Spanish Journal of Statistics

VOLUME 1, NUMBER 1, 2019

Spanish Journal of Statistics

Published by the Spanish National Statistical Institute
NIPO: 096-20-014-2
ISSN: 2695-9070
DOI: https://doi.org/10.37830/SJS.2019.1

SPANISH JOURNAL OF STATISTICS

VOLUME 1, NUMBER 1, 2019

Contents

SPANISH JOURNAL OF STATISTICS
Vol. **1** No. 1, 2019, Pages 5 – 8
doi: https://doi.org/10.37830/SJS.2019.1.01

EDITORIAL

Official Statistics in Spain: Current status and perspectives

Juan M. Rodriguez Poo

Instituto Nacional de Estadística

1 Introduction

This year Official Statistics commemorates the 75th anniversary of the creation of the National Statistics Institute (INE), which, at the same time, also coincides with the celebration of the 150th anniversary of the Geographical and Statistical Institute. This year also, our journal *Estadística Española* (REE), created already in 1958 is having a significant change: It will appear under a new name, *Spanish Journal of Statistics* (SJS), it will be published in English and a completely new Editorial Board has been appointed. In fact, the main aim of this renovation process is to continue the REE's work, although the international and global perspective of statistics will be strengthened, with a clear vocation to contribute to the development and better understanding of the official statistics through quality scientific articles. It will promote also the development and improvement of statistical methods, including statistical software and algorithms.

Official Statistics has undergone a formidable evolution throughout its long history (see Escribano Ródenas and Fernández Barberis (2009), Ray (2003) and García Villar and De Castro Puente (2011)), from its origins mainly based on extensive demographic accounts to the current state in which multiple statistics are produced, covering a large number of thematic areas. This evolution has been marked by the decisive contribution that the establishment of internationally harmonized methodologies has had, the development of statistical techniques, particularly those related to sampling theory (see Devaud and Tillé (2019)), as well as the wide implementation of information and communication technologies in the statistical production process. Nowadays, official statistics play a role of great relevance in our society, continuously offering comparable information of the highest quality that is then used in evidence-based decision-making processes. Thus, statistics constitutes a public good, available and accessible to all citizens in accordance with the principles contained in the code of good practice for European statistics.

The evolution that we observe in Official Statistics is perfectly reflected in the interests that have guided the INE since its creation and that are reflected in the first numbers of the REE. Thus we can find in these early years articles dedicated to Sampling Theory (see Azorín Poch (1959) and Sánchez-Crespo (1976)), Information Theory (see Arnaiz Vellando (1960)), Econometrics (see Pena Trapero (1960) and Pena Trapero (1974)), Monte Carlo techniques (see Diez de Artazcoz (1963)), Game Theory (see Ferrer Martín (1960)), or Statistical Decision Theory (see Neymann (1963)). Already in the early

 Published by the Spanish National Statistical Institute

seventies we find papers devoted to statistics of extreme values (see Damas Rico (1972)), Bayesian Statistics (see Ibarrola Muñoz and Quesada Paloma (1972) and Quesada (1972)). These references and many others show how new statistical techniques have been progressively incorporated into Official Statistics. However, maintaining the current degree of relevance of the statistics requires a permanent renewal effort. Modernization of statistics is an inherent concern of national statistical offices, in areas that include the institutional environment and necessary resources, statistical production methods, data collection, dissemination and communication, or standards and metadata. The new challenges facing official statistics can be seen more clearly in the three specific cases presented below: globalization statistics; mobility statistics; and indicators of sustainable development.

2 Globalization and Mobility Statistics and Sustainable Development Indicators

Globalization represents one of the main challenges for business statistics. Given that statistics have been usually limited by national borders, the activities of groups of multinational companies such as the outsourcing of activities or foreign direct investment, clearly exceeds the traditional limits of performance of official statistics. In this context, it is necessary to coordinate a form of joint work with the rest of the member countries of the statistical community in order to produce economic indicators of globalization, and furthermore, to be able to assess the profile of multinationals, thus achieving an adequate measurement of this phenomenon. Access to administrative information by statistical authorities must be guaranteed, facilitating collaboration between administrations to prepare these statistics, respecting at all times the confidentiality of data and statistical secrecy.

A second case is the analysis of population mobility. This has been a phenomenon of great interest for the design of public policies at both municipal, regional, national and European levels. This phenomenon has usually been studied through population censuses or through specific surveys, which, although they offer high-value information, they have not provided so far a sufficient level of granularity and timeliness of their data. However, the use of big data in the statistical field as a new data source (see Galeano and Peña (2019)) allows new projects to be launched, such as the mobility study recently undertaken by the INE that offers results very detailed in a short space of time, always in aggregate, using percentages of people who leave their areas of residence towards other areas, thus enabling comparisons with the flows observed on a normal day.

A third case is made up of sustainable development indicators, which are, without doubt, the greatest challenge facing the world statistical system today. The production of indicators that covers the 17 objectives and 169 goals, together with their corresponding breakdowns, implies the availability of a wide set of data from base statistical operations and accounting systems that are subsequently presented jointly to offer an aggregate vision of the degree of compliance with the 2030 Agenda. Currently, almost half of the indicators have at least one series of data available, so that, once again, inter-institutional collaboration supported by the coordinating role of the INE is key to achieving complete all the information requested by the United Nations.

3 Conclusions

The modernization process of official statistics has been parallel to the methodological evolution of statistical science. This is perfectly reflected in the topics dealt with by *Estadística Española* since its creation in 1958. We are at a time when official statistics have to face very important challenges. Among them, perhaps the use of new data sources and the search for new forms of production are the most important. I trust that the new *Spanish Journal of Statistics* will collaborate in this task and be successful in its new phase.

References

Arnaiz Vellando, G. (1960). Information theory. *Estadíst. Española No. 9*, 5–29.

Azorín Poch, Francisco (1959). Sampling of finite populations. Systematic sampling. *Estadíst. Española No. 5*, 34–41 (1959).

Damas Rico, Pedro Manuel (1972). Introduction to the statistics of extreme values. Use in simulation methods. *Estadíst. Española* (57), 29–58.

Devaud, Denis and Yves Tillé (2019). Deville and Särndal's calibration: revisiting a 25-years-old successful optimization problem. *TEST 28*(4), 1033–1065.

Diez de Artazcoz, Vicente Jiménez (1963). The Monte Carlo method and its applications. *Estadíst. Española No. 19*, 12–30.

Escribano Ródenas, Ma. Carmen and Gabriela M. Fernández Barberis (2009). The beginnings of the official statistic in the Iberian Peninsula. *Bol. Estad. Investig. Oper. 25*(2), 129–139.

Ferrer Martín, Sebastián (1960). Theory of games. *Estadíst. Española No. 6*, 44–54.

Galeano, Pedro and Daniel Peña (2019). Data science, big data and statistics. *TEST 28*(2), 289–329.

García Villar, Jaume and Miguel Ángel De Castro Puente (2011). First World Statistics Day and the INE contribution to the achievements of official statistics in Spain. *Bol. Estad. Investig. Oper. 27*(1), 77–84.

Ibarrola Muñoz, Pilar and Vicente Quesada Paloma (1972). Neobayesian statistics. *Estadíst. Española* (56), 5–20.

Neymann, J. (1963). Two breakthroughs in the theory of statistical decision making. *Estadíst. Española No. 18*, 5–28.

Pena Trapero, Bernardo (1960). Econometric model for calculation and prediction of a total population subject to natural change and migration, by regions. *Estadíst. Española No. 9*, 47–61.

Pena Trapero, J. B. (1974). New specifications of multi-equation models: the fixed point and related estimation methods, and the iterative method of instrumental variables. *Estadíst. Española* (64–65), 17–37.

Quesada, Vicente (1972). Empirical Bayes methods applied to classification problems. *Estadíst. Española* (55), 13–22.

Ray, S. (2003). Official statistics: past, present and future. *Stat. Appl. (N.S.) 1*(1-2), 1–12.

Sánchez-Crespo, J. L. (1976). A new sampling scheme: selection with graduated variable probabilities without replacement. *Estadíst. Española* (70-71), 5–12.

SPANISH JOURNAL OF STATISTICS
Vol. **1** No. 1, 2019, Pages 9 – 11
doi: https://doi.org/10.37830/SJS.2019.1.02

EDITORIAL

Spanish Journal of Statistics: Welcome message from the new Editor-in-Chief

José María Sarabia

Department of Economics, University of Cantabria

1 Introduction

Welcome to the first issue of Spanish Journal of Statistics (SJS). This new scientific journal replaces Estadística Española, edited and published by the National Statistics Institute of Spain (INE) for more than 60 years. This journal has always been highly regarded by the Spanish scientific community and continues now with this trajectory of excellence, with a more international focus. The issues published can be found on the web-site of the SJS.

The papers to be considered in SJS must contain original theoretical contributions of direct or potential value in applications. Practical applications of methodological aspects are also welcome. The standards of innovation and impact are crucial in the papers published in SJS. Among the topics to be considered are: official Statistics; theory and methods; computation and simulation studies that develop an original methodology; critical evaluations and new applications; development, evaluation, review, and validation of statistical software and algorithms; reviews of methodological techniques; letters to the editor. All the information about the submission process can be found at: `https://www.ine.es/sjs`

2 New challenges for SJS

In the framework of today's digital society, the research agenda, both for modern statistical science and for official statistics, has substantially expanded with new topics and challenges. These new topics and aspects of statistical research include novel methodologies using different sources of information. In this sense, machine learning techniques, together with the corresponding computational tools, have successfully addressed many current data analysis problems.

To the traditional data sources of the surveys and administrative records, we have to add the widespread access to big data. Salgado (2017) discusses the challenges posed by big data within official statistics, including institutional access to data, the new statistical methodology for the treatment of this data and the changes in technological infrastructure.

The role of national statistical institutes and official statistics has become more prominent in recent years. According to Allin (2019): *National statistical offices are not only providers of statistics; they should*

also provide answers to questions raised in society, recognizing that they have to compete for the attention of users. In this context, INE has tackled three important projects within the field of experimental statistics: the household income distribution atlas, the estimate of the number of weekly deaths during the COVID-19 outbreak and the analysis of the mobility of the population during the lockdown by COVID-19. Since these three statistical experiments are part of the strategy of the European Statistical System, their quality has already been assessed by different national statistical offices, besides Eurostat.

A relevant aspect of these projects is the use of the three types of sources: surveys, administrative records and big data. In an interesting contribution, Hand (2018) identifies some statistical challenges in the field of administrative and transaction data, providing a stimulating debate on how to improve the analysis of these types of sources. Torrecilla and Romo (2018) discuss the collection, storage, preprocessing and visualization of huge batches of data, using the terminology *data learning* for this scientific discipline.

3 The new editorial board

Our new editorial board consist of an experimented team of academics and researchers around the world. We currently have associated Editors from Australia, Brazil, Canada, Chile, Spain, Switzerland and the USA. The editorial team combines both specialists in statistical methodologies and official statistics.

4 The first issue

The first issue of this volume consists of two introductory notes and five papers.

The paper “Recovering income distributions from aggregated data via micro-simulations", by Ignacio Moral-Arce, Antonio de las Heras Pérez and Stefan Sperlich proposes a new methodology for the imputation of income densities corresponding to the observed grouped data. The paper introduces a method of density estimation from grouped data. Small sample properties and two empirical examples are presented.

The next four papers are devoted to contributions presented in a session organised by the Spanish Statistical Office, at the XII Public Statistics held in Alcoy (Alicante, Spain) in September 2019. The round table was focused on the “Spanish Living Conditions Survey" (ES-SILC) integrated into the European Statistics on Income and Living Conditions (EU-SILC). First, the paper by Agustín Cañada introduces the session, explaining its objectives within the “peer review" processes of the European Union, by which official statistical systems are subject to control and supervision of the degree of compliance with the European Statistics Code of Practice.

The paper by José María Méndez describes the main characteristics of the ES-SILC and the evolution of its methodology, from the initial version entirely based on sampling surveys, to the current methodology, which combines surveys with administrative sources.

Next, the paper by Alejandro Arias, Albert Arcarons and Amparo González compiles the applications of this survey by the Office of the High Commissioner to estimate child poverty in Spain. Moreover, the authors propose some recommendations for possible improvements of the EU-SILC survey, such as the introduction of labour market trajectories or information on daily living expenses, like education expenditures.

Finally, Jorge Onrubia presents some possibilities of improving the Living Conditions Survey prepared by INE, in view of its upcoming reform, conceived within the updating process of the EU-SILC project promoted by Eurostat. The discussant reviews the inclusion of information from administrative registers, especially those from tax sources, as well as others from social security records.

References

Allin, P. (2019). Opportunities and challenges for official statistics in a digital society. *Contemporary Social Science*, DOI: 10.1080/21582041.2019.1687931.

Hand, D.J. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society A*, 181, 555–605.

Salgado, D. (2017). Big data en la Estadística Pública: retos ante los primeros pasos. (in Spanish) *Economía Industrial*, 405, 121–219.

Torrecilla, J.L., Romo, J. (2018). Data learning from big data. *Statistics and Probability Letters*, 136,15–19.

SPANISH JOURNAL OF STATISTICS
Vol. **1** No. 1 2019, Pages 13–29
doi: https://doi.org/10.37830/SJS.2019.1.03

REGULAR ARTICLE

Recovering income distributions from aggregated data via micro-simulations

Ignacio Moral-Arce[1], Antonio de las Heras Perez[2], Stefan Sperlich[3]

[1] Instituto de Estudios Fiscales - Madrid, ignacio.moral@ief.hacienda.gob.es
[2] Departamento de Economía - Universidad de Cantabria, antonio.heras@unican.es
[3] University of Geneva - GSEM, stefan.sperlich@unige.ch

Received: April 11, 2019. Accepted: January 9, 2020.

Abstract: For the studies of wealth, inequality and poverty, the analysis of income distribution of the individuals is a crucial issue. In practice, however, only aggregated data are available, either in groups or as a few quantiles of the distribution. To perform counterfactual exercises, it is desirable to generate samples of micro income data corresponding to the same population structure. This method serves also for the imputation of income densities corresponding to the observed grouped data. This work introduces a method of density estimation from grouped data. Small sample properties and two empirical examples are delivered.

Keywords: income distribution, grouped data, micro simulation, inequality, nonparametric density estimation

MSC: 62P25, 62P20, 91-08

1 Introduction

The study of income distributions is a crucial issue in the analysis of welfare, inequality and poverty, and can be a major concern for economists, governments or different international institutions. It is well known that any welfare measure and determinant of poverty or inequality can be derived from either the density or the cumulative distribution function. Besides these aspects, the calculation and simulation of actual and potential income distribution functions, respectively, as well as their evolution over time according to different scenarios, it is useful to analyze social mobility, the impact of a crisis, re-distributive policies, market opening (globalization), or poverty and inequality reduction (e.g., Fuentes (2005)). This might be of interest at different levels, may it be the regional, national or global one.

A frequently studied issue in applied economics is the calculation of income aggregate functions derived from subgroups (Griffiths et al. (2005); Chotikapanich et al. (2007)). A typical example is the estimation of the global income distribution (see Milanovic (2006); or Sala-i Martin (2006)) by integrating the income distribution of all countries. In order to compute the world income, often the

countries are considered as the units of the population. If, however, the households or citizens are the units of interest, then one has to account for the different population sizes of the countries: so it is necessary to integrate over the countries' income distributions, weighted accordingly to the population size of each country. Differently from what one can find in the literature, the information about the mean income and population size of each country is not sufficient for obtaining a reasonable distribution estimate because the disregard of the national dispersions, asymmetries, kurtosis, etc. will greatly underestimate the corresponding moments of the international income distribution. Clearly, any subsequent inference related to them, like for example the derivation of poverty or inequality measures, is then biased too. The here presented method allows for aggregation with or without any kind of weighting.

The estimation of income distribution functions essentially depends on the data available to the researchers. Such data may be obtained through various sources: administrative records, censuses, samples, surveys, panels, etc. In many cases, however, the information available to researchers is limited to grouped data or quantiles of income from household surveys or administrative records. Moreover, grouped data are the only source of information on income distributions in many countries or regions playing therefore an important role in the determination of poverty and inequality at the worldwide level. The process of assembling the data can be described as follows: income information of a large number of individuals is summarized through the use of clusters, say intervals, organized by an ascending order of income levels. This grouping may be symmetrical (referring to equidistant quantiles, i.e., the number of individuals in each of the intervals being the same), or asymmetrical (income intervals with therefore different numbers of individuals associated to each interval).

In the case of estimating the actual income distribution for each region or country of interest, one would like to have a method that allows for both, recovering the whole income distribution on the one hand but also recovering the variability or say, uncertainty of the obtained result given the lack of information when provided only with grouped data. The same is true if the objective is rather the simulation of distributions that happen to produce grouped data as those we observed. This is an essential ingredient of micro-simulation studies. In all mentioned situations the correct interpretation of 'uncertainty' depends on the underlying model or procedure used for the estimation and/or simulation. From a statistical point of view this translates to the question of whether a (pre-specified) parametric or a nonparametric distribution is considered. The choice between them depends on how researchers use the available information: in either a fixed or a more flexible manner. Nonparametric methods give more importance to the information provided purely by the data, whereas the parametric approach gives more weight to the model specification emerging from some hypotheses about the data generating process. In these cases, the estimable 'uncertainty' refers exclusively to the statistical part, i.e., the standard errors of the (few) parameters being estimated, but taking the model as being 'certain'.

The economic literature has proposed different approaches to obtain estimates of income distributions from grouped data. In the past, some of the most popular ones have been based on the parametric estimation of Lorenz curves: see Kakwani and Podder (1976) for an explicit parametric Lorenz curve estimator for grouped data; Rasche et al. (1980) for an early review; and Cheong (2002) for a more recent one.

A second approach, which is very popular in the current literature, involves the non-parametric estimation of the income distribution. It is typically just the direct estimate of a density function through the use of kernels (for details see Silverman (1986)). Like the Lorenz curve approach it can be applied to various types of research such as the study of poverty and inequality, cf. Ackland et al. (2013); Chotikapanich et al. (2007); Pinkovskiy and Sala-i Martin (2009); Minoiu and Reddy (2009); or Sala-i Martin (2006). The accuracy of the results depends essentially on the data and bandwidth used

in the calculation of the density, especially when grouped data are the source of information (Minoiu and Reddy (2014); Wu and Perloff (2003)). These non-parametric techniques perform well when the number of observations available to researchers is high. Unfortunately, in these kinds of studies, the available data are often very limited, e.g., to five figures (quintiles). This combination of "limited structure" and "limited data" produces results that are, in turn, of limited value[1] . An econometric solution to this problem are the so-called semiparametric procedures. They impose structure where prior knowledge is offered or where the impact of misspecification is less crucial, but maintain all the nonparametric flexibility elsewhere. In other words, they keep the best part of each. The aim of this paper is the simulation and estimation of income distributions on the base of grouped data which may either represent quantiles or refer to (different) income intervals. Imagine we want to estimate the income distribution of Africa but are only provided with different quantiles for each single African country. In a first step we propose to adapt a parametric regression model to the grouped data of each country. In a second step these models are used to predict (or to randomly draw if simulation is the objective) as many individual incomes as wished for each country. From these one can recover (e.g., by the use of nonparametric kernel density estimators) the income distribution for each country separately as well as the income distribution of any kind of aggregation (e.g., West-Africa). It should finally be mentioned that our procedure can certainly be used for recovering any other continuous distribution (e.g., expenditures) for which only such limited information is available.

2 Data problem and proposed method

The decision about what an appropriate method is depends crucially on how the information is available and grouped. Often researchers have data that are grouped in intervals: you may imagine different income levels of individuals in ascending order. A data source can be household surveys or administrative records. If the information originates from a survey, then the available information is typically given in quantiles, whereas in administrative files, you have prefixed income intervals that contain different numbers of individuals. A representation of grouped data can be thought of as shown in Table 1, where the x_j denote the boundaries of the income intervals. The mean income for each interval is rarely provided but if so, it could be used to improve estimation and prediction procedures. Obviously, one has equidistant quantiles if $n_j = n_k$ for all j, k, i.e., if all intervals contain the same number of individuals. In any case we can obtain some quantiles but often not equidistant ones. Interestingly, most theoretical contributions on the analysis of grouped data (need to) assume to have the information provided in equidistant quantiles. Papers that allow for asymmetric information are quite rare. For our proposal we simply assume to be provided with the information given in Table 1 for the population of interest or for each sub-population of a partition of the target population.

Income intervals	0 to x_1	x_1 to x_2	...	x_{j-1} to x_j	...	x_{J-1}	total support
Number of individuals	n_1	n_2	...	n_j	...	n_J	N
Cumul. proportion of pop.	$P_1 = \frac{n_1}{N}$	$P_1 = \frac{n_1+n_2}{N}$	...	$P_j = \frac{n_1+...+n_j}{N}$	...	$P_J = 1$	100%

Table 1: Income Grouped and Relative Accumulated Data.

[1]It should be mentioned that the existing procedures often exhibit several additional drawbacks. For example, apart from an inadequate bandwidth selection which in fact renders the estimates rather incomparable than comparable, the method proposed in Sala-i Martin (2006) makes only sense when the grouped data are provided in form of quantiles, and if the true underlying density is indeed symmetric.

We consider two scenarios regarding the available information: (A) the data are census based and therefore the information on cumulative proportions p_j (or quantiles) is exact; (B) the data are only survey based and consequently subject to sampling variation. In case (A) you would like to exactly calibrate the further analysis to these cumulative proportions (quantiles), no matter how wiggly the resulting distributions look like; in case (B) you have a deconvolution problem, so you would rather prefer to smooth the income data than performing a calibration along some cumulative proportions (or quantiles) that suffer from sampling errors themselves.

