Instituto Complutense de Análisis Económico

Improving the representativeness of a simple random sample: an optimization model and its application to the Continuous Sample of Working Lives

Vicente Nuñez-Antón
Department of Applied Economics III (Econometrics and Statistics),
Faculty of Economics and Business,
University of the Basque Country UPV/EHU, Bilbao. (Spain)

Juan Manuel Pérez-Salamero González
Department of Financial Economics and Actuarial Science, Faculty of Economics, University of Valencia, Valencia. (Spain).

Marta Regúlez-Castillo
Department of Applied Economics III (Econometrics and Statistics), Faculty of Economics and Business, University of the Basque Country UPV/EHU, Bilbao. (Spain)

Carlos Vidal-Meliá
Department of Financial Economics and Actuarial Science, Faculty of Economics, University of Valencia, Valencia (Spain) and research affiliation with the Instituto Complutense de Análisis Económico (ICAE), Complutense University of Madrid (Spain) and the Centre of Excellence in Population Ageing Research (CEPAR), UNSW (Australia)

Abstract

This paper develops an optimization model for selecting a large subsample that improves the representativeness of a simple random sample previously obtained from a population larger than the population of interest. The problem formulation involves convex mixed-integer nonlinear programming (convex MINLP) and is therefore NP-hard. However, the solution is found by maximizing the "constant of proportionality" – in other words, maximizing the size of the subsample taken from a stratified random sample with proportional allocation – and restricting it to a p-value high enough to achieve a good fit to the population of interest using Pearson's chi-square goodness-of-fit test. The beauty of the model is that it gives the user the freedom to choose between a larger subsample with a poorer fit and a smaller subsample with a better fit. The paper also applies the model to a real case: The Continuous Sample of Working Lives (CSWL), which is a set of anonymized microdata containing information on individuals from Spanish Social Security records. Several waves (2005-2017) are first examined without using the model and the conclusion is that they are not representative of the target population, which in this case is people receiving a pension income. The model is then applied and the results prove that it is possible to obtain a large dataset from the CSWL that (far) better represents the pensioner population for each of the waves analysed.

Keywords optimization, subsampling, chi-square test, *p*-value, Continuous Sample of Working Lives.

JEL Classification C61; C81; C12; H55, J26.

Working Paper nº 1920
May, 2019

UNIVERSIDAD COMPLUTENSE MADRID

ISSN: **2341-2356**

Improving the representativeness of a simple random sample: an optimization model and its application to the Continuous Sample of Working Lives[*].

Vicente Nuñez-Antón[†], Juan Manuel Pérez-Salamero González[‡], Marta Regúlez-Castillo[§] and Carlos Vidal-Meliá[**]

Abstract

This paper develops an optimization model for selecting a large subsample that improves the representativeness of a simple random sample previously obtained from a population larger than the population of interest. The problem formulation involves convex mixed-integer nonlinear programming (convex MINLP) and is therefore NP-hard. However, the solution is found by maximizing the "constant of proportionality" – in other words, maximizing the size of the subsample taken from a stratified random sample with proportional allocation – and restricting it to a p-value high enough to achieve a good fit to the population of interest using Pearson's chi-square goodness-of-fit test. The beauty of the model is that it gives the user the freedom to choose between a larger subsample with a poorer fit and a smaller subsample with a better fit. The paper also applies the model to a real case: The Continuous Sample of Working Lives (CSWL), which is a set of anonymized microdata containing information on individuals from Spanish Social Security records. Several waves (2005-2017) are first examined without using the model and the conclusion is that they are not representative of the target population, which in this case is people receiving a pension income. The model is then applied and the results prove that it is possible to obtain a large dataset from the CSWL that (far) better represents the pensioner population for each of the waves analysed.

Keywords: optimization, subsampling, chi-square test, *p*-value, Continuous Sample of Working Lives.

JEL: C61; C81; C12; H55, J26.

---

[*] We gratefully acknowledge the financial support from the Ministerio de Economía y Competitividad (Spain) and the Basque Government for projects ECO2015-65826-P and IT 793-13 respectively, and Ministerio de Economía y Competitividad, Agencia Estatal de Investigación (AEI), Fondo Europeo de Desarrollo Regional (FEDER), the Department of Education of the Basque Government (UPV/EHU Econometrics Research Group), and Universidad del País Vasco UPV/EHU under research grants MTM2016-74931-P (AEI/FEDER, UE), IT-642-13 and UFI11/03. We would also like to thank Peter Hall for his English support. We are grateful for the comments received at the Seventh International Conference MAF 2016 and the 1st Workshop on Pensions and Insurance, plus all those made by the various anonymous reviewers of previous versions of this article. Any errors are entirely our own.

[†] Department of Applied Economics III (Econometrics and Statistics), Faculty of Economics and Business, University of the Basque Country UPV/EHU, Bilbao. (Spain) (e-mail: vicente.nunezanton@ehu.eus).

[‡] Department of Financial Economics and Actuarial Science, Faculty of Economics, University of Valencia, Valencia. (Spain). (e-mail: juan.perez-salamero@uv.es).

[§] Department of Applied Economics III (Econometrics and Statistics), Faculty of Economics and Business, University of the Basque Country UPV/EHU, Bilbao. (Spain) (e-mail: marta.regulez@ehu.eus)

[**] (Corresponding author). Department of Financial Economics and Actuarial Science, Faculty of Economics, University of Valencia, Valencia (Spain) and research affiliation with the Instituto Complutense de Análisis Económico (ICAE), Complutense University of Madrid (Spain) and the Centre of Excellence in Population Ageing Research (CEPAR), UNSW (Australia). (e-mail: carlos.vidal@uv.es).

1. Introduction.

In practice, the success of any statistical analysis usually depends on asking the right questions or defining the right problem to be analysed, and this includes accurately defining the population that is going to be used as a source of information. The researcher needs to carefully and completely define this population before collecting the sample and give a description of the members to be included. If the sample that the researcher has to work on is drawn by simple random sampling from a population larger than the target population (i.e. a subset of the previous set), it might not be representative of the target population as far as the variables of interest are concerned. This situation can be improved by using a sample obtained by stratified random sampling. If it is possible to obtain a subsample that is more representative of the population of interest than the simple random sample from which it is to be extracted, then all efforts should be directed towards obtaining such a subsample with the aim of achieving results of the highest possible quality.

However, a smaller sample might not be strong enough to identify any relevant characteristics that may be present in the population of interest, so it is desirable to have a large subsample, although this still needs to be of a manageable size. It is therefore vital for the sample selected to be both representative enough for analysis and representative of the target population as regards the main characteristics. A number of papers deal with the problem of selecting representative samples, including Ramsey & Hewitt (2005), Grafström & Schelin (2014), Kruskall & Mosteller (1979a, 1979b, 1979c and 1980) and Omair (2014).

The aim of this paper is to develop an optimization model for selecting a large subsample that improves the representativeness of a simple random sample previously obtained from a population larger than the population of interest. The researcher in this process does not have to have all the data on the population of interest, but must be able to classify the population into homogeneous groups or strata as they would if they were using a stratified random design. Simple random sampling can be vulnerable to sampling error because the randomness of the selection may result in a sample that does not reflect the makeup of the population. A subsample designed using systematic and stratified techniques will fit the population of interest better than the original simple random sample. This method is more efficient than simple random sampling because it ensures the adequate representation of elements across strata as far as the variables of interest are concerned.

The problem formulation involves convex mixed-integer nonlinear programming (convex MINLP) and is therefore NP-hard (Bonami *et al*., 2012 and D'Ambrossio & Lodi, 2013). However, the optimization model we propose finds a global solution by solving a nonlinear programming problem with just one decision variable that is a real positive number. Thus the subsample is selected by maximizing the so-called "constant of proportionality" – i.e. maximizing the size of the subsample taken from a stratified random sample with proportional allocation – and restricting it to a p-value high enough to achieve a good fit to the population of interest using Pearson's chi-square goodness-of-fit test. Doing this also ensures that the subsample will be contained within the initial simple random sample as well as in the population of interest.

In this paper we present an enumeration algorithm for finding the optimal global solution to the problem. By means of a simulation, we analyse the performance of the subsample selection method and the efficiency of the algorithm in solving the problem depending on whether the original simple random sample had a "bad" fit or a "fine" fit

to the population. As we will see later, when the procedure is applied to many cases, it usually finds the optimal global solution within a reasonable time. This resolution time depends on the size of the original simple random sample and how good a fit it is to the target population.

Finally, we show the usefulness of the mathematical optimization model and how its resolution procedure works by applying it to a real case using the Continuous Sample of Working Lives (CSWL, Muestra Continua de Vida Laboral)[1]. The CSWL comprises matched anonymized social security, income tax and census records for a simple random sample containing 4% of Spanish contributors, pensioners and unemployment benefit recipients. Starting from 2004, an edition of the CSWL dataset has been released every year. The application process for obtaining the CSWL data is simple, and approved users are allowed to work with the data on their own computers (MESS, 2018). Pérez-Salamero *et al.* (2017) examined several waves (2005–2013) and concluded that the CSWL was not representative as regards population with a pension income. The real example we develop in this paper extends the analysis carried out by those authors by examining the most recent waves (2014-2017). Applying our model provides us with larger datasets that are much more representative of the retired population[2] in terms of pension type, gender and age.

The structure of the paper is as follows. Section 2 develops a convex MINLP model designed to select subsamples and presents the mathematical formulation used to solve the problem. Section 3 presents the algorithm for finding the global solution and verifies its effectiveness by means of a simulation. Section 4 shows the real-life application of our methodology to the CSWL. For the sake of clarity, after we give a brief description of the CSWL we divide this section into two subsections. The first presents the optimization model tailored to suit the CSWL, while the second analyses the results and the main implications. The paper ends with our conclusions, possible directions for future research and three appendices. In Appendix 1 we prove the convexity of the chi-square statistic function. Appendix 2 contains CSWL pension data for 2005 to 2017 supplied by Spain's National Social Security Institute (INSS), plus the pension distribution for the optimal subsamples obtained. Finally, with the aim of showing the growing interest in the CSWL and the relevance of the journals that have published such research, Appendix 3 lists a selection of papers that have used this dataset.

