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

Published by the Spanish National Statistical Institute

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, $k = 1, \dots, L$).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 $\|f''\|_2^2$.

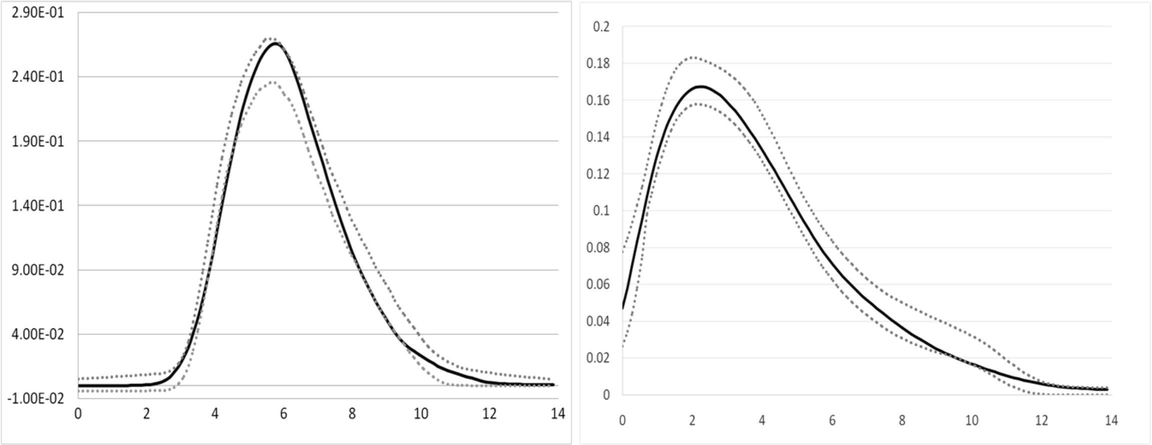

Figure 1: True (solid) and 95% SCI of density estimates (dashed) for the Log-Normal (left) and the Weibull (right) distribution.

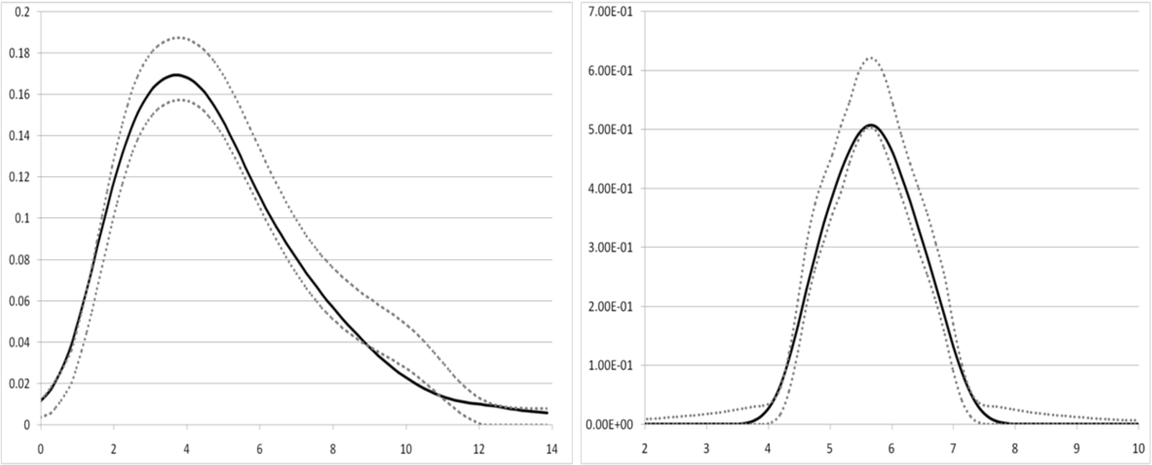

Figure 2: True (solid) and 95% SCI of density estimates (dashed) for the generalized Gamma (left) and the Beta (right) distribution.

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.

