Papeles de Trabajo

AN ESTIMATION OF INCOME DISTRIBUTION USING GROUPED DATA: A GENERAL TWO-STEP

Authors: *Ignacio Moral-Arce*
Instituto de Estudios Fiscales
Antonio de las Heras
Universidad de Cantabria

P.T. n.º 6/2011

N.I.P.O.: 602-11-022-5

INSTITUTO DE ESTUDIOS FISCALES

N. B.: Las opiniones expresadas en este documento son de la exclusiva responsabilidad de los autores, pudiendo no coincidir con las del Instituto de Estudios Fiscales.

Edita: Instituto de Estudios Fiscales
I.S.S.N.: 1578-0252
Depósito Legal: M-23772-2001

INDEX

ABSTRACT

A general method to estimate income distribution functions using information provided in grouped data is presented in this paper. The aggregation may be symmetric (the number of individuals being the same in each interval) or asymmetric, (the number of individuals being different in each interval). This technique enables us to build a worldwide (or national) density function on the basis of national (regional) data which is much more accurate than standard approaches in this type of literature. Our general method is developed in two stages: in the first stage we obtain a Lorenz curve, in the second, using the results of the previous stage an income density function is derived through non-parametric techniques. In addition, several Monte Carlo experiments that prove the good sample properties of our estimator are shown. Finally, two empirical applications of our method of estimation are proposed for both symmetrical and asymmetrical groups.

Keywords: Income distribution, kernel function, nonparametric density estimation, grouped data.

JEL classification: D31, D63, H23.

1. INTRODUCTION

The study of income distribution is a crucial issue in the analysis of inequality and poverty, from a political and socio-economic view, and it is a major concern for individuals, economists, and the government. The determinants of poverty and inequality, and the actions needed to reduce them are equally important. Besides these aspects, the calculation of income distribution functions, and evolution of such distribution over time, is useful to analyze social mobility, the impact of redistributive policies, etc, at different levels (personal, regional, national and global).

A current subject in the applied economic literature focuses on the calculation of income aggregate functions derived from subgroups, i.e. the estimation of global (national) income distribution by integrating the income distribution of countries (regions). When thinking of the world as a unit of analysis, it is interesting to obtain the income of citizens in the whole planet. To compute world income countries are used as units of study. But, if the researcher takes into account the different population size of these units (two countries as China and Denmark have different weights in the calculation of a hypothetical global income function), it is not appropriate to treat each country as a "unit" of analysis (obviously each country does not have the same weight in the aggregation). Therefore, it is necessary to use income distributions weighted according to the population of every country. But this may not be enough to obtain good estimates of poverty and inequality because these methods can obviate the existence of within- country dispersion, therefore implying that all individuals in a country have the same level of income. As a consequence, biased results are obtained. A general method may be necessary, that will allow the aggregation of regional (domestic) functions through the aggregation of "micro" income values generated by each income function in the regions. For more details of global income distribution, see Milanovic (2002) and Sala-i-Martin (2002).

The estimation of income distribution functions essentially depend on the data available to researchers. Such data are obtained through various means: administrative records, censuses, samples, surveys, panels, etc. However, in many cases, the information available to researchers is limited to grouped data (quantiles) of income from household surveys or administrative records. Moreover, grouped data are the only source of information on income distributions in most countries (or regions) playing an important role in the determination of poverty and inequality at 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 (intervals) organized by ascending order of income level. This grouping may be symmetrical (also called quantiles), the number of individuals in each of the intervals being the same, or asymmetrical, the number of individuals associated with each interval being different.

From a theoretical point of view the estimation of the density function can be performed in two different ways: the parametric approach and nonparametric approach. The choice between them depends on how researchers use available information, which is indicated by the specification of the model (structure and parameters) and the data used. Nonparametric methods give more importance to the information from data, whereas the parametric approach gives more weight to the specification of the model, i.e. the hypotheses of the model: the income distribution is symmetrical, linear parameters, etc, and the data must be adapted to this fixed format.

According to the above paragraph, economic literature has proposed two approaches to obtain estimates of density functions or Lorenz curves from grouped data. The first approach is based on parametric estimates of density functions and Lorenz curves. Two of the most widely used methods for estimating this curve are the Quadratic Lorenz curve [proposed by Villasenor and Arnold (1989)], and the Beta Lorenz curve [proposed by Kakwani (1980)]. Both methods perform relatively well in the case of uni-modal distributions, however, the method by Villasenor and Arnold presents a clear advantage to researchers over that by Kakwani, because its computer implementation is much simpler (Datt, 1998). Other noteworthy studies on parametric estimation methods include Ryu and

Slottje (1996), Ravallion and Huppi (1989), and Cheong (2002), in which the most widely-used Lorenz functional forms are compared, and Kakwani and Podder (1976), Rasche, *et al*. (1980), Ortega, *et al*. (1991), Griffiths, *et al*. (2005), and Minoiu and Reddy (2006a). The main drawback of the parametric approach is that if there is a misspecification of the model (i.e., the assumptions regarding the density function underlying the grouped data are not true) the performance of the estimator is poor. This implies that inferences based on the model's results will be questionable in terms of accuracy.

The second approach, and the most widely followed by the current literature involves the non-parametric estimation of the income distribution, which is the direct estimate of the distribution functions through the use of kernels (for details see Silverman, 1986). This approach can be applied to various types of research such as the study of poverty and inequality in Sala-i-Martin (2002, 2006), Ackland, Dowrick and Freyens (2006), and Fuentes (2005), etc. The accuracy of the results depends essentially on the kernel functions and bandwidth used in the calculation of the density function underlying the grouped data utilized (see Minou and Reddy, 2006b, Wu and Perloff , 2003, and Milanovic, 2006 for the study of global income distribution). These non-parametric techniques perform well when the number of observations available to researchers is high. However, typically in these kinds of studies, the available data are limited to five figures (quintiles). The combination of "limited structure" and "limited data" produces results that are, in turn, of limited value (nonparametric bias estimator).

From a practical point of view, the nonparametric estimator of the density function with grouped data performs well if two conditions hold. The first condition is that the grouped data be available in quantiles (in fact, grouped data presented in the literature are in the form of quintiles) since these values provide very relevant information about the underlying distribution. The second condition concerns the shape of the underlying distribution. Specifically, it is necessary to assume symmetric functions, as in the case of the log-normal distribution. In asymmetric situations (due to the underlying distribution), the goodness-of-fit of the estimated density function is poor.

The aim of this paper is the estimation of the income distribution on the basis of grouped data through a two-step method. In the first step, the Lorenz curve is estimated by applying some of the most widely used methods proposed in the literature. In the second step, using the first stage estimates, the nonparametric function of income distribution is obtained. Furthermore, the robustness and accuracy of our two stage method is analyzed. Several Monte Carlo simulations are performed and two applications for different types of grouped data are presented. Our general method allows the aggregation of global (national) functions (domestic) through the aggregation of "micro" income values produced by each function of income in a country (region). It is important because this technique does not add income functions, but the microdata produced by each function.

The paper is organized as follows. In the next section, the characteristics of the data are discussed. In the third section, the econometric specification of our estimation method is presented and the properties of the estimators are described. The sample behavior of the estimator is shown in section 4, using Monte Carlo simulations. In Section 5 two applications of our estimation method for different types of grouped data are proposed. Finally, the conclusions are shown in Section 6.

2. DATA DESCRIPTION

As noted in the introduction, the estimation of density functions, the measures of inequality and poverty depend crucially on how the information is available and has been grouped. Generally, researchers have data grouped in intervals: An interval represents different income levels of individuals grouped in ascending order. A data source can be household surveys or administrative records. Usually, the available information is in quantiles if the information origin is a survey, whereas if the source is administrative files, income intervals often contain different numbers of individuals in each interval. A general representation of grouped data is given in the following table.