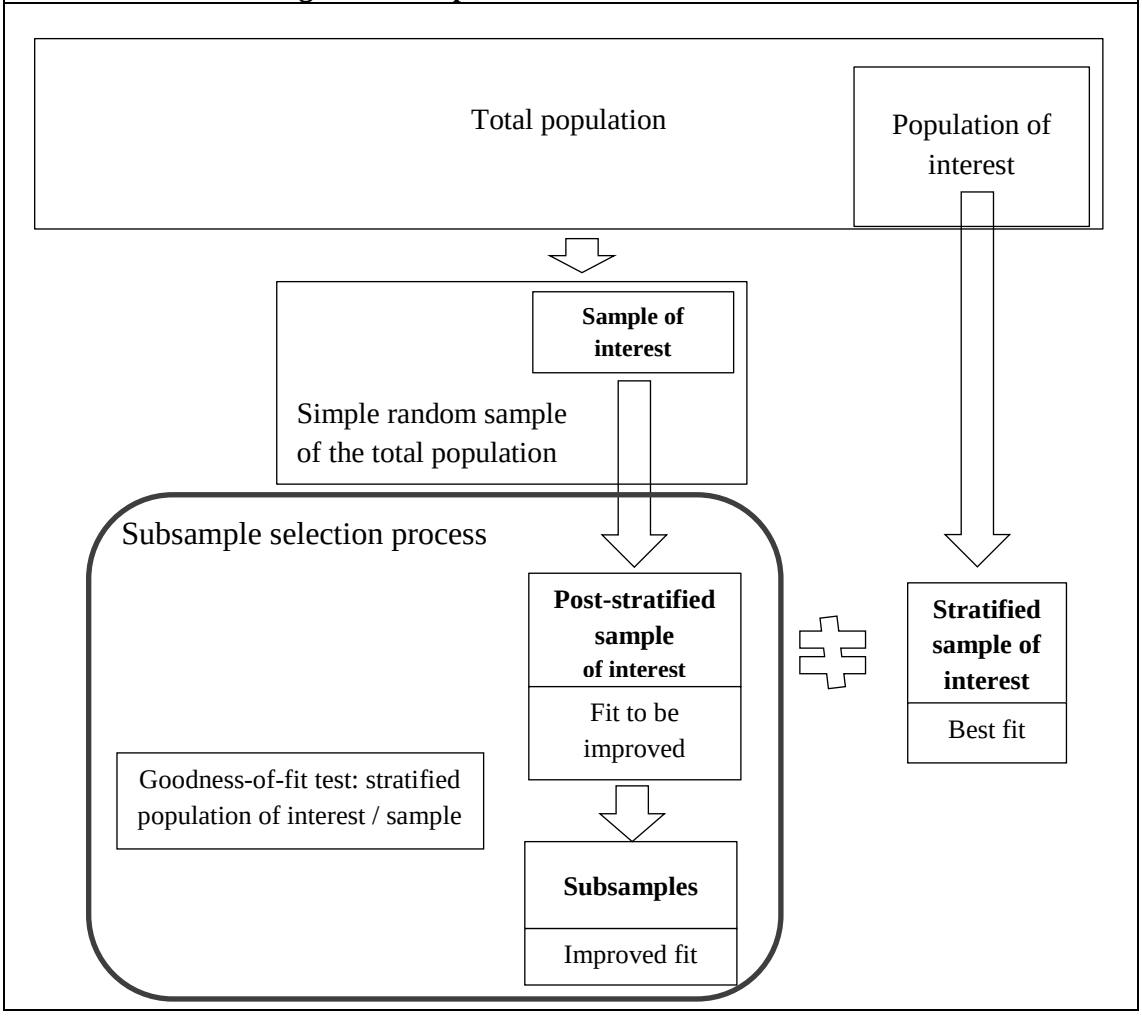
2. The optimization model for improving the representativeness of a simple random sample.

To solve the problem, we need to obtain from the initial simple random sample, which was drawn from a population larger than the population of interest, a large subsample that is more representative of the target population according to a statistical goodness-of-fit criterion, after carrying out a poststratification process (see Figure 1).

[1] Also referred to by several authors as the Continuous Working Life Sample (CWLS) or the Continuous Survey of Working Lives (CSWL), the dataset comprises extensive information on each individual's working life, including their spells of employment, periods of receiving social security benefits, occupational category and date of death where applicable.

[2] The analysis could be replicated for other groups, e.g. contributors, those receiving unemployment benefit, immigrants, native population, etc. However, this is clearly beyond the scope of this paper.

Figure 1.- Graphical illustration of the Problem



The selection process has to take into account the following requirements:

1. The subsample must be as large as possible (R_1 in Approach 1). We do not want to lose valuable information during the selection process, so need to keep as many records as possible.

2. The subsample must form part of both the target population and the original simple random sample (R_2 in Approach 1). This sounds an obvious requirement, but constraints need to be included so as to avoid outliers.

3. The elements to be included in each stratum of the subsample must take the form of a natural number (i.e. a non-negative integer) (R_3 in Approach 1).

4. The fit or representativeness with regard to the population under study must be improved (R_4 in Approach 1). The optimization model should therefore include a goodness-of-fit test for the distribution by strata. It should also make it a requirement that the value of the statistic be smaller than the critical value given a predetermined significance level so as to avoid rejecting the null hypothesis that the subsample has the same distribution as the population of interest. We use Pearson's goodness-of-fit test with the statistic:

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

where $\mathbf{O_i}$ is the observed values (those chosen from the simple random sample to build the subsample), $\mathbf{E_i}$ is the expected or theoretical values (those obtained from the distribution of the population of interest) and **k** is the number of strata for the variable of interest.

Approach 1 shows the mathematical formulation for our subsample selection procedure aimed at improving the representativeness of a simple random sample and fulfilling the requirements listed above, for the case of univariate stratification[3].

The first requirement (R_1 in Approach 1) for the subsample selection procedure involves the objective function **[1]**, i.e. maximize the size of the subsample. Constraints **[4]** and **[5]** take into account the second requirement (R_2 in Approach 1). The requirement to improve the goodness of fit (R_4 in Approach 1) is incorporated with constraints **[2]** and **[3]**, using the chi-square goodness-of fit test once the significance level (α) has been chosen. Finally, constraints **[4]**, **[5]** and **[6]** incorporate the requirement that each stratum of the subsample must be a non-negative integer (R_3 in Approach 1).

Approach 1.- Mathematical formulation for a selection criterion aimed at improving the representativeness of a simple random sample. Maximizing the size of the subsample.

$$\overbrace{\underset{n_i^{SUB}}{\text{Max}}\left\{n^{SUB} = \sum_{i=1}^{k} n_i^{SUB}\right\}}^{R_1} \quad \textbf{[1]}$$

subject to constraints:

$$\overbrace{\chi^2(n_1^{SUB}, \cdots, n_k^{SUB}) = \sum_{i=1}^{k} \frac{(n_i^{SUB} - n_i^{exp})^2}{n_i^{exp}} \leq \chi^2_{(\alpha,\ r)}}^{R_4} \quad \textbf{[2]}$$

$$\overbrace{n_i^{exp} = \frac{N_i}{N} \cdot n^{SUB} = \frac{N_i}{N} \cdot \sum_{i=1}^{k} n_i^{SUB}}^{R_4} \quad \textbf{[3]}$$

$$\overbrace{0 \leq n_i^{SUB} \leq N_i}^{R_2, R_3} \quad \textbf{[4]}$$

$$\overbrace{0 \leq n_i^{SUB} \leq n_i^{SRS}}^{R_2, R_3} \quad \textbf{[5]}$$

$$\overbrace{n_i^{SUB} \in Z}^{R_3} \quad \textbf{[6]}$$

$$\forall i = 1, 2, \cdots, k$$

where:

[3] For multivariate stratification, the mathematical approach could be adapted using as many summation terms and sub-indices as the stratifying variables to be considered.

n^{SUB}: the size of the subsample.

n_i^{SUB}: the size of the i[th] stratum in the subsample (observed values).

k: the number of strata for the variable of interest.

$\chi^2(n_1^{SUB}, \cdots, n_k^{SUB})$: the chi-square goodness-of-fit test statistic.

n_i^{exp}: the expected size of the i[th] stratum in the subsample. This depends on the population relative frequency $\frac{N_i}{N}$ and the size of the subsample n^{SUB}.

$\chi^2_{(\alpha,\ r)}$: the tabulated value of the chi-square distribution with r degrees of freedom and a significance level of α.

N_i: the size of the i[th] stratum in the target population.

N: the size of the target population.

$r = k - 1$: the degrees of freedom equal to the number of strata minus 1, given that in this case there are no parameters to be estimated because the population distribution is known.

n_i^{SRS}: the size of the i[th] stratum in the simple random sample.

Z: the set of integer numbers.

The objective function **[1]** is concave and linear, so it is continuous in a compact set, closed (the constraints are not strict) and bounded (all the decision variables are bounded below and above, **[4]** and **[5]**). Therefore, if the set is not empty, i.e. if we establish a goodness-of-fit improvement criterion that can be satisfied using the data from the original sample, then a solution to the optimization problem will exist, given that the function is bounded (Weierstrass Theorem). We just need to apply a method of finding it. Given constraints **[2]** and **[6]**, this is a nonlinear integer programming problem. In addition, considering the objective function and the constraints, it is a convex programming problem with decision variables bounded both above and below. The constraints, with the exception of **[2],** are linear inequalities, so they define convex sets. Without considering integer constraint **[6]**, the function that calculates the value of the test statistic that leads to constraint **[2]** is a convex function in R^k constrained with **[4]** and **[5]**, because it can be shown that the associated Hessian matrix is positive semidefinite (see Appendix 1).

The goodness-of-fit improvement constraint **[2]** can be rewritten equivalently as **[2a]** using a function that calculates the p-value using Pearson´s chi-square test statistic and the number of strata. This results in the same convex set given by **[2]** for a fixed $\overline{pValue_{min}}$ equal to α. The value of α must be chosen in such a way that the problem is feasible, that a subsample that is a solution to the problem actually exists in the original simple random sample. The p-value is the probability that the test statistic will take on a value at least as extreme as the observed value, assuming that the null hypothesis is true, i.e. that the subsample has the same distribution as the population of interest. If the p-value is less than α, say 0.05, the null hypothesis is rejected. If it is greater than α, then the null hypothesis is not rejected.

$$pValue[\chi^2(n_1^{SUB}, \cdots, n_k^{SUB}); r] \geq \overline{pValue_{min}} \qquad \textbf{[2a]}$$

where

$$pValue[\chi^2(n_1^{SUB},\cdots,n_k^{SUB});r] =$$

$$= 1 - \int_0^{\chi^2(n_1^{SUB},\cdots,n_k^{SUB})} \frac{\left(\frac{1}{2}\right)^{\frac{r}{2}}}{\text{Gamma}\left[\frac{r}{2}\right]} x^{\left(\frac{r}{2}-1\right)} e^{-\frac{x}{2}}\, dx$$

Therefore, excluding the integer requirement for the variables, the opportunity set of the optimization problem is convex and thus can be framed in a wider class: convex mixed-integer nonlinear programming (*convex MINLP*).