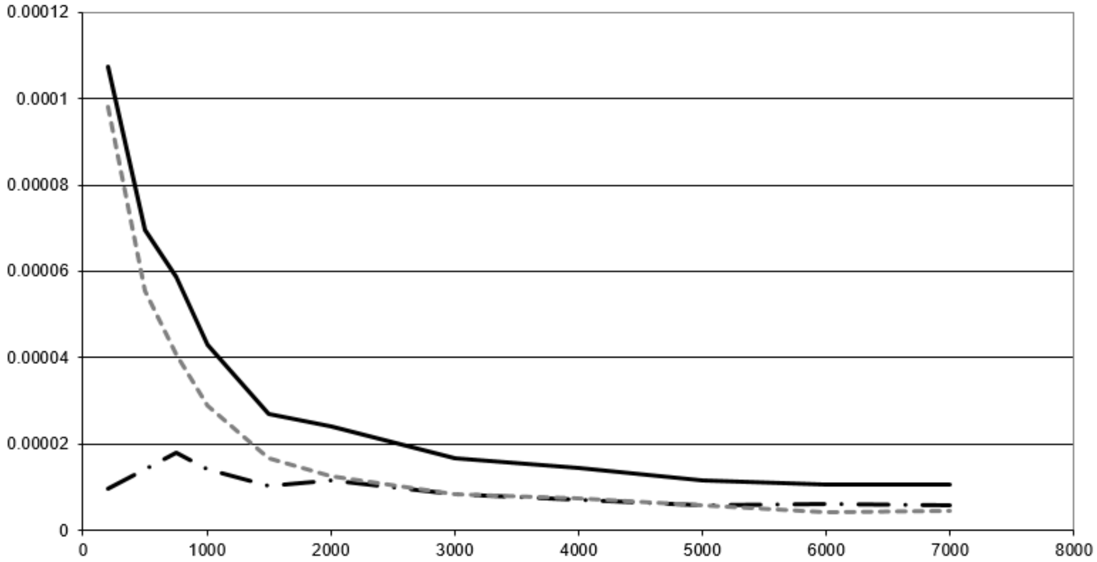

Figure 3: SsD (solid line), Average Variance (dashed grey line) and squared average deviance for increasing sample size when the true density is a Gamma.

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.

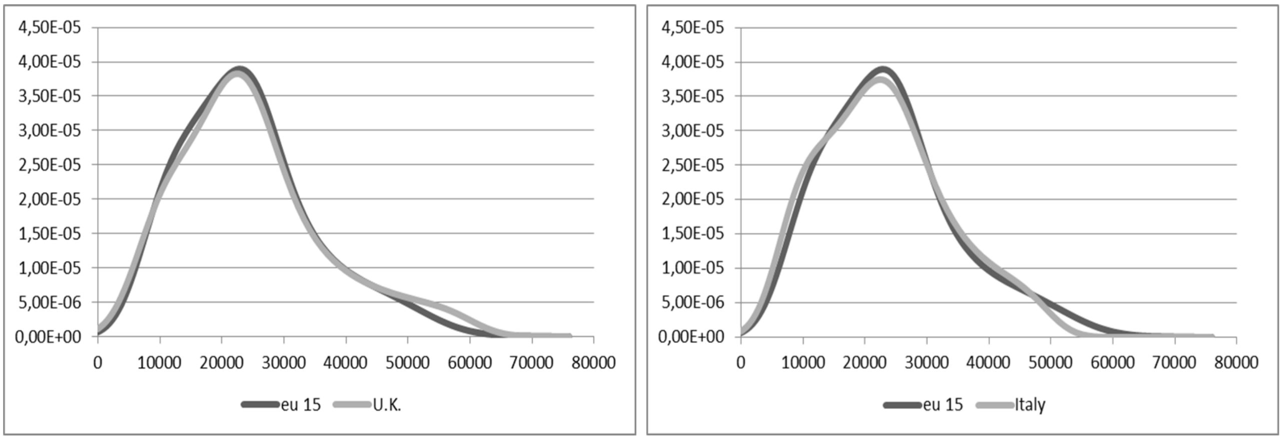

Figure 4: Density Function Estimations of EU Countries and U.K. (left) and Italy (right) in 2001.

