representativeness

Juan Manuel Pérez-Salamero González1 · Marta Regúlez-Castillo2 · Carlos Vidal-Meliá3

Received: 14 February 2016 / Accepted: 4 February 2017 / Published online: 3 March 2017 © The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract This paper studies the representativeness of the Continuous Sample of Working Lives (CSWL), a set of anonymized microdata containing information on individuals from Spanish Social Security records. We examine several CSWL waves (2005–2013) and show that it is not representative for the population with a pension income. We then develop a methodology to draw a large dataset from the CSWL that is much more representative of the retired population in terms of pension type, gender and age. This procedure also makes it possible for users to choose between goodness

We gratefully acknowledge financial support from Ministerio de Economía y Competitividad (Spain) and from the Basque Government via projects ECO2015-65826-P and IT 793-13 respectively. We would also like to thank seminar participants at the Universities of the Basque Country, Barcelona, Valencia and Granada, and Chris Pellow and Peter Hall for their help with the English text. Comments and suggestions made by Prof. Guner and the anonymous referees were extremely helpful in improving the paper. Any errors are entirely due to the authors.

Electronic supplementary material The online version of this article (doi:10.1007/s13209-017-0154-0) contains supplementary material, which is available to authorized users.

Marta Regúlez-Castillo
marta.regulez@ehu.eus
Juan Manuel Pérez-Salamero González
juan.perez-salamero@uv.es
Carlos Vidal-Meliá
carlos.vidal@uv.es

1 Department of Financial Economics and Actuarial Science, University of Valencia, Avenida de los Naranjos s.n, 46022 Valencia, Spain
2 Department of Applied Economics III, University of the Basque Country (UPV/EHU), Avda. Lehendakari Aguirre 84, 48015 Bilbao, Spain
3 Department of Financial Economics and Actuarial Science, University of Valencia, Avenida de los Naranjos s.n, 46022 Valencia, Spain
of fit and subsample size. In order to illustrate the practical significance of our methodology, the paper also contains an application in which we generate a large subsample distribution from the 2010 CSWL. The results are striking: with a very small reduction in the size of the original CSWL, we significantly reduce errors in estimating pension expenditure for 2010, with a $p$ value greater or equal to 0.999.

**Keywords** Continuous Sample of Working Lives · Public pension system · Subsample selection · Stratified sampling · Chi-square test · $p$ value

**JEL Classification** C81 · H55 · J26

1 Introduction

Selecting a representative sample from the population is a very important factor in quantitative research given that results obtained from a wrongly selected sample which is not properly in tune with the object of study cannot be generalized to the whole population. Moreover, smaller than appropriate sample may not have enough diversity to enable significant differences or associations potentially present in the target population to be identified. It is fundamental to make the right choice and decide on the right way to deal with the dataset, always making sure that it is the best one for the purposes of the research. Hence, it is important to check that the sample selected is representative of the target population in terms of demographic characteristics as well as any others that can affect the results of the study to be conducted.1

The Continuous Sample of Working Lives (CSWL) is a random sample (RS) of around 1.2 million people, i.e. 4% of the reference population. It contains administrative data on working lives, which provide the basis for this sample taken from Spanish Social Security records, and comprises anonymized microdata with detailed information on individuals. Izquierdo et al. (2009) point out that this database provides a unique dataset with very rich information about labour market histories and personal characteristics, such as nationality, date and country of birth (or province if Spanish), gender and place of residence when the individual first entered the Social Security system, along with additional information about the composition of the household and labour market variables.

The first wave covers people who had an economic relationship with the Social Security system in 2004, and provides the entire working history (employment, unemployment and retirement) of the sample population. The sample is updated every year using information from the variables selected from the Social Security system dating back to when computerized records began, and from other administrative data sources which record additional information on individuals. Apart from the details given by

1 There are many papers on the representative sample selection issue, including Kruskall and Mosteller (1979a,b,c, 1980)), Ramsey and Hewitt (2005), Grafström and Schelin (2014), and Omair (2014).
the institutions responsible for generating the CSWL,\textsuperscript{2} the data available to researchers date from 2004 to 2015.\textsuperscript{3}

The sample reference population\textsuperscript{4} is defined as individuals who have had some connection (through contributions, pensions or unemployment benefits) to the Social Security system at any time during the year of reference. This population of reference makes it possible to select people who on a particular date in that year had no relationship with Social Security, but who did at some time before or after that date in the same year. This means that the CSWL will contain details about a person who may have had various relationships with Social Security (unemployed, unaffiliated, contributor, pensioner), unlike other datasets that only show a single relationship because they refer to the situation on one specific date in the year.

Individuals are selected from the reference population each year according to whether their identification codes contain certain randomly-generated figures in the correct positions. Each year individuals who figured in the previous version of the CSWL and continue to have a relationship with Social Security remain in the selection, while new individuals are incorporated if they meet the requirement of having identification codes containing the randomly-generated figures described above. The detailed information available on the individuals selected includes work trajectories (from 1967), contribution bases (from 1980) and/or pensions (from 1996), as long as it is contained in the Social Security administrative records. Those individuals who for any reason have no connection to Social Security in a particular year will not figure in the CSWL.

The CSWL is a dataset that has been used in a considerable number of studies, especially on labour economics (Treviño et al. 2008; Benavides et al. 2010; Vall Castello 2012; Bonhomme and Hospido 2013, 2016; Solé et al. 2013; Agliari et al. 2014; Arranz and García-Serrano 2014; Barra et al. 2014 and Nagore García and van Soest 2016) and the Spanish public pension system (Antón Pérez et al. 2007; Argimón et al. 2007; Boado-Penas et al. 2008; Moral Arce et al. 2008; Vidal Meliá et al. 2009; Cairó Blanco 2010; Peinado Martínez 2011; Devesa et al. 2012; Domínguez Fabián et al. 2012; Meneu Gaya and Encinas Goenechea 2012; Vicente Merino et al. 2012; Conde Ruiz and González 2013, and Vegas Sánchez et al. 2013). Other studies give detailed descriptions of its characteristics, advantages and limitations.\textsuperscript{5}

\textsuperscript{2} MTAS (2006); Ministerio de Empleo y Seguridad Social (MESS), (2016).

\textsuperscript{3} Researchers can request versions of the CSWL by post from the Dirección General de Ordenación de la Seguridad Social at the Spanish Ministry of Employment and Social Security (Ministerio de Empleo y Seguridad Social). A separate request must be made for each version. Requests consist of a user profile describing the project being carried out and a document accepting the CSWL’s conditions of use. These are available from MESS (2016) at the following address: http://www.seg-social.es/Internet_1/Estadistica/Est/Muestra_Continua_de_Vidas_Laborales/SolicitarM/index.htm.

\textsuperscript{4} The definition of the population from which the CSWL was taken can be checked in Ministerio de Trabajo y Asuntos Sociales (MTAS), (2006), pp. 25–29.

\textsuperscript{5} Argimón and González (2006), Durán and Sevilla (2006), Toharia Cortés et al. (2007), Durán (2007), García Segovia and Durán (2008), Paxton et al. (2009), Izquierdo et al. (2009), Lapuerta (2010), Arranz and García-Serrano (2011), López Roaldán (2011), Alonso Domínguez (2012), Muñoz de Bustillo et al. (2011), Arranz et al. (2013), and Pérez-Salamero González et al. (2016).
Other countries such as the USA and Germany have similar databases but they use stratification. For example, the US Social Security Administration (SSA) has been compiling a Continuous Work History Sample (CWHS) since the late 1930s\(^6\), which is a one-percent stratified cluster probability sample of all possible Social Security numbers. The population, or sampling frame, from which the CWHS cases are selected consists of the 1 billion possible nine-digit Social Security numbers (SSNs). These digits represent the geographical area of each number allocated, a group for the date of issue and a random serial number [Smith (1989)]. The numbers are thus stratified geographically by place of application for SSN and chronologically by the process of allocation of numbers within each stratum. The information source is the so-called MEF (Master Earnings File—Olsen and Hudson 2009), which is used to determine pensions for retirement, permanent disability and widowhood in the US public Social Security system (OASDI).

Similarly in Germany (Himmelreicher and Stegmann 2008), there is the so-called sample of insured persons and their insurance accounts (Versicherungskontenstichprobe, VSKT), which provides longitudinal data that have a high potential for analyses of employment biographies and pension claims in old age. These data are process-produced, contain very large samples and allow for differentiated analyses of a variety of social groups. The VSKT was initially sampled in 1983 as a stratified\(^6\) random sample with disproportional selection probabilities and since then has continued as a panel containing monthly information on the individuals included in the sample. It represents 1% of the contributing population.

One important branch of research in pension systems is the problem of global ageing and the sustainability of public pension systems, referred to hereafter as PPS. As mentioned above, the CSWL is one of the datasets considered for the purposes of this research in Spain. In order to analyze Spanish PPS, it is necessary to have information on relevant variables such as age and gender for each year of reference as considered in previous studies. These last two factors are essential for correctly estimating life expectancy, so any study that makes estimates of future benefits should select a sample that is representative of the population in terms of age and gender as well as type of pension.

With all this in mind, the first objective of the paper is to analyze whether all the information given by the CSWL on the benefit recipients makes up the best replica of the study population that researchers can have. To solve this question of how representative the CSWL is of the population of pensioners, it is important to carry out a statistical analysis to determine whether there are significant differences between the CSWL sample distribution and the population distribution. Even though the researcher does not have the entire population, the distributions of the population of pensioners organized by age, gender and type of pension on the last day of each year are available. Therefore it is possible to carry out a test in order to check whether the CSWL has the same distribution as the population of pensioners. However, given that the CSWL is not a stratified sample, it is advisable to check whether it covers the correct propor-

\(^6\) Stratification is by gender, nationality, insurance branch of the current account holder and age cohort. http://www.jpi-dataproject.eu/Home/Database/153?topicId=2.
tion of the population in each stratum to be considered in the study of the pensioner population categorized by age, gender and type of pension.

In this paper we conduct such an analysis for the CSWL waves for 2005–2013 and confirm that there is a lack of representativeness in most years. Given these results and the fact that it has long been known that a stratified random sample (SR) enables a more efficient selection to be made when one of the variables in the study of interest presents great variability, as is the case of the ages of pensioners in the different types of pension, the first idea that comes to mind is why not use a random sample with stratification to obtain a dataset that better represents the population of pensioners?

The answer is clear: researchers do not have all the data on the population, so they cannot extract such a sample. They would have to use a stratified random sample (SR) contained in the CSWL, but with a considerably smaller size, which would entail a loss of richness in the information on pensioners’ working lives. Hence it is important and advisable to develop other procedures in order to obtain larger subsamples with less reduction in the total number of pensioners than in the SR subsample contained in the CSWL, and at the same time to make the original CSWL more representative in terms of pension benefits.

Hence the second objective of the paper is to provide researchers with a novel methodology for the design of a dataset on pensioners by making the CSWL more representative of the population in terms of type of pension, gender, and age, and by trying to miss as few pension records as possible so as not to overlook diversity in working lives.

Subsample selection is done by finding a feasible solution to a nonlinear optimization problem (NLP) using mixed integer nonlinear programming with just one real non-negative decision variable, the constant of proportionality, $q$, in a stratified sample design with proportional allocation. Maximizing $q$ is equivalent to maximizing the size of the subsample and is subject to constraints implied by the fact that the number of pensioners to be included in each cohort of the subsample has to be a natural number (non-negative integer) and that the subsample obtained must be included in the population as well as in the CSWL. The methodology applied uses a goodness of fit test—Pearson’s chi-square test—in order to make the subsample selected more representative of the population, providing $p$ values close to 1. This methodology enables us to obtain quite a large dataset included in the CSWL which is much more representative of the pensioner population in terms of type of pension, gender, and age, as would be the case with an SR. In addition, this procedure enables users to choose between goodness of fit quality and subsample size.

Finally, in order to illustrate the gains obtained with the selection of the subsample, the methodology is applied to the CSWL for 2010, to gauge the improvement in the estimate of total pension expenditure in different cohorts taking into account age,

---

7 For an analysis of the problems of NLP see Bazaraa et al. (2006), Bertsekas (1999), Griva et al. (2009), Luenberger (2003), Meneu Gaya et al. (1998) or Ruszczynski (2006) amongst many others.

8 We chose 2010 because it was the most up-to-date sample when we began our work in early 2012. We had a long learning curve. It took us nearly two years to program the necessary tests to make an assessment of its representativeness. Then we started to work with the other samples (Sect. 2) and also found a problem of representativeness. Moreover, after applying the procedure to 2010 we did the same with other years and obtained results that are not very different from those for 2010.
gender and type of pension, even though the main objective is not this but to obtain the subsample itself for use in any subsequent analysis of the Spanish public pension system. Given that the lack of representativeness of the CSWL has also been found in other years, our findings suggest that the same procedure might relevantly be applied to select subsamples in the other waves of the CSWL.

The structure of the rest of the paper is as follows: Sect. 2 analyses the representativeness of the CSWL for the years from 2005 to 2013 with respect to pension benefits. Section 3 sets out the distribution by type of pension, gender, and age of a hypothetical CSWL using SR sampling and the distribution of a subsample obtained by SR sampling using the original CSWL. In this section we show the importance of stratification as a sort of backtesting. We check whether stratification matters by looking at the total expenditure estimated using a stratified random sample. Section 4 details the criteria used for subsample selection and the results obtained. The paper ends with conclusions, pointers for future research and two appendices: the first shows all the tables and graphs with the estimates of the total expenditure deviations for 2010, and the second (online) extends the analysis of the goodness of fit of the CSWL to the population (INSS) for the whole period 2005–2013 and summarizes the problem statement whose solution provides the distribution of the number of pensioners in large subsamples which also represent the population better than the CSWL itself.

2 Analysis of the goodness of fit of the CSWL to the population (2005–2013)

In this section we analyse how well the CSWL pension data distribution fits the population distribution of pensioners at December 31st for the years 2005 to 2013\(^9\) by age, gender, and type of pension. The data are available in the statistical reports of the National Social Security Institute (INSS), though it is important to stress that the population from which the sample is drawn comprises all those individuals who have been registered or have received some kind of contributory pension from the Social Security system at any time during the year of reference, regardless of how long they were in that situation. It does not therefore coincide with the figure for the population at December 31st each year or indeed with the population of pensioners, but is larger. However, in our study there is a process of post-stratification of the CSWL with all pensions registered (current) at December 31st being grouped by cohorts for age, gender and type of pension as of that date. We do not add all the pensions that were registered at some time during the year but only those which were recorded as currently registered on December 31st, just as the INSS statistical report for each year does. The composition of the pensioners population is obtained from INSS (2006–2007), INSS (2008–11) and INSS (2012–2014). Pensions deregistered during the year due to the death of the recipient or because the recipient ceased to meet the requirements for receiving the pension are not considered on December 31st. The statistics on the total number of pensioners in the INSS statistical report consider only those

\(^9\) We do not use the first version from 2004 because it does not give the month of birth, so it is not possible to determine actuarial age as in the subsequent versions.
pensions recorded as currently registered on December 31st, so a comparison between the distributions of pensioners (by type of pension, gender and age) in the CSWL as of December 31st and the INSS makes sense, because the moment in time considered is the same.

In short, the aim of this analysis is to determine whether there are any statistically significant differences in the weights of the cohorts between the sample and the population using a goodness of fit test.

To conduct the test it is necessary to conduct a post-stratification of the CSWL once the sample records with no information on gender or date of birth have been deleted. The main theoretical reason why the population of pensioners is stratified by type of pension, gender and age is that there is a different life expectancy for each pensioner depending on the type of benefit received (retirement, permanent disability, widow(er)’s, orphan’s and family responsibilities), whether the pensioner is a man or a woman and whether he/she has a given age are taken into account. So in order to make accurate forecasts when analysing the sustainability of the Spanish public pension system, in which life expectancy plays a crucial role, those differences have to be taken into account. Therefore it is very important to have a sample of individuals that adequately represents the population of pensioners, taking into account these variables.

Another practical reason is that the information available to us about the population of pensioners is the distribution of this population organized by age, gender and type of pension. Moreover, the age cohorts considered in our analysis are also given in the format in which the information is disclosed by the INSS.

Once the data on pensions from the CSWL has been post-stratified at December 31st by type of pension, gender, and age cohorts, we perform a preliminary analysis comparing the distribution obtained from the CSWL with that of the population for 2005–2013. We use the equivalent table from the statistical report of the Spanish National Social Security Institute (INSS), once those population records that provide no information on gender or date of birth have been deleted. We thus compare the number of pensions in each cohort of the sample with the same cohort of the population, to check for differences with respect to the 4% of the population that the sample should in theory represent.

Table 1 shows the percentage of the population represented for each cohort, pointing out those cohorts underrepresented with percentages of less than 3% and those over-represented with percentages greater than 5% for 2010. Similar tables for other years are provided in the Online Appendix. Mere observation of the tables with these calculations suffices to detect the presence of the same kind of problem of overrepresentation and underrepresentation for certain cohorts in almost all the waves analysed.