The optimization model

Applying mathematical programming to sampling has been done before (Cochran (1977); Särndal *et al*. (1992); Valliant & Gentle (1997); Baillargeon & Rivest (2009); Díaz-García & Ramos-Quiroga (2012 and 2014); Gupta *et al*. (2012 and 2014); Valliant *et al*. (2013); de Moura Brito *et al*. (2015)). However, the aim of our optimization model is not to solve problems of optimum sample allocation in surveys like Neyman allocation (Neyman, 1934), cost-constrained optimal and precision-constrained optimal allocations seek to do. In our case the population parameters are known, unlike in those cases in which they have to be estimated (e.g. population size, strata mean, etc.).

The optimization model for solving the problem set out in Approach 1 is based on stratified random sampling and uses proportional allocation to find a large subsample from within the original simple random sample while improving representativeness, in line with other authors such as Kontopantelis (2013). Proportional allocation in stratified sampling dates back to Bowley (1926) and, given its simplicity, is very common in practice. It uses a sampling fraction in each stratum that is proportional to that of the total population. When no other information except stratum size (N_i) is available, allocating a given sample of size n in the different strata is proportional to their sizes. This implies that the sampling fractions are all equal and the same as the global sampling fraction, its value being the constant of proportionality $q = n/N$.

The problem of maximizing the size of the subsample will be solved by maximizing the constant of proportionality that depends on the number of elements in each stratum, after rewriting the constraints appropriately in terms of the new and only decision variable (q):

$$n_i^{SUB}(q) = q \cdot N_i$$

The mathematical formulation of the optimization model in Approach 2 is the result of replacing the vector of decision variables, n_i^{SUB}, by q in the functions of the mathematical formulation of the problem in Approach 1. Maximizing q is equivalent to maximizing the size of the subsample, even though what we are actually maximizing is $\hat{q}$, the adjusted constant of proportionality, given that we have to consider the integer constraints for the number of units in each stratum. Constraint **[10]** guarantees constraints **[4]** and **[6]**, while constraint **[11]** adapts constraint **[5]**.

The optimization model chosen to solve the problem moves from the *MINLP* framework of Approach 1, with a relatively high number of integer decision variables, to a non-differentiable nonlinear programming problem with only one non-negative real decision variable, the constant of proportionality (q), which means that the intermediate variables need to be integers and have non-differentiable functions. Given that this

problem originates from the general problem shown in Approach 1, its mathematical properties guarantee that a solution exists as long as the set of possible solutions is not empty. This can be assured if the *p*-value $\overline{pValue_{min}}$ is chosen appropriately and is lower than that resulting from the stratified random sample contained in the original simple random sample. This solution gives a global maximum to the problem and provides the larger subsample using the data available in the original simple random sample, which is closer to the sample obtained by stratified random sampling with proportional allocation with knowledge of the distribution by strata of the target population.

Approach 2.- Mathematical formulation of the optimization model. Maximizing the constant of proportionality and verifying the chi-square goodness-of-fit test.

$$\overbrace{\underset{q}{\text{Max}}\left\{\hat{q}(q)=\frac{n^{SUB}(q)}{N}=\frac{\sum_{i=1}^{k}n_i^{SUB}(q)}{N}\right\}}^{R_1} \quad \textbf{[7]}$$

subject to constraints:

$$\overbrace{pValue[\chi^2(q);r]\geq\overline{pValue_{min}}}^{R_4} \quad \textbf{[8]}$$

$$\overbrace{\chi^2(q)=\sum_{i=1}^{k}\frac{\left[n_i^{SUB}(q)-n_i^{exp}(q)\right]^2}{n_i^{esp}(q)}}^{R_4} \quad \textbf{[9]}$$

$$\overbrace{n_i^{SUB}(q)=Round[q\cdot N_i]}^{R_2,R_3} \quad \textbf{[10]}$$

$$\overbrace{0\leq n_i^{SUB}(q)\leq n_i^{SRS}}^{R_2,R_3} \quad \textbf{[11]}$$

$$\overbrace{n_i^{exp}(q)=\frac{N_i}{N}\cdot n^{SUB}(q)=\frac{N_i}{N}\cdot\sum_{i=1}^{k}n_i^{SUB}(q)}^{R_2,R_3} \quad \textbf{[12]}$$

$$\overbrace{0\leq n_{max}^{SUBStR}\leq q\leq\frac{n^{SRS}}{N}}^{R_2,R_3} \quad \textbf{[13]}$$

$$\forall i=1,2,\cdots,k$$

where:

$pValue[\chi^2;r]$: a function that calculates the *p*-value given the test statistic, χ^2, and the degrees of freedom, r, as in **[2a]**.

$r=k-1$: degrees of freedom equal to the number of strata minus 1.

$\overline{pValue_{min}}$: minimum *p*-value fixed by the researcher.

$Round$: function that rounds its argument to the nearest integer.

n_{max}^{SUBStR}: maximum size of a potential subsample contained in a simple random sample that would have been obtained using stratified random sampling directly from the target population. It is equal to $\min_{i\in\{1,2,\ldots,k\}}\left\{\frac{n_i}{N_i}\right\}$, but will be 0 if there is an empty stratum in the original sample but not in the population, i.e. $\exists$i such that $n_i = 0,\ N_i > 0$.

N: size of the target population.

n^{SRS}: size of the simple random sample.

Finally, it is important to mention that in some real problems it might be necessary to regroup those strata that do not reach the minimum size required for Pearson´s chi-square goodness-of-fit test to have a good convergence to χ^2 distribution. In such cases the problem will always have an optimal and, in some extreme cases, a trivial solution. In the extreme case of there being a large size reduction, there might be just one stratum of regrouped expected and observed values adding up to the same amount, and these would provide the same theoretical and observed distributions, distorting the initial sense of the test.

3. The algorithm for solving the model and some simulation results.

In this section we develop an efficient method for solving a specific convex MINLP using the optimization model set out in Section 2 (Approach 2) and an algorithm with the following steps:

Preliminary: We verify that the null hypothesis of the chi-square goodness-of fit test (H_0) – i.e. that the original simple random sample has the same distribution as the target population – is rejected. If not, we would not need to proceed. It is worth mentioning here that Pearson´s chi-square depends on the size of the sample, so the impact of effect size on the rejection of the null hypothesis has to be taken into account (Berkson, 1938; Wang, 1993 and Lin *et al*. 2013). Increasing emphasis has been placed on the use of effect size reporting when analysing social science data.

To carry out this preliminary goodness-of fit test we have to:

- **introduce the data** (*Input*) N_i from the target population, n_i^{SRS} from the simple random sample and set the chosen $\overline{pValue_{mín}}$ that will be the significance level for the test.
- **compute the initial values** (*output*): size of the target population, *N*; size of the original simple random sample, n^{SRS}; expected values, n_i^{exp}, given by **[12]**; degrees of freedom, r; and initial value of the constant of proportionality, $q_{start} = \frac{n^{SRS}}{N}$, $q = q_{start}$.
- **compute the test statistic and the apply the decision rule as in [8] and [9]**. The standardized effect size estimates also need to be reported. Cramer's V as interpreted by Cohen (1988) is used for this purpose.

1. **Computing the observed values obtained by stratified sampling.** Using the value for the constant of proportionality, q, which in the first iteration of the algorithm is equal to the ratio between the size of the simple random sample and the size of the target population, $q = q_{start} = \frac{n^{SRS}}{N}$, we calculate the size of each stratum in the subsample using the nearest integer of $q \cdot N_i$ as in **[10]**, $n_i^{SUB}(q) = Round[q \cdot N_i]$. The observed

values obtained will be the same as those obtained using stratified random sampling from the target population with constant of proportionality q.

2. **Fitting the observed values**. We compare the observed values obtained by stratified random sampling with proportional allocation from the target population, $n_i^{SUB}(q)$, which were found in the previous step, with those in the original simple random sample, n_i^{SRS}, using **[11]**. If $n_i^{SUB}(q) > n_i^{SRS}$, we choose $n_i^{SUB}(q) = n_i^{SRS}$ for the subsample instead of the observed value obtained using $Round[q \cdot N_i]$ from the previous step. However, if $n_i^{SUB}(q) < n_i^{SRS}$, then we take the observed value for the subsample to be $n_i^{SUB}(q) = Round[q \cdot N_i]$, as obtained in the previous step.

3. **Fitting the expected values**. After fitting the observed values, we calculate the total size of the subsample by adding up the number of units along the k stratum, $\sum_{i=1}^{k} n_i^{SUB}(q) = n^{SUB}$. We then compute the expected values as $n_i^{exp} = n^{SUB} \cdot \frac{N_i}{N}$ in order to obtain the same sum when adding up the expected values as we obtained with the observed values. If any stratum is found that is smaller than the minimum required by the test, which is usually 5, we will add it to the smallest nearest one until we reach the minimum number of elements[4].

4. **Goodness-of-fit test**. Using **[8]** and **[9],** we now test the null hypothesis, H_0 – i.e. that the subsample has the same distribution as the target population – by comparing the fitted observed values with the fitted expected values obtained in steps 2 and 3 above. If the null hypothesis is not rejected, we can stop the algorithm because we have found the optimal $\hat{q}^*$, i.e. the distribution of the largest subsample contained in the original simple random sample that fits the population of interest. If the null hypothesis is rejected, we will proceed to the next step.

5. **New value of q.** To find the new value of q in order to start the process again, we now aim to obtain a reduction of q that will provide the global optimal solution. The new value of q_j, ${q_j}^{SUB}$, used to start iteration j, is obtained by subtracting from the previous constant of proportionality, ${q_{j-1}}^{SUB}$, a step value (q_{step}) that is obtained in each case from the initial value of the constant of proportionality:

$$q_j^{SUB} = {q_{j-1}}^{SUB} - q_{step} \;\; ; q_{step} = 0.1 \,/N$$

To shorten the time taken to find the solution, we incrementally reduce q as much as possible to the point where it does not reject the null hypothesis in the goodness-of-fit test. We then reverse the process in order to consider the immediately preceding value of q^{SUB}. From that value onwards we are conducting a grid search in a finer set.