Table 1
INCOME GROUPED AND RELATIVE ACCUMULATED DATA

Data	Accumulated Percentage					
Interval	$0-\bar{l}_1$	$\bar{l}_1-\bar{l}_2$	...	$\bar{l}_{j-1}-\bar{l}_j$	More than $\bar{l}_j$	Total
Number of Individuals	n_1	n_2	...	n_j	n_{j+1}	N
Mean Income	$\bar{X}_1$	$\bar{X}_2$		$\bar{X}_j$	$\bar{X}_{j+1}$	$\bar{X}$
Percentage of Accumulated Population	$p_1=\frac{n_1}{N}$	$p_2=\frac{n_1+n_2}{N}$	...	$p_3=\frac{n_1+..+n_j}{N}$	$p_{j+1}=1$	$N=n_1+n_2+\ldots+n_{j+1}$
Percentage of Accumulated Income	$q_1=\frac{\bar{x}_1n_1}{N\bar{X}}$	$q_2=\frac{\bar{x}_1n_1+\bar{x}_2n_2}{N\bar{X}}$	...	$q_3=\frac{\bar{x}_1n_1+\ldots+\bar{x}_jn_j}{NX}$	$q_{j+1}=1$	$N\bar{X}=x_1n_1+x_2n_2+\ldots+x_{j+1}n_{j+1}$

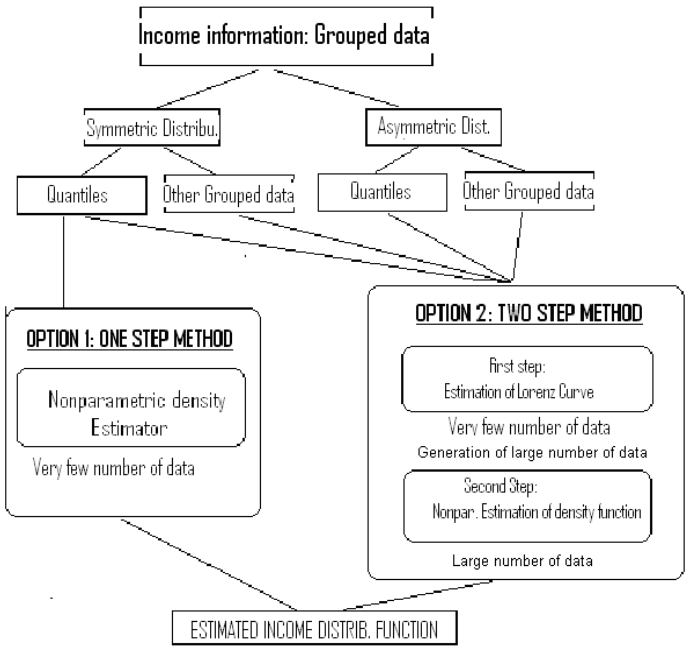
Where $\bar{l}_i$ denotes the exteme values of the interval. The quantiles case is given when $N_i = N_j$ for all i, j, (all intervals have the same number of individuals). If the information comes from administrative sources, the number of individuals of each interval is different, which is a crucial difference to the previous situation. Most works use information provided in quantiles. However, papers that use asymmetric information are quite rare in the literature.

In order to apply our method of estimation, it is necessary to transform the accumulated frequencies information provided in Table 1. This accumulation process is observed in the bottom rows of table 1. These rows show the percentage of accumulated population "p" and the proportion of accumulated income "q" in each interval. Taking into account this representation of the income information, our procedure is developed in the next section.

3. ECONOMETRIC SPECIFICATION

In this section a two-step estimation method of the income density function is proposed. The reasons for our two-step method are outlined in the following diagram:

Figure 1
DIAGRAM OF DENSITY ESTIMATION



The non-parametric estimation method (option 1 of Graph 1) used in Sala-i-Martin (2006) among others has good properties if the information of an underlying symmetric distribution is given in quantiles. In all the other cases such estimates do not work adequately. In addition, non-parametric techniques are appropriate when the number of observations is high. Even in case we have quantiles, the combination of an “unrestricted” estimation method (the non-parametric) and very few data leads to undesirable results.

For these reasons, two steps are proposed in an estimation process. In the first step, the Lorenz curve is estimated using information grouped in intervals (whether they are symmetrical or not). With the estimation of the Lorenz curve "n" values of income are produced. With these values, in the second stage the density function is estimated through a non-parametric technique. This approach allows us to transform the information given by quantiles (insufficient information for the estimation of the nonparametric density function) in a vector of all fictitious values of income needed to estimate the density function more appropriately.

3.1. Lorenz Curve Estimation

The Lorenz curve contains all the information on income inequalities in a population. This curve shows values in relative terms, so the information is not given in absolute values. According to Table 1, a generic representation of the Lorenz curve is given by the following function:

$$q = f(p), \tag{1}$$

Where “p” is the cumulative percentage of the population and “q” is the cumulative percentage of income, f(.) is an unspecified function. Usually, this function is expressed as follows:

$$q = f(p, \pi), \tag{2}$$

Where π is the vector of parameters that specifies the Lorenz curve. There are several studies that estimate the Lorenz curve by providing different types of functional forms, all of these studies can be used in our estimation method.

The non-parametric approach, as developed by Hasegawa and Kozumi (2003) is based on equation (1). The parametric approximation given in (2) allows the use of different approaches. Ryu and Slottje (1996) propose to estimate the Lorenz curve through Bayesian techniques and polynomial functions.

Continuing with the parametric approach, the classical ones estimate the Lorenz curve assuming an explicit form of the underlying income function. In this work, the focus is on the Ryu and Slottje (1996) approach on grounds of flexibility: the assumptions of the estimation method and the simplicity of the estimation procedure[1]. Following the work of Ryu and Slottje (1996), we order the sample from the smallest to the largest values of income $(x_{(0)}, x_{(1)}, \cdots, x_{(1)})$ where “I” is the sample size. The logarithm of $x_{(i)}$ can be approximated by:

$$\log x_{(i)} = \sum_{m=0}^{M} \beta_m p_i^m + u_i, \tag{3}$$

With $p_i = i/I$ and u_i has a symmetric distribution with a zero mean. We use the least squares method to estimate parameters β_m. Let the approximated income of the jth poorest person be:

$$x^*\left(p_j^*\right) = \exp\left[\sum_{m=0}^{M} \hat{\beta}_m p_j^{*m}\right], \tag{4}$$

In order to calculate the Lorenz curve, the final expression is given by:

[1] The objective is to emphasize the flexibility of our estimation method in this stage. In order to obtain the Lorenz curve any of the approaches presented can be used, whether parametric or non parametric. The choice of the approach of Ryu and Slottje (1996) is only used for simplicity reasons. It is advisable you apply different methods of estimation and compare the results.

$$q^*\left(p_j^*\right)=\frac{1}{I\hat{\mu}}\sum_{j=0}^{i}\hat{x}\left(p_j\right)=\frac{1}{I\hat{\mu}}\sum_{j=0}^{i}\exp\left[\sum_{m=0}^{M}\hat{\beta}_m\,p_j^{*m}\right] \qquad (5)$$

Where $\hat{\mu}$ is the estimated mean of the income sample, p^* is the vector of values of cumulative percentage of the population and q* is the vector of accumulated income derived from the estimates of the parameters of the Lorenz curve. For more details about the estimation process see Ryu and Stottje (1996). The data obtained at the end of this stage are arranged in a matrix of size $2X\left(k+1\right)$, where "k" indicates the number of times the distribution functions have been divided ("p*" and"q*(p*)").

3.2. Data transformation and density estimation

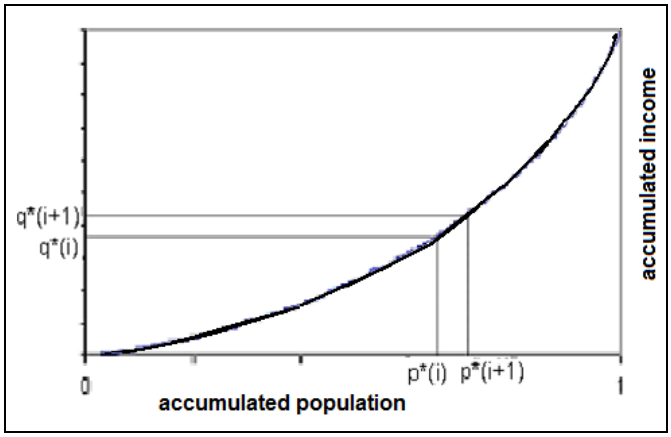
Based on the Lorenz curve obtained in the previous subsection this step focuses on the estimation of the income density function. The information provided in the above step consists of the percentages of accumulated number of individuals $\left(p_i^*\right)$ and accumulated income $\left(q_i^*\right)$. The $2X\left(k+1\right)$ matrix obtained in the previous stage (we consider the case of k = 100) is as follows:

$$\begin{array}{c|ccccc} p_i^* & 0 & p_1^* & p_2^* & \cdots & p_{99}^* & 1 \\ q_i^* & 0 & q_1^* & q_2^* & \cdots & q_{99}^* & 1 \end{array} \qquad (6)$$

The values of p* and q* of the Lorenz curve are expressed in relative and accumulated terms. However, to calculate the income distribution function it is necessary to use absolute values and non accumulated values. Therefore, the next step is the transformation of accumulated and relative values of income and population in their counterpart values (income) and frequencies (population) in absolute terms. Next, this transformation process is shown. For illustration purposes two consecutive values of the Lorenz curve are represented.