Analysing the results for the said ratios, it can be concluded that there are cases in all the years where the figure exceeds 5% or fails to reach 3% of the population, i.e. where it deviates from the percentage of the population (4%) represented by the

---

10 This refers to a special type of survivors’ benefits for family members. This benefit is not compatible with the beneficiary receiving other public pensions.
Table 1  Percentages of pensions in the CSWL out of the total INSS population by age, 2010. Source Authors’ own calculations based on the CSWL 2010 and INSS (2011)

| Age cohorts | Permanent disability |  |  |  |  |  |  |  |  |  |  |  |
|-------------|---------------------|---|---|---|---|---|---|---|---|---|---|---|
|             | Male %  | Female % | Total % | Male %  | Female % | Total % | Male %  | Female % | Total % |
| 15–19       | 2.50     | 0.00      | 0.00     | 2.22     | 884.22    | 884.22   | 0.00     | 0.00     | 0.00    |
| 20–24       | 3.47     | 4.13      | 3.59     | 4.16     | 724.17    | 724.17   | 0.00     | 0.00     | 0.00    |
| 25–29       | 3.78     | 3.74      | 3.73     | 4.16     | 748.48    | 748.48   |
| 30–34       | 4.08     | 3.84      | 4.00     | 3.85     | 748.48    | 748.48   |
| 35–39       | 4.02     | 4.03      | 4.02     | 756.41   | 756.41    |
| 40–44       | 4.02     | 4.03      | 4.02     | 756.41   | 756.41    |
| 45–49       | 4.02     | 3.89      | 3.98     | 773.37   |
| 50–54       | 3.96     | 4.08      | 4.00     | 733.37   |
| 55–59       | 3.96     | 3.89      | 3.98     | 733.37   |
| 60–64       | 4.01     | 4.02      | 4.01     | 740.23   |
| 65–69       | 88.02    | 96.75     | 90.77    | 1023.97  |
| 70–74       | 11.32    | 5.58      | 6.76     | 832.17   |
| 75–79       | 5.79     | 4.33      | 4.64     | 426.54   |
| 80–84       | 4.11     | 4.16      | 4.16     | 375.32   |
| 85 and over | 4.58     | 3.94      | 3.97     | 369.60   |
| Total       | 4.25     | 4.22      | 4.24     | 845.13   |
| Average age | 55       | 57        | 56       | 75       | 76       | 76       |
| Age cohorts | Widow(er)'s | Orphan's |
|-------------|-------------|-----------|
| Male        | Female      | Total     | Male        | Female      | Total     |
| % P. Average| % P. Average| % P. Average| % P. Average| % P. Average| % P. Average|
| 0–4         | 0.00        | 0.00      | 0.00        | 4.58        | 265.96    |
| 5–9         | 0.00        | 0.00      | 0.00        | 4.32        | 263.20    |
| 10–14       | 0.00        | 0.00      | 0.00        | 4.28        | 263.90    |
| 15–19       | Outlier     | 396.36    | 50.00       | 1.042.79    | 3.75      |
| 15–19       | Outlier     | 582.00    | 75.00       | 827.31      | 262.08    |
| 20–24       | 85.71       | 894.40    | 10.68       | 644.19      | 745.44    |
| 25–29       | 73.08       | 642.69    | 8.20        | 711.74      | 678.94    |
| 30–34       | 26.29       | 673.96    | 5.52        | 653.24      | 660.63    |
| 35–39       | 5.89        | 616.53    | 4.52        | 648.78      | 643.71    |
| 40–44       | 4.93        | 594.51    | 4.28        | 625.15      | 620.95    |
| 45–49       | 4.39        | 564.79    | 4.14        | 627.18      | 618.52    |
| 50–54       | 3.59        | 576.48    | 3.88        | 646.28      | 637.47    |
| 55–59       | 3.51        | 570.87    | 3.93        | 638.26      | 630.86    |
| 60–64       | 3.73        | 539.97    | 3.99        | 659.33      | 648.91    |
| 65–69       | 3.70        | 462.44    | 3.95        | 634.15      | 622.96    |
| 70–74       | 4.09        | 400.04    | 4.01        | 605.55      | 593.48    |
| 75–79       | 3.94        | 383.33    | 4.01        | 588.50      | 577.41    |
| 80–84       | 4.05        | 359.75    | 3.97        | 562.96      | 551.40    |
| 85 and over | 3.70        | 333.78    | 3.97        | 515.48      | 504.75    |
| Total       | 3.96        | 436.00    | 3.99        | 582.00      | 572.00    |
| Average age | 72          | 76        | 76          | 3.40        | 352.00    |

Table 1 continued
Table 1 continued

| Age cohorts | Family responsibilities | Total |
|-------------|-------------------------|-------|
|             | Male | P. Average | Female | P. Average | Total | Male | P. Average | Female | P. Average | Total |
|             | %    |           | %      |           | %      |        | %    |           | %      |           |       |
| 0–4         | 0.00 | 0.00      | 0.00   | 0.00      | 0.00   | 0.00   | 4.58 | 265.96    | 3.75   | 262.08    | 4.17  | 264.24 |
| 5–9         | 0.00 | 0.00      | 0.00   | 0.00      | 0.00   | 0.00   | 4.32 | 263.20    | 4.20   | 260.01    | 4.26  | 261.68 |
| 10–14       | 0.00 | 0.00      | 0.00   | 0.00      | 0.00   | 0.00   | 4.26 | 263.90    | 4.36   | 268.06    | 4.31  | 265.94 |
| 15–19       | 2.90 | 318.96    | 3.23   | 243.07    | 3.07   | 276.80 | 3.88 | 271.35    | 4.06   | 267.03    | 3.97  | 270.07 |
| 20–24       | 3.01 | 242.84    | 3.39   | 275.16    | 3.20   | 259.70 | 4.03 | 308.40    | 3.99   | 307.96    | 4.01  | 308.18 |
| 25–29       | 3.38 | 316.99    | 4.21   | 265.04    | 3.85   | 284.83 | 4.91 | 647.26    | 4.46   | 552.51    | 4.76  | 616.63 |
| 30–34       | 5.00 | 237.98    | 2.94   | 281.07    | 3.87   | 255.93 | 4.18 | 690.40    | 4.13   | 607.80    | 4.16  | 659.32 |
| 35–39       | 3.51 | 281.15    | 5.99   | 253.39    | 4.98   | 261.32 | 4.08 | 710.16    | 4.08   | 611.98    | 4.08  | 669.29 |
| 40–44       | 4.19 | 302.81    | 4.35   | 208.14    | 4.27   | 252.69 | 4.06 | 717.12    | 4.13   | 623.99    | 4.09  | 675.10 |
| 45–49       | 5.56 | 352.90    | 3.56   | 417.85    | 4.30   | 386.79 | 4.06 | 734.10    | 4.00   | 645.04    | 4.03  | 692.13 |
| 50–54       | 4.16 | 494.19    | 4.30   | 460.53    | 4.26   | 471.30 | 3.94 | 800.77    | 4.00   | 675.32    | 3.97  | 736.84 |
| 55–59       | 4.53 | 450.86    | 3.48   | 554.48    | 3.79   | 517.08 | 3.94 | 967.80    | 3.95   | 687.47    | 3.94  | 827.84 |
| 60–64       | 3.42 | 451.38    | 4.02   | 486.03    | 3.87   | 478.29 | 3.99 | 1,209.63  | 4.03   | 753.09    | 4.01  | 1,018.27 |
| 65–69       | 5.12 | 490.43    | 4.15   | 530.45    | 4.34   | 521.53 | 4.03 | 1,167.15  | 3.97   | 681.53    | 4.00  | 960.60 |
| 70–74       | 3.83 | 499.33    | 3.79   | 498.91    | 3.79   | 498.97 | 4.05 | 1,007.24  | 4.01   | 609.42    | 4.03  | 822.92 |
| 75–79       | 4.69 | 470.51    | 4.40   | 491.46    | 4.43   | 488.97 | 4.00 | 940.06    | 3.99   | 584.00    | 3.99  | 757.29 |
| 80–84       | 5.12 | 529.06    | 4.09   | 425.30    | 4.20   | 438.06 | 3.97 | 865.29    | 3.95   | 558.28    | 3.96  | 685.23 |
| 85 and over | 3.15 | 467.27    | 3.93   | 466.85    | 3.85   | 466.89 | 3.87 | 758.08    | 3.95   | 510.04    | 3.93  | 586.03 |
| Total       | 4.13 | 442.00    | 4.00   | 474.00    | 4.03   | 467.00 | 4.00 | 975.63    | 3.98   | 599.70    | 3.99  | 783.13 |
| Average age | 61   | 71        | 69     | 70        | 73     | 72     |      |           |        |           |       |       |

Italic values indicate the underrepresented cohorts with percentages less than 3% of the population.
Bold values indicate the overrepresented cohorts with percentages greater than 5% and less than 100%.
Bold italic values indicate the cohorts with outliers or percentages over 100% of the population.
CSWL, with some cohorts being considerably overrepresented in relative terms. In all
the years considered the CSWL also contains age cohorts for some types of pension
that present outliers, while in the population those cohorts have no pensions.

The main mismatches are found in permanent disability and widow(er)’s pensions,
and to a lesser extent retirement pensions, in the case of men. The mismatches are
greatest in 2005, 2006, 2008, 2009, 2010 and 2011, and less significant in 2007, 2012
and 2013, given that the number of cohorts more than one fourth away from 4% is
smaller and where such differences do exist they are smaller. Hence for the years where
the differences are greater, the statistical test is expected to provide results that support
the existence of statistically significant differences not due exclusively to sample size.

Pearson’s chi-squared test ($\chi^2$) is considered as a test of goodness of fit to check
whether the sample follows the same distribution as the population of pensioners as
of 31st December. Goodness of fit tests usually have a given hypothesis as to the
theoretical distribution for the population, which they test using the data observed in
the sample. In the case of the CSWL the distribution of the population of pensioners
by age, gender, and type of pension is known from the statistical report of the INSS
(INSS, 2006–2014).

We use Pearson’s chi-square test, given by:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

in which $O_i$ represents the values observed in the CSWL and $E_i$ the expected or theo-
retical values, i.e. those obtained from the distribution of the population of pensioners
in each of the years considered, 2005–2013, as given by the INSS.

For this test the expected frequency for each cohort needs to be calculated by gender
and type of pension. That is, for a given gender and type of pension we calculate the
relative frequency as the ratio of the number of pensions in each age cohort to the total
pensions for the same gender and type of pension in the population.

To calculate the expected or theoretical values we use the following definition of
the relative distribution of expected frequency for population $\hat{f}$:

$$\hat{f}_{i, j, k} = \frac{N_{i, j, k}}{\sum_{i=1}^{18} N_{i, j, k}} = \frac{N_{i, j, k}}{N_{j, k}}$$

Using this relative frequency, the expected values of the subsample or the absolute
expected frequency for $\hat{e}$ in the subsample are:

$$E_i = \hat{e}_{i, j, k} = \hat{f}_{i, j, k} \cdot n_{j, k}$$

where $i$ is the index for the 18 cohorts into which the variable “age” has been divided;
$j$ is the index corresponding to “gender” (male, female); $k$ is the index for the 5
types of pension (permanent disability, retirement, widow(er)’s, orphan’s and family
responsibilities); $N_{j, k}$: is the number of pension benefits in the population per type of
pension $k$ and gender $j$; and $n_{j, k}$: is the number of pension benefits in the CSWL for
pension type $k$ and gender $j$. 
For large samples, as is the case with the CSWL, which covers more than 300,000 pensioners in every available year, it is very unlikely that the sample will be a perfect fit to the population, so the test statistic will show a rejection of the hypothesis that the sample and the population have the same distribution and conclude that the differences between the two distributions are statistically significant. Those differences could be magnified because of the large size of the sample. To overcome this possible error in the interpretation of the test results (pointed out by Berkson 1938; Wang 1993 and Lin et al. 2013 among others) it is important to ensure that the differences found are not due to the large size of the sample, so there is evidence not only of statistical differences but also of practical ones.

According to Wilkinson (1999), statistical significance refers to whether the effect observed is larger than would be expected by chance, i.e. can the null hypothesis that there is no effect be rejected? This is what is typically addressed by \( p \) values. Practical significance is about whether we should care, i.e. whether the effect is useful in an applied context. Two groups will almost never be exactly the same if thousands or millions of people are tested. That does not mean that every difference is of interest. This is usually associated with effect size measures [e.g. Cohen’s \( d \), (Cohen 1988)].

In statistics, an effect size is a measure of the strength of the relationship between two variables in a statistical population, or a sample-based estimate of that quantity. An effect size calculated from data is a descriptive statistic that expresses the estimated magnitude of a relationship without making any statement about whether the apparent relationship in the data reflects a true relationship in the population. Thus effect sizes complement inferential statistics such as \( p \) values. Sample-based effect sizes are distinguished from test statistics used in hypothesis testing in that they estimate the strength of an apparent relationship rather than assigning a significance level reflecting whether the relationship could be due to chance. The effect size does not determine the significance level or vice-versa. Given a sufficiently large sample size, a statistical comparison will always show a significant difference unless the population effect size is exactly zero.

The best measure of association for the chi-square test is \( \varphi \) (or Cramér’s \( \varphi \)) and Cramér’s \( V \). \( \varphi \) is related to the point-biserial correlation coefficient and Cohen’s \( d \) and estimates the extent of the relationship between two variables \((2 \times 2)\). Cramér’s \( V \) may be used with variables that have more than two levels. \( \varphi \) can be computed by finding the square root of the chi-square statistic divided by the sample size.

\[
\varphi = \sqrt{\frac{\chi^2}{N}} \tag{4}
\]

Similarly, Cramér’s \( V \) is computed by taking the square root of the chi-square statistic divided by the sample size and the length of the minimum dimension.

\[
V = \sqrt{\frac{\chi^2}{N \times (k - 1)}} \tag{5}
\]

with \( k \) = the number of regrouped cohorts.
Once the hypothesis has been tested in the case of the fit of the distribution by age for each gender and type of pension, to determine whether the differences found are statistically significant, the size of the effect is estimated using Cramér’s V, the results of which enable the preliminary analysis of practical significance to be completed.

Depending on the values of a contingency coefficient such as Cramér’s V it is possible to determine the size of the effect (Cohen 1988), and it can be used to help decide whether or not those differences which are statistically significant are due to the large size of the sample.

The results of the test for per type of pension and gender for 2010 are summarized (Table 2) whereas (the results) for the whole period can be found in the Online Appendix (Tables 1.9–1.18). It can be observed that they are almost equivalent in most years, which means that the hypothesis that the CSWL has the same distribution as the population in most pension benefits (permanent disability, retirement and widow(er)s) is rejected. However, in pension benefits for retirement the size of the effect is negligible, so the differences detected can be attributed to the large size of the sample. The causes of these differences can be attributed to the sample design (simple random sampling), to administrative errors and to a reclassification of pensioners older than 65 with permanent disability benefits, who are considered as disabled in the CSWL but as retirement pension beneficiaries in the official population statistics.

In the versions of the CSWL for 2007, 2012, and 2013, the differences found in the distribution of pensioners receiving permanent disability benefits are due to the large size of the sample, given that the size of the effect is negligible. This does not happen in the case of pension benefits for widow(er)s. It has also been detected in those years that the code assigned to most pensioners over 65 on permanent disability benefits is changed to retirement benefits, when in the previous years they continued to be classed as receiving permanent disability benefits. This explains the better fit in 2007, 2012 and 2013. It is worth noting in particular what happens after 2007: the same coding errors appear in the pensions for permanent disability for pensioners older than 65.

Hence it is concluded that the fit is not good for some types of pension benefit given that some cohorts are over or underrepresented in the CSWL with respect to the actual population of pensioners, and contain a number of pensions clearly higher or lower than the figure expected depending on the proportion of the reference population represented by the CSWL (4%). The results seem to suggest that for 2005, 2006, 2008, 2009, 2010 and 2011 the CSWL does not fit the distribution of the population well in terms of type of pension, gender, and age for two types of pension benefit: permanent disability and widowhood. The mismatch is greater in the former, and the poor fit is not attributable solely to the large sample size. In other years the null hypothesis that the CSWL has the same distribution as the population cannot be rejected given the size of the sample, but there is room for improvement in the cases where the null hypothesis cannot be rejected, as will be shown in the next section.

These results must be taken into account when making forecasts on the sustainability of the Spanish public pension system using the CSWL, given that we find that for most years and some types of pension benefit it does not correctly represent the distribution by age of the contributory pensions in the system whose sustainability is to be analysed. It is concluded that the significance of the differences detected goes beyond a single year given that they are found for a considerable number of waves of the CSWL.
| Source | Authors' own calculations |
|---|---|
| Data | Permanent disability | Retirement | Widow(er)'s | Orphan's | Family responsibilities |
| | Male | Female | Male | Female | Male | Female | Male | Female | Male | Female |
| Sample size | 25,960 | 13,694 | 132,151 | 73,116 | 6303 | 85,497 | 5634 | 5302 | 328 | 1184 |
| $\chi^2$ | 29,200 | 16,415 | 57.7 | 54.87 | 1179 | 84.1 | 16.6 | 12.8 | 12.7 | 8.8 |
| Cohorts (k) | 13 | 14 | 8 | 7 | 12 | 13 | 18 | 18 | 14 | 15 |
| $p$ value | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.480 | 0.748 | 0.472 | 0.843 |
| $\chi^2_{\alpha,(k-1)}$ | 21.026 | 22.362 | 14.067 | 12.592 | 19.69 | 21.026 | 27.587 | 27.587 | 22.362 | 23.685 |
| Reject | Yes | Yes | Yes | Yes | Yes | Yes | No | No | No | No |
| V Cramér | 0.306 | 0.304 | 0.008 | 0.011 | 0.130 | 0.009 | 0.013 | 0.012 | 0.055 | 0.023 |
| TE | Large | Large | Negli. | Negli. | Medium | Negli. | Negli. | Negli. | Small | Negli. |

Table 2 Results chi-squared test for 2010. Source Authors’ own calculations
Wang (1993) advocates asking whether these differences are important or significant in practice too. The answer to this question differs from case to case and according to the experience of each research team, as there is no statistic for measuring significance in practice. Hence some researchers may consider an error of 1% to be important while for others it is negligible, depending on the context and goal of each study.