To analyse the performance of this algorithm as applied to solve the optimization model for subsample selection proposed in Section 2, we provide some results from a simulation study. This was carried out using MS Excel, a commonly-used non-specialist software, the advantages of which include the availability of pre-defined functions for calculating $\chi^2_{(\alpha,\ r)}$ and $pValue[\chi^2; r]$ in the goodness-of-fit test and the possibility of incorporating functions defined by the user in Visual Basic for Applications (VBA).

[4] For Pearson's chi-square goodness-of-fit test to be valid, the sample size must be large enough to provide a minimum number of expected elements per category. Núñez-Antón *et al*. (2019) have developed functions for regrouping strata automatically no matter where they are located, thus enabling the goodness-of-fit test to be performed within an iterative procedure. The functions are written in Excel VBA (Visual Basic for Applications) and Mathematica.

We have generated 4,000 populations and their corresponding simple random samples for two scenarios: a "bad" fit of the sample to the population using the chi-square goodness-of-fit test and a "fine" fit with a better fit. The main characteristics of the simulations are:

- The number of strata is a random integer number between 2 and 20.
- The 4,000 populations are generated in 4 blocks of 1,000 as a function of the maximum size of the population strata – 1,000, 10,000, 100,000 and 1,000,000 – given that the minimum size is 1 for all cases.
- The size of the simple random sample stratum is an integer number that results from rounding the product of the size of the corresponding stratum in the population by a percentage. This percentage is randomly selected from an interval that we have set to be [0%, 10%) for the "bad" fit and [3%, 5%) for the "fine" fit.
- For the 4,000 simulations in each group, the optimization model is solved[5] by means of our proposed algorithm with a given significance level, $\overline{pValue_{min}}$, of 5%, which is the most common significance level in practice.

A summary of the simulation results can be seen in Table 1**,** which shows the number of global solutions obtained (row 2) and the average time taken to obtain the solution (row 4) for those cases in which an effect size[6] (row 11) at least exists, even if it is small (row 12).

Towards the end of the table we show the average values for Cramer´s V (row 9) and degrees of freedom (row 10) in those cases for which we have categorized the effect size[7] (rows 11 to 14). These serve to provide general information about the simulated cases that have a global solution. As mentioned earlier, the proposed algorithm always finds the global solution. The total average size (row 5) and the relative average size of the selected subsample with respect to the simple random sample (row 6) from which it is extracted are presented, as is the number of cases in which the subsample is greater than zero (q > 0) (row 8), even though at least one stratum with zero units exists in the original simple random sample, which would prevent us from obtaining the subsample that best fits the population with a constant of proportionality equal to $Min\{n_i^{SRS}/N_i\}$. Finally, we show the size of the obtained subsample with respect to that resulting using as a constant of proportionality $q = Min\{n_i^{SRS}/N_i\}$ when it is not null (row 7).

Looking at the "bad" scenario, considering the 1,000 simulated populations and simple random samples for each of the 4 cases according to maximum strata size, the null hypothesis in the chi-square goodness-of-fit test is rejected in 714, 938, 992 and 996 cases respectively (row 1). Given that the effect size is large in almost all cases (row 14), it could be said that the test provides results to support the existence of statistically significant differences that are not exclusively due to sample size, i.e. the sample does not follow the same distribution as the population.

[5] The optimization model was solved using MsExcel Professional Plus 2016, VBA 7.1, in a computer with an Intel Core i7-2600 Quad-Core Processor 3.4 GHz, 32GB RAM and a Windows 7 Enterprise 64-bit system.

[6] Common practice when interpreting effect sizes is to use the benchmarks for "small," "medium," and "large" effects suggested by Cohen (1988). Effect sizes may inform about practical significance, but they are not inherently meaningful.

[7] Following Cohen's methodology, a table has been constructed that categorizes the effect size as a function of the value of Cramer´s V and the degrees of freedom. This is available to any interested researchers on request to the authors.

Table 1. Summary of the simulation results. Two scenarios per strata size band: "bad": [0%, 10%[and "fine": [3%, 5%). $\overline{pValue_{min}} = 5\%$ Effect size (ES): small, medium, large									
Items		**Cases by maximum strata size**							
		1,000		**10,000**		**100,000**		**1,000,000**	
		Bad	**Fine**	**Bad**	**Fine**	**Bad**	**Fine**	**Bad**	**Fine**
1	**Simple random sample reject cases**[8]	714	-	938	437	992	901	996	989
2	**Global solution cases with ES**	714	-	936	434	977	783	977	840
3	**Cases with regrouped stratum = 1**	13	-	3	0	1	0	0	0
4	**Average time seconds**	3.34	-	23.26	4.19	50.98	11.21	101.15	13.89
5	**Average subsample size**	119.8	-	492.7	1,696.8	3,044.1	10,667.9	23,925.7	97,970.5
6	**Relative average %** $\frac{n^{SUB}}{n^{SRS}}$	65.81	-	41.01	93.49	28.56	88.92	22.90	84.09
7	**Average** $\frac{q^{SUB}}{Min\left\{n_i^{SRS}/N_i\right\}}$	4.64	-	3.39	1.20	2.02	1.12	1.32	1.06
8	$\left\{\begin{matrix} \boldsymbol{Min\{n_i^{SRS}\} = 0} \\ \boldsymbol{n^{SUB} > 0} \end{matrix}\right\}$ **cases**	306	-	107	15	21	2	7	0
9	**Average Cramer's V**	0.21	-	0.20	0.05	0.21	0.05	0.20	0.05
10	**Average df**	8.89	-	9.95	11.49	10.01	11.06	10.19	10.97
11	**Type of ES**	Large	-	Large	Small	Large	Small	Large	Small
12	**Small cases**	8	-	60	434	67	783	68	840
13	**Medium cases**	270	-	304	-	328	-	313	-
14	**Large cases**	436	-	572	-	582	-	596	-
Source: Own									

However, looking at the "fine" scenario for the first column of maximum strata size (1,000), there is no rejection at all (row 1), whereas for the remaining three cases there is an increasing number of rejections of the null hypothesis but with a small effect size (row 12). Those rejections will therefore be mainly due to the size of the sample, although not completely, since some rejections do not even have this small effect size.

If we look at the simulation results for the "bad" scenario we find that there are original simple random samples with null size strata, whereas the corresponding population strata are not null. However, the subsample solution that best fits the population is not the trivial one, $q^* = Min\left\{\frac{n_i^{SRS}}{N_i}\right\}$, because the size reached by the subsample was not 0% of the population (row 8).

For this group of simulations ("bad"), the subsamples obtained have a relative size (row 7) with respect to the smallest non-null size stratum that ranges from 1.32 times larger (32% larger) up to 4.64 times larger (364% larger), so the procedure gives solutions larger than the trivial subsample solution.

[8] Verified cases rejecting the null hypothesis of the chi-square goodness-of fit, i.e. the original simple random sample does not have the same distribution as the target population.

As would be expected, the reduction in size between the original simple random sample and the subsample obtained (row 6) is greater in the "bad" scenario than in the "fine", given the better fit of the original simple random sample to the population in the latter group, and the time it takes to find the global solution is shorter for the "fine" cases than for the "bad". This time scale grows as the size of the original simple random sample increases (row 4). Finally, for the "fine" scenario, the number of solutions obtained with just one regrouped stratum (row 3) is lower than for the "bad" scenario, as is the number of cases with null-size strata in the original simple random sample that result in non-null-size optimal subsamples (row 8).

4.-Applying the model to the Continuous Sample of Working Lives (CSWL).

In this section we apply the methodology developed in the previous sections to the Continuous Sample of Working Lives (CSWL), a set of microdata comprising anonymized information on individuals. This information is administrative data about people's working lives and forms the basis of this sample taken from Spanish Social Security records. The sample reference population is defined as individuals who have had some connection with Social Security (through contributions or receiving some kind of pension or unemployment benefits) at any time during the year of reference. The total number of people involved during the year is higher than the number for any one particular date in that year, and one person can have several different simultaneous or successive relationships depending on their working situation during the year. Those who are outside the Social Security system (certain civil servants and those in the informal economy) are not included in this population. The first wave covers people who had an economic relationship of some sort with Social Security in 2004. However, each wave includes data covering the entire working and pension life of the people selected, starting in 1980. The sample is updated every year using information from the Social Security system, dating back to when computerized records began, and also from other administrative data sources, which record complementary information on individuals. The random sampling method is simple and uses no stratification. The sample provided by the INSS is 4% of the reference population and comprises around 1.2 million people. The population we want to study in this paper is the pensioner population categorized by age, gender and type of pension for the period 2005 to 2017 (DGOSS, 2006-2018).

Data gathered from the CSWL have been widely used by researchers (see selected references in Appendix 3) to investigate a number of issues connected with the Spanish economy and relevant socioeconomic conditions. These include immigrant integration, labour market transitions, the differential impact of the 2008 financial crisis on work safety, the duration of unemployment benefits and repeat claims, earnings inequality, retirement behaviour, labour-market intermittency, transition probabilities between disability states and duration analysis, mortality rates by occupation, and the sustainability of the public pension system.

Despite the widespread use of the CSWL, very little attention has been given to the fact that administrative errors, misclassification problems and the type of sampling used might mean that the data selected are not representative of the population of interest. In studies on global ageing and the sustainability of the public pension system, the CSWL is one of the main datasets considered for research purposes in Spain. It is therefore important to know how representative it is of the pensioner population (Pérez-Salamero *et al*. 2016, 2017). After performing a poststratification process on the CSWL by type of pension, gender and age, Pérez-Salamero *et al.* (2017) revealed that the

sample did not exactly fit the population of interest – with close to zero p-values in many cases and not only because of the size of the sample – for waves 2005 to 2013 on the basis of the INSS (2006-2014) statistics report data (see details in Appendix 2). In the same study, a much less sophisticated version of the procedure proposed in the present paper was applied to generate a large subsample distribution from the 2010 CSWL with striking results. With a very small reduction in the size of the original sample, errors in estimating pension expenditure for 2010 were significantly reduced. We have therefore decided to apply the methodology shown here to extend the analysis carried out by Pérez-Salamero *et al*. (2017) up to and including the last wave available (2017). Thus the contribution will not only be a real case study aimed at illustrating the procedure, it will also be a broad update of the investigation into the representativeness of the CSWL as regards the pensioner population. We simultaneously consider the distribution by age, gender and type of pension because this is a more general case than looking separately at the weight of each age cohort or the weight of one gender within a particular type of pension.