As indicated in the introduction, your objectives could be various: estimate an income distribution from Table 1, simulate[2] an income distribution with proportions equal (if scenario A) or similar (if scenario B) to the observed ones. Furthermore, one might face a partition of a population in L subpopulations, being provided with some information as in Table 1 for each subpopulation k (with potentially different J_k and N_k, $k = 1, \ldots, L$).You could be interested in estimating the joint income distribution. It may be that for each problem and situation there exists one particular sophisticated optimal solution, but what we propose here is one simple and straightforward method for dealing with all these problems in a unified way.

More specifically, we propose a method to generate arbitrarily large samples whose distribution follows the distribution of the real observations to the extent they provide us with information about this distribution[3]. To keep notation simple, at this stage we neglect the subindex k ; in other words you may only want to estimate or simulate one population ($k = 1$). Ryu (1993) and Ryu and Slottje (1996) explain why estimating the inverse of the cumulated distribution of income can be done by regressing the logarithm of income x_j on p_j with zero-mean deviations u_j, i.e.,

$$log\ x_j = \sum_{m=0}^{M} \beta_m p_j^m + u_j \quad \text{with the } \ x_j, p_j \ \text{ taken from Table 1.} \tag{1}$$

Along our experience, setting $M = 3$ (if $J > 3$) gives quite satisfying results, but M can certainly be increased accordingly to the increase of J (like in the method of sieve regression). For scenario A you basically want to interpolate and choose $M = J-1$. In any of these cases the parameter estimates of β_m can be calculated by the ordinary least squares method.

The next step is to generate N observations from an income distribution that coincides with the information you have. In order to respect the income distribution according to Table 1 and equation (1), one has to take N equidistant quantiles $q_1, ..., q_N$ covering the open interval $(0, 1)$, i.e., $q_1 = 1/(N + 1) = 1 - q_N$, and generate

$$y_i = \sum_{m=0}^{M} \hat{\beta}_m q_i^m \quad \text{for } \ i = 1, ..., N. \tag{2}$$

Note that y_i are the predictions of log $x(q_i)$, where the coefficients are the estimates from regression model (1). This generates an artificial sample (or population) $\{y_i\}_{i=1,,N}$ which follows the wanted income distribution. Even if this might not be your main objective, you will see its usefulness below.

In case you are interested in the simulation of (various) populations or samples along model (1) and the information contained in the grouped data at hand, you can use a kind of wild bootstrap approach[4]. Specifically, you proceed as before but generating now

[2]This is of particular interest if you use this method in the context of micro-simulations.

[3]One may say that the simulated populations are calibrated to the observed quantiles.

[4]This idea is borrowed from resampling strategies in nonparametric statistics.

$$y_i = \sum_{m=0}^{M} \hat{\beta}_m q_i^m + v_i, \quad v_i \sim N(0, \sigma_u^2(q_i)), \text{ for } i = 1, ..., N. \tag{3}$$

That is, for each individual you add a random error v that reflects the deviation u in (1), i.e., the deviation of the model from the observed data. Like the wild bootstrap itself and discussed above, this is either done for simulation reasons or because you want to account for the sampling and modeling error, too. In practice, the variance of u also has to be estimated, and in case of heteroscedasticity even as a function of p, respectively q[5]. The data generating process (3) allows you to generate arbitrarily many populations or samples which are all different but follow in their distribution equation (1) and thus respect the information provided in Table 1.

Until now, we have proposed only relatively simple (parametric) models, because it is supposed to have only little information, say a small J. Now, if $L > 1$, then the two steps, namely (1) and (2) or (3) respectively, have to be done for each (sub)population separately, creating samples of size N_k for the k-th (sub)population, $k = 1, ..., L$. Imagine now you are also interested in the distribution of the entire population. For example, imagine you have the grouped data for each region of Spain but you are also interested in estimating the income distribution for entire Spain. Another, completely different but important example is when the information in Table 1 is stratified along some (individual) characteristics that might be important for income. Therefore, you might have the quantiles for domestic workers and immigrants separately but you need the entire income distribution. One could interpret the strata representing different subpopulations in which the population is partitioned. Certainly, the joint distribution can only be revealed if the size of each subpopulation (respectively strata) or its proportion of the total population is known. In either case, the size N_k has to be chosen accordingly to the proportions of the subpopulations, i.e., such that $N_k/N(N = N_1 + N_2 + N_3 + ... + N_L)$ is the proportion of subpopulation k in the total population.

Based on the L samples, the joint log income distribution density $f(y)$ is estimated locally at point y by a nonparametric kernel density estimator with bandwidth h and kernel $K(\cdot)$;

$$\hat{f}_h(y) = \frac{1}{hN} \sum_{i=1}^{N} K(\frac{y - y_i}{h}). \tag{4}$$

For details on non-parametric kernel density estimation, see Silverman (1986). The choice of the kernel is unimportant but not so the choice of the bandwidth, see Härdle et al. (2004). There exist many selection methods, see Heidenreich et al. (2013) for a recent review. Today, almost all statistic or econometric software packages provide this estimator as a standard routine, including an automatic choice of h. If wanted, you can also estimate a density for each subpopulation k separately, simply by using $N = N_k$ (adapting h accordingly) for each.

3 Method check by Monte Carlo Simulations

The following non-negative distributions are considered: log-normal, Weibull, generalized Gamma and the Beta distribution. These are some of the most commonly used when modeling income distributions, see Minoiu and Reddy (2009, 2014). The first goal is to see whether our finally resulting distribution estimator fits well the true underlying distribution. This is achieved by calculating the

[5]In our simulations and our applications we use an ordinary least square regression of $\hat{u}_2 = \gamma_0 + \gamma_1 p + \gamma_2 p^2 + \epsilon$ but you may use any existing method for estimating scedasticity functions.

mean, standard deviation and deciles but later on also by looking at figures of confidence intervals. The study works as follows:

1. A sample of observations (of size 4000) is drawn accordingly to the underlying density function (log-normal, Weibull...): $x_1, x_2, x_3, ..., x_{4000}$. The information from all 4000 observations is summarized in a similar way to that of the first rows of Table 1.
2. Using only the figures of that table, the density is calculated as in (4) with either predictions as in (2) or simulations as in (3) using $M = 3$, $N = 4000$, the kernel $K(\cdot)$ being the standard normal density, and the bandwidth of Park and Marron (1990)[6].
3. Several descriptive statistical measures of the estimated density function are calculated and compared to the actual values of the original data generating density.

This was repeated 1000 times. The averages of the results are shown in Table 2. The quantities represent the ratio between the estimated and true values. The accuracy of our method is quite high except for some values of the Weibull distribution.

In addition to the comparison of position and dispersion measures, the adjustment of our estimator versus the underlying function is illustrated in Figures 1 and 2 which show the 95 simulated confidence intervals (SCI) of the density estimates together with the true data generating one. The solid lines represent the true density functions.

Statistics	**log-normal**	**beta**	**weibull**	**gamma**
mean	1.0016	0.9926	1.0019	1.0162
Std. Deviation	0.9855	0.9598	1.0660	1.0325
Deciles				
0.1	1.0176	1.0045	1.0919	1.2079
0.2	1.0058	0.9967	1.0298	1.0554
0.3	1.0027	0.9898	0.9633	1.0223
0.4	1.0008	0.9860	0.9292	0.9961
median	1.0039	0.9761	0.9021	0.9652
0.6	0.9937	0.9720	0.9023	0.9508
0.7	0.9876	0.9895	0.8949	0.9491
0.8	0.9901	0.9783	0.9128	0.9520
0.9	1.0024	0.9896	0.9831	0.9947

Table 2: Statistical Summary with estimated values divided by true value.

These figures confirm the results in Table 2. The first conclusion from these figures is the good fit of our estimation method. The adjustment on the Weibull distributions and log-normal is very high,

[6]The objective when choosing a bandwidth h is to minimize the mean integrated squared error (MISE):

$$MISE(\hat{f}_h) = \int E\{\hat{f}_h(x) - f(x)\}^2 \, dx \approx \frac{1}{Nh}\|K\|_2^2 + \frac{h^4}{4}\{\mu_2(K)\}\|f''\|_2^2$$

where the approximation holds as h goes to zero, N and Nh to infinity. Minimization with respect to h gives:

$$h_{opt} = \left(\frac{\|K\|_2^2}{\|f''\|_2^2\{\mu_2(K)\}^2 n}\right)^{1/5}.$$

The terms $\|K\|_2^2$ and $\{\mu_2(K)\}^2$ are constants depending only on the kernel function, and are therefore known. However, although $\|f''\|_2^2$ denotes a constant, it depends on the second derivative of the unknown density f. Park and Marron (1990) estimate it by $\frac{1}{Ng^3}\sum_{i=1}^n K''\left(\frac{x-X_i}{g}\right)$. They propose an optimal g and a bias correction for $\|f''\|_2^2$.

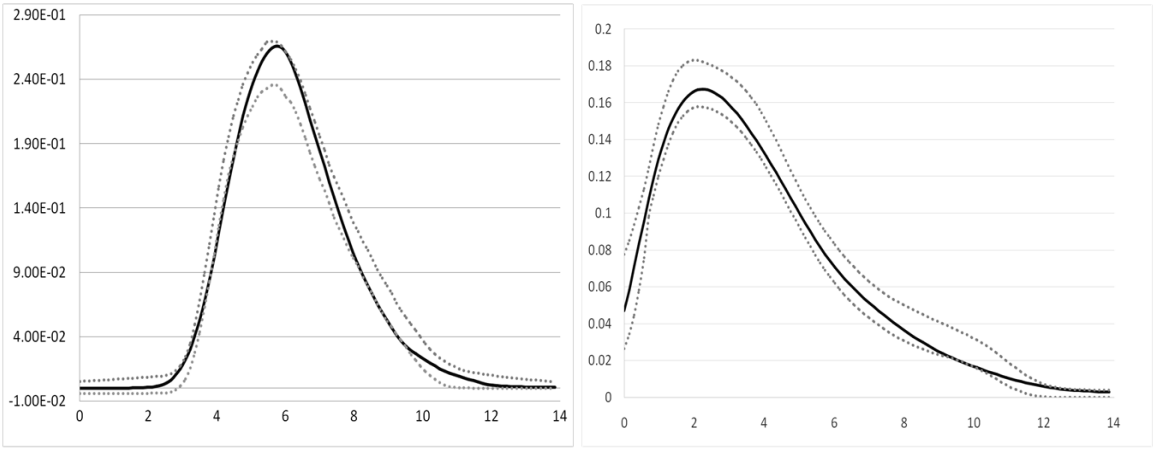


Figure 1: True (solid) and 95% SCI of density estimates (dashed) for the Log-Normal (left) and the Weibull (right) distribution.

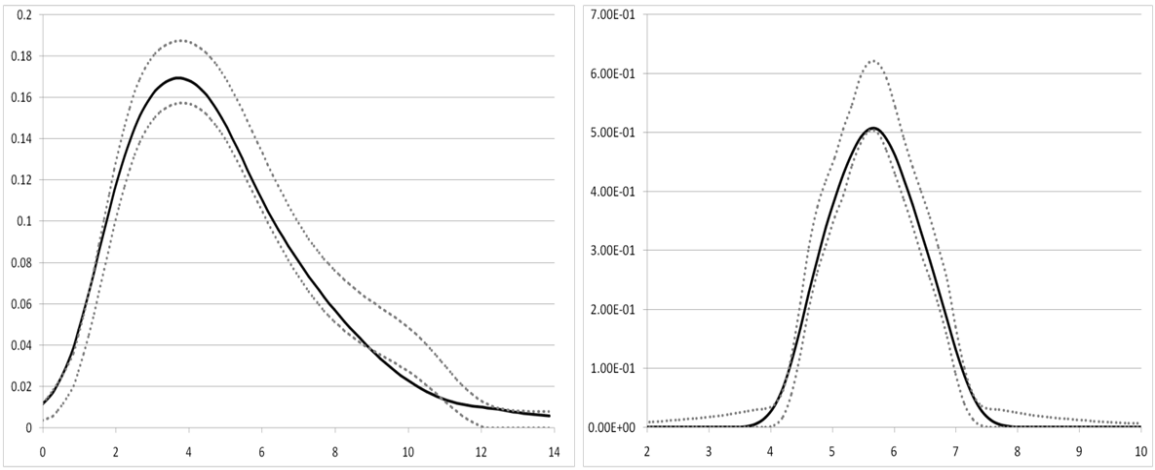


Figure 2: True (solid) and 95% SCI of density estimates (dashed) for the generalized Gamma (left) and the Beta (right) distribution.

the bias can be considered negligible. The asymmetric Gamma distribution presents a good fit with some bias in the right tail of the distribution (similar to the Weibull distribution). The reason lies in the fact that the standard kernel density estimators suffer from a boundary effect in two ways: Case 1 (the standard boundary effect of kernel estimators) occurs when the true density has a boundary, say 0 on the left hand side, and we have some data y_i very close to zero, say $y_i < \varepsilon$. Then a density estimator with bandwidth h predicts a positive density around $\varepsilon - h$ even if this is smaller than zero, i.e., falls outside the true support. This explains why the estimates for the Beta distribution have heavier tails than they should. Case 2: A problem that occurs with long tails when only quantiles are given is that the kernel density must integrate to one but can't predict a positive density outside the interval $(y_{min} - h, y_{max} + h)$. Moreover, the only information we get for the last quantile is its starting point but not its end. When using equation (2), then the density estimator will be zero for values larger than $y_N + h$ and pass all the mass of the last quantile to the interval $(yN - h, yN + h)$. This

produces the upward biases around the value 10 when the true density was Weibull or generalized Gamma.

It is clear that our method is consistent for J going to infinity. But as it has been developed right for the situation where J is mall, such kind of convergence study is irrelevant. However, it could be interesting to see, whether and how the method improves for increasing sample sizes n and N. To this end, 400 random samples of size $n = 250, 500, 750, 1000, ..., 7000$ of a Gamma distribution have been drawn. Let $\hat{f}^{(j)}$ be the two-step estimator of the density f from above of the j-th sample. The measures of discrepancy we consider are the squared expected average deviance (SAvD), the average variance (AvV) and their sum (SsD), namely

$$SAvD_n(\hat{f}) = \left(\frac{1}{n} \sum_{i=1}^{n} \left\{ \left[\frac{1}{400} \sum_{j=1}^{400} \hat{f}^{(j)}(X_i) \right] - f(X_i) \right\} \right)^2$$

$$AvV_n(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{400} \sum_{j=1}^{400} \left\{ \hat{f}^{(j)}(X_i) - \left[\frac{1}{400} \sum_{j=1}^{400} \hat{f}^{(j)}(X_i) \right] \right\}^2$$

$$SsD_n(\hat{f}) = SAvD_n(\hat{f}) + AvV_n(\hat{f})$$

Using the Gaussian kernel and the bandwidth of Park and Marron (1990) in the estimation, these values are calculated for different sample sizes n. The results are shown in Figure 3. A bit surprisingly, the values of these quantities tend to zero as the sample size increase, but with $J = 10$ constant. This is certainly excellent news.

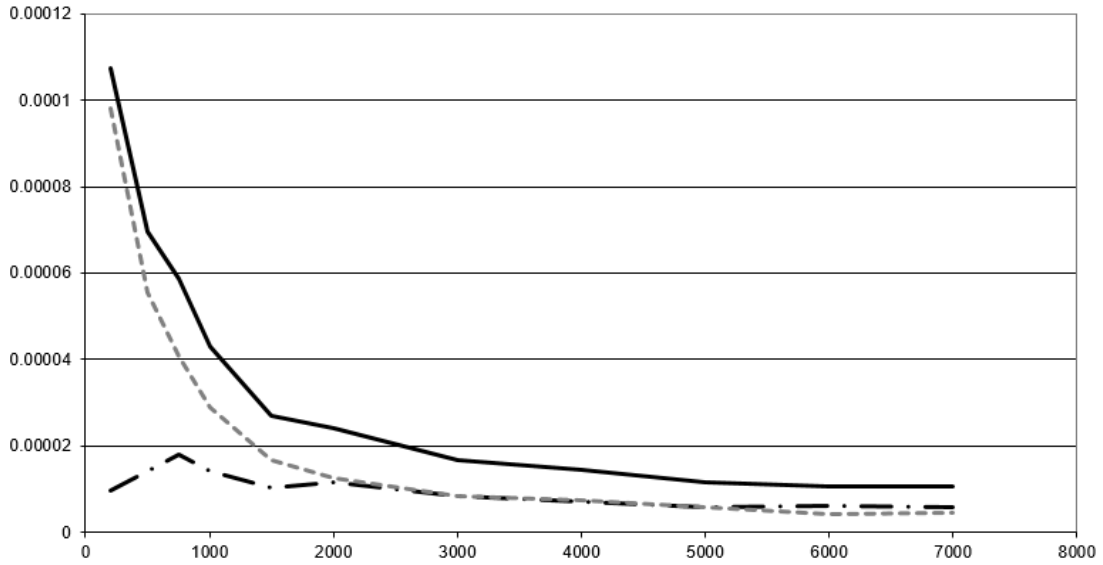


Figure 3: SsD (solid line), Average Variance (dashed grey line) and squared average deviance for increasing sample size when the true density is a Gamma.

4 Two empirical examples

In this section the focus is on the degree of adaptability of our estimation method to any kind of grouped data, avoiding the problems highlighted in the introduction. To this aim we consider two data examples for which we can (at least partly) counter-check the results we obtain from our method. The first example looks at recovering the income distributions and inequality measures for EU member states, and the second at recovering the income distribution from Spain when we are only provided with income quantiles from the various regions.

We start with considering data that are grouped into deciles, provided for the income distributions of the member states of the European Union before the big enlargement in 2001. So we only use information given in Table A1 in the appendix. Based on these symmetrically grouped data we recover the individual and the joint income distribution of the 15 states, and derive various inequality and poverty measures. The density of each country is obtained by using our two-step estimation method: the first step estimates equation (1) and draws samples from (3). The second step is the non-parametric estimation of the income density function based on the generated fictitious samples for each country. In the first step we apply a third grade polynomial ($M = 3$) in equation (1). The adjusted R^2s (not shown) were always higher than 0.97 indicating almost perfect calibration. All calculations of the second stage are performed with the Gaussian kernel and the bandwidth of Park and Marron (1990). Consequently, each country has a different data-adaptive bandwidth.

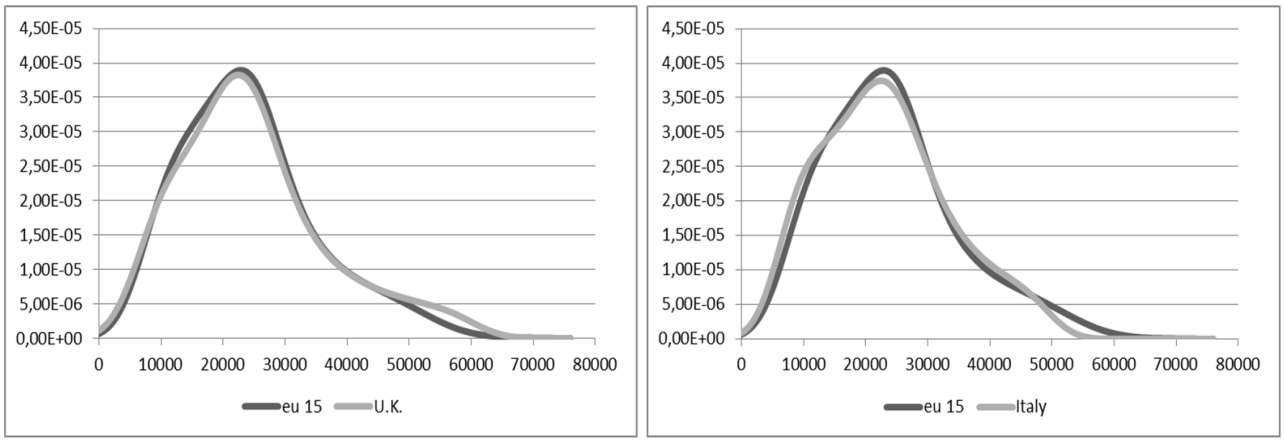


Figure 4: Density Function Estimations of EU Countries and U.K. (left) and Italy (right) in 2001.

Figures 4 to 11 show the corresponding income densities for the considered 15 EU members in 2001 together with the aggregated one. Some countries' distribution is very close to the joint income distribution like for the UK, Italy, Belgium, Netherlands, France and Finland (Figures 4 to 6); some are more concentrated on the left though with long tails on the right such as for Spain, Portugal and Greece (Figures 7 and 8); and finally we have distributions shifted to the right like for Austria, or generally more spread (Figure 9 to 11) such as for Luxembourg. Actually, Greece and Luxembourg are those that reflect the most opposite figures: The minimum modal value of the distributions is the Greek one with a value around 10,500 Euros, while the maximum mode belongs to Luxembourg with a value of about 42,000 Euros.