To estimate the significance in practice of the differences in the distribution of the number of pensions by cohorts, the total annual expenditure on pensions by cohorts could be estimated using the CSWL and compared with the estimate obtained using the population, which is known. In addition, if the estimate provided by a sample obtained using SR can be compared, it would reveal how much margin for improvement there is in these estimates and hence the significance of the differences found. We seek to answer these questions below. In particular, if one of the objectives is to use the data from the CSWL to make forecasts about the sustainability of the Spanish public pension system, it is advisable to have more representative subsamples based on using an SR, the characteristics of which are described in the following section.

3 Checking whether stratification matters

When deciding what sampling design is most appropriate for studying pension benefits and pension expenditure, it is important to know whether it is relevant to divide the population into levels and groups. To show the real importance of stratification for the case of the CSWL, in this section a sort of backtesting is carried out. This is a process widely used in finance, demographics and insurance among other fields. In our case we want to test our methodology on prior time periods. Instead of applying the methodology for the time period forward, in which case its effectiveness could take years to check, our procedure is applied to relevant past data in order to gauge its usefulness.

We have the information on the distribution of pensioners organized by age, gender and type of pension and we know the number of pensions and the mean pension expenditure in each group, so in this section we check whether stratification matters by looking at the total expenditure estimated using a stratified random sample. Given that we do not have the entire population, we obtain the distribution of pensions of a hypothetical stratified random sample extracted from the population and estimate the total annual pension expenditure by cohort at 31st December 2010. This is then compared with the CSWL to check for any improvement in the estimate and forecasts for pension expenditure for 2010 using the hypothetical sample obtained from the population by stratified random sampling (SR) with respect to the CSWL. An improvement can be expected because we find that for most years and some types of pension benefit the CSWL does not correctly represent the distribution by age of the contributory pensions in the system.

It is clear that stratification is indeed relevant because, for example, the age of pensioners is a variable that has an important influence when it comes to forecasting expenditure on pensions. As stated above, these variables are also important in analysing the sustainability of the public pension system, where pensioners’ life expectancy plays a very important role.
One of the main goals of stratification is to give a better cross-section of the population so as to increase relative precision. There are several reasons to use stratified random sampling. Stratification ensures adequate representation of various groups of the population which may be of interest or importance. The use of stratification increases the accuracy with which a characteristic of a population can be estimated. It is achieved by dividing a heterogeneous population into sub-populations, each of which is homogeneous within itself. When there are extreme values in the population, stratification is more powerful because individual strata will be more homogeneous and separate estimates obtained from them can be combined into a precise estimate for the whole population by taking a relatively smaller sample (Singh and Chaudhary 1986).

Once the strata or groups to be considered are established, the next step is to decide the sample size in each stratum, i.e. the method of allocation. In this study we have opted for proportional allocation. This allocation was originally proposed by Bowley (1926), and it is very common in practice because of its simplicity. When the only information available is \( N_i \), the total number of units in the \( i \)-th stratum, as is the case here because the number of pensioners in the population in each age cohort by gender and type of pension is known, a given sample of size \( n \) is allocated in the different strata in proportion to their sizes, i.e. in the \( i \)-th stratum:

\[
\frac{n_i}{n} = \frac{N_i}{N} \rightarrow n_i = n \left( \frac{N_i}{N} \right) = n \left( w_i \right) = N_i \left( \frac{n}{N} \right) \quad (6)
\]

This means that the sampling fraction is the same in all strata and coincides with the overall sampling fraction, whose value is the constant of proportionality, \( q = n/N \).

The technique of post-stratification consists of classifying the population and the selected sample into a given number of strata or groups after selection of the sample as a simple random sample. This is done here by taking the CSWL (simple random sample) and grouping the data by type of pension, gender and age. The problem of post-stratification was discussed originally by Hansen et al. (1953), and the technique is normally applied when the value of the variable used to build the strata is not known but might be known after the simple random sample has been obtained. This is not the case with the CSWL, given that the variables used to build the strata or cohorts (type of pension, gender and age) are known. Following Tryfos (1996), considering the selection of a simple random sample of individuals, the post stratified estimator for the population mean of the variable of interest \( Y \) once the sampled individuals have been classified into the \( M \) groups is:

\[
\tilde{Y}_{ps} = w_1 \tilde{Y}_1 + w_2 \tilde{Y}_2 + \cdots + w_M \tilde{Y}_M \quad (7)
\]

where \( w_i = N_i/N \), and \( \tilde{Y}_i \) is the mean of the variable \( Y \) of the sampled individuals in group \( i \). It is calculated exactly like the stratified estimator \( \tilde{Y}_S \) but is based on the results of a simple -not a stratified- random sample. The weights \( w_i \) are assumed to be known. When the size of the simple random sample, \( n \), is large, the proportion of sampled elements that fall into a given group can be expected to be approximately equal to the proportion of elements of that group or stratum in the population, that is, \( n_i/n \approx N_i/N \). So when \( n \) is large, the post-stratified estimator based on a simple random sample can be expected to behave like a stratified estimator based on a proportional stratified sample.
These considerations on the data from the population are important in studying the average age and average pension of the population relative to the same data in the age cohorts or strata of the CSWL and for the subsequent post-stratification. However, taking into account that the average means and the average pensions of all the population are known for each type of pension, gender and age cohort, the interest is focused on whether \( n_i/n \approx N_i/N \), even though the CSWL size may be considered large. The first section of the paper has questioned whether the proportions in the strata or cohorts of the CSWL are similar to the corresponding ones in the population. Table 3 presents the distribution by type of pension, gender and age that a sample should have using proportional stratified sampling if it is to be representative of the population of pensioners with a constant of proportional allocation of, approximately, \( q = 3.992\% \). This constant of proportional allocation is the result of dividing the number of pension benefits at December 31st contained in the CSWL, \( n_{CSWL} = 349,169 \) (source: Authors’ Own Calculations based on the CSWL for 2010), by the total number of pensions in the population taken from INSS (2011), \( N_{INSS} = 8,747,470 \). The results in Table 3 total 3 pensions more than in the CSWL because of rounding.

To correctly estimate the dimension of the problems found in our analysis it is worth highlighting the differences in the relative importance of each type of pension: retirement, permanent disability and death. For example, in 2010 (see Table 3) retirement pensions account for 59.48% of the total number of beneficiaries. Gender really matters: for men the figure is 78.43% whereas for women it is only 41.50%. Widow(er)s account for 26.31% of the total number of pensions, but for women the figure is 47.72%, which is even higher than for retirement pensions. In the case of men the weight of this contingency is much lower at just 3.73% of the total number of beneficiaries.

Next we estimate total annual pension expenditure by cohort at 31st December 2010 for the population and for the CSWL (Table 4). To isolate the effect of the differences in the distribution of average pension amounts between the CSWL and the actual population (INSS), we estimate total expenditure for each cohort by taking the average pension published in the 2010 INSS statistical report. To obtain pension expenditure for the population (INSS), we multiply the number of pensions by the average pension for each cohort and by the coefficient (between 12 and 14) for adjusting the total amount of pensions at December 31st to recognized expenditure in financial year 2010 on each type of pension, as shown in the fourth column (No. INSS adjusted payments) in Table 4.

To obtain figures for the seventh column in the table [CSWL expenditure (year)], for each cohort we multiply the number of pensions by the average pension, by the raising factor (the ratio of population to sample size, \( 1/q \)) and by the same coefficient that adjusts the monthly pension to the pension recognized as expenditure in 2010. For the case of the SR we proceed in the same way as with the CSWL.

In Appendix 1, Tables 12, 13, 14, 15, 16, 17, 18, 19, 20 and 21 show the expenditure calculated per cohort using the CSWL and the SR hypothetical sample for each type of pension together with Figs. 1, 2, 3 based on these tables. With them we seek to show the distortion in the estimation of pension expenditure for each type of contingency, gender and age cohorts caused by using RS instead of SR. Graphically, the improvement in the estimate of total pension expenditure for all types of pensions and for both men
Table 3 Distribution of pensions at 31/12/2010 in a hypothetical sample extracted from the population using stratified random sampling. Source Authors’ own calculations

| Age cohorts | Permanent disability | Retirement | Widow(er)’s |
|-------------|----------------------|------------|-------------|
|             | Male | Female | Total | Male | Female | Total | Male | Female | Total |
| 15–19       | 2    | 2      | –     | –    | –      | –     | –    | –      | –     |
| 20–24       | 64   | 14     | 78    | –    | –      | –     | –    | 3      | 31     | 34   |
| 25–29       | 260  | 88     | 348   | –    | –      | –     | 15   | 127    | 142   |
| 30–34       | 657  | 269    | 926   | –    | –      | –     | 49   | 341    | 390   |
| 35–39       | 1233 | 583    | 1816  | –    | –      | –     | 49   | 341    | 390   |
| 40–44       | 2029 | 964    | 2993  | –    | –      | –     | 111  | 804    | 915   |
| 45–49       | 2973 | 1472   | 4445  | 3    | 1      | 4     | 229  | 1507   | 1736  |
| 50–54       | 4111 | 2247   | 6358  | 33   | 4      | 37    | 384  | 2457   | 2841  |
| 55–59       | 5524 | 2841   | 8365  | 466  | 11     | 477   | 495  | 3578   | 4073  |
| 60–64       | 7426 | 3533   | 10,959| 10,260| 3849   | 14,109| 560  | 5473   | 6033  |
| 65–69       | 70   | 32     | 102   | 33,134| 17,436 | 50,570| 576  | 7732   | 8308  |
| 70–74       | 4    | 16     | 20    | 28,575| 14,782 | 43,357| 647  | 10,565 | 11,212|
| 75–79       | 10   | 101    | 111   | 27,946| 14,334 | 42,280| 925  | 15,887 | 16,812|
| 80–84       | 17   | 330    | 347   | 19,425| 11,708 | 31,133| 1002 | 16,967 | 17,969|
| 85 and over | 19   | 450    | 469   | 13,468| 12,244 | 25,712| 1351 | 20,038 | 21,389|
| Total       | 24,399| 12,940| 37,339| 133,310| 74,369 | 207,679| 6347 | 85,511 | 91,858|
| Total%      | 6.99 | 3.71   | 10.69 | 38.18 | 21.30  | 59.48 | 1.82 | 24.49  | 26.31 |
| Gender%     | 14.35| 7.22   | –     | 78.43 | 41.50  | 0.00  | 3.73 | 47.72  | –     |
| Age cohorts | Orphan's | | | Family responsibilities | | | | Total | | |
| | Male | Female | Total | Male | Female | Total | Male | Female | Total |
| 0–4 | 83 | 81 | 164 | | | | | | |
| 5–9 | 316 | 297 | 613 | | | | | | |
| 10–14 | 644 | 605 | 1247 | 2 | 1 | 3 | | | |
| 15–19 | 1203 | 1163 | 2366 | 6 | 6 | 12 | | | |
| 20–24 | 691 | 715 | 1406 | 15 | 14 | 29 | | | |
| 25–29 | 73 | 51 | 124 | 9 | 12 | 21 | | | |
| 30–34 | 145 | 100 | 245 | 6 | 7 | 13 | | | |
| 35–39 | 251 | 166 | 417 | 5 | 7 | 12 | | | |
| 40–44 | 365 | 252 | 617 | 8 | 8 | 16 | | | |
| 45–49 | 452 | 315 | 767 | 16 | 27 | 43 | | | |
| 50–54 | 425 | 324 | 749 | 38 | 79 | 117 | | | |
| 55–59 | 345 | 290 | 635 | 54 | 124 | 178 | | | |
| 60–64 | 254 | 263 | 517 | 51 | 152 | 203 | | | |
| 65–69 | 168 | 205 | 373 | 30 | 131 | 161 | | | |
| 70–74 | 95 | 140 | 235 | 17 | 112 | 129 | | | |
| 75–79 | 61 | 120 | 181 | 19 | 148 | 167 | | | |
| 80–84 | 26 | 66 | 92 | 18 | 160 | 178 | | | |
| 85 and over | 9 | 38 | 47 | 24 | 194 | 218 | | | |
| Total | 5606 | 5191 | 10,795 | 318 | 1183 | 1501 | | | |
| Total% | 1.61 | 1.49 | 3.09 | 0.09 | 0.34 | 0.43 | | | |
| Gender% | 3.30 | 2.90 | 6.20 | 0.19 | 0.66 | 3.30 | | | |
Table 4  Estimate of total expenditure on pensions from expenditure at 31st December. Source (Instituto Nacional de la Seguridad Social, INSS 2011). In millions of euros Figures have been rounded up or down as appropriate

| Type of pension      | INSS (month) | Recognized expenditure | No. INSS adjusted payments | l/q | CSWL (month) | CSWL expenditure (year) | % diff. |
|----------------------|--------------|------------------------|----------------------------|-----|--------------|-------------------------|---------|
| Permanent disability | 799          | 11,156                 | 13.96                      | 25.05 | 33.51        | 11,721                  | −5.06   |
| Retirement           | 4647         | 63,268                 | 13.61                      | 182.88 | 62,371       | 18,072                  | 1.42    |
| Widow(er)’s          | 1321         | 18,142                 | 13.73                      | 52.55 | 18,072       | 1317                    | 0.38    |
| Orphan’s             | 95           | 1313                   | 13.82                      | 3.80  | 1317         | 244                     | −0.27   |
| Family responsibilities | 17       | 239                    | 13.77                      | 0.71  | 244          |                         | −1.74   |

INSS and CSWL 2010
and women can be seen, as expected. The fit is worse in the case of SR, where the size of the sample is small due to the loss of elements in the cohorts, so this originates greater differences in the forecast of total expenditure for those cohorts.

Table 5 shows the improvement indicators in the estimate of total pension expenditure using the hypothetical SR sample obtained from the population of pensioners. It shows large reductions in the various measures of error in the estimate of total pension expenditure using the CSWL compared to the population of pensioners.

The indicators used to measure the improvement are defined as follows:

- **R_SRMQE**: reduction in the square root of the mean quadratic error.
- **R_SDAV**: reduction in the sum of the differences in absolute value of estimated expenditure by cohort.
- **MDC**: maximum difference in the estimate of expenditure by cohort as a percentage of total expenditure by gender and type of pension for the case of the SR sample.
- **CSWL_MDC**: maximum difference in the estimate of expenditure by cohort as a percentage of total expenditure by gender and type of pension for the case of the original CSWL.
- **CE \( \leq 0.01\%\)**: percentage of cohorts with an error in the estimate of expenditure of less than 0.01% of total expenditure, in absolute value, for the case of the SR sample.
- **CSWL_CE \( \leq 0.01\%\)**: percentage of cohorts with an error in the estimate of expenditure of less than 0.01% of total expenditure, in absolute value, for the case of the original CSWL.

The SR sample obtained seems to be the best at representing the population of pensioners by age, gender and type of pension, with the larger one being better. The aim of this research is to extract a large subsample obtained from the CSWL, with a view to improving its representativeness, and to compare the reductions in the various measures of error of the estimate of total pension expenditure with those resulting from the hypothetical SR sample obtained from the population (which researchers cannot access).

However, the hypothetical SR sample seems not to be contained in the CSWL. For a research team to obtain a sample of data on pensions such as the one that would result from the hypothetical subsample using SR obtained from the population, but using the CSWL instead, the constant of proportionality \( q = n/N \) would have to be reduced.

We have developed a procedure in Excel VBA (Visual Basic for Applications) to reduce the constant of proportionality \( q = n/N \), using stratified sampling with proportional allocation to obtain the largest subsample contained in the CSWL that could be extracted from the population if available. The result is a sample size of 25% of the original CSWL sample, up to 87,472 pension benefits representing 0.99999% of the population of pensioners. This is a considerable loss in pension data on working lives (previous contributions as well as benefits). Table 6 shows the distribution of this subsample. In addition, this reduction gives a slightly worse fit than the hypothetical sample obtained from the population using SR sampling, as can be seen in Table 7.