4.1-The optimization model adapted to the CSWL

The optimization model considers the results obtained from adapting the mathematical approach laid out in Approach 2 and applies them to this real case involving the pensioner distribution in the CSWL. The procedure will take into account that the value of the test for the subsample to be selected is such that it does not reject the null hypothesis (that the subsample has the same distribution as the pensioner population at 31st December) against the alternative hypothesis (that the subsample does not have the same distribution as the pensioner population at 31st December). The procedure should therefore include a goodness-of-fit test on the distribution of the number of pensioners by age (18 cohorts covering 5-year intervals except for the last one, which represents 85 years and over), gender (male and female) and type of pension (permanent disability, retirement, widow(er)'s, orphan's and family responsibilities[9]), and this includes taking into account the associated *p*-values.

In order to guarantee that the subsample distribution fits the population of interest by pension type, gender and age, constraint **[8]** in Approach 2 is tailored specifically to this case, producing 10 constraints that require that the null hypothesis not be rejected for each of the 10 combinations of pension type and gender. The mathematical approach to this real application is shown in Approach 3**.**

[9] This is a special type of survivor benefit for family members included in the public Spanish Social Security System. It is not compatible with the beneficiary receiving other public pensions.

Approach 3.- Mathematical formulation of the optimization model applied to the CSWL.

$$\overbrace{\underset{q}{\text{Max}}\left\{\hat{q}(q)=\frac{n^{SUB}(q)}{N^{INSS}}\right\}}^{R_1} \quad \textbf{[13]}$$

subject to constraints:

$$\overbrace{pValue_{j,k}\left[\chi^2_{j,k}(q);r_{j,k}(q)\right]\geq \overline{pValue_{min}}}^{R_4} \quad \textbf{[14]}$$

$$\overbrace{\chi^2_{j,k}(q)=\sum_{i\in I_{j,k}(q)}\frac{\left(n^{SUB}_{i,j,k}(q)-n^{exp}_{i,j,k}(q)\right)^2}{n^{exp}_{i,j,k}(q)}}^{R_4} \quad \textbf{[15]}$$

$$\overbrace{n^{SUB}_{i,j,k}(q)=Round\left[q\cdot N^{INSS}_{i,j,k}\right]}^{R_2,R_3} \quad \textbf{[16]}$$

$$\overbrace{0\leq n^{SUB}_{i,j,k}(q)\leq n^{SRS}_{i,j,k}}^{R_2,R_3} \quad \textbf{[17]}$$

$$\overbrace{n^{exp}_{i,j,k}(q)=\frac{N^{INSS}_{i,j,k}}{N^{INSS}}\cdot n^{SUB}(q)}^{R_4} \quad \textbf{[18]}$$

$$\overbrace{0\leq\ n^{SUBStR}_{max}\leq q\leq\frac{n^{CSWL}}{N^{INSS}}}^{R_2,R_3} \quad \textbf{[19]}$$

$$\forall i=1,2,\cdots,18;\ \forall j=1,2;\ \forall k=1,\cdots,5$$

where i is the index for the 18 cohorts into which the "age" variable has been categorized; j is the index corresponding to "gender" (male, female); and k is the index for the 5 types of pension benefit (permanent disability, retirement, widow(er)'s, orphan's and family responsibilities),

where:

$$n^{SUB}(q)=\sum_{k=1}^{5}\sum_{j=1}^{2}\sum_{i=1}^{18}n^{SUB}_{i,j,k}\ (q)$$

$N^{INSS}=\sum_{k=1}^{5}\sum_{j=1}^{2}\sum_{i=1}^{18}N^{INSS}_{i,j,k}$: total number of beneficiaries. $N^{INSS}_{i,j,k}$ is the number of pensioners in the population, with the sub-indices representing the corresponding groups by age, gender and type of benefit. These data are obtained from (INSS, 2006-2018).

$pValue_{j,k}\left[\chi^2_{j,k}(q);r_{j,k}(q)\right]$: a function that depends on the chi-square statistic and the degrees of freedom, both of which in the end also depend on the constant of

proportionality and, above all, the values estimated and observed after regrouping (where applicable).

$r_{j,k}(q) = g\left(n_{1,j,k}^{exp}(q), \cdots, n_{18,j,k}^{exp}(q)\right) - 1$: a function that returns the degrees of freedom in each iteration once the goodness-of-fit test is calculated for each type of pension and gender. It is equal to the expected number of regrouped cohorts minus 1, given that there are no parameters to estimate because the population distribution is already known.

$\overline{pValue_{min}}$: a pre-established α level of statistical significance which is the criterion for the subsample to improve the goodness of fit to the population. This pre-established minimum p-value will be the same for all 10 cases (5 types of pension/2 genders) and has to be high in order to guarantee a better fit to the population than the value given by the CSWL.

$\chi_{j,k}^2(q)$: the sample value for the chi-square statistic calculated in each iteration for each gender and type of pension (10 cases). It evaluates the difference between the regrouped observed values and the expected values for cohort indices with 5 or more elements, avoiding those indices in which the regrouped cohort has no elements.

$\bar{I}_{j,k}(q) = \{i \in I_{j,k}(q) / \bar{n}_{i,j,k}^{exp}(q) \geq 5\}$: a set of indices for regrouped age cohorts that contain 5 or more elements, for each type of pension and gender.

$I_{j,k}(q) = \{1, 2, 3, \ldots, 18\}$: a set of indices for age cohorts by type of pension and gender, which in all cases has 18 age cohorts.

$Round$: a function that returns the nearest integer to its argument.

n_{max}^{SUBStR}: the maximum size of a hypothetical subsample contained in the simple random sample that would have been obtained directly from the population using stratified random sampling, $\min\left\{\frac{n_{i,j,k}}{N_{i,j,k}}\right\}$. It will be equal to 0 it there is any empty cohort in the original simple random sample but not in the population, $\exists i / n_{i,j,k} = 0\,, N_{i,j,k} > 0$.

$n^{CSWL} = \sum_{k=1}^{5}\sum_{j=1}^{2}\sum_{i=1}^{18} n_{i,j,k}^{CSWL}$: the total number of pensioners in the CSWL.

4.2-Main results.

In this section we provide the results of applying the above procedure to the CSWL with the aim of finding a large subsample design to improve the fit to the distribution of the pensioner population. Table 2 shows the global optimal solutions for 2005 to 2017, which range in size from 14.4% of the CSWL in 2013, associated with a $\overline{pValue_{min}}$ of 0.95, to 99.39% in 2012, associated with a $\overline{pValue_{min}}$ of 0.05. We have verified that all the solutions are global.

What immediately attracts our attention in Table 2 is the small size of the subsamples for 2013 and 2014 compared to all the other waves analysed. Indeed, feasible solutions for the 2013 wave were found to range in size from 43.59% of the CSWL, associated with a minimum p value of 0.05, to 14.4%, associated with a p value of 0.95. Similar results are reported for the 2014 wave.

The explanation for this apparent anomaly in 2013 is that there are some cohorts (permanent disability males (0.21%) and females (1.67%), group 65-69 years) that are underrepresented in the subsample with respect to their real weight in the population (4%). For 2014 the explanation is the same, with some cohorts being underrepresented

due to administrative errors. This would include the fact that pensioners over 65 with permanent disability benefits have not been reclassified, so are considered as retirement pension beneficiaries in the CSWL but as disabled in the official population statistics. Since our procedure relies on maximizing the constant of proportionality that depends on the number of units in each stratum, the resulting percentages in the subsample for 2013 and 2014 are coherent.

For all the CSWL waves considered, the value of $\min\left\{\frac{n_{i,j,k}}{N_{i,j,k}}\right\}$ is 0.00%, and therefore the subsamples obtained by our proposed procedure are much larger than the sample contained in the original CSWL that would have been obtained by stratified sampling using a constant of proportionality equal to $q = \min\left\{\frac{n_{i,j,k}}{N_{i,j,k}}\right\}$. The effect size obtained, following Cohen (1988), can be classified as medium for the 2005 to 2011 waves with the exception of 2007, which is small. From 2012 to 2017 the size of the effect is negligible, so the rejection of the hypothesis that the CSWL has the same distribution as the population might be attributed to the large size of the sample.

Using Solver, a Microsoft Excel add-in program, with an Intel® Core™ i7-2600 Processor (32GB RAM, up to 3.40 GH) to solve the convex MINLP problem shown in Table 4, the time needed to solve each of the 39 cases (13 years x 3 significance levels $\overline{pValue_{min}}$) ranged from 1.638 to 11.076 seconds. This is a reasonable time for these types of problem with quite large dimensions, as would be the original problem of maximizing the size of the subsample instead of considering the constant of proportionality.

Using the 2005 to 2017 waves, Table 3 shows the p-values for the 10 cases considered (5 types of pension/2 genders) for the goodness-of-fit test for pensions in the subsample obtained using this procedure compared with those obtained for pensions in the CSWL. Overall we find a lack of representativeness in the CSWL for total pensions in all the waves analysed. Looking at the range of different pensions, the results seem to suggest that, for most of the waves analysed, the CSWL does not fit the distribution of the population well in terms of pension type, gender and age for two types of pension benefit: permanent disability and widow(er)'s.

This supports the notion that the findings in Pérez-Salamero *et al*. (2017) still apply for 2014 to 2017, especially as far as males are concerned. Once our proposed procedure is carried out, and after adjusting the distribution of total pensions with respect to the population using Pearson's goodness-of-fit test, the p-values obtained move towards or become equal to 1. The procedure therefore provides subsamples with a better fit and many more observations than would be attained by a stratified random sample taken from the CSWL.

For each of the years considered, Appendix 2 shows the distribution of pensions by pension type, gender and age of the optimal-design larger subsample for the three significance levels considered ($\overline{pValue_{min}}$ 5%, 50% and 95%). The procedure will enable users to choose, by means of the selected $\overline{pValue_{min}}$, between goodness of fit and subsample size.