Among them, Germany exhibits a very narrow but large middle class. Greece, Portugal and Spain have two characteristics in their income distributions: they are the most asymmetrical ones with a significant tail on the right side. In addition, having the smallest modal values reflects that

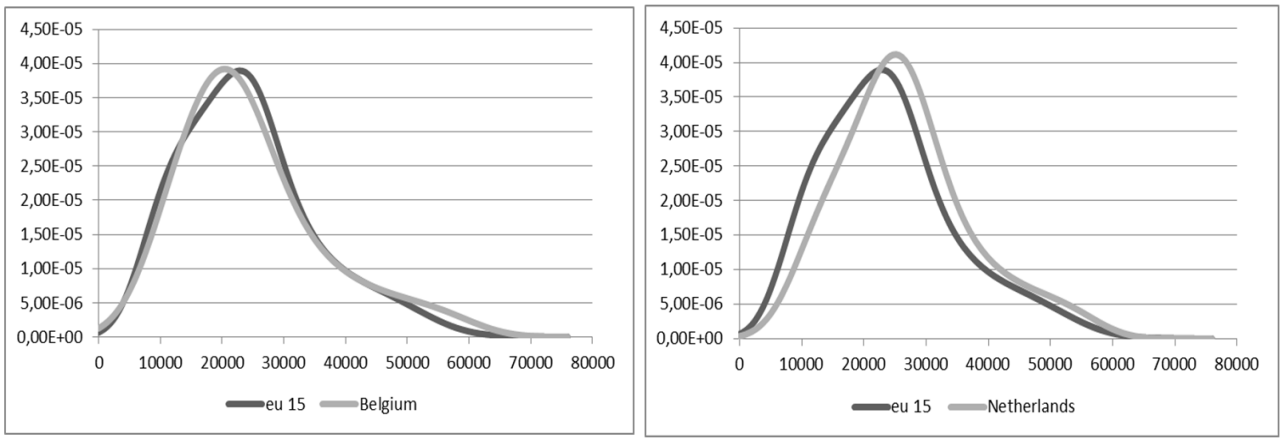


Figure 5: Density Function Estimations of EU Countries and Belgium (left) and Netherlands (right) in 2001.

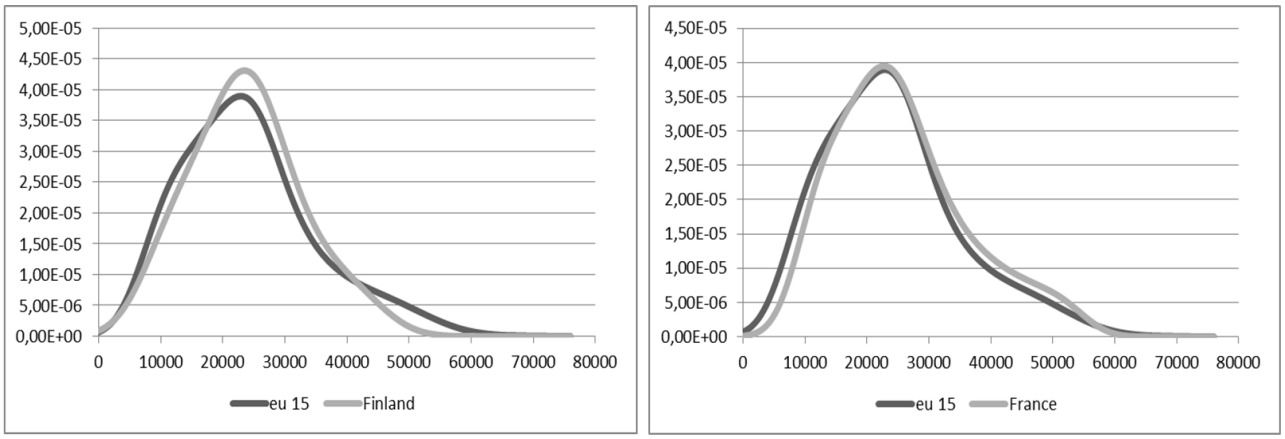


Figure 6: Density Function Estimations of EU Countries and Finland (left) and France (right) in 2001.

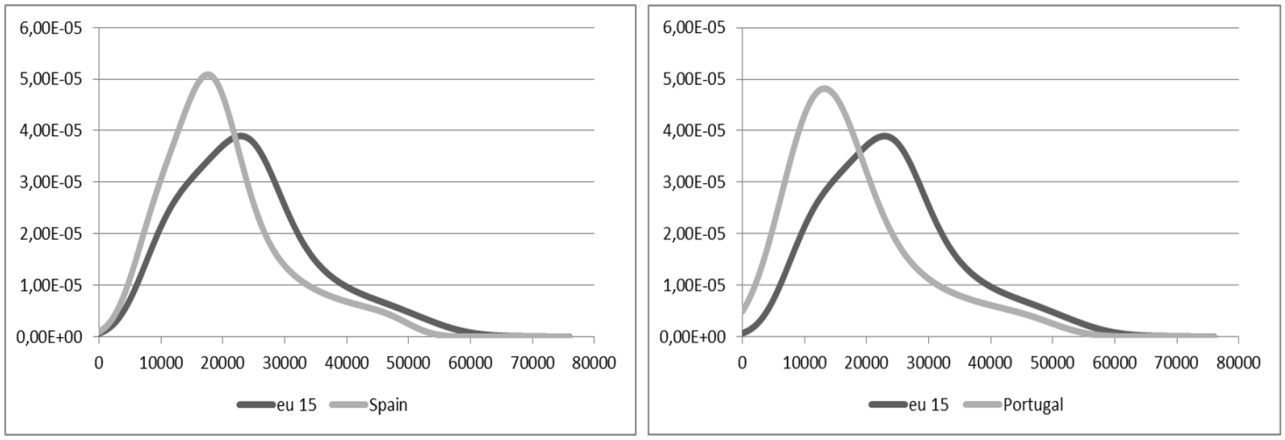


Figure 7: Density Function Estimations of EU Countries and Spain (left) and Portugal (right) in 2001.

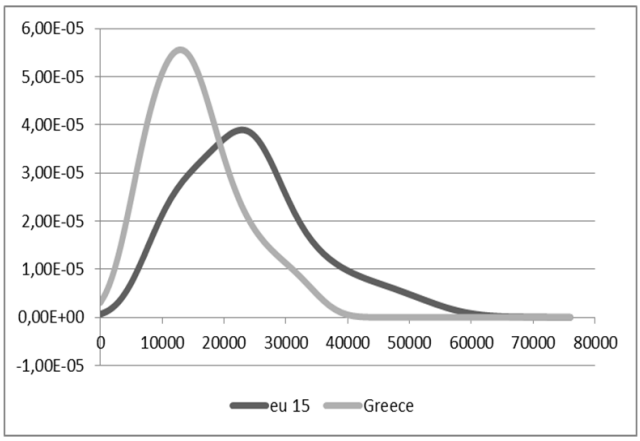
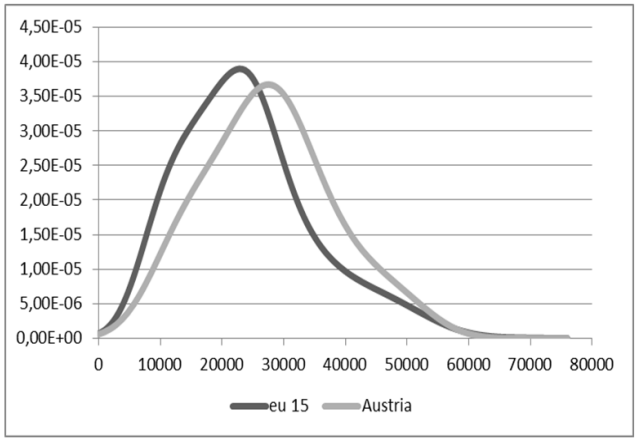


Figure 8: Density Function Estimations of EU Countries and Greece in 2001.



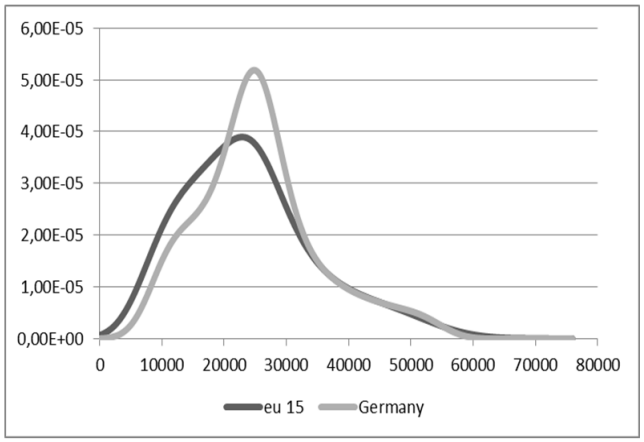
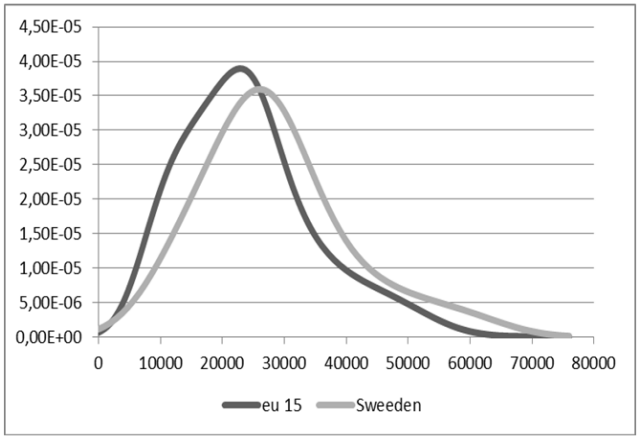


Figure 9: Density Function Estimations of EU Countries and Austria (left) and Germany (right) in 2001.



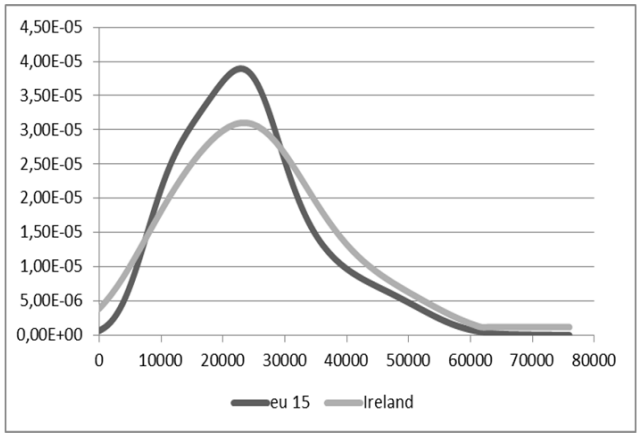


Figure 10: Density Function Estimations of EU Countries and Sweden (left) and Ireland (right) in 2001.

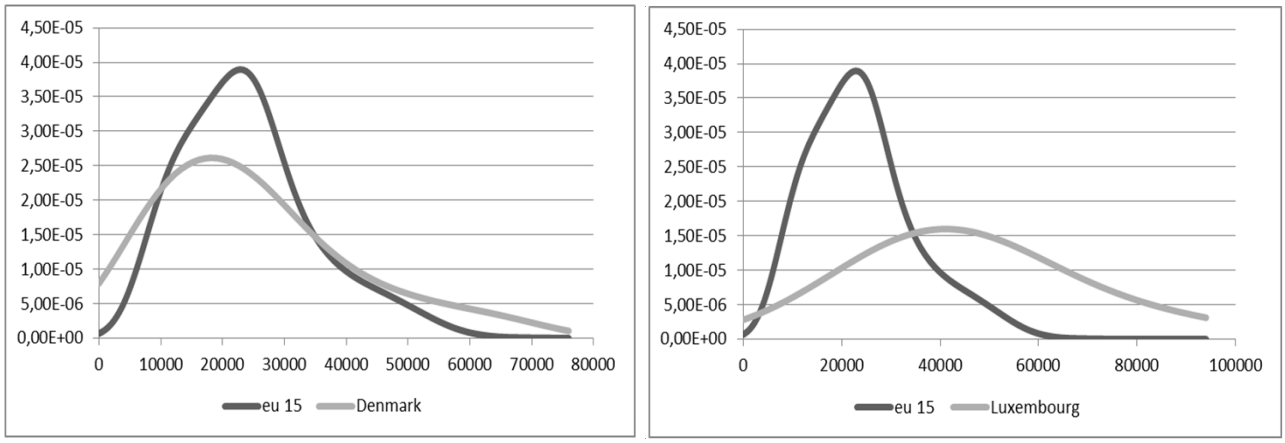


Figure 11: Density Function Estimations of EU Countries and Denmark (left) and Luxembourg (right) in 2001.

these countries have the lowest income level and the highest inequality. On the other end, for Sweden, Denmark, Austria and Germany, our method detects the most equitable income (i.e., the most symmetrical distributions) and the highest mean.

Different measures of poverty and inequality are calculated, see Table 3. For estimating the poverty rates, we chose the threshold which is the most frequently used by Eurostat, i.e., 60% of the median of the households' disposable income. In the estimation of Atkinson's index, we set the aversion parameter equal to 0.5. Note that the Gini indexes calculated by our method are quite similar to those presented by the European Commission for 2001 (Eurostat, 2005). Also for the other indexes, there is a clear consistency with those values published by that reference.

The inequality values like the Gini support the above comments on the shape of the density functions. The smallest values of inequality refer to Nordic and Central European countries; Austria, Germany, the Netherlands, Denmark, Finland and Sweden. On the contrary, in the countries of the Mediterranean area (Spain, Portugal, Greece and Italy) the Gini is substantially higher. The same can be said in the case of inequality values measured in terms of Atkinson indexes. Similarly, relative poverty, i.e., when using the European threshold, had the lowest values in Central and Northern Europe, namely Austria, Germany, the Netherlands, Finland, Sweden and Denmark. The highest values of relative poverty could be found again in Portugal, Spain and Greece, but also for Belgium and the United Kingdom, which had higher levels of average and median income but high inequality.

In our second example we are provided with asymmetrically grouped data from tax records of the Spanish Tax Agency (AEAT, Table A2 in the appendix) for each region (Comunidad Autónoma, CA henceforth) separately. This information was used to impute the income distributions in each CA and for entire Spain. We focused on Spain's 2003 tax information on the common fiscal territory. The key feature making this example different from the previous one was that this information is available only in asymmetrical income intervals, so it is not possible to directly estimate the density function based on quantiles, as done for example in Sala-i Martin (2006).

We need to make two assumptions: firstly, the "taxable income" of individuals is a good proxy of disposable income before income tax; and secondly, the number of claimants in income tax is a good proxy for the number of "individuals" in each interval. The latter assumption is less obvious since the income tax return can be personal or not and therefore the AEAT does not provide the actual number of "individuals" in each interval. However, the number of tax returns will be treated like the number of individuals. It is clear that the "taxable income" is not the equivalent to the "gross

	Population	Mean	Median	<60% Med	Poverty gap	Atkinson	Our Gini	Gini ES*
Austria	7,764	27,591.2	25,125.3	6.144	0.088	0.037	0.232	0.24
Belgium	9,555	26,060.7	23,311.2	16.285	0.183	0.054	0.300	0.28
Finland	4,963	23,242.3	21,018.9	6.226	0.094	0.037	0.233	0.27
France	55,868	26,041.7	23,306.0	13.050	0.126	0.049	0.271	0.27
Germany	76,272	26,515.3	24,479.8	8.518	0.135	0.036	0.249	0.25
Greece	10,337	14,548.5	12,276.5	14.434	0.162	0.071	0.322	0.33
Ireland	3,622	26,656.9	22,052.5	11.320	0.126	0.059	0.293	0.29
Italy	54,672	23,263.7	19,905.3	11.103	0.105	0.061	0.289	0.29
Luxembourg	455	45,175.9	41,949.0	10.549	0.132	0.037	0.263	0.27
Netherlands	14,910	27,472.9	25,179.6	10.718	0.134	0.039	0.252	0.27
Portugal	9,330	18,605.2	15,402.5	21.683	0.242	0.088	0.377	0.37
Spain	37,315	20,891.2	18,456.2	16.530	0.185	0.060	0.316	0.33
U.K.	54,503	26,020.8	22,647.2	14.355	0.183	0.063	0.311	0.35
Denmark	8,295	27,191.9	20,368.2	12.168	0.171	0.031	0.228	0.22
Sweden	4,975	30,139.8	28,334.4	10.794	0.162	0.038	0.264	0.24

Table 3: Measures of poverty in the considered 15 UE countries in 2001. * ES = Eurostat: Differences between our estimates and that of ES might be due to the different income concept used by Eurostat (disposable family income), equivalent in terms of national accounts to the income account of institutional households, while the concept used here is an income equal to GDP, see also Milanovic (2006). Further, note that the Eurostat indexes are just estimates, typically based on samples and certain assumptions on the distribution.

income" available to households. However, this fact is irrelevant for the goal pursued by this study, but can produce negative incomes, cf. Ayala and Onrubia (2001).

For the sake of brevity we skip the presentation of the densities for the 16 CAs and concentrate directly on the second goal of this application: the problem of generating the income aggregate from subgroups, i.e., the estimation of the Spanish national income distribution by integrating the income distributions of CAs. In practice, this is especially interesting for (world) regions where direct information about the aggregated area is not available. In our illustration, however, we have this direct information (the deciles for entire Spain, first line of Table A2) so that we can compare the density estimates that result from our aggregation method when using only the quantiles of the CAs with an estimator based on the quantiles for entire Spain. The fact that both estimates, shown in Figure 12, are virtually identical proves that our aggregation method of the regional information works pretty well. Note that N_k was set for each CA k equal to the number given in the last column of Table A2 as this corresponds to its proportion of the entire population.

Our (aggregation) method works even if the available information is different for each region (symmetric for some, asymmetric for others, different quantiles, different income intervals, etc.); actually, in this example we did not use the fact that all CAs provided their information for the same income intervals. Take as a different example the case where you want to calculate the joint income distribution for West Africa. For each country the information is provided in different terms. While this would create a problem for all the other presently existing nonparametric density estimation methods, our method can be applied straightforwardly. Obviously, the same holds true for calculating the world income distribution.

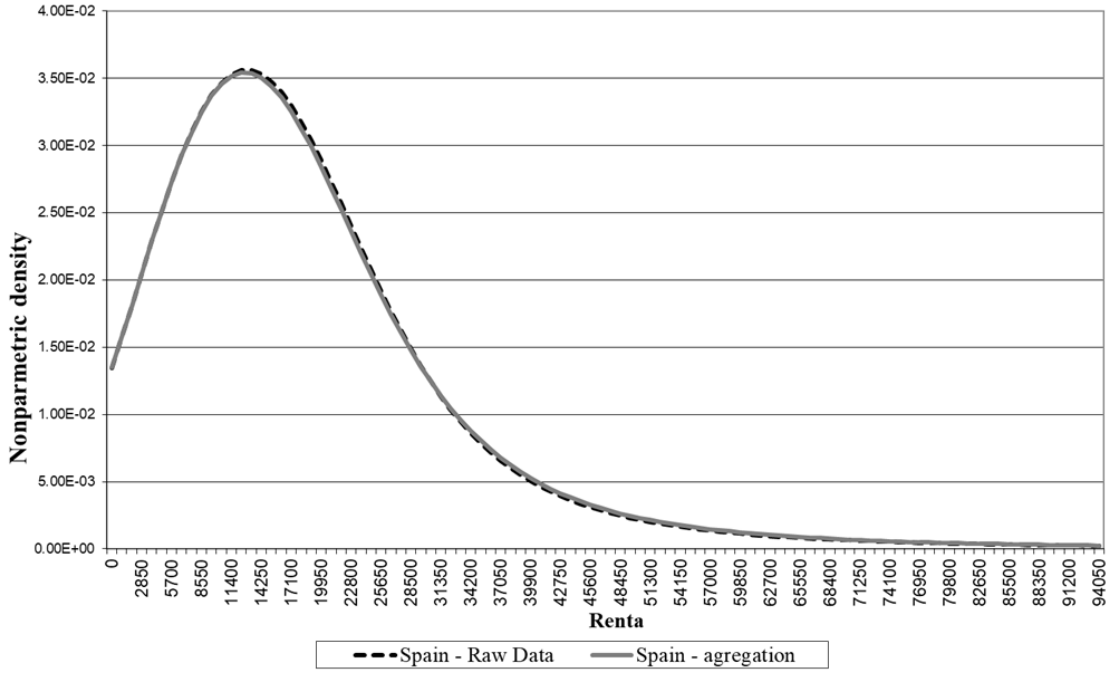


Figure 12: Comparison- Direct estimation of National Income Distribution âĂŞ Raw data - vs estimation through Aggregation of Regional Income Distributions.

5 A Nonparametric alternative and Conclusions

Readers that are more familiar with complex nonparametric estimation problems might, at least at a first glimpse, feel uncomfortable with the idea of first estimate the log-income almost parametrically, generate data from that model to use a nonparametric kernel estimator afterwards. We say here "almost" because it is open to the practitioner to replace (1) by an arbitrarily complex regression model. The important point is here, however, that this is a method for grouped data, and especially when only few information is available (typically not more than percentiles, so maybe 10 points but often even less). Directly applying a kernel estimator without further information does obviously not make much sense then.

An alternative way, though quite technical, is sketched in Dai et al. (2013). They apply spline regression to get an unrestricted estimate of the first derivative of the Lorenz curve. This is used to derive a convex estimate of the Lorenz curve along the steps of Birke and Dette (2007). It is well known how to calculate then the income distribution or various interesting derivatives like e.g., the Gini coefficient. Although the procedure looks quite elegant as it is based on a persistently nonparametric procedure, it has to be admitted that it is also somehow cumbersome. First we use the spline estimator of a derivative from very few data, followed by a kernel smoothing over the predictions obtained from this estimator, a numerical integration over the kernel, then a numerical inversion, and finally another numerical integration of that inverse. Thanks to today's computer and software facilities the procedure has proven to be quite stable and fast (given the few data points), but still strongly dependent on the choice of the spline smoothing method. In practice it does unfortunately not provide an improvement compared to the here presented simple method. Finally, for the calculation of income functions of merged populations one would need to develop another method to obtain the weighted average of the density estimates.