Figure 2

TWO CONSECUTIVE POINTS OF LORENZ CURVE



Starting with the x-axis, the variable p^* (percentage of number of individuals) represents the N units (total individuals) grouped into k intervals of "accumulated" income.

$$\left[0,p_1^*\right],\left(p_1^*,p_2^*\right],\ldots,\left(p_{98}^*,p_{99}^*\right],\left(p_{99}^*,1\right] \text{ where } 0<p_1^*<\ldots<p_{99}^*<1$$

The maximum of this variable is "1", f_i^* is considered as the relative frequency of individuals with an income in the interval $\left(p_i^*,p_{i+1}^*\right]$ p^* reflecting the accumulated frequency. The difference between two consecutive values $\left(p_i^*,p_{i+1}^*\right)$ gives the relative frequency in this interval $f_i^*=p_{i+1}^*-p_i^*$. The next step is to obtain the absolute frequency. The number of individuals in the interval is given by $n_i^*=f_i^*xN$.

The y-axis represents the different values of income (the variable q^*). Two calculations are made: First, the total income of the population is obtained $RT = \overline{X} x N$, where $\overline{X}$ is the mean income of the population; and N is the total number of individuals. Secondly, the difference of two consecutive values of the accumulated income $\left(q_i^*, q_{i+1}^*\right]$ is calculated. This quantity indicates the percentage of income existing in the range, i.e., $r_i = q_{i+1}^* - q_i^*$. This value is bounded between 0 and 1. By multiplying it by the value of total income (RT), R_i is obtained. As a result, the amount of total income (RT) accumulated in the interval, is: $R_i = ri \text{ x } RT$. Now, $\left(2Xk\right)$ series are produced:

$$\begin{array}{c|cccccc} R_i & R_1 & R_2 & \cdots & R_{98} & R_{99} & R_{100} \\ n_i^* & n_1^* & n_2^* & \cdots & n_{98}^* & n_{99}^* & n_{100}^* \end{array} \qquad (7)$$

These series indicate that there are n_i^* individuals in an interval with a total income of R_i. Therefore, the value of the income of each individual in the interval is obtained as follows: $Xi = \frac{R_i}{n_i^*}$, which leads to the following series:

$$\begin{array}{c|cccccc} X_i & X_1 & X_2 & \cdots & X_{98} & X_{99} & X_{100} \\ n_i^* & n_1^* & n_2^* & \cdots & n_{98}^* & n_{99}^* & n_{100}^* \end{array} \qquad (8)$$

Now, there are n_i^* individuals with an income (each unit) of X_i, i.e., the value of the income X_i is repeated n_i^* times in the population. This procedure allows us to transform the information given in Table 1 in a simulated vector of $\left(X_i, n_i^*\right)$ values that represents the same underlying income function.

According to these series, the income function is estimated through non-parametric methods. One of the features of the non-parametric estimator of the density function $f(x)$ is its calculation, no assumptions about the actual density function being made (i.e. this function belongs to a family of parametric functions). The estimator of the density function is given by the following expression:

$$\hat{f}_h(x) = \frac{1}{n^* h} \sum_{i=1}^{n^*} K\left(\frac{x - X_i}{h}\right), \qquad (9)$$

Where "x" is the value of income, where the density function being evaluated; and with "n^*" is the total number of available data. We define "h" as the size of bandwidth and $K(\ )$ as a kernel function. For more details on the non-parametric approach see Härdle (1991) and Härdle *et al.* (2004). The kernel density estimation given in equation (9) requires two parameters: the kernel function "K" and the bandwidth parameter "h". In practice, the choice of the kernel is not nearly as important as the choice of the bandwidth. The econometric theory cannot provide us with a method to select the bandwidth that is applicable in practice and theoretically desirable. We introduce two of the most frequently used methods of bandwidth selection: the plug-in and the cross-validation method. The optimal bandwidth for a kernel density estimate is typically calculated on the basis of an estimate of the mean integrated squared error of the density function. For detailed overviews on plug-in and cross-validation methods of bandwidth selection, see Härdle (1991), Park and Turlach (1992), Sheater and Jones (1991) and Turlach (1993). Practical experiences are available from the simulation studies in Marron (1989), Jones *et al.* (1991), Park and Turlach (1992), and Cao *et al.* (1992). In the next section, we analyze the behaviour of our estimator in finite samples.

4. FINITE SAMPLE ANALYSIS

In this section the sample properties of the estimation method described in section 3 are analyzed. To this effect, Monte Carlo experiments are performed to study the properties of our two-step estimator. The following non-negative distributions are considered: log-normal, Weibull, generalized gamma and

beta, which are some of the most commonly used in literature, for details see Minou and Reddy (2006a).

The first goal is to evaluate whether our density estimator for grouped data fits when compared to the underlying distribution. This is achieved by calculating the mean, median, standard deviation and deciles. Then the simulation experiment is described in detail:

1. A sample of observations (size 4000)is drawn according to the underlying density function (log-normal, Weibull, ...): $x_1, x_2, x_3, \cdots\cdots, x_{4000}$. The information from all 4,000 observations is summarized in a similar way to that of the first rows of Table 1.

2. With the values in the previous table, the estimator of the density function is calculated under section 3. Through the Gaussian kernel, and the approach of Park and Marron (1990)[2] the bandwidth is obtained.

3. Finally, several descriptive statistical measures of the estimated density function are calculated and compared against the actual values.

The results are shown in Table 2. The quantities represent the ratio between the estimated and true values.

Table 2

STATISTICAL SUMMARY. ESTIMATED VALUE / TRUE VALUE

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

The accuracy of our estimation method is quite high except for some values of the Weibull distribution. The goodness of fit of our method is observed when comparing the majority of estimates with respect to the true values, the quantities being very close to 1.

In addition to the comparison of position and dispersion measures, the adjustment of our estimator versus the underlying function is shown in different cases. Figures 3, 4, 5 and 6 show the true and estimated density functions and the bias.

[2] The mean integrated squared error of the density estimator is given by: $\mathrm{MISE} = \int \mathrm{E}\left\{\hat{f}_h(x) - f(x)\right\}^2 dx$. Under the asymptotic conditions $h \to 0$ and $nh \to \infty$, it is possible to approximate $\mathrm{MISE}\left(\hat{f}_h\right) \approx \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\{\mu_2(K)\}^2\|f''\|_2^2$. The terms $\|K\|_2^2$ and $\{\mu_2(K)\}^2$ are constants depending on the kernel function K. Analogously, $\|f''\|_2^2$ denotes a constant depending on the second derivative of the unknown density f. Optimizing the last equation with respect to h yields the following optimal bandwidth: $h_{opt} = \left(\frac{\|K\|_2^2}{\|f''\|_2^2\{\mu_2(K)\}^2 n}\right)^{1/5}$. According to Park and Marron (1990) $\|f''\|_2^2$ is estimated by means of $\|f''\|_2^2 = \frac{1}{ng^3}\sum_{i=1}^{n} K''\left(\frac{x - x_i}{g}\right)$ where "g" is the bandwidth. These authors propose an optimal bandwidth "g" and a bias correction for $\|f''\|_2^2$. For more details see Park and Marron (1990).