Table 2. Summary of results by $\overline{pValue_{min}}$%														
Year	n^{SUB}			$\hat{q} = \left(\frac{n^{SUB}}{n^{INSS}}\right)$%			$\left(\frac{n^{SUB}}{n^{CSWL}}\right)$%			**Time (sec.)**			**D.I.U**	**Effect Size**
	0.05	**0.50**	**0.95**	**0.05**	**0.50**	**0.95**	**0.05**	**0.50**	**0.95**	**0.05**	**0.50**	**0.95**		
2005	240,702	227,849	177,856	2.971	2.812	2.195	74.65	70.67	55.16	5.570	5.180	8.814	$7.188 \cdot 10^{339}$	Medium
2006	315,634	308,116	204,716	3.836	3.745	2.488	95.83	93.55	62.16	7.051	8.752	7.629	$3.564 \cdot 10^{339}$	Medium
2007	319,612	310,649	300,913	3.835	3.727	3.610	96.09	93.40	90.47	8.439	4.727	8.222	$7.216 \cdot 10^{335}$	Small
2008	329,204	321,031	311,082	3.887	3.790	3.673	97.37	94.95	92.01	8.939	7.738	6.864	$6.468 \cdot 10^{338}$	Medium
2009	335,665	330,298	315,426	3.897	3.835	3.662	97.76	96.20	91.87	8.549	10.359	8.487	$1.083 \cdot 10^{338}$	Medium
2010	339,831	335,482	318,778	3.885	3.835	3.644	97.33	96.08	91.30	7.613	7.191	5.585	$2.204 \cdot 10^{336}$	Medium
2011	343,317	334,590	300,129	3.871	3.772	3.384	97.30	94.82	85.06	8.860	8.767	6.848	$4.670 \cdot 10^{332}$	Medium
2012	358,046	354,395	339,223	3.975	3.935	3.766	99.39	98.38	94.17	1.638	7.910	7.519	$5.332 \cdot 10^{330}$	Negligible
2013	158,486	92,024	52,502	1.732	1.005	0.574	43.59	25.31	14.44	3.791	3.994	5.819	$4.939 \cdot 10^{327}$	Negligible
2014	186,109	112,455	66,322	2.005	1.212	0.715	50.48	30.50	17.99	6.646	3.307	5.648	$4.666 \cdot 10^{325}$	Negligible
2015	371,174	369,381	364,563	3.969	3.950	3.898	99.20	98.72	97.43	6.568	11.076	8.377	$4.588 \cdot 10^{326}$	Negligible
2016	376,017	372,166	309,241	3.973	3.932	3.267	99.22	98.20	81.60	9.313	10.967	9.719	$1.497 \cdot 10^{324}$	Negligible
2017	378,463	352,488	278,254	3.954	3.682	2.907	99.13	92.32	72.88	9.329	9.032	9.453	$1.012 \cdot 10^{321}$	Negligible
D.I.U.:-Dimension of the Integer Unrestricted problem. Number of combinations of integer values that the pensioner strata may take, within their bounds but without requiring the constraint that the null hypothesis of the statistical test is not rejected.														
Source: Own														

Table 3. Goodness-of-fit test (population/samples), p-value (CSWL and subsamples) M: Male; F: Female													
Type of pension	**CSWL**												
	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Permanent Disability M	0.000000	0.000000	0.261722	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Permanent Disability F	0.000000	0.000000	0.022240	0.000000	0.000000	0.000000	0.000000	0.000000	0.187779	0.042312	0.000000	0.000000	0.000021
Retirement M	0.000000	0.000004	0.000001	0.000000	0.000000	0.000000	0.000000	0.005268	0.140715	0.013956	0.000743	0.001314	0.002751
Retirement F	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.001217	0.000120	0.000052	0.000011	0.000065	0.013070
Widower's M	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Widow's F	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000014	0.000006	0.000451	0.009670	0.008186	0.208145
Orphan's M	0.000000	0.005082	0.202030	0.261090	0.591337	0.462422	0.837630	0.944315	0.593409	0.370949	0.849506	0.000039	0.112380
Orphan's F	0.000000	0.118497	0.164789	0.141561	0.393848	0.561802	0.296694	0.117598	0.106684	0.285731	0.000101	0.000000	0.070073
Family Responsibilities M	0.002755	0.111863	0.115631	0.396466	0.061782	0.428490	0.208140	0.323662	0.463862	0.327626	0.834403	0.830553	0.915809
Family Responsibilities F	0.003573	0.249222	0.154051	0.021609	0.689061	0.841654	0.960454	0.333362	0.156821	0.659514	0.886466	0.344878	0.758455
Total Pensions	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**	**0.000000**
Type of pension	**Subsample with p-value ≥ $\overline{pValue_{min}}$ = 5%**												
	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Permanent Disability M	0.999062	0.997165	1.000000	1.000000	0.999995	0.999884	1.000000	1.000000	0.050022	0.050023	0.999079	0.999962	1.000000
Permanent Disability F	0.050167	1.000000	0.999935	1.000000	0.999449	0.999896	0.998295	0.986718	1.000000	0.952540	0.998444	0.999978	0.990695
Retirement M	1.000000	0.992407	0.999452	0.674773	0.105610	0.113748	0.050031	0.136429	0.999905	0.999937	0.433788	0.050327	0.050202
Retirement F	0.999948	0.999994	1.000000	0.734123	0.050021	0.050184	0.237496	0.536713	0.999268	0.999864	0.050727	0.068528	0.495790
Widower's M	1.000000	0.050289	0.050099	0.050117	0.094137	0.289216	0.361538	0.205785	1.000000	1.000000	0.907820	0.903889	0.926686
Widow's F	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.997186	1.000000	1.000000	0.886394	0.946287	0.999344
Orphan's M	1.000000	0.973781	0.974972	0.996229	0.997206	0.999660	0.999868	0.999343	1.000000	1.000000	0.998689	0.982116	0.968999
Orphan's F	1.000000	0.994200	0.992617	0.999344	0.996990	1.000000	0.999810	0.999604	1.000000	1.000000	1.000000	1.000000	0.999870
Family Responsibilities M	0.485665	0.452965	0.870183	0.960915	0.834992	0.984285	0.739576	0.894736	1.000000	0.999996	0.984555	0.992232	0.987172
Family Responsibilities F	0.592029	0.958494	0.915957	0.995480	0.997873	0.997951	0.994521	0.839826	0.999999	0.999998	0.999575	0.999999	0.999967
Total Pensions	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**0.999999**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**
Type of pension	**Subsample with p-value ≥ $\overline{pValue_{min}}$ = 50%**												
	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Permanent Disability M	0.999824	0.998681	1.000000	1.000000	1.000000	0.999995	1.000000	1.000000	0.500135	0.500169	0.999890	1.000000	1.000000
Permanent Disability F	0.500237	1.000000	1.000000	1.000000	0.999988	0.999996	0.999946	0.999053	1.000000	0.999968	0.999604	0.999998	1.000000
Retirement M	1.000000	0.999888	1.000000	1.000000	0.993192	0.500103	0.500086	0.742335	0.999991	0.999987	0.834558	0.500022	0.500002
Retirement F	0.999996	1.000000	1.000000	1.000000	0.974744	0.831733	0.999998	0.998129	0.999696	0.999979	0.500139	0.888871	0.999492
Widower's M	1.000000	0.500484	0.500491	0.500446	0.500167	0.704613	0.973377	0.500940	1.000000	1.000000	0.974067	0.994449	1.000000
Widow's F	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.999922	1.000000	1.000000
Orphan's M	1.000000	0.993038	0.990978	0.999867	0.999881	0.999983	1.000000	0.999930	1.000000	1.000000	0.999604	0.999410	1.000000

Table 3. Goodness-of-fit test (population/samples), p-value (CSWL and subsamples) M: Male; F: Female													
Orphan's F	1.000000	0.998723	0.998609	0.999976	0.998971	1.000000	0.999987	0.999933	1.000000	1.000000	1.000000	1.000000	1.000000
Family Responsibilities M	0.596180	0.516585	0.926654	0.981715	0.865071	0.991481	0.815663	0.908726	1.000000	0.999994	0.986995	0.995072	0.996236
Family Responsibilities F	0.730443	0.983661	0.965339	0.998284	0.999231	0.999351	0.997099	0.879598	1.000000	1.000000	0.999852	1.000000	1.000000
Total Pensions	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**
Type of pension	**Subsample with p-value ≥ $\overline{pValue_{min}} = 95\%$**												
	2005	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Permanent Disability M	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.975219	0.950021	1.000000	1.000000	1.000000
Permanent Disability F	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.999985	1.000000	1.000000
Retirement M	1.000000	1.000000	1.000000	1.000000	1.000000	0.950005	0.999992	0.996452	0.999993	0.999999	0.950024	0.950063	0.950000
Retirement F	1.000000	1.000000	1.000000	0.999996	1.000000	0.999897	0.999986	0.999985	0.999986	0.999988	0.962697	1.000000	0.999905
Widower's M	1.000000	1.000000	0.950055	0.950193	0.999535	0.999988	1.000000	0.999689	1.000000	1.000000	0.999129	1.000000	1.000000
Widow's F	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
Orphan's M	1.000000	1.000000	0.995183	0.999997	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.999970	1.000000	1.000000
Orphan's F	1.000000	1.000000	0.999779	1.000000	0.999919	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
Family Responsibilities M	0.950041	0.950010	0.962938	0.992756	0.950023	0.999119	0.950023	0.950006	0.999987	0.998926	0.989983	0.999760	0.999937
Family Responsibilities F	0.987717	0.999995	0.984731	0.999267	0.999880	0.999992	0.999793	0.950267	0.999999	1.000000	0.999985	1.000000	1.000000
Total Pensions	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**1.000000**
Source: Own													

6. Concluding comments and future research.

This paper contains our proposal for an optimization model for improving the representativeness of an already available simple random sample obtained from a bigger population than the population of interest, and the development of a novel methodology for selecting large subsamples can be considered the main contribution of this research. Neither the criterion used to select the subsamples nor the optimization model has been comprehensively explored in the literature.

The optimization model chosen to solve the problem moves from an MINLP framework with a relatively high number of integer decision variables to a non-differentiable nonlinear programming (NLP) problem with only one non-negative real decision variable: the constant of proportionality. We develop a method to efficiently solve a specific convex MINLP using the proposed optimization model by means of an algorithm that we have proved always finds the global solution. Using a simulation study, the procedure has been shown to work well in different scenarios as regards the goodness of fit of the simple random sample to the target population with an efficient use of time.

Another important contribution of this research is the real application of the procedure to the CSWL, a dataset widely used by a broad range of social science researchers comprising a simple random sample obtained from Spanish Social Security records. This dataset has become a baseline for researchers as it provides invaluable information about working lives and enables in-depth studies to be made of many aspects of the Spanish pension system that were previously overlooked. The methodology developed in this paper is applied to extend the analysis carried out by Pérez-Salamero *et al*. (2017) up to 2017, which was the last wave available at the time of writing (April 2019). Thus, the contribution broadly updates the analysis of the representativeness of the CSWL with respect to the pensioner population. We find that, overall, the CSWL lacks representativeness when all pensions in all the waves analysed are considered. Looking at the different types of pension, the results seem to suggest that for most of the waves analysed the CSWL does not fit the distribution of the population well in terms of pension type, gender and age for two types of pension benefit: permanent disability and widow(er)'s. This endorses that the findings in Pérez-Salamero *et al*. (2017) are still true for 2014 to 2017.