Here we have presented an easy-to-handle method for micro-simulations to recover income distributions from grouped data even when only (very) few data points are available. As has been seen, the extension to also obtain corresponding distributions of merged populations like e.g., the one for the EU calculated from quintiles of its member states is straight forward. The method is particularly helpful for countries or years for which more detailed information (e.g., micro data) is rarely available. The excellent performance of the method has been proven in simulations, and its practical use has been illustrated in two application examples.

References

Ackland, R., S. Dowrick, and B. Freyens (2013). Measuring global poverty: why ppp methods matter. *Review of Economics and Statistics 95*(3), 813–824.

Ayala, L. and J. Onrubia (2001). La distribución de la renta en españa según datos fiscales. *Papeles de Economía 88*, 89–112.

Birke, M. and H. Dette (2007). Estimating a convex function in nonparametric regression. *Scandinavian Journal of Statistics 34*(2), 384–404.

Cheong, K.S. (2002). A comparison of alternative functional forms for parametric estimation of the lorenz curve. *Applied economics letters 9*(3), 171–176.

Chotikapanich, D., W.E. Griffiths, and D.S. Prasada Rao (2007). Estimating and combining national income distribution using limited data. *Journal of Business & Economic Statistics 25*(1), 97–109.

Dai, J., I. Moral-Arce, and S. Sperlich (2013). Calibrated estimation of a nonparametric income distribution from a few percentiles. In *Proceedings 59th ISI World Statistics Congress, HongKong*, pp. 4352–4357. ISI.

Eurostat (2005). Regional indicators to reflect social exclusion and poverty. Technical report.

Fuentes, R. (2005). Poverty, pro-poor growth and simulated inequality reduction. Technical Report occasional paper no. 11, Human development report office.

Griffiths, W.E., D. Chotikapanich, and D.S. Prasada Rao (2005). Averaging income distributions. *Bulletin of Economic Research 57*(4), 347–367.

Härdle, W., M. Müller, S. Sperlich, and A. Werwatz (2004). *Nonparametric and Semiparametric Models*. Berlin, Heidelberg: Springer Verlag.

Heidenreich, N.B., A. Schindler, and S. Sperlich (2013). Bandwidth selection methods for kernel density estimation: a review of fully automatic selectors. *AStA - Advances in Statistical Analysis 97*(4), 403–433.

Heston, A., R. Summers, and B. Aten (2005). Penn World Tables. Technical report, University of Pennsylvania.

Kakwani, N.C. and N. Podder (1976). Efficient estimation of the lorenz curve and associated inequality measures from grouped observations. *Econometrica 44*(1), 137–148.

Milanovic, B. (2006). Global income inequality: A review. *World Economics Journal 7*, 131–157.

Minoiu, C. and S. Reddy (2009). The estimation of poverty and inequality through parametric estimation of lorenz curves: an evaluation. *Journal of income distribution 18*(2), 160–178.

Minoiu, C. and S. Reddy (2014). Kernel density estimation on grouped data: the case of poverty assessment. *Journal of economic inequality 12*(2), 163–189.

Park, U. and J.S. Marron (1990). Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association 85*(409), 66–72.

Pinkovskiy, M. and X. Sala-i Martin (2009). Parametric estimation of the world distribution income. Technical Report 15433, NBER working paper.

Rasche, R.H., J. Gaffney, A.Y.C. Koo, and N. Obst (1980). Functional forms for estimating the lorenz curve. *Econometrica 48*(4), 1061–1062.

Ryu, H.K. (1993). Maximum entropy estimation of density and regression functions. *Journal of econometrics 56*(3), 379–440.

Ryu, H.K. and D.J. Slottje (1996). Two flexible functional form approaches for approximating the lorenz curve. *Journal of econometrics 72*(1–2), 251–274.

Sala-i Martin, X. (2006). The world distribution of income: falling poverty and convergence period. *Quarterly Journal of Economics 121*(2), 351–397.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman & Hall/CRC.

Wu, X. and J. Perloff (2003). Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics 115*(2), 347–354.

Appendix

	Pop. in T	GDP p.c.	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Austria	8096.25	26,999.77	4.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00	14.00	19.00
Belgium	10303.88	24,661.91	4.00	5.00	6.00	7.00	8.00	9.00	10.00	12.00	14.50	24.50
Finland	5176.53	22,740.69	4.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00	13.50	19.50
France	59278.01	25,044.54	4.00	5.00	7.00	7.00	8.00	10.00	11.00	12.00	14.50	21.50
Germany	82344.43	25,061.34	4.00	6.00	7.00	8.00	9.00	9.00	10.00	12.00	14.00	21.00
Greece	10975.02	13,982.39	3.00	4.00	6.00	7.00	8.00	9.00	11.00	13.00	15.50	23.50
Ireland	3801.38	24,947.55	3.00	5.00	6.00	7.00	9.00	10.00	11.00	12.00	15.00	22.00
Italy	57714.84	22,487.21	3.00	5.00	6.00	7.00	9.00	1.00	11.00	13.00	14.50	21.50
Luxembourg	435.23	48,217.27	4.00	6.00	7.00	7.00	8.00	9.00	11.00	12.00	14.50	21.50
Netherlands	15897.51	26,293.09	4.00	6.00	7.00	8.00	8.00	9.00	11.00	12.00	14.00	21.00
Portugal	10225.09	17,323.14	3.00	4.00	5.00	6.00	7.00	8.00	10.00	12.50	15.50	29.00
Spain	40717.22	19,536.38	3.33	4.90	5.96	6.94	7.90	8.95	10.22	11.95	14.58	25.27
U. K.	58669.74	24,666.41	3.00	5.00	6.00	7.00	8.00	9.00	11.00	12.00	15.00	24.00
Denmark	8900.87	25,860.69	1.70	3.70	4.60	5.80	7.10	8.70	10.90	13.60	16.60	27.30
Sweden	5359.98	28,551.14	4.10	5.90	6.70	7.50	8.50	9.30	10.20	11.50	13.50	22.80

Table A1: Used data from the EU in 2001. Obtained from www.wider.unu.edu/research/database "The world income inequality data base" and from the Penn Word Tables 3.1 on pwt.econ.upenn.edu, (Heston et al., 2005)

Region (CAs)	- 1.5	1.5 - 6	6 - 12	12 - 21	21 - 30	30 - 60	60 - 150	>150	Total
España	919,000	2,737,612	4,174,720	4,223,910	2,085,731	1,444,135	307,429	41,325	15,933,862
Andalucía	194,168	513,390	720,178	642,465	302,188	181,310	30,256	3,285	2,587,240
Aragón	34,586	108,684	145,811	170,768	80,111	51,807	9,170	1,039	601,976
Asturias	33,105	76,828	105,730	131,949	72,023	39,760	5,924	789	466,108
I. Baleares	16,194	63,958	116,142	98,130	43,418	33,456	7,720	1,086	380,104
Canarias	33,153	109,819	178,811	150,878	79,798	52,630	9,804	1,321	616,214
Cantabria	14,660	36,523	60,421	65,824	31,641	20,294	3,872	443	233,678
C. La Mancha	45,973	148,669	208,889	170,952	70,195	41,557	6,205	586	693,026
C y León	69,260	206,408	279,787	280,044	133,233	82,225	11,841	1,044	1,063,842
Cataluña	121,560	417,572	694,504	877,293	443,030	318,902	80,146	10,465	2,963,472
C. Valenciana	109,833	353,501	539,153	466,755	207,752	138,361	27,430	3,416	1,846,201
Extremadura	32,380	95,002	117,184	82,061	35,751	20,368	3,216	223	386,185
Galicia	74,966	219,536	286,590	237,712	112,262	70,355	12,057	1,559	1,015,037
La Rioja	7,983	23,433	37,360	38,937	16,561	11,232	2,111	227	137,844
Madrid	102,772	279,369	548,049	696,341	407,189	350,849	92,271	15,205	2,492,045
Murcia	28,407	84,920	136,111	113,801	50,579	31,029	5,406	637	450,890

Table A2: Numbers of Taxpayers (www.aeat.es)

Official statistics

Round-table session: ES-SILC/ EU-SILC. XII Public Statistics Conference (XII Jornadas de Estadística Pública). Alcoy, 6 September, 2019.

This section of the Spanish Statistical Journal is devoted to contributions presented in a round table discussion organised by the Spanish Statistical Office (INE), within the XII Public Statistics Conference (Jornadas de Estadística Pública) held in Alcoy (Alicante – España) on September 2019. The round table was focused in the "Spanish Living Conditions Survey" (ES-SILC) integrated in the European Statistics on Income and Living Conditions (EU-SILC).

Three experts participated in this session: José María Méndez, head of the survey at the INE, as the main speaker; and as discussants, Amparo González, from the Office for the Fight against Child Poverty, and Jorge Onrubia, a lecturer at the Complutense University of Madrid. Agustín Cañada, director of INE's Quality Unit, moderated the session. This section of the Journal contains the presentations made during the session by the three speakers, preceded by an introductory note from the moderator. Firstly, the paper by Agustín Cañada intends to put the session in context, explaining its objectives within the "Peer review" processes of the European Union, by which official statistical systems are subject to control and supervision of the degree of compliance with the European Statistics Code of Practice. The paper also justifies the selection of the ES-SILC (EU-SILC), under the perspective of the quality management of these statistics in the Spanish and European spheres.

The central paper by José María Méndez, describes the main characteristics of the ES-SILC and the evolution of its methodology, from the initial version entirely based on sampling surveys, to the current methodology, which combines surveys with administrative sources. It also outlines characteristics of some of the main indicators obtained from these surveys, such as the poverty line or the AROPE indicator. It concludes with a description of future changes in the methodology, agreed at the European level, an analysis that the author takes advantage of precisely to evaluate some of the suggestions made by the discussants at the round table.

Next, within the contributions of the discussants, the work of Amparo González (co-authored with Alejandro Arias and Albert Arcarons) reflects the perspective of a public body that uses the ES-SILC/EU-SILC as a support for public policies. She presents some of the comparative analyses between European countries, mostly in the field of relative poverty lines. She follows an eminently practical approach when evaluating the ES-SILC statistics, by suggesting a greater disaggregation in variables such as the inclusion of employment trajectories, or demographic data of the household members.

In the contribution of the second discussant, Jorge Onrubia, proposals are made regarding the revision of the survey that is planned for the near future: on the one hand, he make suggestions of including in the survey greater details of financial data of the families; on the other, details of the income components (social benefits and their public or private origin, compensation for dismissal, unemployment benefit, etc.). The proposals on these breakdowns are linked to the suggestion that the survey should integrate further information from administrative registers, such as that from tax sources.

Spanish Journal of Statistics
Vol. 1 No. 1 2019, Pages 33–40
doi: https://doi.org/10.37830/SJS.2019.1.04

Official Statistics

Peer Reviews of the European Statistics and the Involvement of users in the Quality assurance of official statistics: An example focused in the Spanish Living Conditions Survey

Agustín Cañada Martínez
Quality Unit Director, Instituto Nacional de Estadística

Abstract: Nowadays, the national statistical offices are moving a step forward from traditional quality assessment done by the offices themselves, towards a more complete quality system that involves external experts and other stakeholders. The INE has launched several initiatives in this context, one of them being the one described in this section of the SSJ: the organization of seminars with experts (researchers and academia) focused in a specific statistic. In this paper, this initiative is described, by putting it in the context of the European Union's Peer Review, processes for auditing the quality of the activity of the statistical Offices. The article also justifies the selection of the "Spanish Living Conditions Survey" (ES-SILC) integrated in the European Statistics on Income and Living Conditions (EU-SILC) for this action, based on the inherent importance of the subjects studied by this statistic (income inequality, poverty...), and on the attention paid to the quality assessment of their data, both in the European and the Spanish environments.

Keywords: European Statistics Code of Practice; peer review; role of users and stakeholders in quality management; living conditions surveys

MSC: 62–02, 62–06, 62P25

1 Introduction

One of the guiding principles of INE's work is to serve its users, which is today the fundamental yardstick for judging the quality of a statistical institution and its products. In recent years, however, INE and the producers of official statistics have been confronted with a paradox: precisely to benefit users, they have been given free access to statistical information. However, this ease of access, which is considered a quality feature, means that at the same time, the NSIs are finding it increasingly difficult to get to know their users, to contact them and, in short, to gather their opinions and needs.

The INE of Spain - like other statistical offices - tries to overcome these limitations by setting up mechanisms and tools to assess user satisfaction with its products and services, e.g. user surveys or the procedures for drawing up statistical plans and programmes.

The INE has launched several initiatives in this context, one of them being the organization of seminars with experts (researchers and academia) focused in a specific statistic, within the yearly "Public Statistics Conference", jointly organised by INE and a scientific academic association. In the last of these conferences the round table was focused in the Spanish Living Conditions Survey (Spanish version of European Union Statistics on Income and Living Conditions (EU-SILC)).

In this paper, this initiative is described, putting it in the context of the European Union's "Peer Review" an auditing process of the degree of compliance of the NSI with the European Statistics Code of Practice. During the last round of those PR exercises, reviewers call for a greater involvement of user and stakeholders in the activity of INE; thus, one of the action launched by the INE to respond to this recommendation, was the organisation of the aforementioned round table.

The paper is organised as follows: section 2 describes, as a frame for the round-table session, the objectives and scope of the European Union's Peer Review methodology; section 3 outlines the main elements of INE's quality management system, emphasising procedures for including external users and other stakeholders in the assessment of INE's products, such as the new round table initiative; the paper concludes in section 4 by justifying the selection of the ES-SILC/ EU-SILC for the PR action, based in the inherent importance of the subjects studied by these statistics (income inequality, poverty…), and in the attention paid to its quality assessment, both in the European and in the Spanish environments.

2 Peer Review and the European common quality framework

In 2005 the European Statistical System (ESS) adopted the European Statistics Code of Practice (ESCoP). The ESCoP sets out a group of principles (dimensions) and related indicators of good practices for the production and dissemination of European official statistics. This constitutes a comprehensive approach which builds upon a common definition of quality in statistics. There have been two further updating of the Code in 2011 and 2017.

The structure of the ESCoP according to the last version (2017) can be seen in Table 1.

Institutional environment	**Statistical processes**	**Statistical output**
P.1: Professional Independence	P.7: Sound methodology	P.11: Relevance
P.1bis: Coordination and cooperation	P.8: Appropriate statistical procedures	P.12: Accuracy and reliability
P.2: Mandate for data collection and access to data	P.9: Non-excessive burden on respondents	P.13: Timeliness and punctuality
P.3: Adequacy of resources	P.10: Cost effectiveness	P.14: Coherence and comparability
P.4: Commitment to quality		P.15: Accessibility
P.5: Statistical confidentiality		
P.6: Impartiality and objectivity		

Table 1: The European Statistics Code of Practice (ESCoP) (version of 2017): Principles (dimensions) of quality.

The ESCoP consists of a set of "Principles" concerning the quality of the products (statistics must be accurate, timely, accessible…) of the production processes (they must use "sound methodology"

and "appropriate statistical procedures") and of the framework in which the statistical activity takes place (professional independence, impartiality and objectivity must be guaranteed, etc.). The Code includes a set of best practice indicators and standards for each of the Principles.

The European countries are committed to fully complying with the CoP and are working towards its full implementation. Nevertheless, the ESCoP does not have the character of a mandatory regulation, but it is a code of self-regulation of the countries. Therefore, from the very moment of the adoption of the ESCoP, a system of periodic assessments review was established to evaluate the degree to which the different countries complied with the Code. The ultimate goal of these assessments was to strengthen the efficiency of the statistical systems at national and European level, and to reassure stakeholders of both the quality and trustworthiness of European statistics.

The first round of these assessments was launched in 2006-2008 by across the Member States, European Free Trade Association (EFTA) countries and Eurostat.

In this first round of assessments it was agreed that the system would be based on the so-called "Peer review" (PR) approach. Among the different alternatives for evaluating systems or institutions, the PR formula represents a balance between internal self-evaluation procedures and external audit-type procedures. It is a kind of evaluation conceived as collaborative and non-compulsory between countries/ institutions with common working methods and objectives. (Cañada, 2015))

In this case, each country (statistical office) would be evaluated by experts from other countries (statistical offices are the "peer" reviewers).

The summarised outcomes of the process can be seen in a 2008 Commission report. In that document, the Commission concluded that, PR had been shown an "overall high compliance levels with ESCoP" in the EU countries; nevertheless, it admits that "full compliance with the Code remains a challenge for the European statistical system". Thus, the report envisaged another round of peer reviews "within the next five years" to reinforce the process for producing reliable and credible European statistics.

At the same time, the ESCoP had been amended in 2011 following the adoption of Regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics.

Consequently, a new round of peer reviews was launched at the end of 2013. This round of peer reviews differed in many respects from the previous one, trying to improve it: firstly, the review encompassed all the Principles of the Code (the first round was focused only in some specific principles); secondly, in addition to the NSIs, a number of other national authorities (ONAs) responsible for producing European statistics, were assessed; and thirdly, and above all, in order to gain an independent view, evaluations were externalised and an audit-like approach was applied.

Upon this new approach, review in each country was conducted by independent reviewers (three statistical experts out of the ESS) and it follows the usual stages of an audit: completion of self-assessment questionnaires (SAQ) by the country, being all the answers supported by evidence; an assessment of SAQ by the reviewers; a visit to the NSI by the reviewers, to obtain direct information and clarification of SAQ; the compilation by the reviewers of reports on the compliance to the CoP, which also include a set of recommendations of areas to be improved; and finally, in response to these recommendations, countries (NSI) developed a multiannual plan (2015-2019) of improvement actions, yearly monitored by Eurostat.

Although it is beyond the scope of this paper to go into detail on the subject, we just do not want to close this section without indicating that preparatory work for a new round has started in 2019: After the last revision of the ESCoP and its adoption by the ESSC in November 2017, compliance of

the national statistical systems with the principles of this revised Code needs to be re-examined in a third round of peer reviews scheduled for 2021-2022.[1]

3 The report for Spain of the second round of PR: towards a better integration of qualified users in the Quality management in the INE

As it has been explained, the PR assessment resulted in reports for each country (NSI) in which the reviewers made a set of recommendations of areas in which some improvements for a better compliance to the ESCoP were needed.

In the case of Spain, the reviewers stated that "statistical activities of INE are in all main respects in compliance with the CoP" (Eurostat, 2015 pp. 4). Nevertheless, within their recommendations, they call for a greater involvement of user and stakeholders in the activity of INE; specifically they suggest that the INE should "develop and implement actions to increase the involvement of external experts in regular reviewing of key statistical outputs".

To put the recommendation into context, it is useful to briefly explain some of the INE's overall quality management system and the involvement of stakeholders and users in the system.

In short, the Quality management System of INE is based in the ESCoP and consists of several types of elements, corresponding to three broad categories (a detail of all the instruments can be seen in Cañada, 2015): a quality policy, stated in some documents approved by the board of Directors, in which the INE global objectives on quality and the guidelines to reach these objectives, are specified; an administrative structure devoted to quality management, composed of a specialised administrative unit - quality unit -; and a quality management collegiate body - the Quality Committee- for monitoring quality policies and instruments; and the quality assessment and monitoring system, made up of several instruments and methods such as: the application and follow-up of a "Barometer of priority quality indicators" (Eurostat standard); the compilation of INE Quality reports for all statistical operations (disseminated freely at the INE website); application of a common standard for the production process; good practices inventories; evaluation procedures, etc.

Within this framework, the Q management of INE includes several mechanisms for attending users (and other "stakeholders") criteria in different areas.

These include: committees for the development of statistical programmes and plans; working groups and university partnerships in various domains... But the main tool for checking needs and satisfaction of users as ex-post assessors of the quality of the production and services provided by INE is the User Satisfaction Surveys (USS). The INE conducts those surveys in order to know their opinion and degree of satisfaction with statistics and dissemination services, and detect new information needs. Surveys have been regularly conducted (every three years) since 2007. The last one was launched in 2019.

Alongside the USS, some additional tools and indicators provide information on the use and impact of INE production, such as Web analytics data on visits and downloads from the INE's website, or an assessment of impact of INE data on the media according to a specific methodology (Cañada et al., 2018).

Nevertheless, one of the recommendation of the PR was that INE should be intensify these kind of activities, in particular by involving "outside experts in regular reviewing of key statistical outputs". INE interpreted the recommendation mostly in terms of improving the continuity and regularity of

[1] A noteworthy feature of the new round is that the envisaged methodology will be a combination of the previous rounds: among other aspects, reviewers teams will be composed of statisticians from the Statistical Offices (peers review approach as in the first round) and external experts (audit approach, as in the second round).

these actions. Therefore, it seemed appropriate to create other supplementary mechanisms to those already in place.

In this respect, INE designed improvement action (IA no. 10) (Eurostat, 2015) broken down in two complementary tasks: firstly, a new procedure to evaluate routinely the statistical operations by the statistical system collegiate bodies, specifically the High Council on Statistics (HCS); secondly, the organization of seminars with experts focused in specific statistical domains.

In the first case, it is worth noting here that the HCS is an advisory body gathering producers and stakeholders of official statistics in a balanced representation. It brings together institutional users, trade unions, business and consumer associations, media and scientific community, and thus it includes all the main stakeholders of the NSI. The HCS has a crucial role in the Spanish statistical system, because one of their functions is to participate in the making, analysis and approval of the National Statistical Plans of Spain.