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

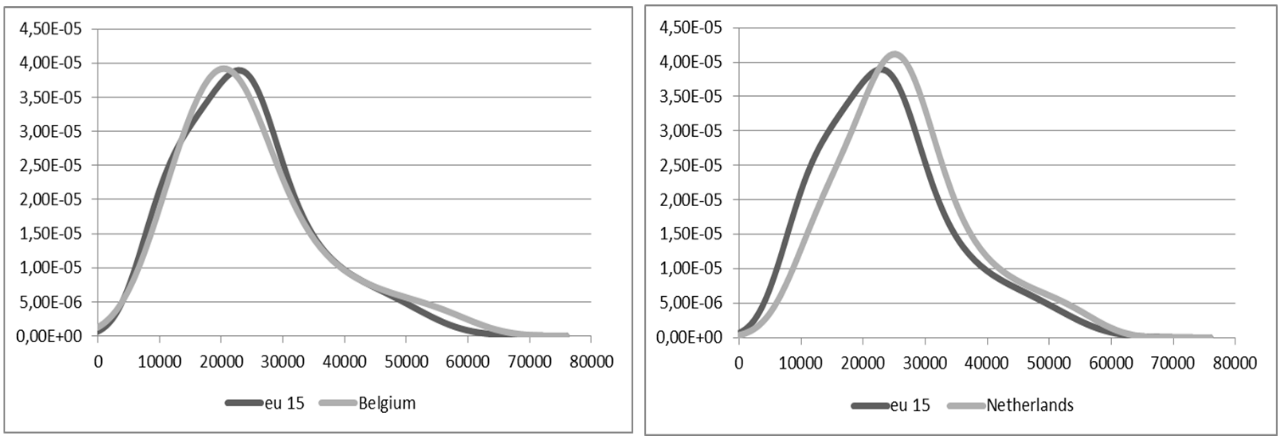

Figure 5: Density Function Estimations of EU Countries and Belgium (left) and Netherlands (right) in 2001.

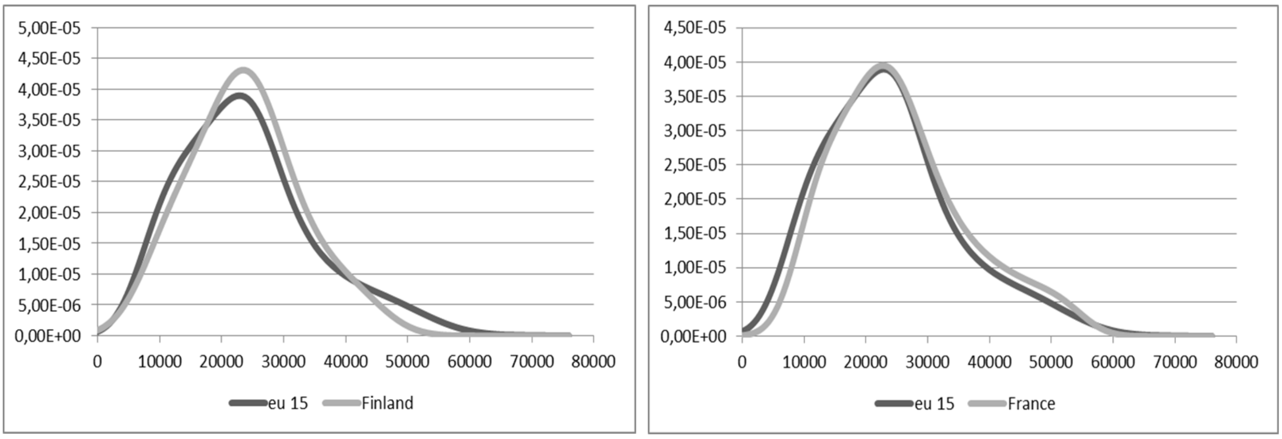

Figure 6: Density Function Estimations of EU Countries and Finland (left) and France (right) in 2001.