In short, backtesting is carried out in this section to show the real importance of stratification for the case of the CSWL. We show, looking at the estimate of total expenditure, that with an SR sampling a better fit can be obtained by age cohorts, gender and type of pension. In order to be able to obtain an SR sample from the
Table 5  Improvement indicators in the estimate of total pension expenditure in % in the hypothetical SR sample with respect to the CSWL. *Source* Authors’ own calculations

| Improvement indicators | Permanent disability | Retirement | Widow(er)’s | Orphan’s | Family responsibilities |
|------------------------|----------------------|------------|-------------|----------|-------------------------|
|                        | Men  | Women | Men  | Women | Men  | Women | Male | Female | Men  | Women |
| R_SRMQE                | 99.95 | 99.84 | 99.92 | 99.83 | 99.41 | 99.53 | 96.11 | 93.75 | 85.58 | 97.44 |
| R_SDAV                 | 99.87 | 99.57 | 99.88 | 99.77 | 99.35 | 99.53 | 95.62 | 94.31 | 82.80 | 96.28 |
| MDC                    | 0.00  | 0.00  | 0.00  | 0.00  | 0.01  | 0.00  | 0.01  | 0.01  | 0.15  | 0.03  |
| CSWL_MDC               | 6.17  | 5.18  | 0.97  | 1.35  | 1.60  | 0.11  | 0.33  | 0.20  | 1.40  | 1.50  |
| CE ≤ 0.01%             | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 66.67  | 82.35  |
| CSWL_CE ≤ 0.01%        | 80.00  | 60.00  | 55.56 | 55.56 | 57.14 | 66.67 | 27.78 | 38.89 | 61.11  | 47.06  |
Table 6  Distribution of pensions at 31-12-2010 in a subsample selected from the CSWL using SR sampling.  Source Authors’ own calculations based on CSWL 2010

| Age cohorts | Permanent disability | | | Retirement | | | Widow(er)’s | | |
|-------------|----------------------|---|---|----------------------|---|---|----------------------|---|---|
|             | Male | Female | Total | Male | Female | Total | Male | Female | Total |
| 20–24       | 16   | 4     | 20    | –    | –      | –     | 1    | 1      | 1     |
| 25–29       | 65   | 22    | 87    | –    | –      | –     | 1    | 8      | 9     |
| 30–34       | 165  | 67    | 232   | –    | –      | –     | 4    | 32     | 36    |
| 35–39       | 309  | 146   | 455   | –    | –      | –     | 12   | 85     | 97    |
| 40–44       | 508  | 242   | 750   | –    | –      | –     | 28   | 201    | 229   |
| 45–49       | 745  | 369   | 1114  | 1    | –      | 1     | 57   | 377    | 434   |
| 50–54       | 1030 | 563   | 1593  | 8    | 1      | 9     | 96   | 616    | 712   |
| 55–59       | 1384 | 712   | 2096  | 117  | 3      | 120   | 124  | 896    | 1020  |
| 60–64       | 1860 | 885   | 2745  | 2570 | 964    | 3534  | 140  | 1371   | 1511  |
| 65–69       | 18   | 8     | 26    | 8301 | 4368   | 12,669| 144  | 1937   | 2081  |
| 70–74       | 1    | 4     | 5     | 7159 | 3703   | 10,862| 162  | 2647   | 2809  |
| 75–79       | 3    | 25    | 28    | 7001 | 3591   | 10,592| 232  | 3980   | 4212  |
| 80–84       | 4    | 83    | 87    | 4866 | 2933   | 7799  | 251  | 4251   | 4502  |
| 85 and over | 5    | 113   | 118   | 3374 | 3067   | 6441  | 339  | 5020   | 5359  |
| TOTAL       | 6113 | 3243  | 9356  | 33,397 | 18,630 | 52,027| 1590 | 21,422 | 23,012 |
Table 6 continued

| Age cohorts | Orphan's Family responsibilities | Total |
|-------------|----------------------------------|-------|
|             | Male | Female | Total | Male | Female | Total | Male | Female | Total |
| 0–4         | 21   | 20     | 41    | –    | –      | –     | 21   | 20     | 41    |
| 5–9         | 79   | 74     | 153   | –    | –      | –     | 79   | 74     | 153   |
| 10–14       | 161  | 152    | 313   | –    | –      | –     | 161  | 152    | 313   |
| 15–19       | 301  | 291    | 592   | 1    | 2      | 3     | 302  | 293    | 595   |
| 20–24       | 173  | 179    | 352   | 4    | 4      | 8     | 193  | 188    | 381   |
| 25–29       | 18   | 13     | 31    | 2    | 3      | 5     | 86   | 46     | 132   |
| 30–34       | 36   | 25     | 61    | 4    | 4      | 8     | 206  | 126    | 332   |
| 35–39       | 63   | 42     | 105   | 2    | 3      | 7     | 206  | 126    | 332   |
| 40–44       | 91   | 63     | 154   | 2    | 2      | 4     | 206  | 126    | 332   |
| 45–49       | 113  | 79     | 192   | 2    | 3      | 11    | 206  | 126    | 332   |
| 50–54       | 106  | 81     | 187   | 4    | 7      | 11    | 206  | 126    | 332   |
| 55–59       | 87   | 73     | 160   | 13   | 31     | 44    | 206  | 126    | 332   |
| 60–64       | 64   | 66     | 130   | 13   | 38     | 51    | 464  | 332    | 797   |
| 65–69       | 42   | 51     | 93    | 8    | 33     | 41    | 851  | 339    | 14,910 |
| 70–74       | 24   | 35     | 59    | 4    | 28     | 32    | 735  | 641    | 13,767 |
| 75–79       | 15   | 30     | 45    | 5    | 37     | 42    | 725  | 766    | 14,919 |
| 80–84       | 6    | 16     | 22    | 4    | 40     | 44    | 513  | 732    | 12,454 |
| 85 and over | 2    | 9      | 11    | 6    | 49     | 55    | 372  | 825    | 11,984 |
| Total       | 1402 | 1299   | 2701  | 78   | 298    | 376   | 42,580 | 44,892 | 87,472 |
Table 7  Improvement indicators in the estimate of total pension expenditure in % of the SR subsample (SRSS) with respect to the CSWL. *Source* Authors’ own calculations

| Improvement indicators | Permanent disability | Retirement | Widow(er)’s | Orphan’s | Family responsibilities |
|------------------------|----------------------|------------|-------------|----------|-------------------------|
|                        | M        | F        | M        | F        | M        | F        | M        | F        | M        | F        |
| R_SRMQE                | 99.78    | 99.52    | 99.59    | 99.57    | 97.85    | 97.77    | 80.75    | 76.98    | 41.70    | 88.12    |
| R_SDAV                 | 99.36    | 98.90    | 99.36    | 99.43    | 97.80    | 97.83    | 77.59    | 79.28    | 36.37    | 83.78    |
| MDC                    | 0.01     | 0.01     | 0.00     | 0.00     | 0.03     | 0.00     | 0.05     | 0.05     | 0.65     | 0.13     |
| CSWL_MDC               | 6.17     | 5.18     | 0.97     | 1.35     | 1.60     | 0.11     | 0.33     | 0.20     | 1.40     | 1.50     |
| CE ≤ 0.01%             | 100.00   | 93.33    | 100.00   | 100.00   | 64.29    | 100.00   | 44.44    | 44.44    | 44.44    | 52.94    |
| CSWL_CE ≤ 0.01%        | 80.00    | 60.00    | 55.56    | 55.56    | 57.14    | 66.67    | 27.78    | 38.89    | 61.11    | 47.06    |
CSWL available and not from the population, the subsample reduces the size to 25% of the original. In the following section we apply a procedure to select large SR subsamples, thus improving the fit of the CSWL to the population of pensioners, bearing in mind that this improvement will not be as great as that obtained using the hypothetical but unfeasible SR sample from the population, but it may at least be as good as the improvement offered by the SR subsample of 87,972 pensions extracted from the CSWL using stratification. The methodology developed has the property of allowing the user to choose the relationship between the desired goodness of fit to the population and the size of the subsample.

4 Selection of a large subsample distribution from the CSWL: results

In this section we explain the criteria for the design of a large subsample to be selected using proportional allocation stratification from the 2010 CSWL data to improve its representativeness with respect to the number of pension benefits in the population. We explain the procedure developed for this and the distribution of pensions in a subsample selected from the 2010 CSWL obtained with this procedure that will improve the fit with a high \( p \) value without missing as many registers as with the SR contained in the CSWL. Besides that, in order to illustrate the practical significance of our findings we show the results obtained in its application to estimate total expenditure by age, gender and type of pension and compare it with those that use the SR hypothetical sample and the subsample with respect to the CSWL.

The aim is to find a subsample for the more general case of fit to the distribution of the CSWL to the pensioner population by age, gender and type of pension.

The selection criteria for finding a subsample with the necessary characteristics are the following:

(a) It must be more representative of the population under study. The procedure should therefore include a goodness of fit test on the distribution of the number of pensioners by age, gender and type of pension that takes into account the associated \( p \) values.

(b) The total number of pensioners needs to be relatively high so as to be bigger than the number that would result from a stratified sample from the CSWL, approximately 1% of the population of pensioners. Hence the requirement is to maximize subsample size, with 1% of the pensioner population being the lower limit.

(c) The subsample obtained must be included in the population as well as in the CSWL. These two requirements might seem obvious, but constraints have to be introduced to avoid the outliers found in the CSWL but not in the population, i.e. those cohorts for which the number of pensions is greater than zero in the CSWL but zero in the population of pensioners. It is also important to have a number of pensioners in each cohort of the subsample that is lower than or equal to that of the corresponding cohort of the population.

The details of the method used to find the subsample design are explained in the Online Appendix. This is a nonlinear programming problem with just one real non-negative decision variable: the constant of proportionality, \( q = n/N \). When the only
information is $N_i$, i.e. the number of elements in each stratum, the allocation or assignment to each stratum in the sample is proportional to the size of the stratum, i.e. $n_i = \left(\frac{n}{N}\right) N_i = q N_i$.

Maximizing $q$ is equivalent to maximizing the size of the subsample. Without taking into account the integer constraints we have:

$$\max_q \left\{ q \cdot N^{INS} \right\} \rightarrow \max_q \left\{ n^{SUB} (q) \right\}$$  \hspace{1cm} (8)$$

If the integer constraints with respect to the number of pensions in any cohort are taken into account, what we are really considering when we maximize $q$ is the maximization of $\hat{q}$, the adjusted constant of proportionality, so this is what is finally considered:

$$\max_q \left\{ q \right\} \leftrightarrow \max_q \left\{ \hat{q} \right\}$$

with $\hat{q} = \frac{\sum_{k=1}^5 \sum_{j=1}^2 \sum_{i=1}^{18} n_{i,j,k}^{SUB} (q)}{\sum_{k=1}^5 \sum_{j=1}^2 \sum_{i=1}^{18} N_{i,j,k}^{INS} (q)} = \frac{\sum_{k=1}^5 \sum_{j=1}^2 \sum_{i=1}^{18} Trunc \left[ q \cdot N_{i,j,k}^{INS} (q) \right]}{\sum_{k=1}^5 \sum_{j=1}^2 \sum_{i=1}^{18} N_{i,j,k}^{INS} (q)}$  \hspace{1cm} (9)$$

In order to make the selected subsample more representative of the population, a goodness of fit test has to be considered. The procedure takes into account that the value of the test for the subsample to be selected is such that it does not reject the null hypothesis (the subsample has the same distribution as the pensioner population at 31st December 2010), by contrast with the alternative hypothesis (the subsample does not have the same distribution as the pensioner population at 31st December 2010). The goodness of fit test used is Pearson’s chi-squared test as explained in Sect. 2 of the paper.

Table 8 shows the results of applying the above procedure to the CSWL. A large subsample design is generated, improving the fit to the distribution of the population.
Table 9  Distribution of pensions at 31-12-2010 in a subsample selected from the CSWL with a $p$ value of $\geq 0.999$

| Age cohorts | Permanent disability | | Retirement | | Widow(er)'s | |
|-------------|----------------------|--|-----------|--|-----------|----------|
|             | Male | Female | Total | Male | Female | Total | Male | Female | Total |
| 15–19       | 1    | 1      | 1     | –    | –      | –     | –    | –      | –     |
| 20–24       | 55   | 12     | 67    | –    | –      | –     | 2    | 26     | 28    |
| 25–29       | 225  | 75     | 300   | –    | –      | –     | 12   | 109    | 121   |
| 30–34       | 570  | 233    | 803   | –    | –      | –     | 42   | 295    | 337   |
| 35–39       | 1069 | 505    | 1574  | –    | –      | –     | 96   | 697    | 793   |
| 40–44       | 1760 | 836    | 2596  | –    | –      | –     | 198  | 1307   | 1505  |
| 45–49       | 2580 | 1277   | 3857  | 2    | –      | 2     | 332  | 2132   | 2464  |
| 50–54       | 3567 | 1950   | 5517  | 28   | 3      | 31    | 429  | 3104   | 3533  |
| 55–59       | 4793 | 2465   | 7258  | 395  | 9      | 404   | 485  | 4749   | 5234  |
| 60–64       | 6443 | 3065   | 9508  | 8903 | 3339   | 12242 | 499  | 6709   | 7208  |
| 65–69       | 60   | 27     | 87    | 28752| 15130  | 43882 | 802  | 13786  | 14588 |
| 70–74       | 3    | 14     | 17    | 24796| 12827  | 37623 | 802  | 13786  | 14588 |
| 75–79       | 8    | 87     | 95    | 24250| 12438  | 36688 | 802  | 13786  | 14588 |
| 80–84       | 15   | 286    | 301   | 16856| 10160  | 27016 | 869  | 14723  | 15592 |
| 85 and over | 16   | 390    | 406   | 11687| 10624  | 22311 | 1172 | 17388  | 18560 |
| Total       | 21,165| 11,222 | 32,387| 115,669| 64,530 | 180,199 | 5499 | 74,195 | 79,694 |
| Age cohorts | Orphan's Family responsibilities | Total |
|-------------|----------------------------------|-------|
|             | Male | Female | Total | Male | Female | Total | Male | Female | Total |
| 0–4         | 71   | 70     | 141   | –    | –      | –     | 71   | 70     | 141   |
| 5–9         | 274  | 257    | 531   | –    | –      | –     | 274  | 257    | 531   |
| 10–14       | 557  | 525    | 1082  | –    | –      | –     | 557  | 525    | 1082  |
| 15–19       | 1043 | 1009   | 2052  | 4    | 5      | 9     | 1048 | 1014   | 2062  |
| 20–24       | 599  | 620    | 1219  | 11   | 12     | 23    | 665  | 647    | 1312  |
| 25–29       | 63   | 44     | 107   | 8    | 10     | 18    | 298  | 155    | 453   |
| 30–34       | 126  | 86     | 212   | 4    | 5      | 9     | 712  | 433    | 1145  |
| 35–39       | 217  | 144    | 361   | 3    | 5      | 8     | 1331 | 949    | 2280  |
| 40–44       | 316  | 218    | 534   | 6    | 7      | 13    | 2178 | 1758   | 3936  |
| 45–49       | 392  | 273    | 665   | 13   | 23     | 36    | 3185 | 2880   | 6065  |
| 50–54       | 368  | 281    | 649   | 33   | 68     | 101   | 4328 | 4434   | 8762  |
| 55–59       | 299  | 251    | 550   | 46   | 107    | 153   | 5962 | 5936   | 11,898 |
| 60–64       | 220  | 228    | 448   | 44   | 131    | 175   | 16,095 | 11,512   | 27,607 |
| 65–69       | 145  | 177    | 322   | 26   | 113    | 139   | 29,482 | 22,156   | 51,638 |
| 70–74       | 82   | 121    | 203   | 14   | 96     | 110   | 25,456 | 22,225   | 47,681 |
| 75–79       | 53   | 104    | 157   | 16   | 128    | 144   | 25,129 | 26,543   | 51,672 |
| 80–84       | 22   | 56     | 78    | 15   | 138    | 153   | 17,777 | 25,363   | 43,140 |
| 85 and over | 6    | 32     | 38    | 19   | 168    | 187   | 12,900 | 28,602   | 41,502 |
| Total       | 4853 | 4496   | 9349  | 262  | 1016   | 1278  | 147,448 | 155,459  | 302,907 |

*Source* Authors’ own calculations
Table 10  Comparison of results of the goodness of fit test: $p$ value. *Source* Authors’ own calculations

| Type of pension                  | CSWL | SS   | SR subsample | SR sample |
|----------------------------------|------|------|--------------|-----------|
| Permanent disability male        | 0    | 1    | 1            | 1         |
| Permanent disability female      | 0    | 1    | 1            | 1         |
| Retirement male                  | 0    | 0.9999191 | 1         | 1         |
| Retirement female                | 0    | 0.9995441  | 1         | 1         |
| Widower’s                        | 0    | 1    | 1            | 1         |
| Widow’s                          | 0    | 1    | 1            | 1         |
| Orphan’s male                    | 0.4624223 | 1    | 1            | 1         |
| Orphan’s female                  | 0.5618020  | 1    | 1            | 1         |
| Family responsibilities male     | 0.4084868 | 0.9990003 | 0.9999445  | 1         |
| Family responsibilities female   | 0.8416536 | 0.9999963 | 1         | 1         |
| Total pensions                   | 0    | 1    | 1            | 1         |

Subsample (SS) with $p$ value $\geq 0.999$

of pensioners, with high $p$ values of the test. There are feasible solutions with sizes ranging from 93.074% of the CSWL, associated with a minimum $p$ value of 0.8, to 70.87% of the CSWL with a $p$ value of 0.9999.

It has been proved that the subsample obtained is included in the population and in the CSWL. Therefore, given the values of the goodness of fit test and the fact that the subsample is well over 1% larger than the subsample obtained using stratified sampling, it is sure to meet the objective of finding bigger subsamples that are more representative than the CSWL.

As an example, Table 9 shows the theoretical distribution of the subsample that provides a goodness of fit test with a $p$ value of $\geq 0.999$.