The PR action involving the HCS stakeholders' representatives consists of the following procedure: firstly, the Quality Unit selects a statistic out from the European Statistics, [2] based on the relevance of the product and the availability of information on quality; next, the Quality Unit compiles a document summarising quality aspects of the selected European Statistics; and thirdly, this document is assessed by some members of the HCS, providing suggestions for improving these statistics.

Up to now, two group of exercises have been developed under this approach, focused in Labour Statistics and Research & Development Statistics.

The second action proposed by INE was to organize periodical seminars with researchers, academia and other experts focused in a specific statistic. In order to give them continuity, it was decided that these seminars would take place at the Public Statistics Conference, an annual event organised by the INE with the cooperation of a scientific statistical association.[3] A monographic session has been included in these conferences, focusing on a specific INE's statistical domain and arranged as a "round table session" with the participation of key experts in the selected statistical operation. Round table discussions would allow to the INE to collect qualified opinions and would constitute a subsidiary element for the evaluation of statistics.

For each conference, the Quality unit proposes first the statistical product to be evaluated. After the necessary approval by the Board of Directors, the Q Unit, in close cooperation to the unit in charge of the product, specifies the details of the round table (selection of the discussants).

Furthermore, in order to guarantee the principles of transparency and accessibility of these actions, the INE committed itself to publishing the papers presented during that session in one of the scientific journals of dissemination within official statistics. This is precisely the reason for this section of the Spanish Journal of Statistics.

4 The EU-SILC & ES-SILC: Some quality management issues

The product to be evaluated in the last round of these seminars (2019) was the (Spanish) Statistics on Income and Living Conditions (ES-SILC) implemented by INE, which is a Spanish version of the European statistics (EU-SILC), with which it is fully compatible.

Different reasons justifies the selection of ES-SILC/ EU-SILC for this exercise.

[2] "European Statistics" follow a regulation, which implies that countries draw up and send regularly to Eurostat a quality report detailing all information relating to quality measures and indicators. Based upon such reports, the Commission (Eurostat) compile a summarized report on the quality for the European Union as a whole. See the example of EU-SILC in Figure **??** and Section 4 of this paper.

[3] The Statistics and Operational Research Society, an association of Spanish researchers/academics in Statistics.

On the one hand, the social and economic importance of these statistics. They can be illustrate by a simple representative example: If the term EU-SILC is entered into Google, more than 800.000 references appear (data obtained in March 2020). This is obviously due to the widening imbalances in the distribution of income in the real world, the growing concern about these issues by policy makers, researchers and the general public, and the pressing need for statistical information on these problems.

In addition, the specific importance of these statistics in the European area must be added, as the survey contributes to the collection of a large number of key indicators for the European Union's social and economic policies.

The EU-SILC instrument is the main source for comparable indicators for monitoring and reporting on living conditions and social cohesion at the EU level. It has a main role in the Europe 2020 strategy (EU2020) the EU's agenda for growth and jobs. Specifically, EU-SILC is the data source selected for assessing one of the EU2020 headline targets, which is the "reduction of the number of people under poverty and social exclusion".

EU-SILC data also provide quantitative evidence for other EU policies and strategies such as monitoring the implementation of the social protection and inclusion dimension of the "European Pillar of Social Rights", (https://ec.europa.eu/commission/priorities/deeper-and-fairer-economic-and-monetary-union/european-pillar-social-rights_en) and provide data for the Social Protection Performance Monitor (SPPM) (https://ec.europa.eu/social/main.jsp?catId=758).

In the case of the ES-SILC, is clear its importance for policy makers and for researchers and institutions involved in the phenomena of social inequality at the Spanish level; to avoid extending ourselves, we can take as representative examples of the relevance and usefulness of the statistic, the projects summarised by the two discussants of the round table included in this issue of the Statistical Journal: among other uses, we can outline that ES-SILC is a main statistical tool for the monitoring of the strategy of the Spanish Government on the reduction of inequalities and poverty.

Beyond the importance of these statistics regarding its "relevance" (using the ESCoP terminology), the other noteworthy feature of this statistical operation, is the role given to its Quality assurance, both at European and national levels. From the very moment it was launched, constant attention has been paid to assessing the quality issues of the statistic. This attention entails that all countries are obliged to the regular (yearly) compilation of quality reports of this statistic.

Quality reports are documents which, for each specific statistic, detail all information relating to quality measures and indicators. Table 2 shows the structure of the Quality reports for the EU-SILC, which follows the EU standard ESQRS ("European Statistics Standard for Quality Reports Structure").[4]

The report focuses in the statistical process and elements of EU-SILC that have an impact on the quality of data. As can be seen in the table, to give them the widest possible scope, the reports are structured according the most relevant principles set out in the Code of Practices, providing information and indicators on the degree to which each principle is met.

Many headings correspond exactly to dimensions/principles of the ESCoP (see Table 1): there are items to provide information on the compliance with the "Output quality dimensions" (as relevance, accuracy and reliability, timelines and punctuality, coherence and comparability, accessibility and clarity); other items on "Process quality dimensions" (statistical processing, cost and burden); and even some items concern the principles of the "Institutional environment" (confidentiality, quality

[4]For simplicity's sake, the 3 digit structure has not been included in the table. As an example of these details, the breakdown of item 6.3 (Non-sampling error) is the following: 6.3.1 Coverage error ; 6.3.2 Measurement error; 6.3.3 Non response error; 6.3.4 Processing error; 6.3.5 Model assumption error.

management). A specific subject is the detail included in item 2 "Statistical presentation", which shows potential deviations from standard definitions and concepts; this is a transversal topic to different Principles of ESCoP (clarity, coherence, etc.).

1. Contact	6. Accuracy and reliability (Cont.)
2. Statistical presentation	6.4 Seasonal adjustment
2.1 Data description	6.5 Data revision - policy
2.2 Classification system	6.6. Data revision - practice
2.3 Sector coverage	7. Timeliness and punctuality
2.4 Statistical concepts and definitions	7.1 Timeliness
2.5 Statistical unit	7.2 Punctuality
2.6 Statistical population	8. Coherence and comparability
2.7 Reference area	8.1 Comparability - geographical
2.8 Time coverage	8.2 Comparability - over time
2.9 Base period	8.3 Coherence: cross domain
3. Statistical processing	8.4 Coherence: sub-annual/ annual
3.1 Source data	8.5 Coherence - National Accounts
3.2 Frequency of data collection	8.6 Coherence - internal
3.3 Data collection	9. Accessibility and clarity
3.4 Data validation	9.1 News release
3.5 Data compilation	9.2 Publications
3.6 Adjustment	9.3 Online database
4. Quality management	9.4 Microdata access
4.1 Quality assurance	9.5 Other
4.2 Quality assessment	9.6 Documentation on methodology
5. Relevance	9.7 Quality documentation
5.1 User Needs	10. Cost and Burden
5.2 User Satisfaction	11. Confidentiality
5.3 Completeness	11.1 Confidentiality - policy
6. Accuracy and reliability	11.2 Confidentiality - data treatment
6.1 Accuracy - overall	12. Comment
6.2 Sampling error	
6.3 Non-sampling error (*)	

Table 2: . Structure of the quality reports of the EU-SILC (according to ESQRS - V2.0). Source: Eurostat, 2016.

The national reports provide useful insight into countries issues. By integrating the information contained in the national reports, Eurostat compiles a global quality report to evaluate the survey from a European perspective; it allows to set up between-country comparisons of some of their key quality dimensions.

The European Union publishes on its website all the detailed quality reports of this survey, both for the different countries and for the Union as a whole.

To this European system for quality assessment, we can add in the case of the Spanish ES-SILC, the other quality requirements and tools established by INE's own internal quality control system. As it has been explained above (see section 3 of this paper), among other tools, all the Statistics have to compile a report according to the European ESMS scheme, all of which are publicly available on the INE website for all statistical operations, including the ES-SILC, from 2014 onwards.

Therefore, ES-SILC/ EU-SILC are among the most detailed European statistics in terms of the methodological information available; using the terms and concepts of the ESCoP, we could say

that they show a high level of "clarity" ("transparency") of sources, methods and even of the quality assessment procedures.

In short, both types of reasons, the relevance of the product on the one hand, and the attention paid to controlling and reporting the quality of the product, on the other, as well as the convenience of making this valuable meta-information known, amply justify the selection of the SILCs for this INE activity.

Let us take the opportunity to make an additional comment in this respect: in spite of the availability of all this information on methods and quality controls of these statistics, they are quite unknown for most of their users. Then, an additional aim of the round table sessions (and of the own dissemination through these papers in the SJS) is to contribute to a better knowledge of the available meta-information on methods and quality on this statistics.

We would like to highlight this last idea, because "in a world experiencing a growing trend of instant information which often lacks the necessary proof of quality" (Eurostat, 2017), it is important to show our users that concern for quality assurance - and the availability of information on this subject - are differential features of official statistics.

References

Cañada, A. (2015). Quality guidelines. Technical report, INE. Madrid.

Cañada, A. and Muñoz, L. and Piñán, A. (2018). Quality reviews of official statistics and the role of the external stakeholders: some initiatives from the NSI of Spain. In *European Conference on Quality in Official Statistics Q2018*. June 2018, Kraków.

Eurostat (2015). Peer review report on compliance with the Code of Practice and the coordination role of the National Statistical Institute of Spain. Technical report.

Eurostat (2016). EU statistics on income and living conditions (EU-SILC) methodology – data quality, 2016. Technical report.

Eurostat (2017). European Statistics Code of practice, 2017. Technical report.

SPANISH JOURNAL OF STATISTICS
Vol. **1** No. 1 2019, Pages 41–56
doi: https://doi.org/10.37830/SJS.2019.1.05

OFFICIAL STATISTICS

The Spanish Survey of Living Conditions (ES-SILC). Characteristics and methodological development

José María Méndez Martín
National Statistical Institute of Spain (INE)

Abstract: The *Encuesta de Condiciones de Vida* (Spanish SILC Survey) is an annual survey integrated in the European Statistics on Income and Living Conditions (EU-SILC). The primary aim of this survey is the regular production of statistics on household income and living conditions, collecting key variables like the total household income and the income components. Initially income information was obtained using exclusively the questionnaires. Nevertheless, access to administrative records offers a good opportunity to improve the quality of income data and allows the use of a more efficient collection method. In 2013 a new methodology was adopted in ES-SILC for the production of the income variables based on the use of administrative files (data from the Spanish Tax Agency and the Social Security system), in combination with the information available in the questionnaires. This paper offers a comparative overview of both methodologies assessing the impact on the main indicators. Also the future scheme of this statistical operation in 2021 is introduced. The new design of the contents will be based on a series of rotating modules that will cover different dimensions of the living conditions of the household.

Keywords: Survey of Living Conditions, household income, administrative data

MSC: 62P20, 62P25, 91B82

1 Introduction

This paper is elaborated in the frame of a peer review process that is part of the European Statistical System (EES) strategy to implement the Code of Practice (CoP). In the last peer review to INE Spain there is a recommendation (Recomm.10) to carry out the assessment of key statistical operations by external experts. To fulfil this recommendation a session has been organized in the context of the XXXVIII Spanish Conference on Statistics and Operational Research, and the XII Conference on Public Statistics to present the methodology to the Spanish Survey of Living Conditions (ES-SILC), having two discussants from the Complutense University of Madrid and the Office of the High Commissioner against Child Poverty.

The structure of the paper is as follows: It first presents a description of the characteristics and objectives of the survey and the treatment of one of the key variables of the survey, the household

income; the following sections describe the evolution of the ES-SILC methodology, from the initial version entirely based on sampling surveys, to the current methodology, which combines surveys with administrative sources; next, some of the main indicators obtained from these surveys are commented on, such as the poverty line or the AROPE indicator, as well as some effects that changes in the methodology have had on these indicators. In a final section future novelties planned for this survey are described, an analysis that the author takes advantage of precisely to evaluate some of the suggestions made by the discussants in the round table.

2 Background

The Spanish Survey of Living Conditions (ES-SILC), integrated in the European Statistics on Income and Living Conditions (EU-SILC), belongs to the set of harmonized statistical operations for the countries of the European Union. The fundamental objective pursued with EU-SILC is to have a reference source on comparative statistics of income distribution and social exclusion in Europe.

EU-SILC is not the first harmonized statistical operation of income and living conditions that is carried out in Europe. Between 1994 and 2001, the European Community Household Panel (ECHP) survey met these needs. ECHP was configured as a pure panel in which the households of a sample selected in 1994 were followed for 8 years, without performing any sample rotation during those years. However, since it was necessary to update its content according to the new demands, it was decided to replace the ECHP survey with a new instrument.

The Commission developed a community action program to promote cooperation between Member States in the fight against social exclusion, which was presented in 2000 to the European Parliament and the Council, to promote the collection and dissemination of comparable statistics in the Member States and at the community level.

In this framework it was decided to replace ECHP from 2002, by another survey that adapted its contents to the new information needs coming from public and private sectors and to the aim of improving the quality of the information, in particular regarding the timeliness of the data. After several studies and preparatory meetings, a pilot test was conducted in 2002, and in 2004 the final survey was initiated in most countries, including Spain (INE, 2005).

3 General description

This statistical operation is legally supported by a set of Regulations published in the Official Journal of the European Union, which determine the commitments acquired by the Member States and Eurostat in relation to the survey.

Regulation (EC) 1177/2003 of the European Parliament and of the Council concerning Community statistics on income and living conditions is the framework regulation that provides legal support to the EU-SILC. This Regulation specifies the objectives of the new statistical instrument, the areas included, the time reference, the calendar of data availability and some quality requirements such as the effective sample size. It clearly defines the responsibilities of the Member States and Eurostat, resulting in the quality of the new instrument.

While the Regulation of the Parliament and of the Council identifies the general legal basis of the new operation, the Commission Regulations detail the target variables that the statistical source includes, the fieldwork aspects, the sampling, etc. Concerning the definitions the survey follows the recommendations for income data of the International Expert Group on Household Income Statistics (United Nations, 2011).

Although the priority in this survey is the production of cross-sectional information (data produced year by year) with a high degree of quality in terms of its timeliness and comparability, the survey also collects longitudinal information referring to the same sample at different years over time (in the Spanish case, the follow-up is carried out over four years). The inclusion of a longitudinal component makes the survey more complex since the households in the sample have to be followed in the different years.

ES-SILC has a final sample size of about 13,000 households. The mode of data collection is CAPI and the interviews are conducted normally between April and July (although exceptionally there have been years in which it has been collected after the summer), publishing the results around May of the following year.

3.1 Main areas of study

The areas of study included in ES-SILC can be classified according to the periodicity of the data collection. The primary areas are those whose collection is carried out annually, while the secondary areas contain the variables that are included in the modules that can change every year. Currently, the part corresponding to the primary areas represents the largest part of the survey, while the part dedicated to the annual module is very small. The primary areas covered by ES-SILC are the following, depending on the type of unit of observation:

- Household
  - Total household income and household income components.
  - Arrears on payments.
  - Non-monetary indicators of deprivation (difficulties in making ends meet, level of indebtedness, enforced lack of basic needs, etc.).
  - Physical and social environment.
  - Dwelling type, tenure status, housing conditions.
  - Housing cost.
- Persons (adults)
  - Demographic data.
  - Total personal income and personal income components.
  - Education.
  - Basic labour information on current activity, including information on last main job for previously active people.
  - Calendar of activities during the income reference period.
  - Health and access to health care.
- Persons (children)
  - Demographic data.
  - Childcare.

The areas of study were reduced compared to those considered in ECHP, which results in a higher quality of data.

Since the 2005 survey a module is introduced every year addressing a specific topic of interest. The modules that have been included during these years are:

- 2005: intergenerational transmission of poverty.
- 2006: social participation.

- 2007: housing conditions.
- 2008: over-indebtedness and financial exclusion.
- 2009: material deprivation.
- 2010: intra-household sharing of resources.
- 2011: intergenerational transmission of disadvantages.
- 2012: housing conditions.
- 2013: well-being.
- 2014: material deprivation.
- 2015: social participation.
- 2016: access to services.
- 2017: health.
- 2018: well-being and housing difficulties.
- 2019: intergenerational transmission of disadvantages.
- 2020: over-indebtedness, consumption and wealth.

3.2 The integrated design

The integrated design combines in a single operation the cross-sectional component (year-to-year results) and the longitudinal component (follow-up of households over a period of time). It was the design recommended by Eurostat in those cases in which it was decided to develop a new survey, being the model adopted by Spain.

The integrated design of ES-SILC consists of a rotating panel survey. Therefore, being a panel, the same units are investigated over the years, but unlike ECHP, which was a pure panel and the panel units were followed for eight years, in ES-SILC the panel units are followed only for four years.

The sample consists of four panel subsamples (four rotating groups), so that each year one of them is replaced by a new subsample. Each of the subsamples remains in the survey for four years, being afterwards replaced by another subsample.

An important advantage of this model is that most of the sample used to obtain the longitudinal component is derived from the cross-sectional sample: the cross-sectional and longitudinal statistics are obtained from the same set of units avoiding duplication in the response burden of the survey.

3.3 Quality in ES-SILC

The ES-SILC is integrated in the Spanish Statistical System, and in the European Statistical System, within which there are various quality control mechanisms.

3.3.1 Quality management at the Spanish National Statistical Institute

ES-SILC is carried out by the Spanish National Statistical Institute (NSI), and belongs to the set of statistics included in the National Statistical Plan. The tasks of the Spanish NSI in relation to statistical production are "the production, within the indicated deadlines, of adequate, reliable and consistent statistics, as well as making available to users the statistical information necessary to facilitate decision-making".

The Spanish NSI, on its website, dedicates a special section to the Quality and Code of Good Practices, showing the importance given to this aspect in official statistics. In this section, the organization established for quality management in the Spanish NSI is presented, based on three basic elements:

- An administrative structure. The Spanish NSI has a Quality Unit with this objective, as well as a Quality Committee, where the lines to be followed are discussed.
- A quality evaluation and monitoring system. The instruments are the preparation of quality reports and indicators, user surveys and external evaluation (Peer Review).
- Other quality components. Among others, there are the dissemination, confidentiality and review policies, and the quality assessment and monitoring policies of the Spanish Public Administration.

The implementation of the Code of Practice in the National Statistical Plan from the European Statistical System (ESS) has a special role in quality management.

3.3.2 The quality of statistics in the ESS

The quality at EU level is based on the adoption by the Statistical Program Committee (SPC, current ESSC) of the European Statistics Code of Practice, in 2005. In its current version (2017) the code sets out 16 key principles, which are reviewed periodically. Its objective is to establish a standard for the development, production and dissemination of European statistics, as well as to ensure the quality and credibility of the data.

It is structured in three sections covering the institutional environment, statistical processes and statistical outputs.

All statistics produced within the European Statistical System are subject to the Code of Practice. But in the case of ES-SILC, in addition, its production is legally supported by various regulations that specify all methodological and procedural aspects.

The framework Regulation of this operation is the above-mentioned Regulation (EC) 1177/2003 of the European Parliament and of the Council, which establishes the legal basis of the production of the EU-SILC. On the other hand, there are several Commission Regulations that address each and every one of the harmonized methodological aspects.

Among these legislation there is also a specific regulation dedicated to the quality of the EU-SILC. Commission Regulation (EC) 28/2004 describes in detail the intermediate and final quality reports of the EU-SILC that all Member States must send to Eurostat. Eurostat subsequently generates a comparative quality report where you can see the weaknesses and strengths of the survey. In addition Eurostat carries out a review of the survey data before the results are published.

4 Income in ES-SILC

One of the main objectives of the survey is the analysis of household income received during the reference period. The income reference period adopted in ES-SILC is annual from January to December of the calendar year preceding the interview.

This statistical operation collects personal income of the household members and also household income components of those sources that are difficult to individualize. Thus, this survey collects personal income in individual questionnaires (persons aged 16 years or older). On the other hand, in the household questionnaire other income components are collected (property income, housing assistance, etc.) that are more typical of the household as a whole. All this aggregate information allows to build the total household income.

The total disposable household income is calculated as follows (net of tax on income at source and social contributions):

+ Monetary employee income

+ Non-monetary employee income (company car)
+/- Profits / losses from self-employment
+ Capital income
+ Property income
+ Social benefits (unemployment, retirement, etc.)
+ Pension from individual private plans
+ Regular inter-household cash transfers received
+ Income received by children under 16
− Regular inter-household cash transfers paid
− Repayments/receipts for tax adjustments
− Regular taxes on wealth

4.1 Gross and net income

In EU-SILC, the income target variables include the collection of gross amounts at the level of individual and income component. This allows a better comparability between Member States to analyze income components, because it will not depend on the tax system or the social security contributions of the country. It should not be forgotten that the objective of EU-SILC is not only to obtain information at country level, but also to obtain comparable data among Member States of the European Union, which requires a standardized use of definitions and methodologies.

However, at the aggregate level, the key variable is the total disposable household income, in particular for studies of income distribution and poverty.

4.2 Monetary and non-monetary income

ES-SILC collects some non-monetary income components, although they are not always included in the definition of total disposable household income:

- Company car for employees
- Other non-monetary employee income, such as luncheon vouchers
- Employer's social insurance contributions
- Imputed rent
- Value of goods produced for own consumption (value of food and beverages produced and consumed within the same household)

Currently, in the definition of the total disposable household income, only the company car for employees is included.

5 Methodology based on Sampling Survey (base 2004)

At the beginning of ES-SILC from 2004, the data collection methodology was carried out through sampling survey procedures and the use of questionnaires. Since 2013 the methodology was changed complementing the data sources with the use of administrative files.