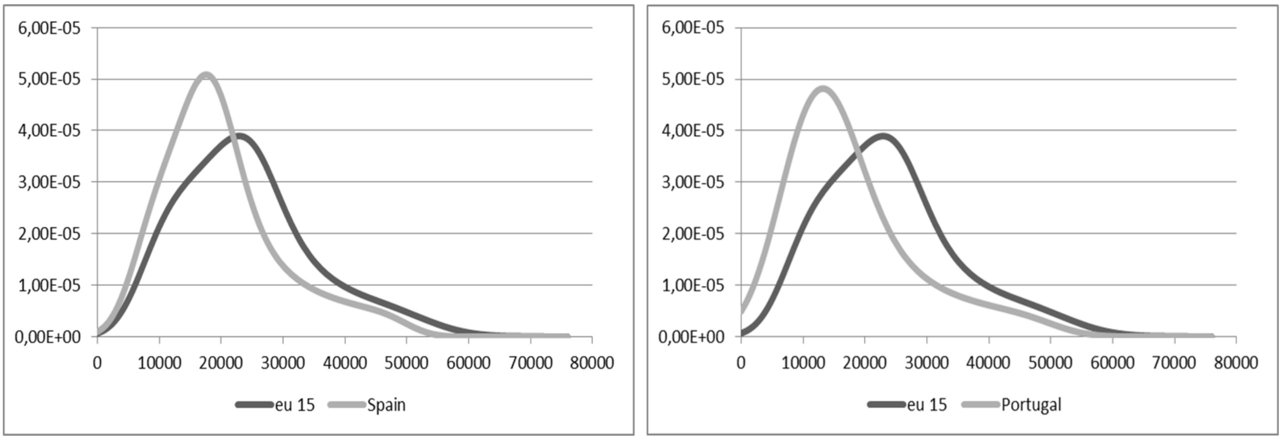

Figure 7: Density Function Estimations of EU Countries and Spain (left) and Portugal (right) in 2001.

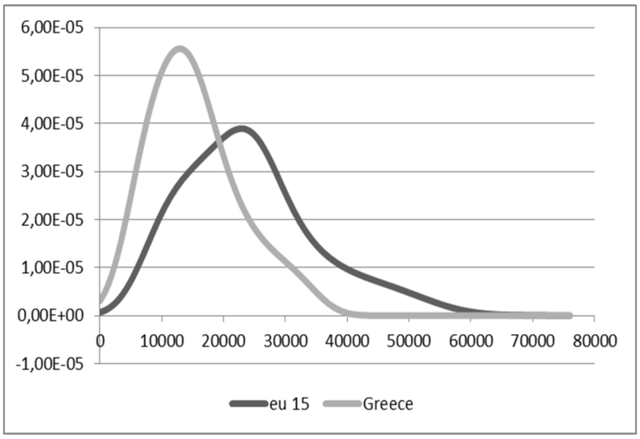

Figure 8: Density Function Estimations of EU Countries and Greece in 2001.