Tables 22, 23, 24, 25 and 26 in Appendix 1 show the differences in the estimated total expenditure on pensions between the population and a large size subsample obtained for a $p$ value greater or equal to 0.999 for each type of pension, gender, and age cohort. A major reduction in errors can be seen in the estimation of pension expenditure with a very small reduction in the size of the original CSWL.

Table 10 shows the $p$ values for the 10 cases considered (5 types of pension by 2 genders) for the goodness of fit test for pensions in the subsample obtained using this procedure (SS) compared to the pensions in the CSWL, along with the subsample design selected from the CSWL using stratification (SR subsample, Table 6) and the hypothetical sample extracted from the population using stratified random sampling (SR sample, Table 3). Obviously the values in the subsample obtained by the procedure designed are lower than those obtained with the SR sample and subsample, but the differences with the latter are almost non-existent. It is therefore possible to find subsamples contained in the CSWL and in the population that have a better fit and many more observations than would be provided by a stratified random sample taken from the CSWL. Overall, the distribution of total pensions is adjusted to the population using Pearson’s goodness of fit test using a $p$ value of 1, as in the SR sample and
Table 11 Improvement indicators in the estimate of total pension expenditure in % with respect to the CSWL: subsample (SS) with \( p \) value \( \geq 0.999 \), SR subsample (SRSS) and SR sample (SRS). *Source* Authors’ own calculations

| Indicators | Case | Permanent disability | Retirement | Widow(er)’s | Orphan’s | Family responsibilities |
|------------|------|----------------------|------------|-------------|----------|------------------------|
|            |      | M   | F   | M   | F   | M   | F   | M   | F   | M   | F   | M   | F   | M   | F   |
| R_SRMQE    | SS   | 99.82 | 99.69 | 98.17 | 99.07 | 98.54 | 96.14 | 87.27 | 88.20 | 58.34 | 90.89 |
|            | SRSS | 99.78 | 99.52 | 99.59 | 99.57 | 97.85 | 97.77 | 80.75 | 76.98 | 41.70 | 88.12 |
|            | SRS  | 99.95 | 99.84 | 99.92 | 99.83 | 99.41 | 99.53 | 96.11 | 93.75 | 85.58 | 97.44 |
| R_SDAV     | SS   | 99.50 | 99.24 | 97.41 | 98.46 | 98.54 | 96.79 | 88.98 | 89.58 | 53.93 | 86.77 |
|            | SRSS | 99.36 | 98.90 | 99.36 | 99.43 | 97.80 | 97.83 | 77.59 | 79.28 | 36.37 | 83.78 |
|            | SRS  | 99.87 | 99.57 | 99.88 | 99.77 | 99.35 | 99.53 | 95.62 | 94.31 | 82.80 | 96.28 |
| MDC        | SS   | 0.01 | 0.01 | 0.00 | 0.00 | 0.03 | 0.01 | 0.06 | 0.03 | 0.63 | 0.10 |
|            | SRSS | 0.01 | 0.01 | 0.00 | 0.00 | 0.03 | 0.01 | 0.05 | 0.05 | 0.65 | 0.13 |
|            | SRS  | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.01 | 0.01 | 0.15 | 0.03 |
| CE ≤ 0.01% | SS   | 100.0 | 100.0 | 88.89 | 100.0 | 57.14 | 100.0 | 61.11 | 55.56 | 5.56 | 5.88 |
|            | SRSS | 100.0 | 93.33 | 100.0 | 100.0 | 64.29 | 100.0 | 44.44 | 44.44 | 44.44 | 52.94 |
|            | SRS  | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 66.67 | 82.35 |
subsample. More important in the improvement in forecasts for pension expenditure are the differences found by type of pension and gender.

Table 11 shows the improvement indicators in the estimate of total pension expenditure using the subsample obtained with a \( p \) value greater than 0.999 as well as those that use the SR hypothetical sample and subsample with respect to the CSWL.

It can be seen that the subsample (SS) obtained by the procedure greatly reduces the errors in the estimation of total expenditure in each cohort of the CSWL considered and that in many cases it improves on the reductions given by the SR subsample (SRSS) contained in the CSWL, which was 1\% of the population. Obviously it does not reach the results obtained with the best possible sample of similar size to the CSWL, which would be obtained using an SR sampling from the population.

The estimation errors in pension expenditure by contingency, gender and age cohort of the SR subsample with a size of 87,472 benefits, and those of the subsample obtained with a \( p \) value greater than 0.999, which numbers 302,907 benefits, are much smaller than those obtained with the original CSWL, but somewhat lower than those from the sample obtained by the procedure proposed. However, this last sample is 3.462 times bigger, which is a great advantage for forecasting given the diversity in working lives for any subsequent analysis of the sustainability of public pension systems.

5 Summary, conclusions and future research

The CSWL is a set of anonymized microdata taken from Spanish Social Security records. The availability of this data has marked a turning point for studies on the Spanish public pension system since it has given researchers access to very valuable information about individuals that enables them to examine in depth numerous aspects of the pension system that had previously been ignored. However, the CSWL is obtained using a simple random sample (RS), so its fit to the population in terms of age, gender and type of pension is worse than would be obtained using stratified random sampling (SR) with proportional allocation. In this paper we examine how representative the CSWL data is of pension benefits. The results show that it does not fit the distribution of the population of pensioners by age and gender in two types of benefit: permanent disability and widower’s benefits for the case of men; there is also a mismatch to a lesser extent regarding retirement pensions. These mismatches are bigger for 2005, 2006, 2008, 2009, 2010 and 2011, and smaller in 2007, 2012 and 2013 due to the correction of a codifying error in permanent disability benefits for persons over 65. It is hard to understand why the error reappears after the correction in 2007.

We check the effects of this poor fit of the CSWL to the pensioner population in terms of pension benefits by estimating the annual total pension expenditure by cohorts obtained from the CSWL and the figure that would have resulted from a hypothetical SR. Given that researchers cannot access full data on the population, they are unable to obtain an SR sample and must resort to an SR subsample smaller than the CSWL and contained within it. The problem is that the reduction in size is considerable, so richness in pension types is lost.
For the reasons indicated, it can be said that it is advisable to use other procedures to select large size subsamples, so the representativeness of the original CSWL is improved with less reduction in the number of pensions. With the methodology developed, large subsamples can be obtained that pass the $\chi^2$ test giving $p$ values near 1, so it can be concluded not only that it is feasible to find a large dataset selected from the CSWL that better represents the population of pensioners but also that it is possible to draw up a procedure for extracting subsamples from the CSWL with a good fit to the population of interest by type of pension, gender and age.

Apart from the problems described, the CSWL is a powerful dataset whose size enables representative large subsamples to be extracted which emulate the whole population of pensioners. This is taken into account in formulating and solving optimization problems to obtain feasible solutions (contained in the population and in the CSWL) that more or less meet the goodness of fit requirements in terms of a $p$ value chosen by the researcher, with sizes ranging from 70 to 93% of the original.

According to our findings, it can be said that studies conducted using subsamples of data on pension benefits from the CSWL obtained using SR sampling with proportional allocation taking into account the known distribution of the population should not give rise to doubt as to how well the CSWL represents the population. However, those which use data on pension benefits selected from the CSWL without checking that they are representative may produce conclusions based on age cohorts which are overrepresented or underrepresented in the subsamples with respect to their real weight in the population.

Last but not least, we would like to emphasize that this research meets the conditions for reproducibility. Boyland (2016) establishes the guidelines for a paper to be recognized as “reproducible”: the most important are that the data used in the analysis should be accessible to other researchers and that the algorithms or methods of analysis should be specified in the manuscript in sufficient detail to allow the results to be reproduced. This is our case: any research team can replicate the procedure for selecting more representative large subsamples from the CSWL.

A further line of research directly related to the results of this paper would be to consider another variable in the design and selection procedure for large size subsamples plus the number of pensions, such as the amount of the pensions selected. The objective would be to find a dataset that fits the population as well as possible, taking into account not only the number of pensions but also the average pension amount per cohort of the population, the data for which is also known.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
6 Appendix 1. Estimate of total expenditure on pensions from expenditure at 31st December INSS and CSWL 2010

In Tables 12, 13, 14, 15, 16, 17, 18, 19, 20 and 21 the information provided is represented by the following terms:

SRMQE: Square root of the mean quadratic error in the differences in pension expenditure obtained by comparing the population of pensioners (INSS) and the sample (CSWL).

SDAV: Differences in absolute value of expenditure estimated by cohort obtained by comparing the population of pensioners (INSS) and the sample (CSWL).

(SDAV /INSS)%: Differences in absolute value of expenditure estimated by cohort as a percentage of total pension expenditure for the population of pensioners for the case of the CSWL.

SRMQE *: Square root of the mean quadratic error in the differences in pension expenditure obtained by comparing the population of pensioners (INSS) and the stratified random sample (SR).

SDAV *: Differences in absolute value of expenditure estimated by cohort obtained by comparing the population of pensioners (INSS) and the stratified random sample (SR).

(SDAV */INSS)%: Differences in absolute value of expenditure estimated by cohort as a percentage of total pension expenditure for the population of pensioners for the case of the SR.

In Tables 22, 23, 24, 25 and 26 the information provided is different because the three last indicators are replaced by the following:

SRMQE **: Square root of the mean quadratic error in the differences in pension expenditure obtained by comparing the population of pensioners (INSS) and the large subsample (Sub-m).

SDAV **: Differences in absolute value of expenditure estimated per cohort obtained by comparing the population of pensioners (INSS) and the large subsample (Sub-m).

(SDAV **/INSS)%: Differences in absolute value of expenditure estimated per cohort as a percentage of total pension expenditure for the population of pensioners for the case of the Sub-m.
Fig. 1 Comparing CSWL and SR. Total pension expenditure error (€) for permanent disability and retirement, 2010. Source Authors’ own calculations

Fig. 2 Comparing CSWL and SR. Total pension expenditure error (€) for widow(er)’s and orphan’s. 2010 Source Authors’ own calculations

Fig. 3 Comparing CSWL and SR. Total pension expenditure error (€) for family responsibilities, 2010. Source Authors’ own calculations
Table 12 Differences in pension expenditure for CSWL/SR (permanent disability, male, 2010). *Authors’ own calculations*

| Age Cohorts | Items | INSS | CSWL | Dif. INSS-CSWL | % total | % coh. | SR | Dif. INSS-SR | % total | % coh. |
|-------------|-------|------|------|----------------|---------|--------|----|--------------|---------|--------|
| 15–19       |       | 305,359 | 191,248 | 114,111 | 0.00 | 37.37 | 382,493 | −77,134 | 0.00 | −25.26 |
| 20–24       |       | 14,786,632 | 12,868,830 | 1,917,803 | 0.02 | 12.97 | 14,707,107 | 79,525 | 0.00 | 0.54 |
| 25–29       |       | 66,209,567 | 71,298,075 | −5,088,508 | −0.06 | −7.69 | 66,204,787 | 4780 | 0.00 | 0.01 |
| 30–34       |       | 176,911,083 | 167,459,431 | 9,451,652 | 0.12 | 5.34 | 176,880,869 | 30,213 | 0.00 | 0.02 |
| 35–39       |       | 338,371,808 | 345,500,577 | −7,128,768 | −0.09 | −2.11 | 338,362,630 | 9178 | 0.00 | 0.00 |
| 40–44       |       | 568,653,256 | 572,532,114 | −3,878,858 | −0.05 | −0.68 | 568,603,856 | 49,400 | 0.00 | 0.01 |
| 45–49       |       | 855,446,886 | 862,229,288 | −6,782,402 | −0.09 | −0.79 | 855,317,200 | 129,686 | 0.00 | 0.02 |
| 50–54       |       | 1,239,548,604 | 1,230,802,261 | 8,746,343 | 0.11 | 0.71 | 1,239,535,674 | 12,930 | 0.00 | 0.00 |
| 55–59       |       | 1,923,519,010 | 1,928,343,775 | −5,824,765 | −0.07 | −0.28 | 1,923,549,829 | −75,819 | 0.00 | 0.00 |
| 60–64       |       | 2,684,969,401 | 2,698,738,227 | −13,768,827 | −0.17 | −0.51 | 2,684,975,602 | −6,201 | 0.00 | 0.00 |
| 65–69       |       | 23,152,655 | 510,528,427 | −487,375,772 | −6.17 | −2.105.05 | 23,055,924 | 96,731 | 0.00 | 0.42 |
| 70–74       |       | 607,041 | 1,721,631 | −1,114,590 | −0.01 | −1.183.61 | 573,872 | 33,169 | 0.00 | 5.46 |
| 75–79       |       | 1,341,538 | 1,946,441 | −604,903 | −0.01 | −45.09 | 1,297,616 | 43,922 | 0.00 | 3.27 |
| 80–84       |       | 2,208,409 | 2,273,656 | −65,247 | 0.00 | −2.95 | 2,147,323 | 61,086 | 0.00 | 2.77 |
| 85 and over |       | 2,312,153 | 2,654,879 | −342,726 | 0.00 | −14.82 | 2,292,380 | 19,322 | 0.00 | 0.84 |
| Total       |       | 7,898,343,401 | 8,409,579,860 | −511,236,459 | −6.47 | | 7,897,932,613 | 410,788 | 0.01 |
| SRMQE       |       | 551,696,275 | 6.98 | | | | 60,552 | 729,096 | 0.01 |
| SDAV        | (SDAV/INSS)% | 125,979,362 | SRMQE * | SDAV * | (SDAV*/INSS)% |
Table 13: Differences in pension expenditure for CSWL/SR (permanent disability, female, 2010). Source: Authors’ own calculations.

| Age cohorts | Items | INSS | CSWL | Dif. INSS-CSWL | % total | % coh. | SR | Dif. INSS-SR | % total | % coh. |
|-------------|-------|------|------|---------------|---------|-------|----|---------------|---------|-------|
| 15–19       |       | 34,536 | 0 | 34,536 | 0.00 | 100.00 | 0 | 34,536 | 0.00 | 100.00 |
| 20–24       |       | 3,081,629 | 3,190,155 | −108,526 | 0.00 | −3.52 | 2,977,452 | 104,177 | 0.00 | 3.38 |
| 25–29       |       | 18,866,018 | 17,672,673 | 1,193,345 | 0.04 | 6.33 | 18,965,633 | −99,615 | 0.00 | −0.53 |
| 30–34       |       | 63,016,409 | 56,942,976 | 6,073,433 | 0.19 | 9.64 | 63,035,098 | −18,689 | 0.00 | −0.03 |
| 35–39       |       | 140,961,788 | 135,488,475 | 5,473,313 | 0.17 | 3.88 | 141,051,968 | −90,180 | 0.00 | −0.06 |
| 40–44       |       | 238,294,736 | 240,660,760 | −2,366,023 | −0.07 | −0.99 | 238,187,864 | 106,873 | 0.00 | 0.04 |
| 45–49       |       | 368,676,785 | 359,369,973 | 9,306,812 | 0.29 | 2.52 | 368,632,791 | 43,994 | 0.00 | 0.01 |
| 50–54       |       | 576,453,959 | 588,924,294 | −12,470,335 | −0.38 | −2.16 | 576,350,836 | 103,123 | 0.00 | 0.02 |
| 55–59       |       | 790,117,521 | 787,997,244 | 2,120,277 | 0.07 | 0.27 | 790,215,650 | −98,129 | 0.00 | −0.01 |
| 60–64       |       | 936,367,310 | 942,257,779 | −5,890,469 | −0.18 | −0.63 | 936,418,603 | −51,293 | 0.00 | −0.01 |
| 65–69       |       | 7,253,437 | 175,860,418 | −168,606,981 | −5.18 | −2324.51 | 7,205,487 | 47,950 | 0.00 | 0.66 |
| 70–74       |       | 2,113,208 | 2,955,423 | −842,215 | −0.03 | −39.85 | 2,055,928 | 57,279 | 0.00 | 2.71 |
| 75–79       |       | 12,951,479 | 14,034,344 | −1,082,865 | −0.03 | −8.36 | 13,004,188 | −52,709 | 0.00 | −0.41 |
| 80–84       |       | 42,361,600 | 44,154,717 | −1,793,118 | −0.06 | −4.23 | 42,357,359 | 4241 | 0.00 | 0.01 |
| 85 and over |       | 57,016,186 | 56,288,493 | 727,693 | 0.02 | 1.28 | 57,048,658 | −32,472 | 0.00 | −0.06 |
| Total       |       | 3,257,566,599 | 3,425,797,722 | −168,231,123 | −5.16 | 3,257,507,515 | 59,084 | 0.00 | SRMQE |
| SRMQE       |       | SDAV | (SDAV/INSS)% | SDAV* | (SDAV*/INSS)% | 43,809,541 | 218,089,940 | 6.69 | 71,243 | 945,259 | 0.03 |
Table 14  Differences in pension expenditure for CSWL/SR (retirement, male, 2010). *Source* Authors’ own calculations