Since 2004 the method of data collection has been mainly the personal interview (CAPI) with all the household members of the dwellings surveyed. The interviewer contacts the household and requests the information necessary to complete the questionnaires, making all the necessary interviews to collect the required information. In this way, both the information related to income and the rest of the variables related to the living conditions of the household are collected. The questionnaire

collects income data during the interview through some alternative questions that try to facilitate the response. Taking into account that the reference period is annual, it is usually requested in the question for an amount and for the number of times received. The possibility of providing annualized amounts is also usually offered.

One of the greatest difficulties in collecting income data from the respondents is the lack of knowledge of the gross amounts, or of the taxation of income at source. This has made necessary the use of net-gross models to impute the gross amounts.

In the 2013 survey, a new base was started with the use of administrative files. However, the collection of all the income information in the questionnaires was continued until 2014, simplifying the questionnaire from 2015 due to the suppression of many questions related to income amounts.

6 Methodology based on the use of administrative files (base 2013)

In 2013 a new methodology was adopted in ES-SILC in the production of data on household income based primarily on the use of administrative files, combined with information available on the questionnaires. Access to administrative records has meant an improvement in data quality and efficiency in the method of data collection.

Data related to household income are obtained using a mixed methodology combining the information provided by the respondent to the surveys with the administrative records of the Spanish Tax Administration, Social Security, and the Tax Administration of Navarra, Bizkaia and Gipuzkoa.

Due to the change in methodology, there is a break in the time series that makes the income data not comparable with the data published in previous years. For this reason, retrospective data have been produced since 2008 using the new methodology based on administrative files, which are comparable with 2013 data.

Within the European Union, the increasing use of administrative data in statistical production is an essential element in the process of modernization of social statistics. Many countries use administrative files in this survey for the production of income variables. Apart from the Nordic countries, Netherlands, France, Austria or Slovenia make very important use of administrative files. In other cases, partial use is made of them covering mainly the social benefits variables (Eurostat, 2013).

6.1 Method of data collection

The methodological change involves mainly collecting the income variables using the administrative files. From the Tax Identification Number (NIF, personal identification) of the sample persons, data from the Tax and Social Security Sources are collected and, together with the data collected in the questionnaires, the construction of the income variables is carried out.

In the production of income variables by components from administrative files, in a first approximation the methodology would be to identify the corresponding field or fields of the corresponding administrative forms, and do the appropriate processes to obtain the required target variables.

This procedure is valid in some cases, but in other income components we observe that an exclusive use of administrative files in their construction is difficult. This is due to the problems that arise in relation to the geographical coverage (no tax data are available in Alava), coverage of income recipients (problems in certain groups such as domestic employees, the informal economy, etc.) or problems with the precise identification of the income component in the administrative file. In these cases, complementary information is collected through the personal interview to construct the income target variables.

Finally, combining data from the survey and the administrative files, the variables related to income are constructed. In few cases in which the amount cannot be captured from any of these two sources, mathematical imputation is necessary. The merge of the sample persons with the administrative data is done through the Tax Identification Number (NIF). The capture of the NIF in ES-SILC was first introduced in the 2009 survey, covering around 98% of adults.

6.2 Construction of the target income variables

The methodology of using administrative files is not direct. In an ideal/naive situation it would consist in exchanging amounts (see Figure 1). The amounts that have been collected through the questionnaire, now are collected by taking certain fields of the corresponding forms of the administrative files:

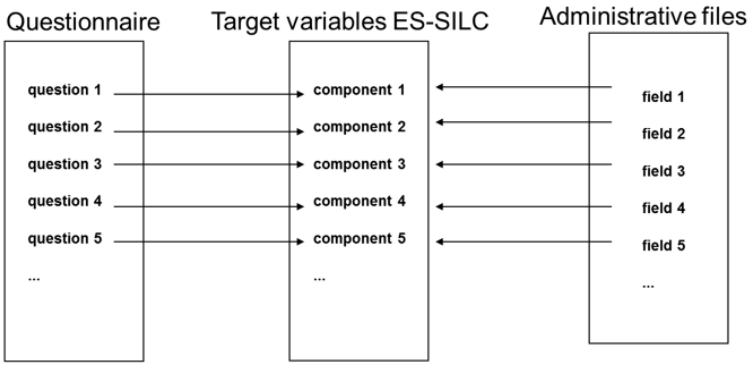


Figure 1: Links among target variables and data provided by sources: Ideal situation.

However the real situation is more complex. From the perspective of the questionnaire, without taking into account the non-sampling errors of any household survey, there can be measurement and classification errors like, for example, having an amount collected in a question corresponding to more than one target income variables of ES-SILC (for example, the question that collects employee income may mistakenly include sickness benefits).

From the perspective of the administrative file the situation is even more complicated. The correspondence between fields of administrative forms and amounts of the target variables is sometimes not direct as it is shown in Figure 2.

This situation makes it necessary to control the risk of duplicity. An amount to be loaded from the questionnaire in a certain target variable may already be included in some cases in other variable in the load from the administrative file.

For this reason, in the construction of the target income variables we will consider two types of variables:

- Individual variables of direct construction. The variable is loaded from the questionnaire or the administrative file. There are no risks of duplicity in the sense of having the amount already collected in another variable. In this situation we have, for example, inter-household transfers (obtained from the questionnaire) or capital income (obtained from the administrative file).
- Group of variables in which the procedure must be in block to avoid risks of duplication. For example, employee income, income from self-employment and some social benefits (sickness, maternity) do not have a clear allocation of amounts. From the perspective of the questionnaire, amounts of one component can be filled in another (for example, income from self-employment

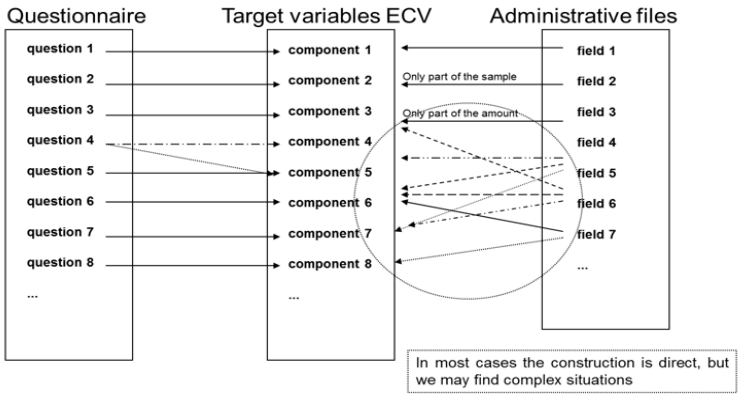


Figure 2: Links among target variables and data provided by sources: Real situation.

can be considered by the respondent as employee income, and vice versa). From the perspective of the administrative file, there may also be problems in separating, for example, sickness benefits from employee income. In these cases, we will take all the amounts either from the administrative file or from the questionnaire. In the case of taking the data from the questionnaire, the information available in the administrative file can be used to improve the construction of the target variables.

The loading strategy is summarized in Figure 3, which shows the origin of the different income components (comp 1, comp 2, etc.) for the different records or sample units:

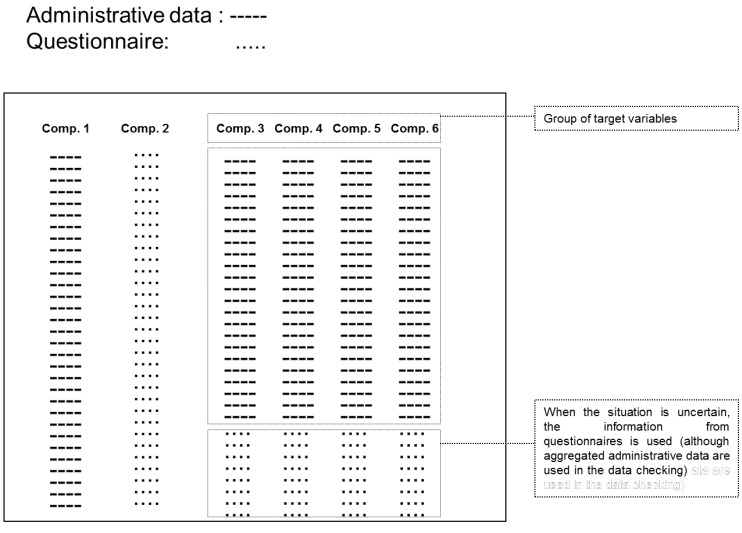


Figure 3: Origin of the different income components (comp 1, comp 2, etc.) for the different records or sample units.

6.3 Administrative sources

6.3.1 Register of Public Social Benefits. Social Security Sources

The Social Security is the managing administration of numerous social benefits having lots of information that covers the entire national territory. However, there are a number of social benefits, mainly non-contributory pensions, that are managed from the Autonomous Communities and that are outside the scope of Social Security management.

However, the Social Security is responsible with the management and operation of the Register of Public Social Benefits. The creation of this registry allows coordinated knowledge and transfer of data between public entities, in order to facilitate the allocation of benefits, as well as avoiding fraud.

Therefore, the Social Security is the owner of this register, with data regarding the benefits paid by management entities of the Social Security System as well as bodies outside this system.

6.3.2 Forms 190 and 100. Tax Sources

Currently there are data are available from the Tax Agency (AEAT), and the tax administrations of Navarra, Bizkaia and Gipuzkoa.

Initially, the information contained in the Income Tax Returns is rich enough to know the different components of income in the households in the sample. However there are some difficulties. In particular, the group of those not obliged to fill the tax returns is quite important and, on the other hand, the possibility of family joint declaration makes difficult the individualization of income required by the ES-SILC.

For this reason it is essential the access to other information available in the Tax Sources. We must keep in mind that the tax returns are in fact an annual adjustment of the taxes that have already been paid during the year in a system of withholdings and taxes paid at source.

Besides the Income Tax Returns, the Tax Administration have extensive information. In particular, there is a form with the annual summary of withholdings and taxes at source (Form 190 of the Tax Agency) that contains individual information on different income components, including income recipients who are not obliged to fill the tax return. Detailed information on it can be found in the regulations regarding this form.

This register contains useful information for constructing the employee income variable and some social benefits (there are some fields that identify the income component or subcomponent).

On the other hand, Form 100 (Tax Agency) contains the Income Tax Returns. Although Form 190 is initially used as the main source of information, in the case of income from self-employment, capital and property income, the valuable information available in Form 100 is used.

In summary, the most important variable in ES-SILC is the disposable household income. This variable is obtained as the sum of income components (salaries, pensions, etc.). In the production of these variables, information from administrative files is sometimes used. When there is no information in the questionnaires or in the administrative files, the mathematical imputation must be used.

7 Main indicators

Most of the studies on poverty and inequality that are obtained from EU-SILC are based on the variable “disposable household income”.

Although this variable is constructed at the household level, many indicators take the person as a unit. When transferring household information to the individuals, it must be taken into account if all household members have the same needs (per capita) or there are economies of scale (equivalence scales). In the case of the indicators of the EU-SILC, the second criterion is usually adopted.

The Spanish NSI, following the Eurostat recommendations, uses the income per unit of consumption (or equivalised disposable income) as the reference variable. This variable is defined as the ratio between the disposable household income and the number of consumption units (equivalised household size). To obtain the number of consumption units, the modified OECD equivalence scale is used:

First adult	1
Other people aged 14 or over	0.5
Every child under 14 years old	0.3

The unit of analysis is usually the person, so once the income per unit of consumption is calculated for each household, it is assigned to each of its members. This income per unit of consumption (or equivalised disposable income) is used in the calculation of indicators of relative poverty and income distribution.

7.1 At-risk-of-poverty threshold

This relative poverty threshold depends on the distribution of the income per unit of consumption, taking the person as the unit (for this reason is a relative threshold, because takes into account the situation of the population to which it belongs). The poverty threshold is recalculated each year, increasing or decreasing depending on the variation of the median of the equivalised disposable income.

This threshold is set at 60% of the median of the equivalised disposable income. Other thresholds such as 40%, 50% or 70% of the median can be considered. It can also be obtained by including or excluding the imputed rent in the definition of disposable household income.

7.2 At-risk-of-poverty rate

The at-risk-of-poverty rate is the percentage of people who are below the at-risk-of-poverty threshold.

Many of the poverty rates are broken down according to classification variables, for example age and gender shown in Table 1.

	Total	**Males**	**Females**
Total	21.5	20.9	22.2
Less than 16 years	26.2	25.5	26.9
From16 to 64 years	22.1	21.3	23
65 years or over	15.6	14.7	16.3

Table 1: ES-SILC 2018. At-risk-of-poverty rate, by age and gender. (Percentages of total population in each category). Source: Own elaboration from ES-SILC 2018.

The meaning of these indicators is the percentage of people at-risk-of-poverty in each group, measuring the impact of relative poverty by groups.

7.3 Income inequality indicators. Gini coefficient and S80/S20 income quintile share ratio

The Gini coefficient, shown in Table 2, is defined as the ratio between the cumulative proportion of the population ordered by equivalised income and the cumulative proportion of income received by them.

On the other hand, the S80/S20 income quintile share indicator, also shown in Table 2, is the ratio between the sum of the equivalised income of the last quintile and the first one. That is, the population are sorted by the equivalised income, taking the group of the poorest 20% and the richest 20%. The sum of the equivalised income is determined for both groups, calculating the ratio between them. It will always be greater than one, having more inequality when the value is higher.

	Gini	**S80/S20**
Value	33.2	6.0

Table 2: ES-SILC 2018. Gini coefficient and ratio S80/S20 (Values). Source: Own elaboration from ES-SILC 2018.

7.4 AROPE indicator. People at risk of poverty or social exclusion (Europe 2020 strategy)

The AROPE indicator (at risk of poverty or social exclusion) is an aggregate indicator that combines three concepts: at-risk-of-poverty, material deprivation and low work intensity (see Table 3). It is defined as the population that are in at least one the following situations:

- At-risk-of-poverty (60% of the median of the equivalised income).
- In a situation of severe material deprivation. The households are lacking in at least four items from a list of nine. The nine items considered are:
  - Cannot afford to go on holidays, at least one week a year.
  - Cannot afford a meal of meat, chicken or fish at least every two days.
  - Cannot afford to keep the home adequately warm.
  - Unable to face unexpected expenses.
  - Arrears in the payment of expenses related to the main dwelling (mortgage or rent, utility bills ...) or on hire purchase instalments.
  - Cannot afford to have a car.
  - Cannot afford to have a telephone.
  - Cannot afford to have a TV.
  - Cannot afford to have a washing machine.
- Households without work or with low work intensity. People living in households with very low work intensity are defined as people aged 0-59 years living in households where the adults worked 20% or less of their total potential during the previous 12 months. This variable does not apply in the case of people 60 and older.

The three concepts included in the concept of the AROPE indicator (at-risk-of-poverty, material deprivation and work intensity) do not always affect the same population, although there are intersections between them, as it is shown in Figure 4.

	Total
At risk of poverty or social exclusion	26.1
At-risk-of-poverty	21.5
Severe material deprivation	5.4
Without work or with low work intensity (from 0 to 59 years old)	10.7

Table 3: ES-SILC 2018. At risk of poverty or social exclusion (AROPE) indicator and its components by gender (Percentages of total population in each category). Source: Own elaboration from ES-SILC 2018.

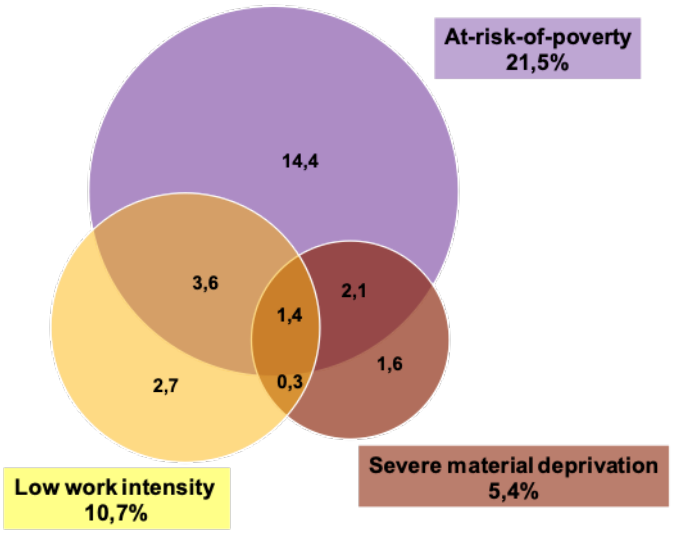


Figure 4: ES-SILC 2018. Intersections between subpopulations of AROPE Indicator. Source: Own elaboration from ES-SILC 2018.

8 Impact of the methodological change in the ES-SILC on indicators

In preliminary comparative studies between the ES-SILC data based on the use of the questionnaires (base 2004) and those data based on the use of administrative files (base 2013), carried out with data from the 2009 - 2012 surveys, the results presented in Table 4 were obtained.

The main conclusion of this study is that the use of administrative files has a very important impact on indicators based on the level of income, increasing significantly their value. On the other hand, there was an apparently lower impact on inequality indicators.

In the analysis of the at-risk-of-poverty indicators, it should be taken into account that the confidence intervals are quite wide (Eurostat initially published this type of indicators in whole numbers). For example, in the case of the total population, the 95% confidence interval of the at-risk-of-poverty rate in ES-SILC is around $\pm$ 1.4.

The comparative studies between the data obtained using questionnaires and data from administrative files are available as working papers in the Spanish NSI website (Vega et al., 2011; 2014).

	2009		2010		2011		2012	
	Survey	**Admin. data**	**Survey**	**Admin. data**	**Survey**	**Admin. data**	**Survey**	**Admin.** data
At risk of poverty or social exclusion	24,5	24,7	26,7	26,1	27,7	26,7	28,2	27,2
Less than 16 years	29,8	31,9	32,1	32,6	32,3	31,6	32,8	31,4
From 16 to 64 years	23,3	23,0	26,7	25,3	28,2	27,0	30,1	29,0
65 years or over	24,3	24,9	21,4	22,9	20,9	21,2	16,6	16,5
At risk of poverty	20,1	20,4	21,4	20,7	22,2	20,6	22,2	20,8
Less than 16 years	26,5	28,9	28,3	28,8	28,7	27,2	28,9	26,9
From 16 to 64 years	17,9	17,5	20,1	18,6	21,3	19,3	22,4	20,9
65 years or over	23,1	23,8	20,5	21,8	19,5	19,8	14,8	14,8
Gini coefficient	33,0	32,9	34,4	33,5	34,5	34,0	35,0	34,2
S80/S20 income quintile share ratio	6,4	5,9	7,2	6,2	7,1	6,3	7,2	6,5
Median equivalised income (euros)	14.483	17.042	14.369	16.922	13.907	16.280	13.885	16.119

Table 4: ES-SILC. Basic indicators. Survey versus administrative files. Source: Own elaboration from ES-SILC 2018.

9 The future of EU-SILC

9.1 Future plans

One of the problems of ES-SILC is the difficulty in the territorial disaggregation of the indicators. Although some basic indicators are published at the autonomous community level (NUTS2), in general the small sample size does not allow robust estimates. In 2013 there was a request from Eurostat for the EU-SILC in order to produce high-quality regional indicators since information on poverty and inequality could be used in the allocation of cohesion funds from 2020.

The external and internal demand for better regional coverage of ES-SILC has led to a process of duplication of the sample size starting in 2019 and applied to the new sample. Being a rotating panel with four rotation groups, the final duplication will be consolidated in 2022.

A revision of EU-SILC is also planned to be implemented in 2021. The future survey will be based on a series of rotating modules that will cover different dimensions of the living conditions of the households. Each year two modules, one included every 3 years and the other every 6 years, will cover demands such as housing, child welfare, access to services, etc. The fixed core of the survey, in comparison with the current survey, will be reduced so as not to increase the total burden on the respondents (see Table 5).

This model will allow the regular inclusion of different topics related with the living conditions broadening the scope of this survey (Argüeso et al., 2013).

Regarding the mode of data collection, ES-SILC has been collected by CAPI since 2005. In 2017 a pilot test was carried out, in which the multi-channel data collection (CAWI-CATI-CAPI) has been tested. The modernization of data collection in ES-SILC will increase the efficiency of the survey, although there are also important challenges due to the specific complexity of ES-SILC and the type of information collected.

		Rolling modules	
Nucleus (all years)	**Year**	**3 year rolling module**	**6 year rolling module**
Income	1	Children	*New policy needs 1*
Material deprivation	2	Health	Quality of life
Economic activity	3	Labour and	Intergenerational transmission
Demography		housing	of disadvantages and Housing
Education		conditions	difficulties. *New policy needs 2*
Child care	4	Children	Access to services
Housing costs	5	Health	*New policy needs 3*
Health	6	Labour and	Over-indebtedness,
Quality of life		housing conditions	consumption and wealth
Miscellaneous			

Table 5: Distribution of the topics among the annual survey and the rolling modules scheduled in the revision of EU-SILC. Source: Own elaboration.

9.2 Demands from users

The revision of ES-SILC in 2021 provides an opportunity for the inclusion of variables widely demanded by users, including additional questions in the questionnaires or collecting them from the administrative files.

The inclusion of additional questions in the questionnaires has the concern of the increase of the overload of the interview. This is particularly relevant in a future multi-channel data collection scenario, in which many interviews will be CAWI, and respondents must self-complete the questionnaire online. Nevertheless, the revised EU-SILC, with the scheme of rotating modules, will contribute to expand the content of this statistical operation addressing more in detail specifics topics like children, access to services, housing etc. widely demanded by users like the Office of the High Commissioner against Child Poverty. Also the reduction of the questionnaire due to the use of administrative files would give room for the inclusion in the future of some additional questions covering topics of national relevance like, for example, the energy poverty.