Figure 3

TRUE VS ESTIMATED DENSITY (SOLID LINE) AND BIAS (DASHED LINE) OF LOG-NORMAL DIST

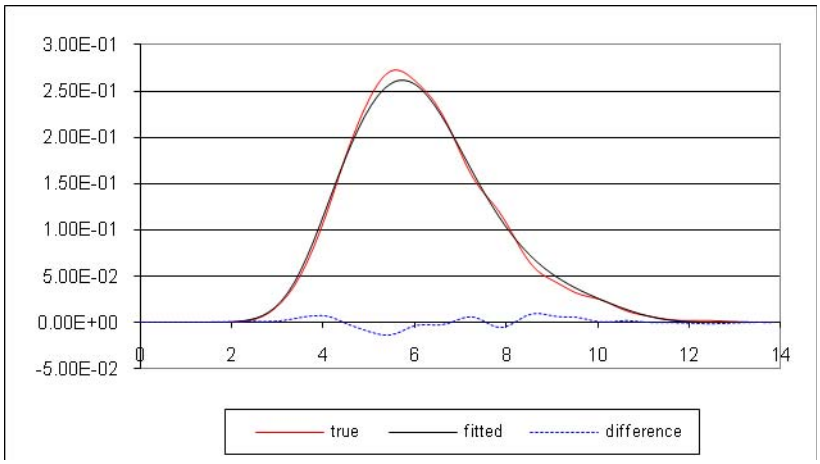


Figure 4

TRUE VS ESTIMATED DENSITY (SOLID LINE) AND BIAS (DASHED LINE) OF WEIBULL DIST

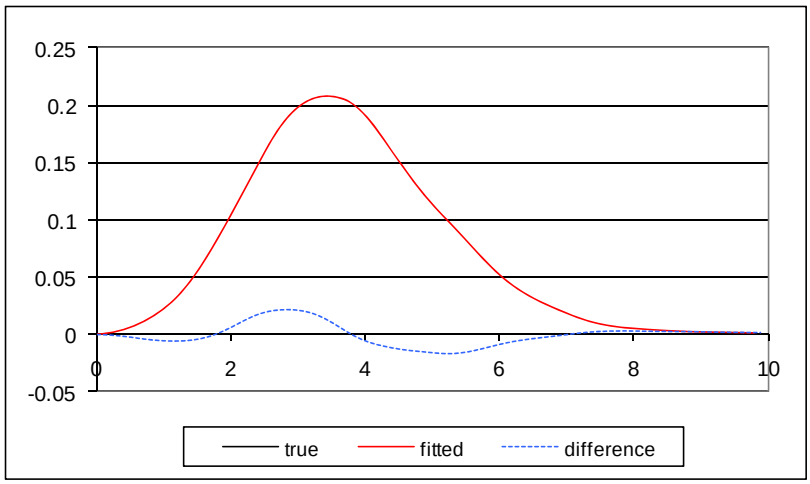


Figure 5

TRUE VS ESTIMATED DENSITY (SOLID LINE) AND BIAS (DASHED LINE) OF GAMMA DIST

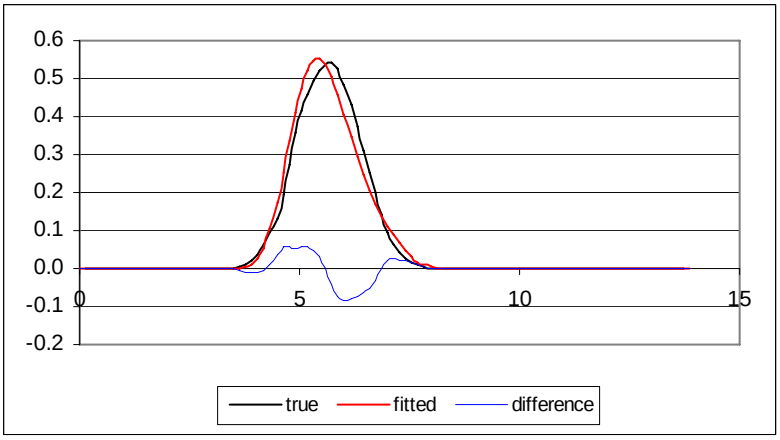
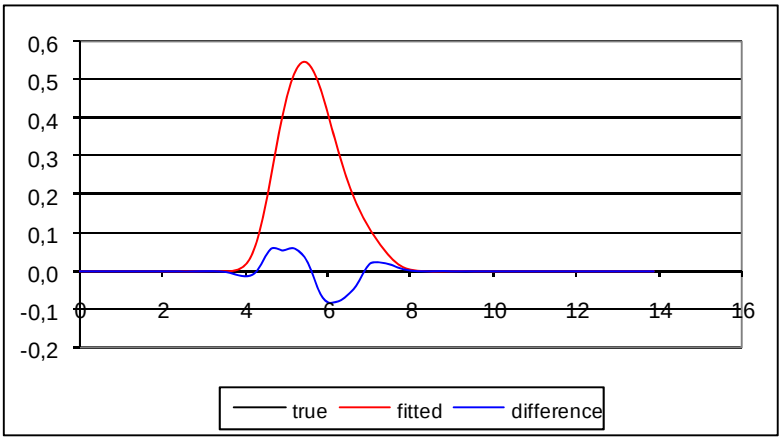


Figure 6

TRUE VS ESTIMATED DENSITY (SOLID LINE) AND BIAS (DASHED LINE) OF BETA DIST



The solid lines represent the true and estimated density functions, the dashed line is the estimation bias (expressed as the difference between the true density function and estimated function). These figures confirm the results in Table 2. The first conclusion from these figures is the good fit of our estimation method. The second is that there are some minor distortions. The adjustment on the Weibull distributions and log-normal is very high, the bias can be considered negligible. The asymmetric Gamma distribution presents a good fit with some bias in the bottom of the distribution. The worst results occur in the Beta distribution, which overestimates the true function at the bottom of the distribution. These results match the information shown in Table 2. It must be recalled that the degree of the bias of the estimated density can affect the calculation of measures of inequality and poverty.

Another relevant feature of our estimator is asymptotic behavior. To this end, 400 random samples of size n = 200, 500, 750, 1000, ... 7000 of a gamma distribution have been drawn. Let $\hat{f}^{(j)}(x)$ be the two-step estimator of the density $f(x)$ developed in section 3 of the j-th sample. The following measures of discrepancy (bias, variance and mean square error) are defined below:

$$\text{Bias}_n\left(\hat{f}(x)\right) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left[\frac{1}{400}\sum_{j=1}^{400}\hat{f}^{(j)}(X_i)\right] - f(X_i)\right\} \tag{15}$$

$$\text{Var}_n\left(\hat{f}(x)\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{400}\sum_{j=1}^{400}\left\{\hat{f}^{(j)}(X_i) - \left[\frac{1}{400}\sum_{j=1}^{400}\hat{f}^{(j)}(X_i)\right]\right\}^2 \tag{16}$$

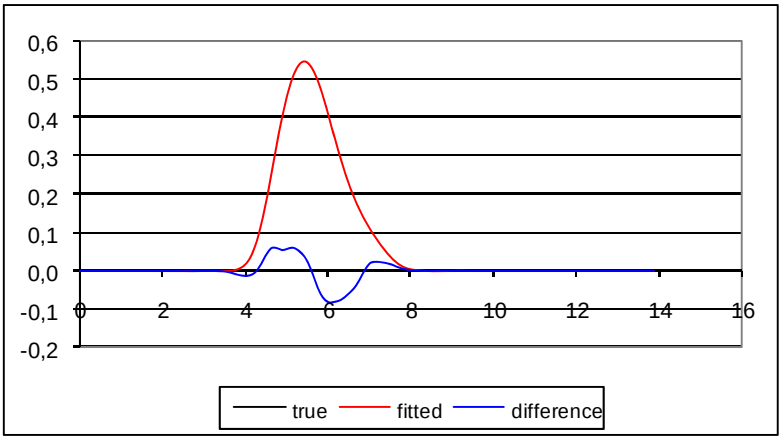
$$\text{MSE}\left(\hat{f}(x)\right) = \text{Sesgo}_n^2\left(\hat{f}(x)\right) + \text{Var}_n\left(\hat{f}(x)\right) \tag{17}$$

Using the Gaussian kernel and the bandwidth of Park and Marron (1990) in the estimations the above values for different sample sizes are calculated. The results are shown in Figure 7.

As expected, the values of these quantities tend to zero as the sample size increases. This is due to the asymptotic behavior of our estimator. However, when the sample is very small the bias may not be negligible. In the next section two applications of our method of estimation are presented.

Figure 7

MEAN SQUARE ERROR, VARIANCE AND BIAS IN THE: Y-AXIS. THE SAMPLE SIZE IS REPRESENTED IN : H-AXIS



5. EMPIRICAL APPLICATIONS

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. First, data grouped into deciles, illustrated with a European Union case are used. Then, asymmetric grouped data from tax records of the Spanish Tax Agency (www.aeat.es) have been applied. Adding to this, our method allow us to generate national (or global) density function aggregating regional (or national) density functions. This feature is shown in the second example.

5.1. Symmetric grouped data: The income distribution in the European Union

In this subsection we analyze the income distribution and poverty and inequality measures of EU member states in 2001. Our estimation process requires information on population (individuals), average income and quantiles of the income distribution. Information from various sources has been necessary: the income series (deciles) for 2001 was obtained using the database of global income inequality (The world income inequality data base, http://www.wider.unu.edu/research/database), that provides information on inequalities in income / expenditure for a panel of developed or developing countries for the period 1950-2001. The population and GDP per capita information were obtained from the Penn Word Tables 3.1 (Heston, Summers and Aten, 2002) (http://pwt.econ.upenn.edu). Table 3 shows the data used in this application.