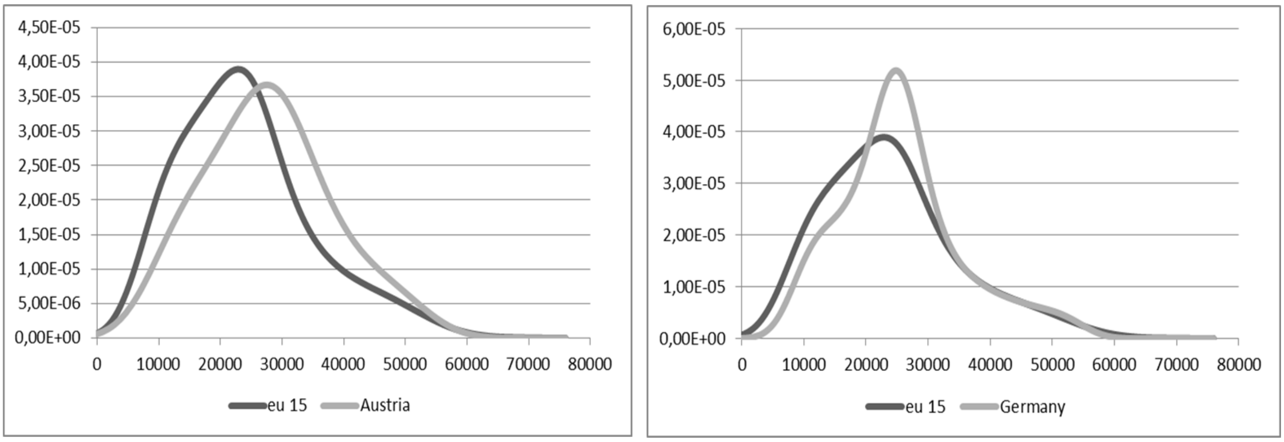

Figure 9: Density Function Estimations of EU Countries and Austria (left) and Germany (right) in 2001.

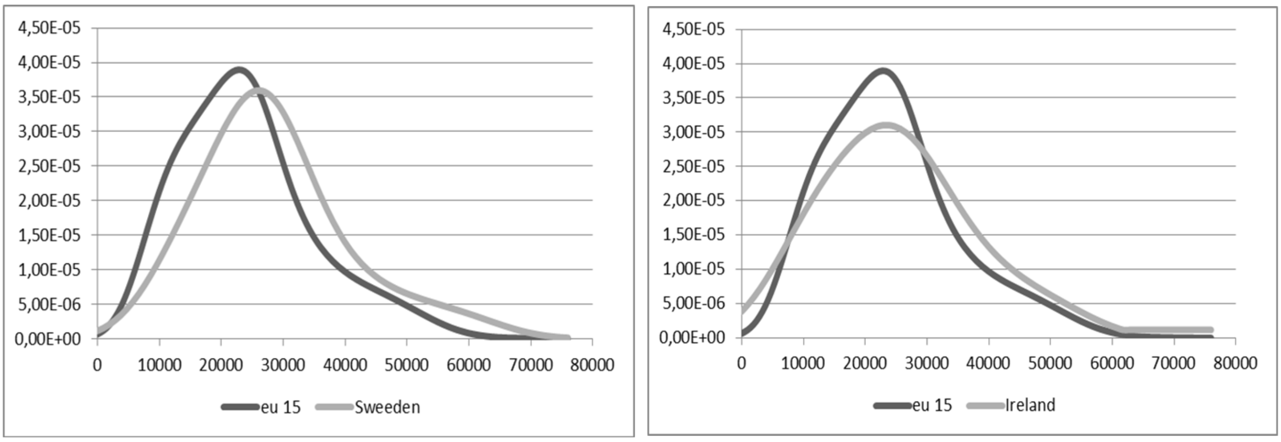

Figure 10: Density Function Estimations of EU Countries and Sweden (left) and Ireland (right) in 2001.