In the case of the collection of additional information from administrative sources, we have the important advantage that the respondent's burden is not increased. The demand of detailed information available in administrative databases is very important from the academic users, although it is not always possible to collect this information at microdata level for the respondents of the survey. In the case of the tax data, the main concern is the access to individual information that is restricted by the Taxation Law when this information is not included explicitly in a variable of a European regulation. Even so, there are still opportunities for the extension of the use of administrative files, in particular in variables related to the mortgage expenses of the dwelling, and in the disaggregation of some social benefits, that are included in the European regulation. Also the use of non-tax administrative databases can be envisaged.

References

Argüeso A. and Escudero T. and Méndez J.M. and Izquierdo M.J. (2013). Alternativas en la construcción de un indicador multidimensional de calidad de vida. Technical report, INE. Madrid.

European Parliament And Council Of The European Union (2003). Regulation of the European Parliament and of the Council of 16 June 2003 concerning Community Statistics on Income and Living

Conditions (EU-SILC) (EC) No. 1177/2003. Technical report.

Eurostat (2013). The use of registers in the context of EU-SILC: challenges and opportunities. Technical report, Eurostat Statistical Working Papers. Edited by Markus Jäntti, Veli-Matti Törmälehto and Eric Marlier.

INE (2005). La Encuesta de Condiciones de Vida. Metodología. Technical report. Madrid.

United Nations (2011). Canberra Group Handbook on Household Income Statistics, 2nd edition. Technical report. New York and Geneva.

Vega P. and Méndez J.M. (2011). Linking data from administrative records and the Living Conditions Survey. Technical report, INE. Madrid.

Vega P. and Méndez J.M. (2014). Comparación de los ingresos del trabajo entre la Encuesta de Condiciones de Vida y las fuentes administrativas. Technical report, INE. Madrid.

Spanish Journal of Statistics
Vol. 1 No. 1 2019, Pages 57–64
doi: https://doi.org/10.37830/SJS.2019.1.06

Official Statistics

Using EU-SILC to design and evaluate policies against child poverty in Spain

Alejandro Arias[1], Albert F. Arcarons[1], Amparo González-Ferrer[1]

[1] Office of the High Commissioner against Child Poverty

Abstract: EU-SILC is an essential tool for research on social policy. This article compiles, on one hand, the applications of this survey by the Office of the High Commissioner against Child Poverty in Spain. This Office used the Spanish version of the survey on fields such as the design of a minimum income scheme, the evaluation of energy poverty from a child perspective, the changing dynamics of poverty according to country of origin, or the impact of housing costs on poverty. On the other hand, the article also proposes some recommendations for possible improvements of the survey, such as the introduction of labour market trajectories or information on daily living expenses, for instance schooling costs.

Keywords: EU-SILC, Spain, poverty, child poverty, public policy, evaluation.

MSC: 62P20, 62P25, 91B82

1 Introduction

Eurostat has periodically published information on at-risk of poverty rates in the EU since 1995. These figures have systematically ranked Spain in the worst positions within the EU, especially when focusing on child poverty. According to the most recent data, in 2018, 26.8 per cent of children in Spain lived in households with income below the poverty line (60% of the median equivalent income). [1] Importantly, such a bad position was not a direct consequence of the crisis (in 2008, child poverty affected 27.3 per cent of children), although it obviously contributed to worsen it. Moreover, child poverty is particularly persistent in Spain: in 2018, approximately 75 per cent of children at risk of poverty had also been in poverty in at least two of the three previous years. In other words, child poverty in Spain is particularly high, much more than expected for a country with its level of

[1]Equivalised disposable income is the total income of a household that is available for spending or saving, divided by the number of household members converted into equivalised adults. Household members are equivalised or made equivalent by the following so-called modified OECD equivalence scale: the first household member aged 14 years or more counts as 1 person; each other household member aged 14 years or more counts as 0.5 person; each household member aged 13 years or less counts as 0.3 person

economic and social development. Besides, poverty seems to be a persistent experience for children, which threatens to leave indelible imprints on their adulthood.

Aware of the previous figures and the enormous challenge ahead, the last Socialist government decided to create the High Commissioner against Child Poverty (ACPI, its acronym in Spanish) in June 2018, to put child poverty at the centre of the political agenda. In other words, the very creation of the High Commissioner was partly due to EU-SILC and the rich information it has made available over the years.

The ACPI's mandate consists of carrying out studies and analyses about child poverty dynamics in Spain, designing and proposing measures to prevent and fight against child poverty to other ministries and bodies from both the administration and civil society, and the follow up and evaluation of actions, programs and policies in this field.

Since its creation, EU-SILC has been a crucial instrument for the realization of the ACPI's mission. Among other things, EU-SILC and, in particular, its Spanish version (Spanish Survey of Living Conditions (ES-SILC) allows, for instance, to estimate the extension and intensity of child poverty in Spain by using different poverty thresholds, to study its distribution by household composition, to explore the multidimensional nature of child poverty, and investigate the main characteristics of households with children in poverty and their poverty dynamics.

In this article, we will summarize some of the most relevant applications of the EU-SILC and ES-SILC done in the context of the ACPI's activities. In the process of analyzing SILC data, we have come across different limitations in the content, format and dissemination of the data. At the same time, comparing SILC with other potentially alternative and complementary data sources, we have confirmed many of its multiple strengths and advantages for a better understanding of poverty dynamics. We will describe some of these findings and their implications for public policy, and provide some proposals for improving the amount and quality of information to fight against child poverty through ES-SILC.

2 Reforming child benefits to reduce child poverty

One of the most relevant applications of EU-SILC when proposing policies aimed at reducing child poverty in Spain consisted in using its data to estimate the cost and impact of extending and increasing child benefits considering different potential scenarios. As the EU-SILC data illustrate, the current system of child allowances in Spain remains extremely inefficient in reducing child poverty, in comparison to other EU countries. Accordingly, ACPI explored the pace at which child benefits need to increase in order to eradicate extreme (25% of the median income) and high (40% of the median income) child poverty, and how to modulate coverage and generosity to make the cost of the reform tolerable for the State in the context of fiscal consolidation.

In Spain, we lack accurate information on individual and household income. Until very recently, tax information was never publicly exploited for this goal and, in any case, microdata from tax registers are not available. Moreover, income information based on tax registers entails certain limitations that particularly affect low-income households, especially when informal economy is important. For all these reasons, EU-SILC remains a crucial instrument when estimating poverty rates, even if its sample size may represent a limitation when examining extreme poverty for relatively small groups such as the child population.

Using EU-SILC information, ACPI estimated the size of child population in Spain living below different poverty lines (60%, 40% and 25% of the median income) and proposed a gradual increase of child benefits over a period of four years, aimed at eradicating both extreme and high child poverty

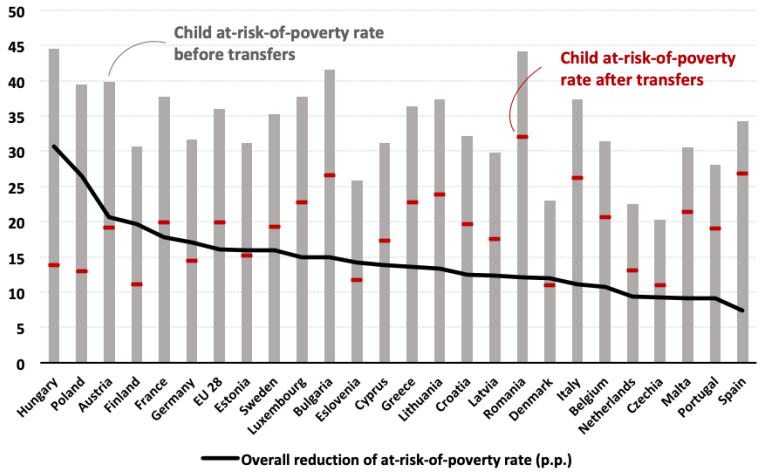


Figure 1: Child at-risk-of-poverty rate and its reduction by transfers EU-28 (2018). Source: Own compilation based on data from EU-SILC (2018).

by the end of the period. Table 1 summarizes the reduction achieved in extreme, high and moderate child poverty in the different years and scenarios considered for the estimations. During the first three years, the planned measures include only successive increases of the child benefits. In the fourth year, a Minimum Living Income for low-income families (regardless of the presence of children) would be deployed. Three different scenarios, in which both the beneficiaries and the amount provided to them vary, were considered. In Scenario 1, all households with annual income below a certain amount will receive a fixed benefit. In Scenarios 2 and 3, beneficiaries will be households with annual income below the 25% and 40% of the median income, respectively, and the amount they will receive should be enough to overcome those income thresholds. The administrative procedure to make the payment also varies across the scenarios.

Child poverty rate in 2017:		**Year 1**	**Year 2**	**Year 3**	**Year 4(1)**	**Year 4 (2)**	**Year 4 (3)**
Moderate (60%)	26.8%	0%	0%	9%	9%	9%	9%
High (40%)	12.4%	8.1%	23.4%	23.4%	26.6%	23.4%	100%
Extreme (25%)	5.2%	21.2%	46.2%	46.2%	75%	100%	100%

Table 1: Child poverty rate reduction (in percentage). Source: Own compilation.

These estimations were the basis of the proposal to double the amount received by children in households below the 25% median income threshold, finally approved in March 2019 by a Royal Decree.[2] People below the 25% threshold largely overlap with people suffering material deprivation, a complementary indicator that informs us about the type of services and goods people cannot access because of economic difficulties.

As Table 2 shows, in any of the three scenarios considered in the original estimations, the proposed increase in the child allowance would benefit between 60 and 70 per cent of children living in households that declared to suffer difficulties to meet needs by the end of the month, go on va-

[2]We also replicated these estimations for year 1 with non-public information from the tax office to double check the results obtained using ES-SILC.

cation once a year, and lack capacity to face unforeseen expenditures, respectively. In other words, setting the threshold at 25% of income poverty to define the beneficiaries of the child benefits increase ensured that the proposed reform focused on extreme child poverty.

However, when analyzing how income poverty overlaps with material deprivation some limitations arise. There are some material deprivation indicators that are clearly outdated, for instance "access to TV, phone and washing machine". In Spain, only 1 per cent of the total population declared impossibility to access these goods, even if they are considered to be in a situation of extreme poverty using any of the other indicators that compose the "material deprivation" measure. Therefore, its discriminatory power in identifying people in extreme poverty is very low. In any case, old indicators should be kept to preserve consistency over time. Nevertheless, future rounds of the survey should include new indicators in order to better capture child poverty, intergenerational transmission of poverty, and community and subjective dimensions of poverty.

	IMV	**Scenario 1**	**Scenario 2**
a. Difficult or very difficult to make ends meet	57.8%	59.6%	58.8%
b. Severe material deprivation (cannot afford at least four of the following items)*	25,0%	25.5%	24.4%
Go on holiday one week a year	69.3%	69.8%	68.9%
Eat meat or proteins regularly	12.5%	13.6%	15.1%
Keep their home adequately warm	25.6%	25.2%	26%
Face unexpected expenses	69.1%	71.6%	70%
Pay their rent, mortgage or utility bills	25.2%	30.9%	26.3%
A car	22.2%	22.2%	20.4%

Table 2: Percentage of beneficiary households by type of severe material deprivation. The 2020 indicator also includes access to a television, telephone, and washing machine, but deprivation on these items is residual (about 1%). Source: Own compilation.

3 Evaluation of the program against energy poverty from a child perspective

The Ministry for Ecological Transition requested the ACPI's support to evaluate potential future reforms of the energy poverty benefits to evaluate, in particular, the coverage of different types of poor households guaranteed by the latest reform passed in October 2018. The analyses concluded that coverage was weaker for poor households composed by two adults with two children and households with more than two adults and at least one child, compared to all the remaining households including single parent families, large families or households composed only by low-pension retirees.

Without the precise information provided by the evaluation using ES-SILC data, typical households are unlikely to be considered as insufficiently covered and, therefore, adequately protected and supported. However, in the evaluation process, several limitations of ES-SILC became evident. The information contained in the survey does not allow to identify several of the special categories of beneficiaries defined in the "Bono Social" regulation such as people with disability equal or higher than 33%, people with specific support needs for basic daily-life activities, low-income retirees, and victims of terrorism and gender violence. Their identification was only possible by linking individuals included in the ES-SILC to specific administrative registers, and this kind of operations are not always easy or possible. Alternatively, some questions could be added to the ES-SILC questionnaire

allowing the interviewees to identify themselves, or not, as belonging to some of these groups of vulnerable people, with an explicit mention about the administrative recognition of their vulnerable condition when applicable.

In relation to this topic of energy poverty, the information collected by ES-SILC questionnaire in its current version remains too limited, especially if one takes into account the increasing relevance of this problem in a country like Spain. The indicator should be referred not only to adequate temperature during winter but also during summer, due to the strong impact that increasingly hot summers are having on particular regions and groups of people whose dwellings are not well prepared for such adverse climate conditions. In addition, also the cost of the service, even if approximate, is a crucial piece of information on this regard in order to facilitate the identification of hidden poverty. More generally, a rotating module on energy poverty could also be designed and implemented.

4 Child poverty dynamics in immigrant and non-immigrant

ACPI also carried out a comparison of child poverty dynamics among immigrant and non-immigrant households. Instead of focusing only on foreign versus national children, we linked child and household information to identify children living in immigrant and non-immigrant households according to their parents' place of birth. Both the foreign and the immigrant origin population are more likely to suffer higher poverty rates than the rest of the population in Spain due to several reasons, among which their occupational segregation and higher unemployment rates play a crucial role. The extent to which this situation translates into child poverty had not been carefully examined. Two important factors expected to affect child poverty by origin of the parents are eligibility and generosity of social benefits, and the age composition of the child population among immigrant and non-immigrant households, since we know that the cost varies depending on the age of the child.

The analysis of ES-SILC data, as summarized in Table 3, showed not only higher poverty rates among children but also, and most importantly, different patterns in the incidence of child poverty depending on the age of the child and their immigrant or non-immigrant origin. Poverty among children in immigrant households was approximately 3 times higher than among non-immigrant ones, before and after the crisis. However, in comparison to their non-immigrant counterparts, the poverty rate worsened much more among immigrant origin children below 4 years old (from 3.2 to 4.5 times higher), while it substantially reduced among adolescents.

	2008		**2017**		**2008**	**2017**
	Native origin	Immigrant origin	Native origin	Immigrant origin	Ratio (native vs. Immigrant)	Ratio (native vs. Immigrant)
Total	19.8%	57.1%	22%	65.4%	2.9	3
$<$ 4 years	17.3%	54.7%	16.5%	73.6%	3.2	4.5
4 - 14 years	20.8%	56.2%	22.2%	64.8%	2.7	2.9
15 - 17 years	19.9%	63.7%	28.5%	53.9%	3.2	1.9

Table 3: Child poverty rate reduction (in percentage). Source: Own compilation.

The identification of these different dynamics among immigrant and non-immigrant origin households was possible thanks to the possibility of linking parents and children in ES-SILC, as well as the inclusion of information on age of the children year by year, and the origin of the parents. However, lacking of information on the actual country of birth beyond the basic distinction EU and non-EU prevents from a more detailed understanding of potential internal differences within the immigrant population. Moreover, the Spanish version of EU-SILC does not provide information

either on whether children living in the surveyed households were born in Spain or abroad, nor their nationality. The omission of these two pieces of information seriously limits the possibility to better depict child poverty dynamics among immigrant origin children and, therefore, the convenience of designing specific policies to fight against it, or not, depending on the results. Finally, the major limitation of ES-SILC information when exploring poverty dynamics for immigrant populations, children or adults, has to do with the lack of information on their length of residence in the country. Length of residence in the country of immigration is the most critical variable in explaining different integration trajectories among immigrants, and its omission largely prevents a proper understanding of changes in living conditions of foreign-born people. For this reason, length of residence should be systematically asked in the survey for all the foreign-born members of the household.

By including information on place of birth for people below 16, and detailed information on years since arrival for all the foreign-born, ES-SILC would be a much better instrument to understand child poverty dynamics in immigrant households. In particular, the inclusion of these two pieces of information would allow to identify the role played by the recent arrival of their parents in Spain, the tougher impact of the crisis on immigrant households, and the different timing of entry into the labour market among immigrant origin adolescents compared to the non-immigrant ones to explain the observed patterns.

Here it is important to emphasize that the limited information collected by the survey in relation to the interviewees' retrospective labour trajectories greatly hampers a better understanding of the role played by this factor in shaping some of the reported differences. In its current version, the questionnaire collects monthly information about some dimensions of labour status but only for the previous year, which remains clearly insufficient to understand phenomena such as the increasing proportion of working poor or the very high rate or persistent poverty in Spain.

One additional issue related to the limitation in ES-SILC to study poverty dynamics in immigrant origin households has to do with the deficient design of the variable HX060 "Type of household". First of all, the category "others" is a black box particularly frequent for specific groups such as immigrant origin households in vulnerable situations. A better understanding of the composition of their households seems convenient. Secondly, regardless of the immigrant origin of the household, this variable classifies as dependent children individuals from 18 to 25 if they are not active in the labour market. Finally, from a user-friendly perspective, it would be desirable to add a different variable that considers categories that specifies single parent, cohabitating parental couple, etc., and always considering the number of children in the household.

Finally, in future versions of the survey, it would be ideal to provide information that allows to identify different family units within the households, an important issue taking into account that multi-family households are much more common among vulnerable groups compared to the average household.

5 Impact of housing prices on child poverty

ACPI also conducted a comparative analysis of the components and dynamics of child poverty in Spain focusing on the role played by the housing market before, during and after the crisis, using ES-SILC data. First, the results clearly showed how the "housing cost overburden rate" (when the cost of housing represents more than 40 per cent of household total disposable income) is systematically associated with poverty in Spain. Approximately, 40 per cent of the poor suffer overburden rate, and this percentage has hardly varied over the period 2007-2017. Second, child population, again, suffers a higher overburden rate than the adult population, regardless of poverty. In addition, overburden

is much more frequent for child population in Spain than in the rest of Europe, especially if they are poor children. Moreover, more importantly, in Spain the housing cost overburden rate is larger among the child population in poverty than among the total population in poverty, while the opposite pattern is observed in the EU, on average. This situation highlights the weak housing policies and their major contribution to poverty in general, and child poverty in particular, in our country.

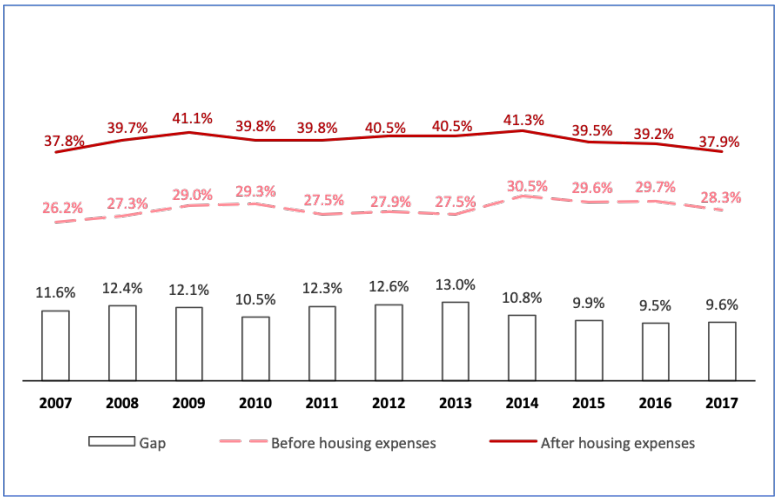


Figure 2: Child at-risk-of-poverty-rate before and after housing expenses in Spain (2007-2017). Source: EU-SILC (2007-2017).

When analyzing these data, it became evident the lack of some relevant information on the dwelling characteristics, as well as regarding the process of housing acquisition and/or renting. Given the high proportion housing expenditures represent over the total household income in Spain, and their major contribution to income poverty, additional information on the registered value of the dwelling, the pending mortgage, and the existence of other real estate properties beyond the usual dwelling, are relevant for a better understanding of the living conditions of the household members and their financial constraints.

6 Module on childcare

Finally, another area with room for improvement to contribute to a better understanding of child poverty and, more generally, inequality in children's living conditions has to do with the battery of questions about childcare included in the EU-SILC questionnaire. The module collects information on the weekly number of hours of different types of care provided to children under 12 in the household (parental care, schooling, early childhood education and care (0-3 years), at-home professional care or grandparental care). Unfortunately, this information is of very limited use in understanding increasing patterns of inequality among children if complementary information on aspects such as the cost of these services, the public or private nature of the child care centre, the distance from the dwelling to the (closest) child care centre, or the reasons why these services are not used, are fully omitted in the survey. This complementary information is crucial in order to identify obstacles and inequality in accessing public education services, for instance, and reasons underlying the decision to take children to childcare and early education, which is known to be crucial (conditional on quality) for the future development and opportunities of the children.

In relation to this topic, better information on school expenditures, with special attention to school meal services and nutrition remains an urgent need for both policy makers and researchers.

7 Other limitations and potential improvements

In this article, we have revised some of the applications that the High Commissioner against Child Poverty made over 2018/2019 using the valuable information contained in EU-SILC and, in particular, ES-SILC for the analysis of multiple dimensions of child poverty in Spain. EU-SILC has become a central instrument for exploring crucial social problems in our societies, with the enormous advantage of its highly standardized quality, periodicity and international comparability. As this article has briefly summarized, its potentialities for better understanding of child poverty and its multidimensional nature are multiple. However, there is room for improvement in relation to areas such as housing, childcare, school expenditures and labour market trajectories, which are strongly linked to poverty dynamics in general, and child poverty in particular.