Table 3

UE DATA (YEAR 2001)

	Populati.	GDP per capita	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Austria	8096.25	26999.77	4.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00	14.00	19.00
Belgium	10303.88	24661.91	4.00	5.00	6.00	7.00	8.00	9.00	10.00	12.00	14.50	24.50
Finland	5176.53	22740.69	4.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00	13.50	19.50
France	59278.01	25044.54	4.00	5.00	7.00	7.00	8.00	10.00	11.00	12.00	14.50	21.50
Germany	82344.43	25061.34	4.00	6.00	7.00	8.00	9.00	9.00	10.00	12.00	14.00	21.00
Greece	10975.02	13982.39	3.00	4.00	6.00	7.00	8.00	9.00	11.00	13.00	15.50	23.50
Ireland	3801.38	24947.55	3.00	5.00	6.00	7.00	9.00	10.00	11.00	12.00	15.00	22.00
Italy	57714.84	22487.21	3.00	5.00	6.00	7.00	9.00	10.00	11.00	13.00	14.50	21.50
Luxembourg	435.23	48217.27	4.00	6.00	7.00	7.00	8.00	9.00	11.00	12.00	14.50	21.50
Netherlands	15897.51	26293.09	4.00	6.00	7.00	8.00	8.00	9.00	11.00	12.00	14.00	21.00
Portugal	10225.09	17323.14	3.00	4.00	5.00	6.00	7.00	8.00	10.00	12.50	15.50	29.00
Spain	40717.22	19536.38	3.33	4.90	5.96	6.94	7.90	8.95	10.22	11.95	14.58	25.27
U.K.	58669.74	24666.41	3.00	5.00	6.00	7.00	8.00	9.00	11.00	12.00	15.00	24.00
Denmark	8900.87	25860.69	1.70	3.70	4.60	5.80	7.10	8.70	10.90	13.60	16.60	27.30
Sweden	5359.98	28551.14	4.10	5.90	6.70	7.50	8.50	9.30	10.20	11.50	13.50	22.80

The density of each country is obtained by using our general method of estimation in two steps: the first step gives the Lorenz curve, evaluated in 100 points. The second step is the non-parametric estimation of the income density function in each of 100 fictitious values for each country. All calculations of the second stage are performed with the Gaussian kernel and the optimal bandwidth of Park and Marron. Clearly, each country has a different bandwidth.

In the first step, equation (4) is estimated by using OLS. We assume a fourth grade polynomial series in this equation. The estimation and adjust R^2 are shown in Table 4. When paying attention to R^2, the good performance of different regressions, with values higher than 0.97 in all of the cases, must be outlined.

Table 4

OLS ESTIMATIONS OF EU COUNTRIES IN 2001, EQUATION (4)

	Populati.	GDP per capita	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Austria	8096.25	26999.77	4.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00	14.00	19.00
Belgium	10303.88	24661.91	4.00	5.00	6.00	7.00	8.00	9.00	10.00	12.00	14.50	24.50
Finland	5176.53	22740.69	4.00	6.00	7.00	8.00	9.00	10.00	11.00	12.00	13.50	19.50
France	59278.01	25044.54	4.00	5.00	7.00	7.00	8.00	10.00	11.00	12.00	14.50	21.50
Germany	82344.43	25061.34	4.00	6.00	7.00	8.00	9.00	9.00	10.00	12.00	14.00	21.00
Greece	10975.02	13982.39	3.00	4.00	6.00	7.00	8.00	9.00	11.00	13.00	15.50	23.50
Ireland	3801.38	24947.55	3.00	5.00	6.00	7.00	9.00	10.00	11.00	12.00	15.00	22.00
Italy	57714.84	22487.21	3.00	5.00	6.00	7.00	9.00	10.00	11.00	13.00	14.50	21.50
Luxembourg	435.23	48217.27	4.00	6.00	7.00	7.00	8.00	9.00	11.00	12.00	14.50	21.50
Netherlands	15897.51	26293.09	4.00	6.00	7.00	8.00	8.00	9.00	11.00	12.00	14.00	21.00
Portugal	10225.09	17323.14	3.00	4.00	5.00	6.00	7.00	8.00	10.00	12.50	15.50	29.00
Spain	40717.22	19536.38	3.33	4.90	5.96	6.94	7.90	8.95	10.22	11.95	14.58	25.27
U.K.	58669.74	24666.41	3.00	5.00	6.00	7.00	8.00	9.00	11.00	12.00	15.00	24.00
Denmark	8900.87	25860.69	1.70	3.70	4.60	5.80	7.10	8.70	10.90	13.60	16.60	27.30
Sweden	5359.98	28551.14	4.10	5.90	6.70	7.50	8.50	9.30	10.20	11.50	13.50	22.80

Figures 8a & 8b show the results for the 15 EU members in 2001[3]. Note the different ways of showing the income functions. It is noted that there are functions highly concentrated around their modal value (or median) such as Greece or distributions much more scattered as Luxembourg. These two countries are those that reflect the most opposite values. The minimum modal value of the distributions is the Greek one with a value around 10,500 Euros, while the maximum mode belongs to a central European country with a value of 42,000 Euros. Another aspect which seems worth noting is the clear difference between the countries of the Mediterranean area and the rest. Greece, Portugal, Spain and Italy have two characteristics in their income distributions. The distributions in these countries are the most asymmetrical, with a significant tail on the right side. In addition, the smallest modal values reflect that these countries have the lowest income level (on average) and the most concentrated distribution (the greatest inequality). On the other end, Sweden, Denmark, Austria and Germany present values referred to the most equitable income (i.e. the most symmetrical distributions) and the highest mean and median values.

In addition to the density functions, different measures of poverty and inequality are calculated, which are shown in Table 5. In our estimations of poverty rates the threshold chosen is that most frequently used by Eurostat: 60 for 100 of the median of the function of households' disposable income. In the estimation of Atkinson's index we assume a value of 0.5 for the aversion parameter.

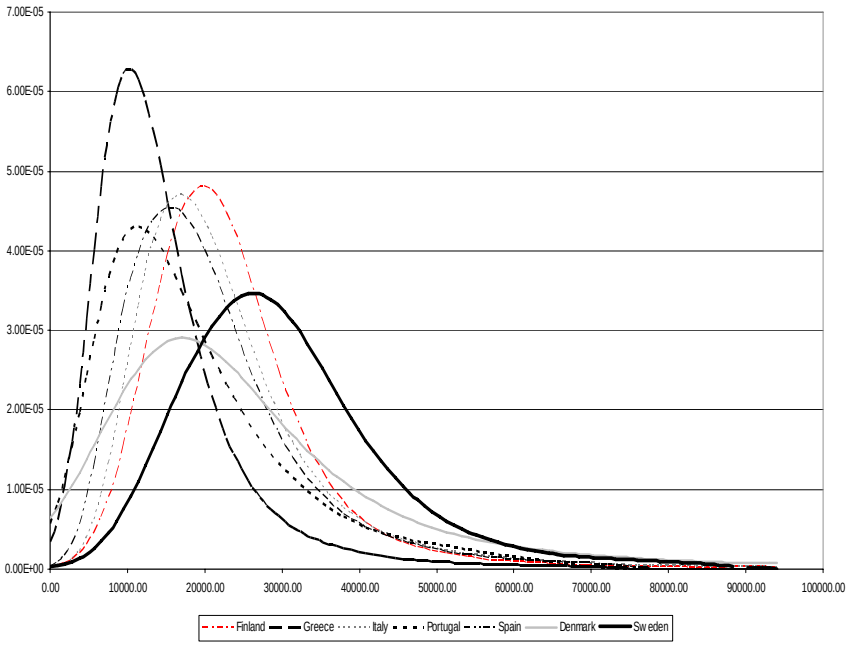
The Gini index values are quite similar to those presented by the European Commission, for 2001, in its publication (Eurostat, 2005). For the other indices there is a high consistency with those values published by previous reference. Differences in rates are due to the different income concept used by Eurostat (disposable family income), equivalent in terms of national accounts to the income account of institutional household, while the concept used here is an income equal to GDP, which is always higher than the previously mentioned[4].

[3] The density function of Sweden is included in both figure 8a and 8b for the sake of comparability.

[4] For more details of a methodological discussion see Milanovic (2006).

Figure 8a

DENSITY FUNCTION ESTIMATIONS OF EU COUNTRIES IN 2001



It is worthwhile to make some observations on some of the estimated indices. Inequality values in terms of Gini indices support the above comments on the shape of the density functions. The smallest values of inequality are referred to Nordic and Central European countries– Austria (0.232),Germany (0.249),the Netherlands (0.252),Denmark (0,228),Finland (0.233) and Sweden (0.264). On the contrary, in the countries of the Mediterranean area such as Spain (0.316), Portugal (0.377), Greece (0.322) and Italy (0.289). The Gini index values are substantially higher.