| Age Cohorts | Items | INSS | CSWL | Dif. INSS-CSWL | % total | % coh. | SR | Dif. INSS-SR | % total | % coh. |
|-------------|-------|------|------|----------------|---------|--------|----|--------------|---------|--------|
| 45–49       |       | 1,884,599 | 4,721,346 | −2,836,747 | −0.01   | −150.52 | 2,023,417 | −138,817 | 0.00 | −7.37 |
| 50–54       |       | 24,498,988 | 22,758,852 | 1,740,136 | 0.00   | 7.10 | 24,226,957 | 272,032 | 0.00 | 1.11 |
| 55–59       |       | 306,141,087 | 259,593,995 | 46,547,092 | 0.10 | 15.20 | 306,252,563 | −111,476 | 0.00 | −0.04 |
| 60–64       |       | 4,914,459,889 | 4,928,735,927 | −14,276,038 | −0.03 | −0.29 | 4,914,324,211 | 135,678 | 0.00 | 0.00 |
| 65–69       |       | 13,406,207,704 | 12,946,954,241 | 459,253,464 | 0.97 | 3.43 | 13,406,065,692 | 142,012 | 0.00 | 0.00 |
| 70–74       |       | 10,077,723,226 | 10,227,331,080 | −149,607,854 | −0.31 | −1.48 | 10,077,708,704 | 14,522 | 0.00 | 0.00 |
| 75–79       |       | 9,162,490,841 | 9,174,926,772 | −12,435,931 | −0.03 | −0.14 | 9,162,389,244 | 101,597 | 0.00 | 0.00 |
| 80–84       |       | 5,929,164,483 | 5,896,272,031 | 32,892,451 | 0.07 | 0.55 | 5,929,186,735 | −22,252 | 0.00 | 0.00 |
| 85 and over |       | 3,706,045,451 | 3,611,096,868 | 94,948,583 | 0.20 | 2.56 | 3,705,999,753 | 45,698 | 0.00 | 0.00 |
| Total       |       | 47,528,616,269 | 47,072,391,111 | 456,225,157 | 0.96 | | 47,528,177,276 | 438,993 | 0.00 | |
| SRMQE       | SDAV  | SDAV (SDAV/INSS)% | SRMQE * | SDAV* (SDAV*/INSS)% |
| 165,304,287 | 814,538,296 | 1.71 | 132,242 | 984,083 | 0.00 |
| Age Cohorts | Items     | INSS | CSWL | Dif. INSS-CSWL | % total | % coh. | SR   | Dif. INSS-SR | % total | % coh. |
|-------------|-----------|------|------|----------------|---------|--------|------|--------------|---------|--------|
| 45–49       |           | 758,668 | 0    | 758,668        | 0.00    | 100.00 | 703,933 | 54,735       | 0.00    | 7.21   |
| 50–54       |           | 2,284,701 | 8,267,554 | −5,982,852 | −0.04   | −261.87 | 2,543,841 | −259,140 | 0.00    | −11.34 |
| 55–59       |           | 7,239,094 | 12,816,651 | −5,577,557 | −0.04   | −77.05  | 7,049,098 | 189,997 | 0.00    | 2.62   |
| 60–64       |           | 1,188,821,446 | 1,218,007,856 | −29,186,410 | −0.19   | −2.46   | 1,188,960,679 | −139,234 | 0.00    | −0.01  |
| 65–69       |           | 4,218,373,005 | 4,005,749,812 | 212,623,193 | 1.35    | 5.04    | 4,218,376,134 | −3,129 | 0.00    | 0.00   |
| 70–74       |           | 3,116,814,648 | 3,122,570,036 | −5,755,389 | −0.04   | −0.18   | 3,116,850,139 | −35,491 | 0.00    | 0.00   |
| 75–79       |           | 2,869,721,202 | 2,850,870,204 | 18,850,998 | 0.12    | 0.66    | 2,869,664,495 | 56,707 | 0.00    | 0.00   |
| 80–84       |           | 2,225,411,510 | 2,183,358,336 | 42,053,175 | 0.27    | 1.89    | 2,225,345,153 | 66,358 | 0.00    | 0.00   |
| 85 and over |           | 2,110,089,457 | 2,075,972,095 | 34,117,361 | 0.22    | 1.62    | 2,110,076,702 | 12,755 | 0.00    | 0.00   |
| Total       |           | 15,739,513,731 | 15,477,612,545 | 261,901,186 | 1.66    |         | 15,739,570,172 | −56,441 | 0.00    |        |
| SRMQE       |           | SDAV (SDAV/INSS)% | SDAV * (SDAV*/INSS)% | SRMQE * | SDAV* (SDAV*/INSS)% |
| 74,123,693  |           | 354,905,603 | 2.25 | 122,332 | 817,544 | 0.01 |

Table 15 Differences in pension expenditure for CSWL/SR (retirement, female, 2010). Source Authors’ own calculations.
Table 16 Differences in pension expenditure for CSWL/SR (widow(er)’s, male, 2010). Source Authors’ own calculations

| Age Cohorts | Widowers | CSWL | Dif. INSS-CSWL | % total | % coh. | SR | Dif. INSS-SR | % total | % coh. |
|-------------|----------|------|----------------|---------|--------|----|---------------|---------|--------|
| 20–24       | 66,012   | 1,417,508 | −1,351,496 | −0.14   | −2047.34 | 0  | 66,012      | 0.01    | 100.00 |
| 25–29       | 540,207  | 9,889,789  | −9,349,583  | −0.97   | −1730.74  | 520,511 | 19,696  | 0.00    | 3.65   |
| 30–34       | 2,755,085 | 18,143,751 | −15,388,667 | −1.60   | −558.56   | 2,805,711 | −50,626 | −0.01   | −1.84  |
| 35–39       | 9,656,863 | 14,254,243 | −4,597,380  | −0.48   | −47.61    | 9,700,721 | −43,858 | 0.00    | −0.45  |
| 40–44       | 21,697,996 | 26,788,117 | −5,090,121  | −0.53   | −23.46    | 21,704,054 | −6,058  | 0.00    | −0.03  |
| 45–49       | 43,628,560 | 47,968,375 | −4,339,816  | −0.45   | −9.95     | 43,589,935 | 38,625  | 0.00    | 0.09   |
| 50–54       | 74,875,391 | 67,320,300 | 7,555,092   | 0.78    | 10.09     | 74,929,777 | −54,385 | −0.01   | −0.07  |
| 55–59       | 97,628,683 | 85,766,306 | 11,862,376  | 1.23    | 12.15     | 97,595,302 | 33,381  | 0.00    | 0.03   |
| 60–64       | 106,978,121 | 99,954,535 | 7,023,586   | 0.73    | 6.57      | 107,024,968 | −46,484 | 0.00    | −0.04  |
| 65–69       | 96,878,531 | 89,871,131 | 7,007,400   | 0.73    | 7.23      | 96,938,814 | −60,283 | −0.01   | −0.06  |
| 70–74       | 96,308,469 | 98,743,786 | −2,435,317  | −0.25   | −2.53     | 96,360,001 | −51,532 | −0.01   | −0.05  |
| 75–79       | 126,986,943 | 125,258,118 | 1,728,825  | 0.18    | 1.36      | 127,042,504 | −55,561 | −0.01   | −0.04  |
| 80–84       | 128,756,951 | 130,783,551 | −2,026,600  | −0.21   | −1.57     | 128,726,907 | 30,043  | 0.00    | 0.02   |
| 85 and over | 156,871,416 | 145,223,841 | 11,647,575  | 1.21    | 7.42      | 156,831,114 | 40,302  | 0.00    | 0.03   |
| Total       | 963,629,228 | 961,383,351 | 2,245,877   | 0.23    | 1.24      | 963,770,319 | −141,091 | −0.01   |
| SRMQE       | SDAV      | (SDAV/INSS)% | SDAV*       | 91,403,834 | 9.49    | 45,459 | 597,210 | 0.06   |
| 7,731,448   |           |             |             |         |         |        |         |        |
Table 17  Differences in pension expenditure CSWL/SR (widow(er)'s, female, 2010).  Source Authors’ own calculations

| Age Cohorts | Items | INSS | CSWL | Dif. INSS-CSWL | % total | % coh. | SR | Dif. INSS-SR | % total | % coh. |
|-------------|-------|------|------|----------------|---------|--------|----|--------------|---------|--------|
| 15–19       |       | 49,299 | 617,531 | −568,231 | 0.00 | −1152.61 | 0 | 49,299 | 0.00 | 100.00 |
| 20–24       |       | 923,659 | 2,471,233 | −1,547,575 | −0.01 | −167.55 | 898,623 | 25,036 | 0.00 | −2.71 |
| 25–29       |       | 6,681,699 | 13,731,344 | −7,049,645 | −0.04 | −105.51 | 6,756,635 | −74,936 | 0.00 | −1.12 |
| 30–34       |       | 28,328,605 | 39,166,412 | −10,837,806 | −0.06 | −38.26 | 28,423,380 | −94,775 | 0.00 | −0.33 |
| 35–39       |       | 75,839,985 | 85,836,319 | −9,996,334 | −0.06 | −13.18 | 75,828,843 | 11,142 | 0.00 | 0.01 |
| 40–44       |       | 174,441,044 | 186,969,507 | −12,528,463 | −0.07 | −7.18 | 174,387,694 | 53,350 | 0.00 | 0.03 |
| 45–49       |       | 329,709,962 | 341,995,880 | −12,285,918 | −0.07 | −3.73 | 329,739,836 | −29,873 | 0.00 | −0.01 |
| 50–54       |       | 546,726,285 | 531,110,215 | 15,616,070 | 0.09 | 2.86 | 546,680,598 | 45,688 | 0.00 | 0.01 |
| 55–59       |       | 800,379,741 | 788,844,067 | 11,535,674 | 0.07 | 1.44 | 800,470,738 | −90,997 | 0.00 | −0.01 |
| 60–64       |       | 1,246,491,319 | 1,245,789,695 | 701,624 | 0.00 | 0.06 | 1,246,462,234 | 29,085 | 0.00 | 0.00 |
| 65–69       |       | 1,690,747,024 | 1,675,089,499 | 15,657,525 | 0.09 | 0.93 | 1,690,819,938 | −72,914 | 0.00 | 0.00 |
| 70–74       |       | 2,213,631,066 | 2,226,236,620 | −12,605,554 | −0.07 | −0.57 | 2,213,645,911 | −14,846 | 0.00 | 0.00 |
| 75–79       |       | 3,221,905,038 | 3,236,094,120 | −14,189,082 | −0.08 | −0.44 | 3,221,870,375 | 34,663 | 0.00 | 0.00 |
| 80–84       |       | 3,299,420,799 | 3,280,493,020 | 18,927,779 | 0.11 | 0.57 | 3,299,327,021 | 93,778 | 0.00 | 0.00 |
| 85 and over |       | 3,543,135,246 | 3,525,870,290 | 17,264,957 | 0.10 | 0.49 | 3,543,168,598 | −33,352 | 0.00 | 0.00 |
| Total       |       | 17,178,410,772 | 17,180,315,751 | −1,904,979 | −0.01 | −69.651 | 17,178,480,423 | 0.00 | 0.00 |
| SRMQE       |       | 12,156,918 | 161,312,238 | 0.94% | SDAV | (SDAV/INSS)% | SRMQE * | SDAV* | (SDAV*/INSS)% |
Table 18  Differences in pension expenditure for CSWL/SR (orphan’s, male, 2010). Source Authors’ own calculations

| Age Cohorts | Items | INSS | CSWL | Dif. INSS-CSWL | % total | % coh. | SR | Dif. INSS-SR | % total | % coh. |
|-------------|-------|------|------|----------------|---------|-------|----|--------------|---------|-------|
| 0–4         |       | 7,803,824 | 7,623,526 | 180,298 | 0.03 | 2.31 | 7,823,838 | −20,014 | 0.00 | −0.26 |
| 5–9         |       | 29,121,284 | 28,784,445 | 336,839 | 0.05 | 1.16 | 29,133,958 | −12,674 | 0.00 | −0.04 |
| 10–14       |       | 58,463,746 | 58,659,352 | −195,605 | −0.03 | −0.33 | 58,465,229 | −1,483 | 0.00 | 0.00 |
| 15–19       |       | 111,301,184 | 113,446,237 | −2,145,053 | −0.32 | −1.93 | 111,303,403 | −2,220 | 0.00 | 0.00 |
| 20–24       |       | 70,741,698 | 68,479,303 | 2,262,395 | 0.33 | 3.20 | 70,770,143 | −28,445 | 0.00 | −0.04 |
| 25–29       |       | 7,948,902 | 7,903,526 | 45,376 | 0.01 | 0.57 | 7,969,815 | −20,913 | 0.00 | −0.26 |
| 30–34       |       | 16,334,864 | 17,390,523 | −1,055,663 | −0.16 | −6.46 | 16,305,890 | 28,974 | 0.00 | 0.18 |
| 35–39       |       | 30,858,497 | 30,124,102 | 734,395 | 0.11 | 2.38 | 30,893,190 | −34,693 | −0.01 | −0.11 |
| 40–44       |       | 47,979,793 | 47,336,676 | 643,117 | 0.10 | 1.34 | 47,990,256 | −10,463 | 0.00 | −0.02 |
| 45–49       |       | 64,821,503 | 64,226,465 | 595,037 | 0.09 | 0.92 | 64,784,434 | 37,068 | 0.01 | 0.06 |
| 50–54       |       | 66,142,871 | 65,181,509 | 961,362 | 0.14 | 1.45 | 66,211,998 | −69,127 | −0.01 | −0.10 |
| 55–59       |       | 56,916,812 | 55,779,767 | 1,137,045 | 0.17 | 2.00 | 56,863,868 | 52,944 | 0.01 | 0.09 |
| 60–64       |       | 43,038,549 | 41,952,331 | 1,086,219 | 0.16 | 2.52 | 42,999,474 | 39,075 | 0.01 | 0.09 |
| 65–69       |       | 29,513,972 | 28,466,807 | 1,047,165 | 0.15 | 3.55 | 29,477,148 | 36,824 | 0.01 | 0.12 |
| 70–74       |       | 17,199,909 | 16,961,833 | 238,076 | 0.04 | 1.38 | 17,228,457 | −28,547 | 0.00 | −0.17 |
| 75–79       |       | 11,423,767 | 11,182,027 | 241,739 | 0.04 | 2.12 | 11,358,169 | 65,597 | 0.01 | 0.57 |
| 80–84       |       | 4,884,678 | 4,674,329 | 210,349 | 0.03 | 4.31 | 4,948,133 | −63,455 | −0.01 | −1.30 |
| 85 and over |       | 1,731,827 | 1,854,025 | −122,199 | −0.02 | −7.06 | 1,705,118 | 26,708 | 0.00 | 1.54 |
| Total       |       | 676,227,678 | 670,026,787 | 6,200,892 | 0.92 | 9.46 | 676,232,522 | −48,43 | 0.00 | 0.00 |
| SRMQE       |       | SDAV | (SDAV/INSS)% | SRMQE * | SDAV* | (SDAV*/INSS)% |
| Total       |       | 971,688 | 13,237,932 | 1.96 | 37,759 | 579,224 | 0.09 |
Table 19 Differences in pension expenditure for CSWL/SR (orphan’s, female, 2010). Source Authors’ own calculations

| Age Cohorts | Items | INSS | CSWL | Dif. INSS-CSWL | % total | % coh. | SR | Dif. INSS-SR | % total | % coh. |
|-------------|-------|------|------|----------------|---------|-------|----|---------------|---------|-------|
| 0–4         |       | 7,753,229 | 7,334,824 | 418,405 | 0.07 | 5.40 | 7,769,368 | −16,138 | 0.00 | −0.21 |
| 5–9         |       | 27,528,214 | 26,739,431 | 788,783 | 0.12 | 2.87 | 27,526,198 | 2016 | 0.00 | 0.01 |
| 10–14       |       | 55,514,468 | 56,168,221 | −653,752 | −0.10 | −1.18 | 55,497,927 | 16,541 | 0.00 | 0.03 |
| 15–19       |       | 107,417,380 | 107,058,933 | 358,447 | 0.06 | 0.33 | 107,397,178 | 20,202 | 0.00 | 0.02 |
| 20–24       |       | 72,979,097 | 73,563,722 | −584,625 | −0.09 | −0.80 | 72,931,079 | 48,017 | 0.01 | 0.07 |
| 25–29       |       | 5,630,385 | 4,933,746 | 696,639 | 0.11 | 12.37 | 5,624,459 | 5926 | 0.00 | 0.11 |
| 30–34       |       | 11,195,443 | 12,089,322 | −893,879 | −0.14 | −7.98 | 11,209,779 | −14,335 | 0.00 | −0.13 |
| 35–39       |       | 20,350,313 | 21,146,602 | −796,289 | −0.12 | −3.91 | 20,299,734 | 50,579 | 0.01 | 0.25 |
| 40–44       |       | 34,027,031 | 33,979,856 | 47,175 | 0.01 | 0.14 | 33,984,576 | 42,455 | 0.01 | 0.12 |
| 45–49       |       | 45,644,604 | 46,321,862 | −677,258 | −0.11 | −1.48 | 45,646,870 | −2,266 | 0.00 | 0.00 |
| 50–54       |       | 51,255,329 | 49,970,189 | 1,285,140 | 0.20 | 2.51 | 51,273,396 | −18,067 | 0.00 | −0.04 |
| 55–59       |       | 48,391,341 | 47,695,568 | 695,773 | 0.11 | 1.44 | 48,465,328 | −73,987 | −0.01 | −0.15 |
| 60–64       |       | 45,588,571 | 45,274,016 | 314,555 | 0.05 | 0.69 | 45,531,116 | 57,455 | 0.01 | 0.13 |
| 65–69       |       | 36,227,961 | 35,317,779 | 910,183 | 0.14 | 2.51 | 36,275,057 | −47,096 | −0.01 | −0.13 |
| 70–74       |       | 25,384,450 | 24,171,404 | 1,213,046 | 0.19 | 4.78 | 25,430,022 | −45,572 | −0.01 | −0.18 |
| 75–79       |       | 22,551,568 | 22,258,188 | 293,380 | 0.05 | 1.30 | 22,515,950 | 35,617 | 0.01 | 0.16 |
| 80–84       |       | 12,465,936 | 12,797,465 | −331,529 | −0.05 | −2.66 | 12,545,105 | −79,169 | −0.01 | −0.64 |
| 85 and over |       | 7,287,001 | 6,637,386 | 649,615 | 0.10 | 8.91 | 7,372,008 | −85,008 | −0.01 | −1.17 |
| Total       |       | 637,192,322 | 633,458,514 | 3,733,808 | 0.59 | | 637,295,152 | −102,831 | −0.02 | |
| SRMQE       | SDAV  | (SDAV/INSS)% | (SDA V/INSS)% | SDAV* | SDAV* | (SDAV*/INSS)% |
| 715,958     |       | 11,608,474 | 1.82 | 44,732 | 660,448 | 0.10 |
Table 20 Differences in pension expenditure for CSWL/SR (family responsibilities, male, 2010). Source Authors’ own calculations