The application of the adapted optimization model to the 2005-2017 waves of the CSWL show that large subsamples can be obtained that will satisfy the chi-square goodness-of-fit test with associated *p*-values close to one. It can therefore be concluded that, for all the waves considered, it is possible to select large subsamples from the CSWL that better represent the pensioner population than the CSWL's own dataset, with a better fit to the distribution of the population's pensioners by type of pension, age and gender. And last but not least, with this procedure the users can choose between the desired goodness of fit and the size of the subsample they want, thereby allowing them a certain degree of freedom to adapt the procedure to the research to be conducted.

From an applied perspective, since the CSWL has been widely used by researchers to investigate various issues in connection with the Spanish economy and its socioeconomic conditions, but without testing its representativeness with respect to the population of interest, the real example developed in this paper could and should be extended to other groups of interest such as contributors, recipients of unemployment benefits, immigrants and/or the native population.

Finally, the model has been implemented using MS Excel, so as a future line of research we would like to use other optimization software and include in it the functions we have developed for regrouping strata automatically and then compare the results.

7. References

Baillargeon, S., & Rivest, L. P. (2009). A general algorithm for univariate stratification. *International Statistical Review*, 77(3), 331-344.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the Chi-Square Test. *Journal of the American Statistical Association*, 33(203), 526-536.

Bonami, P., Kilinç, M., & Linderoth. J. (2012). Algorithms and software for convex mixed integer nonlinear programs. In J. Lee & S. Leyferr (Eds.), *Mixed Integer Nonlinear Programming. The IMA Volumes in Mathematics and its Applications*, vol 154 (pp. 1-39). New York: Springer.

Bowley, A. L. (1926). Measurement of precision attained in sampling. *Bulletin of the International Statistical Institute 22*(1), 6-62.

Cochran, W. G. (1977). *Sampling Techniques*. New York: John Wiley.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale. NJ: Erlbaum.

De Moura Brito, J. A., Do Nascimento Silva, P. L., Silva Semaan, G., & Maculan, N. (2015). Integer programming formulations applied to optimal allocation in stratified sampling. *Survey Methodology*, 41(2), 427-442.

D'Ambrosio, C., & Lodi, A. (2013). Mixed integer nonlinear programming tools: an updated practical overview. *Annals of Operations Research*, 204(1), 301-320.

Díaz-García, J. A., & Ramos-Quiroga, R. (2012). Optimum allocation in multivariable stratified random sampling: stochastic matrix mathematical programming. *Statistica Neerlandica*, 66(4), 492-511.

Díaz-García, J. A., & Ramos-Quiroga, R. (2014). Optimum allocation in multivariable stratified random sampling: a modified Prékopa's approach. *Journal of Mathematical Modelling and Algorithms*, 13, 315-330.

DGOSS (2006-2018). *Muestra Continua de Vidas Laborales 2005-2017*. Madrid: Dirección General de Ordenación de la Seguridad Social. Ministerio de Trabajo, Migraciones y Seguridad Social.

Grafström, A., & Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41, 277 –290.

Gupta, N., Sana Ifthekar, S., & Bari, A. (2012). Fuzzy goal programming approach to solve non-linear bi-level programming problem in stratified double sampling design in presence of non-response. *International Journal of Scientific & Engineering Research*, 3(10), 1-9.

Gupta, N., Ali, I., & Bari, A. (2014). An optimal chance constraint multivariate stratified sampling design using auxiliary information. *Journal of Mathematical Modelling and Algorithms in Operations Research*, 13(3), 341-352.

INSS (2006-14). *Informes Estadísticos 2005-2013*. Madrid: Instituto Nacional de la Seguridad Social. Secretaría de Estado de la Seguridad Social. Ministerio de Trabajo, Migraciones y Seguridad Social.

INSS (2006-18). *Informes Estadísticos 2005-2017*. Madrid: Instituto Nacional de la Seguridad Social. Secretaría de Estado de la Seguridad Social. Ministerio de Trabajo, Migraciones y Seguridad Social.

Kontopantelis, E. (2013). A greedy algorithm for representative sampling: repsample in Stata. *Journal of Statistical Software*, 56, 1-18.

Kruskall, W., & Mosteller, F. (1979a). Representative sampling, I. *International Statistical Review*, 47(1), 13-24.

Kruskall, W., & Mosteller, F. (1979b). Representative sampling, II: scientific literature. excluding statistics. *International Statistical Review*, 47(2), 111-127.

Kruskall, W., & Mosteller, F. (1979c). Representative sampling, III: The Current Statistical Literature. *International Statistical Review*, 47(3), 245-265.

Kruskall, W., & Mosteller, F. (1980). Representative sampling, IV: the history of the Concept in Statistics. 1895-1939. *International Statistical Review*, 48(2), 169-195.

Lin, M., Lucas, H. C., & Shmieli, G. (2013). Research commentary: too big to fail. Large samples and the p-value Problem. *Information Systems Research*, 24(4), 906-917.

MESS (2018). *MCVL. Muestra Continua de Vidas Laborales. Guía del contenido*. Estadísticas. Presupuestos y Estudios. Estadísticas. Muestra Continua de Vidas Laborales. Documentación MCVL.

http://www.seg-social.es/wps/wcm/connect/wss/320b09c6-dc33-42be-b532-08880e618742/MCVLGuia20180725.pdf?MOD=AJPERES&CVID= (accessed 11 Sep 2018).

Neyman, J. (1934). On the two different aspects of the representative method: The method of representative sampling and the method of purposive sampling. *Journal of the Royal Statistical Society*, 97(4), 558-625.

Nuñez-Antón, V., Pérez-Salamero González, J. M., Regúlez-Castillo, M., Ventura-Marco, M., & Vidal-Meliá, C. (2019). Automatic regrouping of strata in the goodness-of-fit chi-square test. *SORT*, 43(1). In Press.

Olsen. A.;Hudson, R. (2009). Social Security Administration's Master Earnings File: background information. *Social Security Bulletin*, 69(3), 29-45.

Omair. A. (2014). Sample size estimation and sampling techniques for selecting a representative sample. *Journal of Health Specialties*, 2(4), 142-147.

Pérez-Salamero González, J. M., Regúlez-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 Española*, 217(2), 67–130

Pérez-Salamero González, J. M., Regúlez-Castillo, M., & Vidal-Meliá. C. (2017). The continuous sample of working lives: improving its representativeness. *SERIEs*, 8(1), 43-95.

Ramsey, C. A., & Hewitt, A. D. (2005). A Methodology for assessing sample representativeness. *Environmental Forensics*, 6, 71–75.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). *Model Assisted Survey Sampling. Springer Series in Statistics*. New York: Springer Verlag.

Smith, C. (1989). The Social Security Administration's Continuous Work History Sample. *Social Security Bulletin*, 52(10), 20–28.

Valliant, R., & Gentle, J. E. (1997). An application of mathematical programming to sample allocation. *Computational Statistics & Data Analysis*, 25(3), 337-360.

Valliant, R., Dever, J., & Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Statistics for Social and Behavioral Sciences, 51. New York: Springer.

Wang, C. (1993). *Sense and Nonsense of Statistical Inference: Controversy, Misuse and Subtlety*. New York: Marcel Dekker.

Zweimüller, J., Winter-Ebmer, R., Lalive, R., Kuhn, A., Wuellrich, J.P., Ruf, O., & Büchi, S. (2009). Austrian Social Security Database. *IEW - Working Papers*, 410. Institute for Empirical Research in Economics - University of Zurich.

Appendix 1. Proof of the convexity of the chi-square statistic function.

Appendix available upon request to the authors.

Appendix 2. Distribution of pensions.

Appendix available upon request to the authors.

Appendix 3. Selection of papers that use the Continuous Sample of Working Lives (CSWL)[1]

Agliari, E., Barra, A., Contucci, P., Sandell, R., & Vernia, C. (2014). A stochastic approach for quantifying immigrant integration: the Spanish test case. *New Journal of Physics*, 16, 103034.

Alonso Domínguez, A. (2012). Labor transitions of Spanish workers: a flexicurity approach. *Revista Internacional de Organizaciones/International Journal of Organizations*,9, 121-143.

Alonso, F., Devesa J.E., Devesa M., Domínguez, I., Encinas, B., Meneu, R., & Nagore, A. (2018). Towards an adequate and sustainable replacement rate in defined benefit pension systems: The case of Spain. *International Social Security Review*, 71 (1), 51-70.

Álvarez de Toledo, P., Núñez, F., & Usabiaga, C. (2011). An empirical analysis of the matching process in Andalusian public employment agencies. *Hacienda Pública Española*, 198 (3), 67-102.

Álvarez de Toledo, P., Núñez, F., & Usabiaga, C. (2013). Análisis "cluster" de los flujos laborales andaluces. *Revista de Estudios Regionales*, 97, 195-221.

Álvarez de Toledo, P., Núñez, F., & Usabiaga, C. (2014). An empirical approach on labour segmentation. Applications with individual duration data. *Economic Modelling*, 36, 252-267.

Álvarez de Toledo, P., Núñez, F., & Usabiaga, C. (2017). ¿Quién se empareja con quién en el mercado laboral español? Un análisis cluster basado en la Muestra Continua de Vidas Laborales. *Investigación económica*, 76(299), 3-182.

Amuedo Dorantes, C., & Borra, C. (2013). On the differential impact of the recent economic downturn on work safety by nativity: The Spanish experience. *IZA Journal of Migration and Development*, 2(1), 1-26.

Anghel, B., Basso, H., Bover, O. *et al*. (2018). Income, consumption and wealth inequality in Spain. *SERIEs*, 9(4), 351-357.

Arranz, J. M., & García-Serrano, C. (2011). Are the MCVL tax data useful? Ideas for mining. *Hacienda Pública Española/Revista de Economía Política*, 199(4), 151-186.

Arranz, J. M., & García-Serrano, C. (2014a). Duration and recurrence in unemployment benefits. *Journal of Labor Research* 35(3), 271-295.

[1] This reference list is not exhaustive but represents some of the most important published papers that have used the CSWL for data.

