The country factor on regional income distributions in Europe: A functional ANOVA approach

Pedro Delicado[1] and Magda Mercader[2]

(First version: July 2005. This version: June 2006)

Abstract

The distribution of regional Gini indices in Europe using the income distribution before taxes and transfers is not explained by the country to which the region belongs, i.e. the dispersion of the Ginis is not significantly reduced when we control for the country variable. On the contrary, there is a clear dependency between the regional Ginis and the country when the distribution of income before taxes and transfers is considered. This evidence is based on EUROMOD a multicountry tax-benefit model of the EU-15 (See Mercader and Levy 2004). We study to what extent this conclusion holds when we consider the complete income distributions instead of a summary inequality measure such as the Gini index. We use functional ANOVA (following Cuevas et al. 2004) in order to study the country explicative power on the dispersion of regional income density functions (estimated non-parametrically) before and after taxes and transfers. Our statistical evidence suggests that regional income distributions in different countries are different, both before and after redistribution takes place. However, the null assumption of equality of mean regional distributions among countries (a factor country equal to zero) is rejected more strongly in the after distribution case.

Acknowledgements

In this paper we use data from the EUROMOD project. We are indebted to all past and current members of the EUROMOD consortium. Our acknowledgment is also for Horacio Levy who helped us with the data. The views expressed in this paper, as well as any errors, are the responsibility of the authors.[3]

[1] Universitat Politècnica de Catalunya. Pedro.Delicado@upc.edu

[2] Universitat Autònoma de Barcelona. Spain. Magda.Mercader@uab.es

[3] In particular, this applies to the interpretation of EUROMOD results and any errors in its use. EUROMOD is continually being improved and updated and the results presented here represent work in progress. The EUROMOD version used in this work is 25a. EUROMOD relies on micro-data from 15 different sources for fifteen countries. These are: the European Community Household Panel (ECHP) User Data Base made available by Eurostat, the Austrian version of the ECHP facilitated by the Interdisciplinary Centre for Comparative Research in the Social Sciences, the Panel Survey on Belgian Households (PSBH) contributed by the University of Liège and the University of Antwerp, the Income Distribution Survey given by Statistics Finland, the Enquête sur les Budgets Familiaux (EBF) made available by INSEE, the public use version of the German Socio Economic Panel Study (GSOEP) supplied by the German Institute for Economic Research (DIW), Berlin, the Living in Ireland Survey contributed by the Economic and Social Research Institute, the Survey of Household Income and Wealth (SHIW95) given by the Bank of Italy, the Socio-Economic Panel for Luxembourg (PSELL-2) facilitated by CEPS/INSTEAD, the Socio-Economic Panel Survey (SEP) made available by Statistics Netherlands through the mediation of the Netherlands Organisation for Scientific Research - Scientific Statistical Agency, the Income Distribution Survey supplied by Statistics Sweden and the Family Expenditure Survey (FES), offered by the UK Office for National Statistics (ONS) through the Data Archive. Material from the FES is Crown Copyright and is used by permission. Neither the ONS nor the Data Archive bear any responsibility for the analysis or interpretation of the data reported here. An equivalent disclaimer applies for all other data sources and their respective providers cited in this acknowledgement.

1. **Introduction**

Countries, through different mechanisms, including their history and political, social and economic institutions, are important factors in distinguishing final personal inequality levels.

One of the key elements in understanding of the shape of a country's final income distribution is its tax-benefit system (See Atkinson, 2000). Cross-country empirical evidence shows how the impact of taxes and transfers critically depends on the country (See Atkinson et al (1995), Wagstaff et al (1999) among others).

This paper aims at providing further evidence on the significance of national tax-benefit systems in explaining the final shape of income distributions. Relying on microdata from EUROMOD, representative of the European income distribution *before* and *after* the operation of tax-benefit systems[4], an experimental statistical approach is adopted. We analyse the role of the country factor in explaining the variability in regional income distributions *shapes* before and after national tax-benefit systems operate. The main questions we aim at answering are: Which is the contribution of the country factor on the observed variability of regional income distributions before taxes and transfers? By how much this contribution increases when the role of taxes and transfers systems has been accounted for?

In a recent paper Mercader and Levy (2004) study similar questions. By means of a simple ANOVA model they show that while the distribution of regional Gini indices in Europe using the income distribution before taxes and transfers is not significantly explained by the country to which the region belongs, i.e. the dispersion of the Ginis is not significantly reduced when we control for the country variable, there is a clear dependency between the regional Ginis and the country when the distribution after taxes and transfers is considered. In this paper we take further their analysis. We study to what extent this conclusion holds when we consider the complete income distributions instead of a summary inequality measure such as the Gini index. In order to take account of the full income distribution, we use Functional Data Analysis techniques.[5] Functional ANOVA (following Cuevas et al. 2004) is used in order to study the country explicative power on the dispersion of regional income density functions (estimated non-parametrically) before and after taxes and transfers.