| Age Cohorts | Items | INSS | CSWL | Dif. INSS-CSWL | % total | % coh. | SR | Dif. INSS-SR | % total | % coh. |
|-------------|-------|------|------|----------------|---------|--------|----|--------------|---------|--------|
| 0–4         |       | 2774 | 0    | 2774           | 0.01    | 100.00 | 0  | 2774         | 0.01    | 100.00 |
| 5–9         |       | 34,908 | 0 | 34,908       | 0.07    | 100.00 | 0  | 34,908       | 0.07    | 100.00 |
| 10–14       |       | 184,744 | 0 | 184,744      | 0.39    | 100.00 | 185,129 | −385 | 0.00    | −0.21 |
| 15–19       |       | 575,030 | 606,059 | −31,029 | −0.07 | 5.40 | 626,334 | −51,304 | −0.11 | −8.92 |
| 20–24       |       | 1,368,435 | 1,220,430 | 148,006 | 0.31 | 10.82 | 1,408,853 | −40,418 | −0.09 | −2.95 |
| 25–29       |       | 917,850 | 1,034,412 | −116,563 | −0.25 | 12.70 | 873,190 | 44,660 | 0.09 | 4.87 |
| 30–34       |       | 495,251 | 458,742 | 36,510 | 0.08 | 7.37 | 531,731 | −36,480 | −0.08 | −7.37 |
| 35–39       |       | 376,278 | 441,309 | −65,031 | −0.14 | 17.28 | 413,444 | −37,166 | −0.08 | −9.88 |
| 40–44       |       | 634,402 | 796,349 | −161,947 | −0.34 | 25.53 | 665,678 | −31,276 | −0.07 | −4.93 |
| 45–49       |       | 2,489,716 | 1,924,184 | 565,532 | 1.20 | 22.71 | 2,520,099 | −30,383 | −0.06 | −1.22 |
| 50–54       |       | 6,337,035 | 6,539,086 | −202,052 | −0.43 | 3.19 | 6,277,538 | 59,497 | 0.13 | 0.94 |
| 55–59       |       | 8,709,729 | 8,361,978 | 347,751 | 0.74 | 3.99 | 8,747,297 | −37,568 | −0.08 | −0.43 |
| 60–64       |       | 8,304,579 | 7,998,722 | 305,857 | 0.65 | 3.68 | 8,244,272 | 60,307 | 0.13 | 0.73 |
| 65–69       |       | 5,075,358 | 5,145,549 | −70,191 | −0.15 | 1.38 | 5,005,829 | 69,529 | 0.15 | 1.37 |
| 70–74       |       | 2,595,976 | 2,873,847 | −277,871 | −0.59 | 10.70 | 2,644,943 | −48,967 | −0.10 | −1.89 |
| 75–79       |       | 2,931,373 | 3,038,375 | −107,003 | −0.23 | 3.65 | 2,975,054 | −43,681 | −0.09 | −1.49 |
| 80–84       |       | 2,773,354 | 3,270,777 | −497,423 | −1.06 | 17.94 | 2,785,316 | −11,962 | −0.03 | −0.43 |
| 85 and over |       | 3,228,353 | 3,886,016 | −657,663 | −1.40 | 20.37 | 3,213,648 | 14,705 | 0.03 | 0.46 |
| Total       |       | 47,035,146 | 47,595,835 | −560,689 | −1.19 | 83,211 | 47,118,357 | 59,497 | 0.13 | 0.94 |
| SRMQE       | SDAV  | (SDAV/INSS)% | SRMQE * SDAV* | (SDAV*/INSS)% |
| 284,045     | 3,812,854 | 8.11 | 40,952 | 655,969 | 1.39 |
Table 21 Differences in pension expenditure for CSWL/SR (family responsibilities, female, 2010). Source Authors’ own calculations

| Age Cohorts | Women |   |   |   |   |   |   |   |   |   |
|-------------|-------|---|---|---|---|---|---|---|---|---|
|             | INSS  | CSWL| Dif. INSS-CSWL | % total | % coh. | SR | Dif. INSS-SR | % total | % coh. |
| 5–9         | 35,740| 0  | 35,740          | 0.02    | 100.00 | 68,874 | −33,134 | −0.02 | −92.71 |
| 10–14       | 138,191| 0  | 138,191         | 0.07    | 100.00 | 96,166 | 42,025  | 0.02  | 30.41  |
| 15–19       | 616,309| 518,756| −97,553      | 0.05    | 15.83  | 597,670 | 18,639  | 0.01  | 3.02   |
| 20–24       | 1,295,902| 1,341,184| −45,281     | −0.02   | −3.49  | 1,283,926| 11,976  | 0.01  | 0.92   |
| 25–29       | 1,110,832| 1,127,638| −16,806    | −0.01   | −1.51  | 1,080,723| 30,109  | 0.02  | 2.71   |
| 30–34       | 574,950| 657,905| −82,955     | −0.04   | −14.43 | 593,092 | −18,142 | −0.01 | −3.16  |
| 35–39       | 545,949| 582,647| −36,699    | −0.02   | −6.72  | 573,293 | −27,344 | −0.01 | −5.01  |
| 40–44       | 690,937| 593,233| 97,704      | 0.05    | 14.14  | 668,962 | 21,975  | 0.01  | 3.18   |
| 45–49       | 4,309,005| 3,877,752| 431,253    | 0.22    | 10.01  | 4,324,381| −15,376 | −0.01 | −0.36  |
| 50–54       | 13,519,574| 12,523,472| 996,102    | 0.52    | 7.37   | 13,547,715| −28,141 | −0.01 | −0.21  |
| 55–59       | 21,753,640| 23,720,665| −1,967,025 | −1.02   | −9.04  | 21,749,807| 3832   | 0.00  | 0.02   |
| 60–64       | 26,154,496| 25,463,463| 691,034    | 0.36    | 2.64   | 26,174,497| −20,000 | −0.01 | −0.08  |
| 65–69       | 23,230,331| 23,912,383| −682,052   | −0.35   | −2.94  | 23,285,795| −55,464 | −0.03 | −0.24  |
| 70–74       | 19,558,689| 19,234,458| 324,230    | 0.17    | 1.66   | 19,599,400| −40,711 | −0.02 | −0.21  |
| 75–79       | 25,022,010| 25,084,791| −62,781   | −0.03   | −0.25  | 25,026,745| −4,735  | 0.00  | −0.02  |
| 80–84       | 25,558,449| 23,464,670| 2,093,779  | 1.09    | 8.19   | 25,566,907| −8,458  | 0.00  | −0.03  |
| 85 and over | 28,329,850| 31,220,883| −2,891,033 | −1.50   | −10.20 | 28,347,904| −18,054 | −0.01 | −0.06  |
| Total       | 192,444,854| 193,323,899| −879,045  | −0.46   | −141,001 | 192,585,855| −141,001 | −0.07 | −0.03  |
| SRMQE       | SDAV   | (SDAV/INSS)% | SRMQE * | SDAV*  | (SDAV*/INSS)% |
| 1,054,156    | 10,690,218 | 5.55 | 27,027 | 398,118 | 0.21 |

| Source | Authors’ own calculations |
Table 22  Differences in pension expenditure INSS/subsample $p \geq 0.999$ (permanent disability, 2010). Source Authors’ own calculations

| Age Cohorts | Male | | | Female | | | |
|-------------|------|---|---|--------|---|---|---|
| | Subsample | Dif. INSS-Sub-m | % total | % coh. | Subsample | Dif. INSS-Sub-m | % total | % coh. |
| 15–19 | 220,457 | 84,902 | 0 | 27.8 | 34,536 | 0 | 100 |
| 20–24 | 14,569,347 | 217,286 | 0 | 1.47 | 2,941,901 | 139,728 | 0 | 4.53 |
| 25–29 | 66,043,284 | 166,283 | 0 | 0.25 | 18,632,710 | 233,308 | 0.01 | 1.24 |
| 30–34 | 176,896,993 | 14,090 | 0 | 0.02 | 62,938,469 | 77,940 | 0 | 0.12 |
| 35–39 | 338,163,805 | 208,004 | 0 | 0.06 | 140,841,964 | 119,823 | 0 | 0.09 |
| 40–44 | 568,552,433 | 100,823 | 0 | 0.02 | 238,110,734 | 184,003 | 0.01 | 0.08 |
| 45–49 | 855,622,316 | −175,430 | 0 | −0.02 | 368,643,986 | 32,799 | 0 | 0.01 |
| 50–54 | 1,239,780,393 | −231,789 | 0 | −0.02 | 576,565,387 | −111,428 | 0 | −0.02 |
| 55–59 | 1,923,966,300 | −447,290 | −0.01 | −0.02 | 790,353,566 | −236,045 | −0.01 | −0.03 |
| 60–64 | 2,685,366,896 | −397,496 | −0.01 | −0.01 | 936,455,134 | −87,824 | 0 | −0.01 |
| 65–69 | 22,780,636 | 372,019 | 0 | 1.61 | 7,008,211 | 245,225 | 0.01 | 3.38 |
| 70–74 | 496,143 | 110,898 | 0 | 18.27 | 2,073,701 | 39,507 | 0 | 1.87 |
| 75–79 | 1,196,648 | 144,890 | 0 | 10.8 | 12,912,527 | 38,952 | 0 | 0.3 |
| 80–84 | 2,184,087 | 24,323 | 0 | 1.1 | 42,316,630 | 44,970 | 0 | 0.11 |
| 85 and over | 2,225,709 | 86,444 | 0 | 3.74 | 56,993,802 | 22,384 | 0 | 0.04 |
| Total | 7,898,065,445 | 277,956 | 0 | | 3,256,788,722 | 777,877 | 0.02 |
| SRMQE*** | SDAV** | (SDAV**/INSS)% | SRMQE*** | SDAV** | (SDAV**/INSS)% |
| 224,800 | 2,781,965 | 0.04 | 134,684 | 1,648,472 | 0.05 |
Table 23 Differences in pension expenditure INSS/subsample \( p \geq 0.999 \) (retirement, 2010). *Source* Authors’ own calculations

| Age cohorts | Male | | | | Female | | | |
|-------------|------|---|---|---|------|---|---|---|
|              | Subsample | Dif. INSS-Sub-m | % total | % coh. | Subsample | Dif. INSS-Sub-m | % total | % coh. |
| 45–49        | 1,554,978 | 329,622 | 0 | 17.49 | 0 | 758,668 | 0 | 100 |
| 50–54        | 23,695,892 | 803,097 | 0 | 3.28 | 2,199,284 | 85,417 | 0 | 3.74 |
| 55–59        | 299,240,941 | 6,900,146 | 0.01 | 2.25 | 6,648,344 | 590,751 | 0 | 8.16 |
| 60–64        | 4,915,672,241 | −1,212,352 | 0 | −0.02 | 1,188,956,870 | −135,425 | 0 | −0.01 |
| 65–69        | 13,409,902,981 | −3,695,277 | −0.01 | −0.03 | 4,219,562,943 | −1,189,938 | −0.01 | −0.03 |
| 70–74        | 10,080,621,078 | −2,897,852 | −0.01 | −0.03 | 3,117,725,815 | −911,167 | −0.01 | −0.03 |
| 75–79        | 9,164,967,450 | −2,476,609 | −0.01 | −0.03 | 2,870,412,299 | −691,727 | 0 | −0.02 |
| 80–84        | 5,930,874,362 | −1,709,880 | 0 | −0.03 | 2,226,068,247 | −656,737 | 0 | −0.03 |
| 85 and over  | 3,707,109,575 | −1,064,124 | 0 | −0.03 | 2,110,537,554 | −448,097 | 0 | −0.02 |
| Total        | 47,533,639,497 | −5,023,228 | −0.01 | | 15,742,111,986 | −2,598,255 | −0.02 | |
| SRMQE**      | 3,019,891 | 21,088,957 | 0.04 | | 691,774 | 5,467,927 | 0.03 | |
Table 24  Differences in pension expenditure INSS/subsample $p \geq 0.999$ (widow(er)’s, 2010).  *Source* Authors’ own calculations

| Age | Male | Female |
|-----|------|--------|
|     | Subsample | Dif. INSS-Sub-m | % total | % coh. | Subsample | Dif. INSS-Sub-m | % total | % coh. |
| 15–19 | 0 | 66,012 | 0.01 | 0 | 0 | 49,299 | 0 | 0 |
| 20–24 | 400,008 | 140,199 | 0.02 | 25.95 | 6,532,392 | 149,308 | 0 | 2.23 |
| 25–29 | 958,494 | 17,688 | 0.02 | 6.09 | 28,120,860 | 207,745 | 0 | 0.73 |
| 30–34 | 2,587,396 | 140,199 | 0.02 | 0.28 | 174,270,014 | 171,030 | 0 | 0.1 |
| 35–39 | 2,587,396 | 140,199 | 0.02 | 0.28 | 329,658,177 | 51,785 | 0 | 0.01 |
| 40–44 | 74,677,780 | 197,611 | 0.02 | 0.13 | 800,491,893 | 127,226 | 0 | 0.01 |
| 45–49 | 97,501,458 | 129,500 | 0.02 | 0.12 | 1,246,768,860 | 277,541 | 0 | 0.02 |
| 50–54 | 96,806,800 | 71,731 | 0.02 | 0.07 | 1,691,193,916 | 446,892 | 0 | 0.03 |
| 55–59 | 96,313,125 | −4,657 | 0 | 0 | 2,214,993,639 | −6,318,265 | 0 | 0.02 |
| 60–64 | 106,848,621 | 129,500 | 0.02 | 0.12 | 1,246,768,860 | −277,541 | 0 | 0.02 |
| 65–69 | 96,806,800 | 71,731 | 0.02 | 0.07 | 1,691,193,916 | −446,892 | 0 | 0.03 |
| 70–74 | 96,313,125 | −4,657 | 0 | 0 | 2,214,993,639 | −6,318,265 | 0 | 0.02 |
| 75–79 | 126,973,117 | 13,826 | 0 | 0.01 | 3,222,808,815 | −903,777 | 0 | 0.03 |
| 80–84 | 128,691,983 | 64,968 | 0.02 | 0.05 | 3,300,249,066 | −828,267 | 0 | 0.03 |
| 85 and over | 156,831,968 | 39,449 | 0 | 0.03 | 3,544,191,488 | −1,056,242 | 0 | 0.03 |
| Total | 962,300,877 | 1,328,350 | 0.14 | 17,181,597,024 | 123 | 1,056,242 | 0 | 0.03 |

| SRMQE** | SDAV** | (SDAV**/INSS)% | SRMQE** | SDAV** | (SDAV**/INSS)% |
|---------|--------|---------------|---------|--------|---------------|
| 112,601 | 1,337,664 | 0.14 | 469,112 | 5,179,630 | 0.03 |
Table 25  Differences in pension expenditure INSS/subsample $p \geq 0.999$ (orphan’s, 2010).  Source Authors’ own calculations