Arranz, J. M., & García-Serrano, C. (2014b). Duration of joblessness and long-term unemployment: is duration as long as official statistics say? In: Malo, M.; Sciulli, D. (Eds) *Disadvantaged Workers. Empirical Evidence and Labour Policies* (pp. 297-320). New York: Springer.

Arranz, J. M., & García-Serrano, C. (2014c). The interplay of the unemployment compensation system, fixed-term contracts and rehirings: The case of Spain". *International Journal of Manpower*, 35(8), 1236–1259.

Arranz, J. M., García-Serrano, C. & Hernanz, V. (2013). How do we pursue "labormetrics"? An application using the MCVL. *Estadística Española*, 55(181), 231-254.

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. *Scientific Reports*, 4, 4174.

Benavides F.G., Duran, X., Gimeno, D., Vanroelen, C., & Martínez, J.L. (2015). Labour market trajectories and early retirement due to permanent disability: a study based on 14 972 new cases in Spain. *European Journal of Public Health*, 25(4), 673–677.

Bentolila, S., García-Pérez, J.I., & Jansen, M. (2017). Are the Spanish long-term unemployed unemployable? *SERIEs*, 8 (1), 1-41.

Boado-Penas, M. C., Valdés-Prieto, S., & Vidal-Meliá, C. (2008). An actuarial balance sheet for pay-as-you-go finance: solvency indicators for Spain and Sweden, *Fiscal Studies*, 29, 89–134.

Bonhomme, S., & Hospido, L. (2017). The cycle of earnings inequality: evidence from Spanish social security data. *The Economic Journal*, 127(603), 1244–1278

Bonhomme, S., & Hospido, L. (2013). Earnings Inequality in Spain: New Evidence Using Tax Data. *Applied Economics* 45(30), 4212–4225

Cairó Blanco, I. (2010). An empirical analysis of retirement behaviour in Spain: Partial versus full retirement. *SERIEs - Journal of the Spanish Economic Association*, 1(3), 325-356.

Carrasco, R., & García Pérez, J. I. (2015). Employment dynamics of immigrants versus natives: evidence from the boom-bust period in Spain, 2000–2011. *Economic Inquiry*, 53(2), 1038-1060.

Castañer-Garriga, A., Pérez-Salamero González, J.M., & Vidal-Meliá, C. (2017), Evaluación de las tarifas de las Pensiones de Accidentes de Trabajo y Enfermedades Profesionales (2011-2015). *Innovar Journal*, 27(66), 153-167.

Cebrián, I., & Moreno, G. (2013). Labour market intermittency and its effect on gender wage gap in Spain. *Revue Interventions Économiques* [Online], 47.

Cebrián, I., & Moreno, G. (2015). The Effects of gender differences in career interruptions on the gender wage gap in Spain. *Feminist Economics*, 21(4), 1-27.

Conde Ruiz, J.I., & González C.I. (2013). Reforma de pensiones 2011 en España. *Hacienda Pública Española*, 204(1), 9–44

Conde-Ruiz, J. I. & González, C. I. (2016). From Bismarck to Beveridge: the other pension reform in Spain. *SERIEs*, 7(4), 461-490.

Cueto, B., & Rodríguez, V. (2014). Sheltered employment sentres and labour market integration of people with disabilities: a quasi-experimental evaluation using Spanish data. In: M. Malo & D. Sciulli (Eds), *Disadvantaged Workers. Empirical Evidence and Labour Policies* (pp. 65-91). New York: Springer.

De la Roca, J., & Puga, D. (2017). Learning by working in big cities. *Review of Economic Studies*, 84(1), 106-142.

De Pedraza, P., Villacampa González, A., & Muñoz de Bustillo Llorente, R. (2012). Immigrants' employment situations and decent work determinants in the Spanish labour market. *International Journal of Humanities and Social Science*, 2(6), 1-19.

Devesa, J.E., 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 Española*, 201, 9–38.

Dudel, C., López Gómez, M.A., Benavides, F.G. *et al*. (2018). The length of working life in Spain: levels, recent trends, and the impact of the financial crisis. *European Journal of Population*, 34(5), 769-791.

Duran, X., Vanroelend, C., Deboosere, P., & Benavides, F.G. (2016). Social security status and mortality in Belgian and Spanish male workers. *Gaceta Sanitaria*, 30(4), 293–295.

García García, M., & Nave Pineda, J. M. (2018). Impacto en las prestaciones de jubilación de la reforma del sistema público de pensiones español. *Hacienda Pública Española*, 224(1), 113-137.

García Pérez, J. I., Jiménez Martín, S., & Sánchez Martín, A. R. (2013). Retirement incentives, individual heterogeneity and labor transitions of employed and unemployed workers. *Labour Economics*, 20, 106-120.

García Pérez, J. I., & Osuna, V. (2014). Dual labour markets and the tenure distribution: reducing severance pay or introducing a single contract? *Labour Economics*, 29, 1-13.

García Pérez, J. I., & Rebollo Sanz, Y. (2009). The use of permanent contracts across Spanish regions: do regional wage subsidies work? *Investigaciones Económicas*, XXXIII (1), 97-130.

García Pérez, J.I., Marinescu, I., & Castelló, J.V. (2018). Can fixed-term contracts put low skilled youth in a better career path? Evidence from Spain. *The Economic Journal*. In Press.

García-Gómez, P., Jiménez-Martín, S., & Castelló, J. V. (2012). Health, disability, and pathways into retirement in Spain. In D. A. Wise (Ed.), *Social security programs and retirement around the world* (pp.127–174). Chicago: University of Chicago Press.

Garda, P. (2013). *Essays on the Macroeconomics of Labor Markets*. Doctoral Dissertation. Barcelona: Universitat Pompeu Fabra.

Gómez Tello, A., & Nicolini, R. (2017). Immigration and productivity: a Spanish tale. *Journal of Productivity Analysis*, 47(2), 167-183.

González, L., & Ortega, F. (2011). How do very open economies adjust to large immigration flows? Evidence from Spanish regions. *Labour Economics*, 18, 57-70.

Jiménez-Martín, S., Juanmartí Mestres, A., & Vall Castelló, J. (2019), Hiring subsidies for people with a disability: do they work? *European Journal of Health Economics*. In Press.

López Gómez M.A., Durán X., Zaballa E., *et al*. (2016) Cohort profile: the Spanish WORKing life Social Security (WORKss) cohort study. *BMJ Open*, 6(3): e008555.

López, M.A., Benavides, F.G., Alonso, J., Espallargues, M., Durán, X., & Martínez, J.M. (2014). The value of using administrative data in public health research: the Continuous Working Life Sample. *Gaceta Sanitaria*, 28(4), 334-337.

López, M.A., Duran, X., Alonso, J., Martínez, J.M., Espallargues, M., & Benavides, F.G. (2014). Estimating the burden of disease due to permanent disability in Spain during the period 2009-2012. *Revista Española de Salud Pública*, 88(3), 349-358.

Marie, O., & Vall Castello, J. (2012). Measuring the (income) effect of disability insurance generosity on labour market participation. *Journal of Public Economics*, 96, 198–210.

Moral Arce I., Patxot, C., & Souto, G. (2008). La sostenibilidad del sistema de pensiones. Una aproximación a partir de la CSWL. *Revista de Economía Aplicada*, XVI(E–1), 29–66

Muñoz de Bustillo, R., De Pedraza, P., Antón, J. I. & Rivas, L. A. (2011). Working life and retirement pensions in Spain: The simulated impact of a parametric reform. *International Social Security Review*, 64(1), 73-93.

Nagore García, A. (2017) Gender differences in unemployment dynamics and initial wages over the business cycle. *Journal of Labor Research*, 38(2), 228-260.

Nuñez-Antón, V., Pérez-Salamero González, J. M., Regúlez-Castillo, M., Ventura-Marco, M., & Vidal-Meliá, C. (2019), Automatic regrouping of strata in the goodness-of-fit chi-square test. *SORT*, 43 (1). In Press.

Patxot, C., Souto, G., & Villanueva, J. (2009). Fostering the contributory nature of the Spanish retirement pension system: An arithmetic micro-simulation exercise using the MCVL. *Presupuesto y Gasto Público*, 57, 7-32.

Peinado Martínez, P. (2011). *Pension System's reform in Spain: a dynamic analysis of the effects on welfare*. Doctoral Dissertation. Bilbao: Universidad del País Vasco.

Peinado Martínez, P. (2014). A dynamic gender analysis of Spain's pension reforms of 2011. *Feminist Economics*, 20(3), 163-190.

Peinado Martínez, P., & Serrano Pérez, F. (2011). A dynamic analysis of the effect of social security reform on Spanish widow pensioners. *Panoeconomicus*, 58(5), 759-771.

Pérez-Salamero González, J.M. (2015). La MCVL como fuente generadora de datos para el estudio del sistema de pensiones. Doctoral Dissertation. Valencia: Universidad de Valencia.

Pérez-Salamero González, J. M., Regúlez-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 Española, 217(2), 67–130

Pérez-Salamero González, J. M., Regúlez-Castillo, M., & Vidal-Meliá, C. (2017). The continuous sample of working lives: improving its representativeness. *SERIEs*, 8(1), 43-95.

Rebollo-Sanz, Y. (2012). Unemployment insurance and job turnover in Spain. *Labour Economics*, 19, 403–426.

Sánchez Martín, A. R., & Sánchez Marcos, V. (2010). Demographic change and pension reform in Spain: An assessment in a two-earner OLG model. *Fiscal Studies*, 31, 405–452.

Solé, M.; Diaz Serrano, L., & Rodríguez, M. (2013). Disparities in work, risk and health between immigrants and native-born Spaniards. *Social Science & Medicine*, 76, 179-187.

Vall Castello, J. (2012). Promoting employment of disabled women in Spain: Evaluating a policy. *Labour Economics*, 19, 82-91.

Vall Castelló, J. (2017). What happens to the employment of disabled individuals when all financial disincentives to work are abolished? *Health economics*, 26(S2), 158-174.

Vegas Sánchez, R., Argimón, I., Botella, M., & González, C. (2013). Old age pensions and retirement in Spain. *SERIEs*, 4, 273-307.

Vidal-Meliá, C. (2014). An assessment of the 2011 Spanish pension reform using the Swedish system as a benchmark. *Journal of Pension Economics and Finance*, 13(3), 297-333.

Vidal-Meliá, C., Boado Penas, M. C., & Settergren, O. (2009). Automatic balance mechanisms in Pay-As-You-Go pension systems. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 34, 287-317.