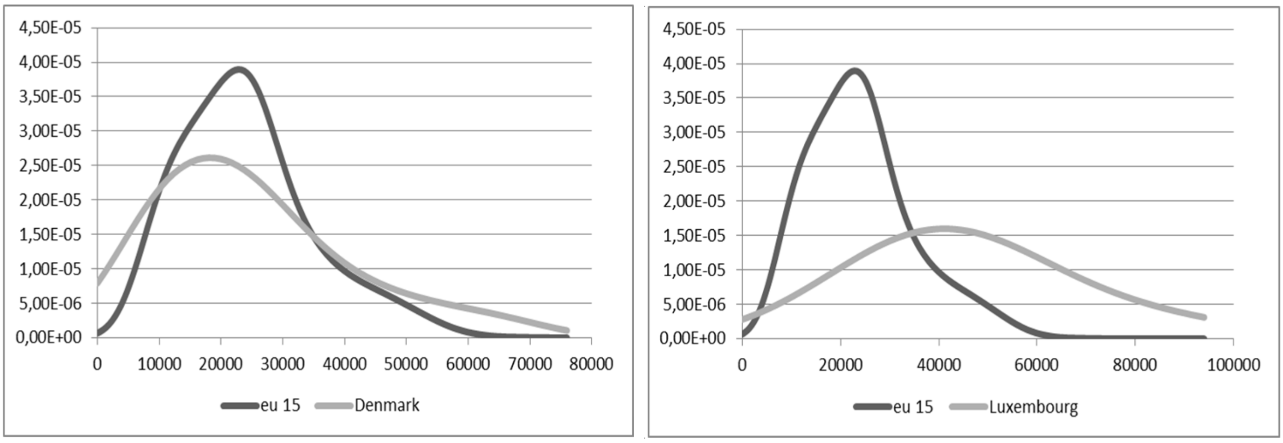

Figure 11: Density Function Estimations of EU Countries and Denmark (left) and Luxembourg (right) in 2001.

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.

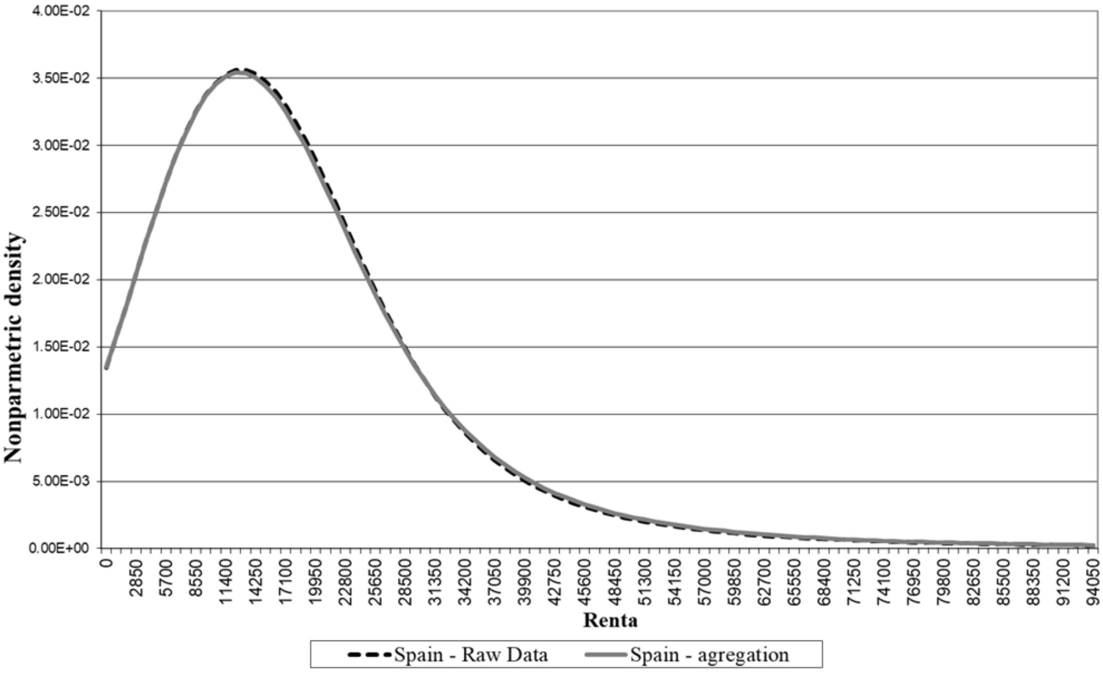

Figure 12: Comparison- Direct estimation of National Income Distribution âĂŞ Raw data - vs estimation through Aggregation of Regional Income Distributions.

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)