Figure 8b

DENSITY FUNCTION ESTIMATIONS OF EU COUNTRIES IN 2001

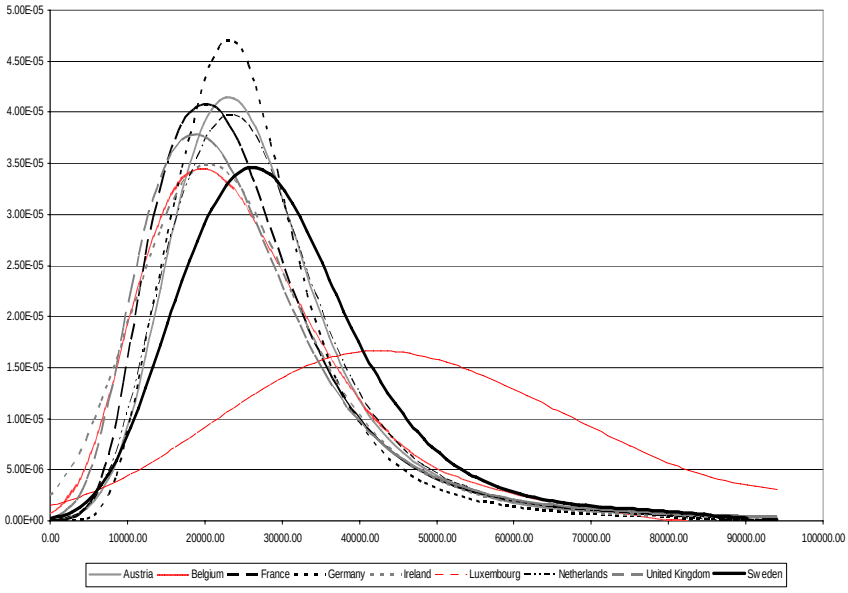


Table 5

MEASURES OF POVERTY IN 15 UE COUNTRIES (2001)

	Populati.	Mean	Median	% < 60%median	poverty gap	I. Atkinson	I. Gini
Austria	7764	27591.2	25125.3	6.144	0.088	0.037	0.232
Belgium	9555	26060.7	23311.2	16.285	0.183	0.054	0.300
Finland	4963	23242.3	21018.9	6.226	0.094	0.037	0.233
France	55868	26041.7	23306.0	13.050	0.126	0.049	0.271
Germany	76272	26515.3	24479.8	8.518	0.135	0.036	0.249
Greece	10337	14548.5	12276.5	14.434	0.162	0.071	0.322
Ireland	3622	25656.9	22052.5	11.320	0.126	0.059	0.293
Italy	54672	23263.7	19905.3	11.103	0.105	0.061	0.289
Luxembourg	455	45175.9	41949.0	10.549	0.132	0.037	0.263
Netherlands	14910	27472.9	25179.6	10.718	0.134	0.039	0.252
Portugal	9330	18605.2	15402.5	21.683	0.242	0.088	0.377
Spain	37315	20891.2	18456.2	16.530	0.185	0.060	0.316
U.K.	54503	26020.8	22647.2	14.355	0.183	0.063	0.311
Denmark	8295	27191.9	20368.2	12.168	0.171	0.031	0.228
Sweden	4975	30139.8	28334.4	10.794	0.162	0.038	0.264

The same can be said in the case of inequality values measured in terms of Atkinson indexes. On the one hand, relative poverty, when using the European threshold, also presents the lowest values in Central European countries , Austria(6.149),Germany (8.,518) y the Netherlands (10.178) –and in the following Nordic countries– Finland (6.26)Sweden(10.794) and Denmark (12.168); which are the countries with the least inequality. On the other, the highest values of relative poverty are referred not only to Mediterranean countries such as Portugal (21.683), Spain (16.530), and Greece (14.434)– all these countries having high levels of inequality but also to Central countries such as Belgium (16,530) and United Kingdom (14.355), which have higher levels of average and median income and nevertheless present high values of inequality.

5.2. Asymmetric grouped data: Income distribution in Spain

In this example we focus on Spain's 2003 tax information on common fiscal territory, which comes from the Spanish Tax Agency. The key feature is that this information is available in asymmetrical income intervals (the number of individuals in each interval is different), so it is not feasible to directly estimate the density function based on quantiles. The information is offered in two separate tables. The first table provides information on the average income in each interval. The second table shows the number of individuals in each he intervals[5]. Such information is provided by Tables 6a and 6b.

There are two goals in this case: Firstly, the distribution functions of taxable income tax are calculated using asymmetrically grouped data. Secondly, and most important, the distribution of national taxable income is obtained by adding the values of estimated income derived from regional information (Autonomous Communities).

[5] In this subsection we are making 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 equivalent to the number of "individuals" in each interval. The latter assumption is obviously inaccurate, since the income tax return can be personal or not and, therefore, the AEAT does not provide us the actual number of "individuals" in each section. However, the number of tax returns can be considered, for the purposes of our estimate, as the number of individuals in each section.

The "taxable income" is not the equivalent in monetary terms to the "gross income" available to households, although this fact is not very relevant for the goal pursued by this stydy. For further discussion see Ayala and Onrubia (2001).

Table 6a

AVERAGE INCOME OF TAXPAYERS/TAXPAYERS' AVERAGE INCOME

Region (CCAA)	negative - 1,5	1,5 - 6	6-12	12-21	21 - 30	30 - 60	60 - 150	More than 150	Total
España	1332.61	6907.31	11589.84	18241.83	27300.49	42136.51	85283.20	239913.35	18796.17
Andalucia	1534.63	6858.35	11543.37	18101.09	27233.48	41689.28	83187.98	294167.39	16632.85
Aragón	1699.79	6906.54	11597.54	18264.97	27232.36	41896.00	84277.42	260223.34	18297.83
Asturias	1539.28	6718.00	11665.89	18446.51	27348.15	41627.14	84080.90	277534.64	18399.99
I.Baleares	854.62	7038.31	11545.98	18074.95	27198.34	42142.33	85649.47	272781.25	18749.95
Canarias	1407.38	7054.44	11434.33	18098.86	27436.55	42019.88	84097.30	281576.42	18165.79
Cantabria	1361.19	6748.81	11659.29	18258.34	27314.65	41999.87	84447.36	298713.09	18609.64
C. La Mancha	1439.26	6865.71	11485.83	18023.90	27168.59	41806.80	82735.87	271998.68	15705.91
C y Leon	96.62	6758.35	11491.46	18167.79	27309.40	41488.19	83378.51	265791.50	16937.91
Cataluña	1392.01	6942.03	11680.08	18308.86	27282.80	42213.67	85670.33	289087.05	21151.76
C. Valenciana	1404.09	7049.12	11648.79	18156.90	27203.47	42065.93	84874.38	309880.93	17473.67
Extremadura	2048.79	6908.01	11332.75	17939.78	27132.69	41714.15	81225.85	247755.62	14653.38
Galicia	914.47	6764.93	11464.84	18159.17	27249.37	41962.53	84390.87	284532.54	16382.15
La Rioja	1650.71	6886.62	11703.73	18096.44	27263.18	41458.51	85040.37	306004.87	18010.03
Madrid	1327.75	6975.34	11706.52	18525.62	27451.71	42751.32	86816.25	313821.84	24221.32
Murcia	1880.96	6980.29	11627.07	18123.64	27279.13	41833.63	84154.57	285683.11	16868.83

Table 6b

NUMBER OF TAXPAYERS (INDIVIDUALS)

Region (CCAA)	negative - 1,5	1,5 - 6	6-12	12-21	21 - 30	30 - 60	60 - 150	More than150	Total
España	919000	2737612	4174720	4223910	2085731	1444135	307429	41325	15933862
Andalucia	194168	513390	720178	642465	302188	181310	30256	3285	2587240
Aragón	34586	108684	145811	170768	80111	51807	9170	1039	601976
Asturias	33105	76828	105730	131949	72023	39760	5924	789	466108
I.Baleares	16194	63958	116142	98130	43418	33456	7720	1086	380104
Canarias	33153	109819	178811	150878	79798	52630	9804	1321	616214
Cantabria	14660	36523	60421	65824	31641	20294	3872	443	233678
C. La Mancha	45973	148669	208889	170952	70195	41557	6205	586	693026
C y Leon	69260	206408	279787	280044	133233	82225	11841	1044	1063842
Cataluña	121560	417572	694504	877293	443030	318902	80146	10465	2963472
C. Valenciana	109833	353501	539153	466755	207752	138361	27430	3416	1846201
Extremadura	32380	95002	117184	82061	35751	20368	3216	223	386185
Galicia	74966	219536	286590	237712	112262	70355	12057	1559	1015037
La Rioja	7983	23433	37360	38937	16561	11232	2111	227	137844
Madrid	102772	279369	548049	696341	407189	350849	92271	15205	2492045
Murcia	28407	84920	136111	113801	50579	31029	5406	637	450890

The distribution of taxable income in each region is estimated through our method[6]. The results are shown in Figures 9a and 9b[7]. The Autonomous Communities of Madrid, Catalonia and the Balearic Islands have the highest levels of income, but the distributions are very asymmetric. In contrast, the regions of Andalusia, Extremadura and Castile La Mancha have the lowest level of taxable income.