SPANISH JOURNAL OF STATISTICS
Vol. **1** No. 1 2019, Pages 65–72
doi: https://doi.org/10.37830/SJS.2019.1.07

OFFICIAL STATISTICS

The forthcoming reform of the Spanish Living Conditions Survey: Extension proposals from an applied perspective

Jorge Onrubia
Universidad Complutense de Madrid (ICEI - UCM) and FEDEA

Abstract: The Spanish Living Conditions Survey (ECV) is the statistical production for Spain of the European Statistics on Income and Living Conditions (EU-SILC), encouraged and coordinated by Eurostat. This article presents some suggestions of improving the Living Conditions Survey elaborated by INE, in view of its upcoming reform, foreseen within the updating process of the EU-SILC project promoted by Eurostat. In particular, this paper reviews the incorporation of information from administrative registers, especially data from tax sources, as well as others from Social Security records. This work includes a series of proposals to improve the quality of currently gathered information in the ECV for Spanish personal income tax (IRPF), social contributions and social benefits. In addition, some extensions are proposed in relation to information on access to the main dwelling and its financing.

Keywords: ECV, EU-SILC, Household Surveys, Microdata, Tax Register Data

MSC: 62P20, 62P25, 68P01

1 Introduction

The European Union Statistics on Income and Living Conditions (EU-SILC) is a survey aiming to collect timely and comparable cross-sectional and longitudinal multidimensional microdata on income, poverty, social exclusion and living conditions, as well as other socioeconomic aspects such as housing, health, education, and benefits Eurostat (2020a). This instrument is anchored in the European Statistical System (ESS).

The EU-SILC project was launched in 2003 on the basis of a "gentlemen's agreement" in six Member States (Belgium, Denmark, Greece, Ireland, Luxembourg and Austria) and Norway. The EU-SILC database was released for the first time in 2004 for the EU-15 (except Germany, the Netherlands, the United Kingdom) and Estonia, Norway and Iceland. For a full description of the EU-SILC country-converage please consult the implementation graph in Eurostat EU-SILC website.[1]

[1]https://ec.europa.eu/eurostat/documents/203647/203704/EU-SILC+implementation+by+country.pdf

The Spanish Living Conditions Survey (Encuesta de Condiciones de Vida, ECV) is the statistical production for Spain of the European Statistics on Income and Living Conditions (EU-SILC), encouraged and coordinated by Eurostat, the European Union Statistics Office. Therefore, the ECV belongs to the set of harmonised statistical operations for European Union countries. The fundamental objective pursued by EU-SILC national sections is to have a comparable source of reference on statistics on income distribution and social exclusion in the scope of the European Union (extended to other countries of the continent, such as Norway, Iceland or Switzerland).

The aim of this article is to present some suggestions to improve the Living Conditions Survey prepared by INE, in view of its upcoming reform, foreseen within the updating process of the EU-SILC project promoted by Eurostat. The content of this article was presented on 6 September 2019, in the round table organised by the Spanish Statistical Office (Instituto Nacional de Estadística, hereinafter, INE), within the XII Public Statistics Conference (Jornadas de Estadística Pública) held in Alcoy.

Essentially, this article reviews the potential that, in my opinion, the incorporation of information from administrative registers may have, particularly data from tax registers. The quality of these registers and the extensive experience in their maintenance provide a great opportunity to enhance a great statistical product such as the ECV produced by the INE since 2004.

2 The possibilities of using administrative records in ECV

A comprehensive guide to the possibilities that can be offered by incorporating individualized statistical information from administrative registers (corresponding to population, taxes, social security data, public social benefits, and health and education activities, as main areas) can be consulted in J´antti et al. (2013). As highlighted in this report, "In the specific context of social statistics, the re-use of existing data and in particular administrative data has been identified by the European Statistical System as a key developing area in the process of modernising and streamlining social surveys". However, we have to take into account the different degree of use of administrative data by countries. In some cases it depends on the quality and development of administrative records. In other cases, it depends on the legal limitations related to the dissemination of personal information. The importance of the availability and usability of such data by national statistical offices, including timeliness of receipt and comparability, should also not be overlooked. Finally, an issue that is sometimes more difficult to achieve is the cultural change both in public administrations that provide the data and in the national statistical offices.

However, between countries, there are still notable differences on how the information is obtained to build the national databases to be integrated into EU SILC, as well as in their degree of intensity when combined. Even in direct surveys there are differences. Most countries carry out face-to-face interviews by Paper And Pencil Interviewing (PAPI), while other use Computer Assisted Personal Interviewing (CAPI). In Germany, the interviews are usually self-administered by respondents. The use of administrative records has been a widely used source of information in the Nordic countries for decades, as well in the Netherlands and Slovenia, and to a limited extend also in France, Ireland and Latvia. Conducting telephone interviews to supplement information obtained through the use of administrative sources is also a method used in some countries. In Spain, as explained José María Méndez in his article in this same volume of the Spanish Journal of Statistics (Méndez, 2019), in the Spanish ECV, since 2005, information has been collected through CAPI, with a pilot test having been carried out in 2017 (which is expected to continue in 2019), in which the so-called "multi-channel collection" CAWI-CATI-CAPI is being trialed. With the modernisation of data collection, greater

efficiency in ECV production is pursed, although it is recognised that this requires major challenges to be faced, given the complexity of ECV itself, and of the information it incorporates.

On consistency over time and on comparisons between countries in EU SILC when combining data provided directly from the respondents' questionnaires with data obtained from administrative records, you can see Krell et al. (2017). This article shows that achieving successful administrative data usage is not an easy task, and that methodological transparency is essential for reaching such an objective.

In the case of Spain, in the Living Conditions Survey 2013, a new methodology was adopted in the production of data regarding household income, consisting of the use of administrative files, although the information available in the questionnaires continues to be used as well. In my opinion, the incorporation of information from administrative registers implies an improvement in the quality of the data. The use of these administrative sources, given the high quality of the statistical records of the Spanish public administrations, represents a significant efficiency gain in the data collection method.

Although the first wave of the ECV that incorporated administrative-source income was that of 2013 (annual income corresponding to 2012), since the Tax Identification Number (NIF) had been available since 2009 (with 98% coverage), retrospectively data have been provided with this source for the waves from 2009 to 2012 (with annual income, respectively, for 2008 to 2011). This duplication of information in those waves has been useful in order to be able to comparatively analyse the differences between both sources and their impact on the measurement of inequality and poverty, among other issues.

Initially, the objective pursued was to enhance the production of household income variables. In this way, household income related data are prepared using a mixed methodology, which combines the information provided by the respondent with the administrative records of the State Tax Administration Agency, Social Security, the Tax Office of Navarra and the Diputaciones Forales of Bizkaia and Gipuzkoa. It is worth noting that the Diputación Foral of Alava did not participate in this contribution of information. In this case, it is simply a question of lack of willingness, under a peculiar interpretation of tax autonomy.

3 Proposals for the extension of the ECV

This section of the article incorporates some proposals with possible extensions of the information currently contained in the ECV, which could be developed mostly from the administrative sources (from tax and social security administrations) already used since the methodological change of the 2013 wave.

3.1 Annual ECV extension modules

The ECV incorporates an annual module pre-defined by Eurostat. These modules complement the basic information coordinated according to Eurostat's EU-SILC project. These ad-hoc modules are developed each year and complement the information provided by the variables permanently collected in EU-SILC, with supplementary variables about unexplored aspects of social inclusion Eurostat (2020b). Since 2005, which included a module dedicated to the Intergenerational transmission of poverty, the ECV has incorporated the following modules: (2006) Social participation, (2007) Housing conditions, (2008) Over-indebtedness and financial exclusion, (2009) Material deprivation, (2010) Intra-household sharing of resources, (2011) Intergenerational transmission of disadvantages, (2012)

Housing conditions, (2013) Wellbeing, (2014) Material deprivation, (2015) Social/cultural participation and material deprivation, (2016) Access to services, (2017) Health and children's health, and (2018) Material deprivation, well-being and housing difficulties. For the ECV 2019, the module established is Intergenerational transmission of disadvantages, household composition and evolution of income. As it can be seen, chosen themes are often recurrent, although variations are introduced.

One possibility, in my opinion, of great value, would be to offer a three-year module linked to the subject matter of the three-yearly Encuesta Financiera de las Familias (EFF) (Bank of Spain, 2020), drawn up by the Bank of Spain and included in the Spanish National Statistical Plan.[2] The main objective of this module would be to enable consistency analysis between the ECV and the EFF, as well as to carry out complementary studies on financial and savings decisions of Spanish households. The design of the module should take into account that the reference year of the annual income collected in this wave of the ECV (that of the previous year), from the personal income tax (IRPF) records, would coincide with the reference year of the EFF.[3]

With regard to the information to be incorporated, the questions asked of the interviewed households should be related to savings (proportion of income allocated to savings), indebtedness (level and destination) and the composition of the portfolio of assets that make up their wealth. As background, we can mention the ECV 2008 module aimed at "Over-indebtedness and financial exclusion", and the one planned for ECV 2020, on "Over-indebtedness, consumption and wealth".

An experience on ad-hoc modules of the EU-SILC with the incorporation of information on consumption and wealth of households, with an orientation to the testing of the HBS and the HFCS, is the one developed by the Statistical Institute of Belgium for the SILC of Belgium (Statistics Belgium, 2018). The experience, as recognized in the assessment study is the high non-response in the case of consumption questions, the underestimation in the case of wealth questions, largely attributed to the breadth and little specificity of the questions. It may also be justified by disincentives for households to transmit reliable information (which we also know it takes place in the HFCS).

3.2 Broadening the information contained in the ECV from tax records

As mentioned above, there is now a broad consensus about the enormous potential role of administrative records in enriching the information provided by household surveys, especially tax records. According to the coordination guidelines established by Eurostat, from 2013 onwards, the information on household income provided in the national surveys on income and living conditions integrated into EU-SILC will come from tax records contained in the files of tax administrations for the management of personal income taxes.

The use of these tax sources has also made it possible to incorporate administrative information on personal income tax (IRPF) payments into these surveys. Likewise, the administrative records of the Social Security allow the incorporation into the ECV of the amounts corresponding to the social contributions paid by employees, and self-employed. Specifically, the Spanish ECV incorporates, in each wave and for each member of the household, the total amount of withholdings charged and payments on account made for IRPF and Social Contributions withheld or paid during year $t-1$ (the year for which the annual income is offered in the ECV). In addition, the amount, positive or negative

[2]The Spanish Survey of Household Finances (EFF) is part of the Eurosystem's Household Finance and Consumption Survey (HFCS), coordinated by European Central Bank, which collects household-level data on households' finances and consumption (European Central Bank, 2020). The HFCS datasets from the first and second wave were released respectively in April 2013 and December 2016.

[3]The last wave released by the EFF has 2017 as reference year (Bank of Spain, 2019).

(depending on whether it is an amount paid or a refund received), of the differential tax liability for year $t-2$ is also incorporated as variable.

However, with the complete information from the IRPF and Social Contributions records, we think it would be useful:

1. To include a variable for the "differential tax liability" of IRPF corresponding to the fiscal year $t-1$ (the reference year of the annual income, and the withholdings and payments on account of the IRPF). This would make it possible to have the "total tax liability" of the IRPF, corresponding to that reference year. Nowadays, users have a "total tax liability" of the IRPF, approximated by a cash criterion: withholdings charged $(t-1)$ + payments on account $(t-1)\pm$ differential tax liability $(t-2)$.
2. To offer separate information on the amounts corresponding to the total amount of employee/self-employed made for IRPF during year $t-1$. This information is provided by the State Tax Administration Agency (AEAT), and by Basque Country Provincial Tax administrations, and Navarre Foral Tax Administration). And, on the other hand, the amount of employee/self-employed Social Contributions provided by the Social Security. Currently, variable HY140G includes the sum of IRPF's withholdings charged and payments on account for $t-1$, and Social Contributions paid in $t-1$, while the differential tax liability for $t-2$ is provided in variable HY145N.

Likewise, it would be advisable to improve the information referring to income from personal entrepreneurial economic activities, identifying the estimation regime applied in the IRPF (direct estimation of incomes and expenditures, objective estimation by means of modules, and objective estimation by coefficients for farmers and stockbreeders). Also, if it is the case, it would be interesting to reflect whether the data included in the survey comes from the tax return or from the response to the interview form. In relation to this question, it would be very useful for studies on compliance and tax evasion, to ask in the interview questionnaire for the amount obtained from this type of income, in order to compare it with the income reported in the IRPF return.

3.3 Detailed information on Spanish personal income tax (IRPF) in the ECV

In the same line, it would be desirable to incorporate, in the ECV's file of members, more detailed information on IRPF variables, available in the annual IRPF's withholding statement of household members. In this case, from the perspective of the ECV user, the objectives would be:

1. To allow greater precision in the analysis of the effects of taxation on personal income;
2. The possibility of using ECV, exploiting the richness of its socio-economic variables, to carry out micro-simulation exercises for alternative IRPF designs. Within this field, this extension of IRPF variables would be very useful to provide greater precision and reliability to the EUROMOD tax "modules".[4]

A proposal of IRPF's variables to be incorporated could be the following:

- Variables of primary interest: type of tax filing (Individual, Joint taxation, only withholdings); general taxable base and savings taxable base; amount of minimum personal and family allowances; State/Autonomous Communities gross tax liabilities; State/Autonomous Communities tax credits; total tax liability; differential tax liability; total refundable tax credits.

[4]EUROMOD is the open-access Tax-Benefit Microsimulation Model for the European Union, for the European Union. EUROMOD allows researchers and policy analysts to calculate, in a comparable manner, the effects of taxes and benefits on household incomes and work incentives for the population of each country and for the EU as a whole. For more information, see EUROMOD website: https://www.euromod.ac.uk

- Other variables with detailed information: amounts of income components by source (labour, financial and non-financial capital incomes, rents, business and professional incomes, capital gains), contributions to pension plans and other social protection systems, details of personal and family allowances, etc.

3.4 Social benefits incorporated in ECV

Another interesting possibility for extending the information offered by the ECV concerns the exploitation of administrative records containing data on social benefits. Firstly, for all benefits, it would be very useful to provide differentiated information on the public or private nature of the entity paying them. At present, this is not the case for benefits of an educational kind.

Another change that would be quite useful, from the analysts' point of view, would be the differentiation of severance payment and unemployment benefits. Both are now included in a single variable, and estimates must be made according to the legal limitation on the amounts of the latter.

Other possibilities to enrich currently provided information in the ECV could be:

1. Incorporation of information on the type of unemployment benefit, depending on whether it is a contributory benefit or family aid for the long-term unemployed. It would also be interesting for the analysis of this benefit to be able to identify the receipt of the subsidy in a single payment.
2. In the case of educational transfers, it would be convenient to incorporate some information on the educational level at which the aid is aimed (childhood education, primary, secondary, high school, and university education).

Likewise, it would be very useful for the territorial analysis of these benefits to be able to differentiate in one variable the Public Administration that makes the payment (State, Social Security, Autonomous Communities, Local Corporations).

3.5 Information on housing acquisition, finance, and rental

The dwelling in which households reside is a field of considerable interest for the expansion of information in a forthcoming ECV reform. It should be borne in mind that, on the expenditure side, the Spanish Household Budget Survey (Encuesta de Presupuestos Familiares, EPP) does not provide information on the house purchase, on considering it to be an investment, nor on other secondary dwellings owned, either for the use of by household members, or for rental.

The incorporation of information related to the dwelling should come from different tax registers, such as those of the Real Estate Tax (IBI), IRPF, the Property Transfer Tax (Impuesto sobre Transmisiones Patrimoniales Onerosas, ITPO) or the Inheritance and Gift Tax. In addition, the information available in the public Real Estate Registries could be used.

Given the volume of information available in the field of housing, the objectives to be achieved with the extension of the ECV should be specified. Thus, in the first place, the inclusion of new variables should be aimed at improving the analysis related to the acquisition, tenure and financing of real estate by households. This is an improvement already planned for 2021 in the action plan of the Spanish National Statistics Institute (INE). Secondly, the information to be included must make it possible to identify whether the main dwelling under the ownership regime was a new construction (first transmission) or was a used dwelling (second transmission), as well as to record the year of construction and the year of acquisition by the household.

It would also be desirable to incorporate the purchase price of the dwelling, the mode of acquisition (purchase, inheritance, donation), as well as the Cadastral Value (gathered in the administrative registers of the AEAT).

As regards house finance, it would be very useful to identify whether there are any outstanding loans, and the amount pending repayment in the reference year. It would also be interesting to be able to differentiate the type of loan involved (mortgage, other loans with financial institutions, loans from relatives), as well as the term, and the interest rate applied.

When it comes to second housing, it would be interesting to incorporate information on other (non-main) dwellings available to the user or rented (the ECV now only has data on the main dwelling of the household).

For main dwellings that are rented, the ECV should incorporate detailed information on the amount paid for rent, differentiated from the rest of the residence expenses that are paid by the tenant. In the case of secondary dwellings, it would be interesting to include information, as well as the main reason and duration of the rental.

4 Concluding remarks

The purpose of this article, as stated in the introduction, was to set out some possibilities for improving the information offered by the Spanish ECV, essentially by incorporating data from administrative registers. The forthcoming reform of the EU-SILC project, promoted by Eurostat, may be a good opportunity to enhance the use of administrative records as a source of rich and reliable statistical information.

In Spain, the quality of these administrative registers and the long experience in their maintenance and continuous improvement can provide an important added value to a valuable and contrasting statistical product such as the ECV carried out by INE. However, I am aware that progress in this area requires sufficient human and material resources to be provided to Spanish Statistical Office (INE). In my opinion, undoubtedly, the widespread use of the ECV over the past 15 years, in practically all areas of applied social sciences, justifies the necessary investment.

Acknowledgments

An earlier version of this paper was presented at the XII Public Statistics Conference (Jornadas de Estadística Pública) (Alcoy, Spain, September 9, 2019). I gratefully acknowledge the useful comments and suggestions received from the participants in the abovementioned meeting. We would also like to thank Andoni Montes for his valuable comments and suggestions after a careful reading of the manuscript. All the remaining errors are entirely our responsibility. I acknowledge the financial support of Spanish Ministry of Science, Innovation and Universities (before Ministry of Economy and Competitiveness), Project ECO2016-76506-C4-3-R.

References

Bank of Spain (2019). Encuesta Financiera de las Familias (EFF) 2017: métodos, resultados y cambios desde 2014. Technical Report 4/2019, Boletín económico/Banco de España.

Bank of Spain (2020). Survey of Household Finances. Technical report, Bank of Spain. https://www.bde.es/bde/en/areas/estadis/estadisticas-por/encuestas-hogar/relacionados/Encuesta_Financi/.

European Central Bank (2020). Household Finance and Consumption Survey (HFCS). Technical report, European Central Bank. https://www.ecb.europa.eu/stats/ecb_surveys/hfcs/html/index.en.htm.

Eurostat (2020a). European Union Statistics on Income and Living Conditions: Ad-hoc modules. Technical report, Eurostat. https://ec.europa.eu/eurostat/web/income-and-living-conditions/data/ad-hoc-modules.

Eurostat (2020b). European Union Statistics on Income and Living Conditions (EU-SILC). Technical report, Eurostat. https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions.

Jäntti, Markus, Veli-Matti Törmälehto, and Eric Marlier (2013). *The use of registers in the context of EU-SILC: challenges and opportunities: 2013 Edition*. Publications Office.

Krell, Kristina, Joachim R Frick, and Markus M Grabka (2017). Measuring the consistency of cross-sectional and longitudinal income information in eu-silc. *Review of Income and Wealth 63*(1), 30–52.

Méndez, José María (2019). The Spanish Survey of Living Conditions (ES-SILC). *Spanish Journal of Statistics 1*(1).

Statistics Belgium (2018). Preparation for the revision of EU-SILC : Testing of rolling modules in EU-SILC 2017 - Final Report. Technical report, Statistics Belgium.

Spanish Journal of Statistics

General Information

The Spanish Journal of Statistics (SJS) is the official journal of the National Statistics Institute of Spain (INE). The journal replaces Estadstica Espaola, edited and published in Spanish by the INE for more than 60 years, which has long been highly influential in the Spanish scientific community. The journal seeks papers containing original theoretical contributions of direct or potential value in applications, but the practical implications of methodological aspects are also welcome. The levels of innovation and impact are crucial in the papers published in SJS.

SJS aims to publish original sound papers on either well-established or emerging areas in the scope of the journal. The objective of papers should be to contribute to the understanding of official statistics and statistical methodology and/or to develop and improve statistical methods; any mathematical theory should be directed towards these aims. Within these parameters, the kinds of contribution considered include:

- Official Statistics.
- Theory and methods.
- Computation and simulation studies that develop an original methodology.
- Critical evaluations and new applications
- Development, evaluation, review, and validation of statistical software and algorithms.
- Reviews of methodological techniques.
- Letters to the editor.

One volume is published annually in two issues, but special issues covering up-to-date challenging topics may occasionally be published.

Author guidelines

The Spanish Journal of Statistics publishes original papers in the theory and applications of statistics. A PDF electronic version of the manuscript should be submitted to José María Sarabia, Editor in chief of SJS via email to sjs@ine.es. Submissions will only be considered in English.

Manuscripts must be original contributions which are not under consideration for publication anywhere else. Its contents have been approved by all authors and. A single-blind refereeing system is used, so the identity of the referees is not communicated to the authors. Manuscripts that exceed 30 journal pages are unlikely to be considered for publication. More detailed information can be found at https://www.ine.es/sjs.

Published by the Spanish National Statistical Institute