| Age Cohorts | Male | | | | | | Female | | | |
|-------------|------|---|---|---|---|---|---|---|---|---|
|              | Subsample | Dif. INSS-Sub-m | % total | % coh. | Subsample | Dif. INSS-Sub-m | % total | % coh. |
| 0–4          | 7,714,898 | 88,926 | 0.01 | 1.14 | 7,739,783 | 13,446 | 0 | 0.17 |
| 05–09        | 29,120,114 | 1170 | 0 | 0 | 27,456,995 | 71,219 | 0.01 | 0.26 |
| 10–14        | 58,471,997 | -8,251 | 0 | -0.01 | 55,515,057 | -589 | 0 | 0 |
| 15–19        | 111,239,038 | 62,145 | 0.01 | 0.06 | 107,407,454 | 9926 | 0 | 0.01 |
| 20–24        | 70,717,834 | 23,864 | 0 | 0.03 | 72,900,145 | 78,952 | 0.01 | 0.11 |
| 25–29        | 7,928,591 | 20,311 | 0 | 0.26 | 5,593,625 | 36,760 | 0.01 | 0.65 |
| 30–34        | 16,333,421 | 1443 | 0 | 0.01 | 11,112,854 | 82,590 | 0.01 | 0.74 |
| 35–39        | 30,787,815 | 70,682 | 0.01 | 0.23 | 20,299,010 | 51,304 | 0.01 | 0.25 |
| 40–44        | 47,893,589 | 86,203 | 0.01 | 0.18 | 33,889,715 | 137,316 | 0.02 | 0.4 |
| 45–49        | 64,766,198 | 55,305 | 0.01 | 0.09 | 45,602,977 | 41,627 | 0.01 | 0.09 |
| 50–54        | 66,088,468 | 54,403 | 0.01 | 0.08 | 51,260,577 | -5,248 | 0 | -0.01 |
| 55–59        | 56,809,189 | 107,622 | 0.02 | 0.19 | 48,354,510 | 36,832 | 0.01 | 0.08 |
| 60–64        | 42,932,107 | 106,442 | 0.02 | 0.25 | 45,500,639 | 87,932 | 0.01 | 0.19 |
| 65–69        | 29,327,449 | 186,523 | 0.03 | 0.63 | 36,104,191 | 123,771 | 0.02 | 0.34 |
| 70–74        | 17,142,206 | 57,703 | 0.01 | 0.34 | 25,335,774 | 48,676 | 0.01 | 0.19 |
| 75–79        | 11,375,866 | 47,900 | 0.01 | 0.42 | 22,494,300 | 57,268 | 0.01 | 0.25 |
| 80–84        | 4,826,372 | 58,306 | 0.01 | 1.19 | 12,270,111 | 195,825 | 0.03 | 1.57 |
| 85 and over  | 1,310,368 | 421,458 | 0.06 | 24.34 | 7,156,197 | 130,803 | 0.02 | 1.8 |
| Total        | 674,785,521 | 1,442,157 | 0.21 | 1,198,408 | 635,993,913 | 1.19 |
| SRMQE**      | 123,743 | 1,458,658 | 0.22 | 1,210,083 | 84,498 | 0.19 |
Table 26  Differences in pension expenditure INSS/subsample $p \geq 0.999$ (family responsibilities, 2010).  *Source* Authors’ own calculations

| Age Cohorts | Male | | | Female | |
|-------------|------|------|------|--------|------|
|              | Subsample | Dif. INSS-Sub-m | % total | % coh. | Subsample | Dif. INSS-Sub-m | % total | % coh. |
| 0–4         | 0     | 2774   | 0.01 | 100    | 0     | 35,740   | 0.02 | 100    |
| 5–9         | 0     | 34,908 | 0.07 | 100    | 0     | 138,191 | 0.07 | 100    |
| 10–14       | 0     | 184,744| 0.39 | 100    | 0     | 72,675   | 0.04 | 100    |
| 15–19       | 481,332| 93,698 | 0.20  | 16.29  | 574,130| 42,179   | 0.02 | 6.84   |
| 20–24       | 1,190,960| 177,475| 0.38  | 12.97  | 1,268,596| 27,306   | 0.01 | 2.11   |
| 25–29       | 894,718| 23,131 | 0.05  | 2.52   | 1,038,157| 72,675   | 0.04 | 6.54   |
| 30–34       | 408,631| 86,621 | 0.18  | 17.49  | 488,342| 86,608   | 0.05 | 15.06  |
| 35–39       | 285,955| 90,323 | 0.19  | 24.37  | 472,040| 73,909   | 0.04 | 13.54  |
| 40–44       | 575,513| 58,888 | 0.13  | 9.28   | 674,745| 16,193   | 0.01 | 2.34   |
| 45–49       | 2,360,321| 129,395| 0.28  | 9.78   | 4,246,373| 62,632   | 0.03 | 1.45   |
| 50–54       | 6,284,197| 52,838 | 0.11  | 0.83   | 13,442,436| 77,138   | 0.04 | 0.57   |
| 55–59       | 8,589,503| 120,226| 0.26  | 1.38   | 21,634,537| 119,103  | 0.06 | 0.55   |
| 60–64       | 8,199,076| 105,503| 0.22  | 1.27   | 26,003,760| 150,736  | 0.08 | 0.58   |
| 65–69       | 5,001,016| 74,342 | 0.16  | 1.46   | 23,154,122| 76,209   | 0.04 | 0.33   |
| 70–74       | 2,510,878| 85,098 | 0.18  | 3.28   | 19,365,383| 193,306  | 0.1  | 0.99   |
| 75–79       | 2,887,961| 43,412 | 0.09  | 1.48   | 24,950,700| 71,310   | 0.04 | 0.28   |
| 80–84       | 2,675,614| 97,741 | 0.21  | 3.52   | 25,419,523| 138,926  | 0.07 | 0.54   |
| 85 and over | 2,932,721| 295,631| 0.63  | 9.16   | 28,298,187| 31,663   | 0.02 | 0.11   |
| Total       | 45,278,397| 1,756,749| 3.73  | 191,031,029| 1,413,825| 0.73   |

**SRMQE** = SDAY**/(SDAY**/INSS)**

**SDAY** = SDAY**/(SDAY**/INSS)**
References

Agliari E, Barra A, Contucci P, Sandell R, Vernia C (2014) A stochastic approach for quantifying immigrant integration: the Spanish test case. New J Phys 16:103034. http://stacks.iop.org/1367-2630/16/i=10/a=103034

Antón Pérez J, Braña Pino J, Muñoz de Bustillo Llorente R (2007) Edad efectiva de jubilación en España: un análisis a partir de la explotación de la Muestra Continua de Vidas Laborales de la Seguridad Social. Jornadas de Usuarios de la Muestra Continua de Vidas Laborales. Madrid, 4 & 5 October 2007. Ministerio de Trabajo y Asuntos Sociales and FEDEA

Argimón I, González CI (2006) La Muestra Continua de Vidas Laborales de la Seguridad Social. Bol Econ Banco de Esp (May) 40–53. http://www.bde.es/f/webbde/SES/Secciones/Publicaciones/InformesBoletinesRevistas/BoletinEconomico/06/May/Fich/art3.pdf

Argimón I, González C, Vegas R (2007) Jubilación entre los 60 y los 65 años. Algunas características. Presup Gasto Público 47:161–184

Arranz JM, García-Serrano C, Hernanz V (2013) How do we pursue "labormetrics"? An application using the CSWL. Estad Esp, 181:231–254. http://www.ine.es/ss/Satellite?L=0&c=INERevEstad_C&p=1254735226759&pagename=ProductosYServicios%2FPYSLayout&_charset_=UTF-8&cid=1259943175448&submit=1r

Arranz JM, García-Serrano C (2011) Are the CSWL tax data useful? Ideas for mining. Hacienda Pública Exp/Rev Econ Política 199(4):151–186

Arranz JM, García-Serrano C (2014) The interplay of the unemployment compensation system, fixed-term contracts and rehiring: the case of Spain. Int J Manpow 35(8):1236–1259

Barra A, Contucci P, Sandell R, Vernia C (2014) An analysis of a large dataset on immigrant integration in Spain. The statistical mechanics perspective on social action. Sci Rep, 4:4174. www.nature.com/scientificreports

Bazaraa MS, Sherali HD, Shetty CM (2006) Nonlinear programming: theory and applications, 3rd edn. Wiley-Interscience, Hoboken

Benavides F, Durán X, Martínez J, Jódar P, Boix P, Amable M (2010) Incidencia de incapacidad permanente en una cohorte de trabajadores afiliados a la seguridad social, 2004–2007. Gac Sanit 24(5):385–390

Berkson J (1938) Some difficulties of interpretation encountered in the application of the chi-square test. J Am Stat Assoc 33(203):526–536

Berkeson J (1938) Some difficulties of interpretation encountered in the application of the chi-square test. J Am Stat Assoc 33(203):526–536

Bertsekas DP (1999) Nonlinear programming, 2nd edn. Athena Scientific, Boston

Boado-Penas MC, Valdés-Prieto S, Vidal-Meliá C (2008) The actuarial balance sheet for pay-as-you-go finance: solvency indicators for Spain and Sweden. Fisc Stud 29:89–134

Bonhomme S, Hospido L (2013) Earnings inequality in Spain: new evidence using tax data. Appl Econ 45(30):4212–4225

Bonhomme S, Hospido L (2016) The cycle of earnings inequality: evidence from Spanish Social Security Data. Econ J. doi: 10.1111/ecoj.12368

Bowley AL (1926) Measurement of precision attained in sampling. Bull Int Stat Inst 22(1):6–62

Boyland JE (2016) Reproducibility. IMA J Manag Math 27(2):107–108. doi: 10.1093/imaman/dpw003

Cairó Blanco I (2010) An empirical analysis of retirement behaviour in Spain: partial versus full retirement. SERIEs—J Span Econ Assoc 1(3):325–356

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Erlbaum, Hillsdale

Conde Ruiz JJ, González CI (2013) Reforma de pensiones 2011 en España. Hacienda Pública/ Rev Public Econ 204(1):9–44

Devesa JE, Devesa M, Domínguez I, Encinas B, Meneu R, Nagore A (2012) Equidad y sostenibilidad como objetivos ante la reforma del sistema contributivo de pensiones de jubilación. Hacienda Pública Esp/Rev Econ Pública 201(9):38–44

Domínguez AA (2012) Labor transitions of Spanish workers: a flexicurity approach. Rev Int Organ/Int J Organ 9:121–143. http://www.raco.cat/index.php/RIO/article/view/281050/368714

Domínguez Fabián I, Devesa Carpio M, Rosado Cebrián B, (2012) La Muestra Continua de Vidas Laborales y su potencial para analizar la solvencia del sistema de pensiones desde la perspectiva del empleo. Madrid: Ministerio de Empleo y Seguridad Social. FIPROS. [Consulted 8-3-2013]. http://www.seg-social.es/prdi00/groups/public/documents/binario/174191.pdf

Durán A (2007) La muestra continua de vidas laborales de la seguridad social. Rev Minist Trab Asun Soc 1:231–240
Durán A, Sevilla M (2006) Una muestra continua de vidas laborales. In: El papel de los registros administrativos en el análisis social y económico y el desarrollo del sistema estadístico. Carmen Marcos García (directora) Colección: Estudios de Hacienda Pública. Instituto de Estudios Fiscales, Madrid. pp. 241–252

García Segovia F, Durán A (2008) Nuevos avances en la información laboral: la Muestra Continua de Vidas Laborales. Economistas 116:228–231

Grafström A, Schelin L (2014) How to select representative samples. Scand J Stat 41:277–290

Griva I, Nash SG, Sofer A (2009) Linear and nonlinear optimization, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia

Hansen MH, Hurwitz WN, Madow W (1953) Sample surveys: methods and theory. Wiley, New York

Himmelreicher RK, Stegmann M (2008) New Possibilities for socio-economic research through longitudinal data from the research data centre of the German Federal Pension Insurance (FDZ-RV). Schmollers Jahrb 128(4):647–660

Instituto Nacional de la Seguridad Social (INSS) (2006-2007) Informes Estadísticos 2005–2006. Madrid: INSS. Secretaría de Estado de la Seguridad Social. Ministerio de Trabajo y Asuntos Sociales

INSS (2008–11) Informes Estadísticos 2007-2010. Madrid: INSS. Secretaría de Estado de la Seguridad Social. Ministerio de Trabajo e Inmigración

INSS (2012–2014) Informes Estadístico 2011-2013. Madrid: INSS. Secretaría de Estado de la Seguridad Social. Ministerio de Empleo y Seguridad Social

Izquierdo M, Lacuesta A, Vegas R (2009) Assimilation of immigrants in Spain: a longitudinal analysis. Labour Econ 16(6):669–678

Kruskall W, Mosteller F (1979a) Representative Sampling I. Int Stat Rev 47(1):13–24

Kruskall W, Mosteller F (1979b) Representative sampling, II: scientific literature, excluding statistics. Int Stat Rev 47(2):111–127

Kruskall W, Mosteller F (1979c) Representative sampling, III: the current statistical literature. Int Stat Rev 47(3):245–265

Kruskall W, Mosteller F (1980) Representative sampling, IV: the history of the concept in statistics, 1895–1939. Int Stat Rev 48(2):169–195

Lapuerta I (2010) Claves para el trabajo con la Muestra Continua de Vidas Laborales. DemoSoc working paper (2010-37)

Lin M, Lucas HC, Shmieli G (2013) Research commentary: too big to fail: large samples and the p-value problem. Inf Syst Res 24(4):906–917

López Roldán P (2011) La Muestra Continua de Vidas Laborales: posibilidades y limitaciones. Aplicación al estudio de la ocupación de la población inmigrante. Metodol Encuestas 13:7–32

Luenberger DG (2003) Linear and nonlinear programming, 2nd edn. Kluwer Academic Publishers, Boston

Meneu Gaya R, Encinas Goenechea B (2012) Valoración de la reforma del sistema de pensiones español de 2011 desde la óptica de la viabilidad financiero-actuarial. Un análisis a través de la CSWL. Madrid: Ministerio de Empleo y Seguridad Social. FIPROS. [Consulted 14-5-2014]. http://www.seg-social.es/prdi00/groups/public/documents/binario/174193.pdf

Meneu Gaya R, Pérez-Salamero González JM, Ventura Marco M (1998) Fundamentos de optimización matemática en Economía. Repro-Exprés, S. L. http://roderic.uv.es/handle/10550/25951

MESS (2016) La Muestra Continua de Vidas Laborales. Guía del contenido. Estadísticas, Presupuestos y Estudios. Estadísticas, http://www.seg-social.es/prdi00/groups/public/documents/binario/190489.pdf. Accessed 22-May-2015

Ministerio de Trabajo y Asuntos Sociales (MTAS) (2006) La Muestra Continua de Vidas Laborales. Colección Informes y Estudios. Serie Seguridad Social, vol 26. Ministerio de Trabajo y Asuntos Sociales, Madrid

Moral Arce I, Patxot C, Souto G (2008) La sostenibilidad del sistema de pensiones. Una aproximación a partir de la CSWL. Rev Econ Apl XVI(1–2):29–66

Muñoz de Bustillo R, De Pedraza P, Antón JI, Rivas LA (2011) Working life and retirement pensions in Spain: the simulated impact of a parametric reform. Int Soc Secur Rev 64(1):73–93

Nagore García A, van Soest A (2016) New job matches and their stability before and during the crisis. CentER discussion paper; vol. 2016-033. Tilburg University: Econometrics

Olsen A, Hudson R (2009) Social security administration’s master earnings file: background information. Soc Secur Bull 69(3):29–45

Omair A (2014) Sample size estimation and sampling techniques for selecting a representative sample. J Health Spec 2(4):142–147
Patxot C, Souto G, Villanueva J (2009) Fostering the contributory nature of the Spanish retirement pension system: an arithmetic micro-simulation exercise using the CSWL. Presup Gasto Público 57:7–32

Peinado Martínez P (2011) Pension system’s reform in Spain: a dynamic analysis of the effects on welfare. Serrano Pérez, F. (director). Doctoral thesis. Universidad del País Vasco. [Consulted 4-3-2014]. https://addi.ehu.es/bitstream/10810/8113/8/peinado.pdf

Pérez-Salamero González JM, Régulez-Castillo M, Vidal-Meliá C (2016) Análisis de la representatividad de la MCVL: el caso de las prestaciones del sistema público de pensiones. Hacienda Pública Esp/Rev Public Econ 217(2/2016): 67–130

Ramsey CA, Hewitt AD (2005) A methodology for assessing sample representativeness. Environ Forensics 6:71–75

Ruszczynski AP (2006) Nonlinear optimization. Princeton University Press, Princeton

Singh D, Chaudhary FS (1986) Theory and analysis of sample survey designs. Wiley Eastern Limited, Hoboken

Smith CM (1989) The social security administration’s continuous work history sample. Soc Secur Bulletin 52:10

Solé M, Díaz Serrano L, Rodríguez M (2013) Disparities in work, risk and health between immigrants and native-born Spaniards. Soc Sci Med 76:179–187

Toharia Cortés L, Moreno G, Muñoz C (2007) La mejora del sistema de información estadística procedente de los registros de la seguridad social. Ministerio de Trabajo y Asuntos Sociales, Madrid

Treviño R, Vidal E, Devolder D (2008) Factores e indicadores de vulnerabilidad en la conciliación del empleo y la familia. Ministerio de Trabajo e Inmigración, Madrid

Tryfos P (1996) Sampling methods for applied research: text and cases. Wiley, Hoboken

Vall Castello J (2012) Promoting employment of disabled women in Spain: evaluating a policy. Labour Econ 19:82–91

Vegas Sánchez R, Argimón I, Botella M, González C (2013) Old age pensions and retirement in Spain. SERIEs J Span Econ Assoc 2013(4):273–307

Vicente Merino A, Calderón Milán M, Martínez Aguado T (2012) Muchos pierden y pocos ganan: efectos de la reforma legislativa sobre el poder adquisitivo del trabajador tras la jubilación. An Inst Actuar Esp 18:77–110

Vidal Meliá C, Boado Penas MC, Settergren O (2009) Automatic Balance Mechanisms in pay-as-you-go pension systems. The Geneva Pap Risk and Insur Issues Pract 34:287–317. doi:10.1057/gpp.2009.2

Wang C (1993) Sense and nonsense of statistical inference: controversy, misuse, and subtlety. Marcel Dekker, New York

Wilkinson L, APA Task Force on Statistical Inference (1999) Statistical methods in psychology journals: guidelines and explanations. Am Psychol 54:594–604