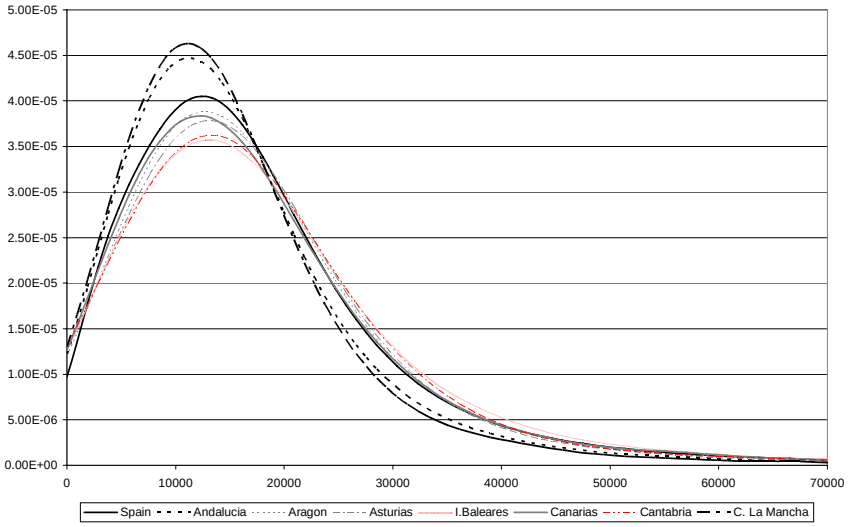
One common feature of all the functions in Figure 9 is the fact that densities cut the y-axis above 0. The explanation for this result comes from the tax information in Tables 6a and 6b, in which there is a significant number of individuals having a taxable income (average) with negative values[8].

[6] In this case, we use the approach of Villasenor and Arnold (1989) for estimating the Lorenz curve, instead of the initially proposed of Ryu and Slottje. According to this variation, we want to show the flexibility of our estimation method, because we can use any existing method in the first step.

[7] The Spanish density function is included in 9a and 9b for the sake of comparability.

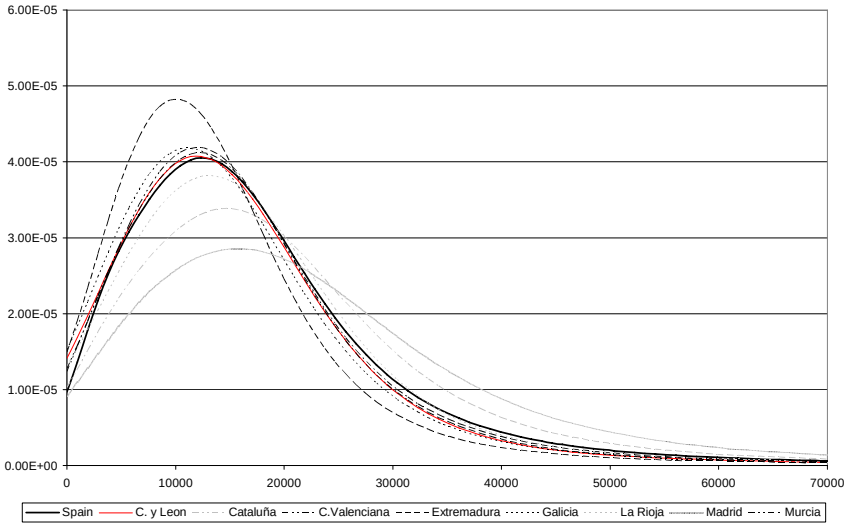
[8] Originally, the AEAT offers a range of negative values of income and another between 0-1.5. In our analysis we have added these two intervals to ensure the theoretical restrictions of the Lorenz Curve.

Figure 9a
INCOME DISTRIBUTION OF SPANISH REGIONS (CCAA)



In addition to the density functions, several measures of relative poverty and inequality are calculated. These results are shown in Table 7. The first column shows the median values of the density function of taxable income in each region. The poverty threshold is defined as the 60 for 100 of the median value of each region. The highest values of relative poverty are found in Madrid, Catalonia, the Balearic Islands, Galicia and Valencia. In the first three cases, the values for the median are significantly higher than the national one. If the Gini index is applied, it is indeed the same: the largest values correspond to Madrid, Catalonia and Galicia. These results seem to confirm the relationship between inequality and relative poverty (European Commission, 2005). Clearly, these results are related to the underlying density function in each case. In most cases, the regions with the highest values of average income are also the regions with the most skewed taxable income distribution[9].

Figure 9b
INCOME DISTRIBUTION OF SPANISH REGIONS (CCAA)



[9] Relative poverty is calculated according to the poverty line in each Region (Autonomous Community), not with respect to the national poverty line. Therefore, it is possible to call it "intra-regional poverty level." Similarly, the relative poverty with respect to a common single threshold (national) can be estimated: this is the "interregional relative poverty."

The second goal of this application focuses on the problem of the generation of income aggregate functions from subgroups, i.e. the estimation of the Spanish national income distribution by integrating the income distributions of regions. Our general method allows the aggregation of regional functions through the aggregation of "micro" income values produced by each function of income in the regions. This is important because our technique does not add regional functions of income, but regional microdata derived from each function (all the different individuals that have been simulated by the density function).

Table 7

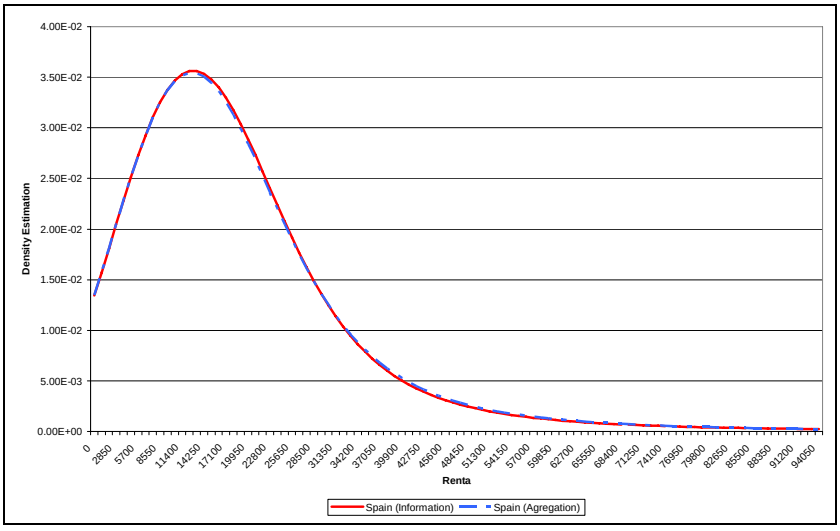
MEASURES OF INCOME AND INEQUALITY IN SPANISH REGIONS (CCAA). COMMON FISCAL TERRITORY

Region (CCAA)	median	% pob<0.6*median	povertygap	Atkinson	Gini
España	15164.372	24.557	0.411	0.135	0.404
Andalucia	13447.659	23.382	0.384	0.130	0.396
Aragón	14687.058	22.892	0.371	0.122	0.388
Asturias	14786.613	22.015	0.348	0.120	0.388
I.Baleares	15493.165	24.810	0.433	0.127	0.403
Canarias	14537.350	24.268	0.404	0.127	0.398
Cantabria	14996.803	23.116	0.387	0.124	0.393
C. La Mancha	12957.376	22.553	0.377	0.118	0.380
C y Leon	13446.678	22.368	0.346	0.126	0.393
Cataluña	17460.513	25.004	0.433	0.133	0.403
C. Valenciana	14274.287	24.581	0.421	0.131	0.397
Extremadura	11789.865	22.654	0.368	0.117	0.385
Galicia	12990.533	24.368	0.399	0.141	0.410
La Rioja	14598.718	22.867	0.393	0.114	0.385
Madrid	19790.010	25.315	0.421	0.138	0.428
Murcia	13829.374	23.063	0.389	0.118	0.382

Figure 10 shows the function of the 2003 taxable income derived from two different approaches. On one approach, the density is calculated using national information provided in the first row of the tables 6a and 6b. On the second approach, the densities of income in each region are computed and the (hypothetical) micro values of income derived from regional densities are obtained. The aggregation of the micro values of all regions gives us the equivalent population (counterfactual) of our country. In our opinion the two settings are almost identical, so our method of estimation and the subsequent aggregation of the regional functions work properly.

Figure 10

COMPARISON- ESTIMATION OF NATIONAL INCOME DISTRIBUTION (N.I.D) VS N.I.D. THROUGH AGGREGATION OF REGIONAL INCOME DISTRIBUTIONS



6. CONCLUSIONS

This paper presents a general method for estimating density functions for any type of distribution and clustering of data. These intervals can represent both quantiles (i.e. The number of individuals within each interval is the same in all intervals) or not. This technique enables us to build a worldwide (or national) density function on the basis of national (regional) data which is much more accurate than standard approaches in this type of literature. We show the good properties of the estimation process in finite samples by means of Monte Carlo experiments. Finally, we present two empirical applications. Firstly, we calculate the income density in the countries of the European Union using data grouped in quantiles for year 2001. Secondly, we calculate the density of Spanish regions based on data asymmetrically grouped. Furthermore, we calculate the density of the income tax of Spain as the aggregation of micro functions that generate revenue for regional densities.

REFERENCES

ACKLAND, R.; DOWRICK, S. and FREYENS, B. (2006): "Measuring global poverty: why PPP methods matter" paper presented at the conference of the international association for Research in Income and Wealth, Cork, Ireland.

AYALA, L. and ONRUBIA, J. (2001): "La distribución de la renta en España según datos fiscales". Monografico sobre la distribución de la renta en España. *Papeles de Economía*, n.º 88, pp. 89-112.

CAO, R.; CUEVAS, A. and GONZALEZ-MANTEIGA, W. (1992): "A comparative study of several smoothing density estimation", manuscript, Universidad de Vigo.

CHEN, S. and RAVILLION, M. (2002): "How did the world's poorest fare in the 19902?" *Review of income and wealth*, vol. 47(3), 283-300.

CHEONG, K.S. (2002): "A comparison of alternative functional forms for parametric estimation of the Lorenz Curve", *Applied economics Letters*, vol 9, 171-176.

DATT, G. (1998): "Computational tools for poverty measurement and analysis". International food policy research institute. Food consumption and nutrition division discussion paper n.º 50.

EUROPEAN COMMISION (2005): "First European quality of life servey: income inequatilities and deprivation".

EUROSTAT (2005): "Regional indicators to reflect social exclusion and poverty".

FUENTES, R. (2005): "poverty, pro-poor growth and simulated inequality reduction". Human development report office occasional paper n.º 11.

GRIFFITHS, W.E.; CHOTIKAPANICH, D. y PRAKASA RAO, D.S. (2005): "Averaging income distributions", *Bulletin of economic research*, 57, 347-367.

HÄRDLE, W. (1991): *Smoothing techniques, with implementations in S*. Springer, New York.

HÄRDLE, W.; MÜLLER, M.; SPERLICH, S. and WERWATZ, A. (2004): *Nonparametric and Semiparametric Models*. Springer Verlag Berlin Heidelberg.

HASEGAWA, H. and KOZUMI, H. (2003): "Estimation of lorenz curves: a bayesian nonparametric approach". *Journal of econometrics*, 115, 277-291.

HESTON A.; SUMMERS, R. and ATEN, B. (2002): Penn World Tables. Unviversity of Pennsylvania.

JONES, M.C.; MARRON, J.S. and PARK, B.U. (1991): "A simple root-n bandwidth selector", *The annals of statistics*, 19,1919-1932.

JONES, M.C.; MARRON, J.S. and SHEATHER, J.S. (1996): "A brief survey of bandwidth selection for density estimation", *Journal of the American statistical association*, vol. 91, 401-407.

KAKWANI, N.C. (1980): "On a class of poverty measures". *Econometrica*, vol 48(2), 437-446.

KAKWANI, N.C. and PODDER, N. (1976): "efficient estimation of the Lorenz curve and associated inequality measures from grouped observations". *Econometrica*, vol 44(1), 137-149.

MARRON, J. (1989): "Comments on a data based bandwidth selector" *Computational statistics & data analysis*, 8, 155-170.

MILANOVIC, B. (2002): "True world income distribution, 1988 and 1993: First calculation based on household surveys alone" *Economic Journal*, 112, 51-92.

— (2006): "global income inequality: What it is and why it matters?". *World Bank policy research working paper*. 3865. World Bank.

MINOIU, C. and REDDY, S. (2006a): "The assessment of poverty and inequality through parametric estimation of lorenz curves: an evaluation". Mimeo, Department of Economics. Columbia University.

— (2006b): "Kernel Density estimation in poverty and inequality analysis: validity and robustness", mimeo, Department of Economics, Columbia University.

ORTEGA, P.; MARTIN, G.; FERNANDEZ, A.; LADOUX, M. and GARCIA, A. (1991): "A new functional form for estimating lorenz curves" *Review of income and wealth*, vol 37. 447-452.

PARK, U and MARRON, J. S. (1990): Comparison of data-driven bandwidth selectors. *Journal of the american statistical association*. Vol 85. 66-72

PARK, B. and TURLACH, B.A. (1992): "Practical performance of several data driven bandwidth selectors", Computational Statistics, 7, 251-270.

RASCHE, R.H.; GAFFNEY, J.; KOO, A.Y.C. and OBST, N. (1980): "Functional forms for estimating the Lorenz curve". *Econometrica*, vol. 48, 1061-1062.

RAVILLION, M. and HUPPI, M. (1989): "Poverty and under-nutrition in Indonesia during the 1980s", Policy, planning and research working n.º 286, Agriculture and rural development deparment. Wolrd Bank.

RYU, H.K. and SLOTTJE, D.J. (1996): "Two flexible functional form approaches for approximating the lorenz curve", *Journal of econometrics*, 72, 251-274.

SALA-I-MARTIN, X. (2002): "the world distribution of income (estimated from individual country distribution)". National bureau of economic research working paper n.º 8933.

— (2006): "the world distribution of income: falling poverty and convergence, period". *Quaterly journal of economics*. Vol. 121. N.º 2. 351-397.

SHEATHER, S.J. and JONES, M.C. (1991): "A reliable data-based bandwidth selection method for kernel density estimation". *Journal of the royal statistical society*. Series B. vol 53(3). 683-690.

SILVERMAN, B.W. (1986): *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability 26. chapman & Hall/CRC.

TURLACH, B.A. (1993): "Bandwidth selection in kernel density estimation: a review" discussion paper 9307, Institut fur statistik und okonometrie, Humboldt-Universitat zu Berlin.

VILLASENOR, J.A. and ARNOLD, B.C. (1989): "Elliptical lorenz curves". *Journal of econometrics*. Vol 40. 327-338.

WAND, M.P. and JONES, M.C. (1995): *Kernel Smoothing*. Chapman and Hall. London.

WU, X. and PERLOFF, J. (2003): "Calculation of Maximum entropy densities with application to income distribution", *Journal of Econometrics*, vol. 115, 347-354.

SÍNTESIS

PRINCIPALES IMPLICACIONES DE POLÍTICA ECONÓMICA

El estudio de la distribución de la renta es un elemento crucial en el análisis de la desigualdad y pobreza, que desde un punto de vista social, económico y político son de gran relevancia y que preocupan a economistas, gobiernos e individuos, teniendo mucha importancia los factores que las determinan así como las acciones necesarias para su reducción. Además de estos aspectos de gran relevancia, el calculo de estas funciones de renta, así como su evolución a lo largo del tiempo, son útiles para estudios sobre movilidad social, impactos de políticas redistributivas, etc; a diferentes niveles: personales, regionales, nacionales, mundiales, etc.

La estimación de las funciones de distribución de renta depende crucialmente de los datos disponibles para los investigadores. La forma en que se recaban estos es múltiple: registros administrativos, censos, muestras, encuestas, paneles, contabilidades entre otros. Sin embargo, en muchos casos, la información disponible para los investigadores consiste únicamente en datos agrupados en porcentajes (cuantiles) de renta procedentes de encuestas a hogares o de registros administrativos. Es más, los datos agrupados son la única fuente de información de las distribuciones de renta en la mayor parte de los países (o regiones) que tienen un importante papel en la determinación de la pobreza y desigualdad a nivel mundial (o nacional), o en el caso de países desarrollados cuando se desea realizar estudios de renta a niveles locales o regionales, en los que la disponibilidad de información relevante vía encuestas, paneles, etc. no suele existir.

La técnica desarrollada en este trabajo consiste en un método bietápico de estimación de la función de distribución de la renta a partir de datos agrupados. En la primera etapa se estima la curva de Lorenz aplicando cualquiera de los métodos conocidos en la literatura. En la segunda etapa, utilizando las estimaciones obtenidas previamente, estimamos de forma no paramétrica la función de distribución de la renta. Además, presentamos dos aplicaciones para distintos tipos de agrupaciones de datos, tanto a nivel nacional como internacional.