If in micro-simulation it is common the evaluation of the effect of a *marginal* policy change on a given population, this work uses a complementary approach. We aim at exploring what can be learned from applying the same system on *marginally* different populations. Regions in a given country offer a particularly attractive natural framework for experimental analysis. Firstly, because regions can be seen as being *distorted pictures* of a given country. Regions belonging to a country often share a number of characteristics such as a common legal reference system, similar labour market and demographic structures and traditions. Secondly, regions are real sub-populations of a given country. Knowing the effect of a country tax-benefit system in one of its regions is therefore of clear relevance, also from a normative perspective. One could, of course, argue that in a European context with 25 countries, having each its own tax-benefit

[4] Only countries from the European Union before May 2004 are considered.

[5] Jenkins and Van Kerm (2004) using similar techniques show the relevance of looking at the whole distribution in accounting for income distribution trends.

system, the ideal experimental structure for our analysis would require a matrix in which we evaluate the 25 tax-benefit-systems on the 25 countries. However, we do not have this information available. Microsimulation research has demonstrate that applying to country *k* the tax-benefit system of country *j* poses a number of difficult issues that cannot be always easily solved (Atkinson et al, 1988). Our experiments, by relying on real populations, avoid adopting arbitrary assumptions that such an ideal experimental structure would require.

The organisation of the paper is as follows. After this introduction, Section 2 includes a description of EUROMOD, the tax-benefit model used, presents the regional data as well as other assumptions underlying the construction of the income distributions studied in this work. Section 3 summarises previous findings based on the univariate case for regional Ginis before and after the role of national tax-benefit systems. Description of both the Functional Data Analysis techniques for dealing with income density functions and the functional ANOVA model for estimating the country effect on regional income distributions is offered in Section 4. This section also explains the methods proposed for testing the null hypothesis of equality of average Ginis among countries. The procedure used for the estimation of regional density functions is offered in Section 5. Results of the functional ANOVA on regional EU density functions are presented in Section 6. We end with a concluding section with some policy implications from our analysis.

2. A description of the data

This analysis relies on EUROMOD. EUROMOD is an integrated tax-benefit micro simulation model for all countries of the old European Union before May 2004.[6] EUROMOD is a source of harmonized micro-data on the different income components. It allows to construct in a comparable manner the income distribution before and after the main taxes and cash benefit programs in place in the different countries. The original databases on which EUROMOD relies are summarised in Table 1.

<Insert Table 1>

Our analysis focuses on the comparison of two income distributions. The distribution of income before public cash transfers are added and taxes are deducted. More precisely, the before taxes and transfers income distribution includes all components of market income: wages and salaries and self-employment income (net of employer insurance contributions and other benefits, but gross of employee contributions to such schemes), property income (interest, rents, dividends) as well as occupational pensions from employers, regular inter-household cash transfers and other sources of income which are not government transfers. The after taxes and transfers income distribution is the before income distribution plus all social transfers in cash minus employee Social Insurance Contributions, personal income taxes and other taxes. Indirect taxes are not considered.

All income distribution estimates presented are based on the relative equivalent income (net or after taxes and transfers) per person in each region. Incomes are *equivalent incomes* in the sense that the household incomes are divided by the equivalent number of adults living in there (using the modified OECD scale) and income *per person*

[6] For details about the EUROMOD's team and project, see its website: http://www.econ.cam.ac.uk/dae/mu/emod. For a detailed description of the model see Sutherland (2001).

indicates that each household income is weighted by their number of members. Incomes are *relative* because in each region the observed equivalent incomes are divided by the median of regional equivalent incomes. The analysis does not look at the variability of income distributions due to differences in income levels but it focuses on the differences in shapes. Finally, the whole analysis refers to 1998 annual incomes.[7]

Regional information in EUROMOD

Table 2 shows the regional information used in our analysis. For most countries, the official classification of regions established by Eurostat (NUTS system) is available. Our analysis uses NUTS1 information for most countries but we keep NUTS2 for countries in which this information is available.[8] Notice that average population size per region significantly diverges among countries, from less than 1 million people in Finland to 5.5 million people in Spain. Also, the number of regions varies greatly. Out of the total number of regions considered, France concentrates 22 of the regions analysed, while Austria and Belgium only 3 regions each and Greece 4. For the other countries, the number of regions is in the range 6 to 12.

<Insert Table 2>

Table 3 shows the sample size available for the different regions. Sample sizes vary also greatly, from 3,404 in Östra Mellansverige in Sweden to Corse in France for which only 38 observations are available.

<Insert Table 3>

3. Previous findings

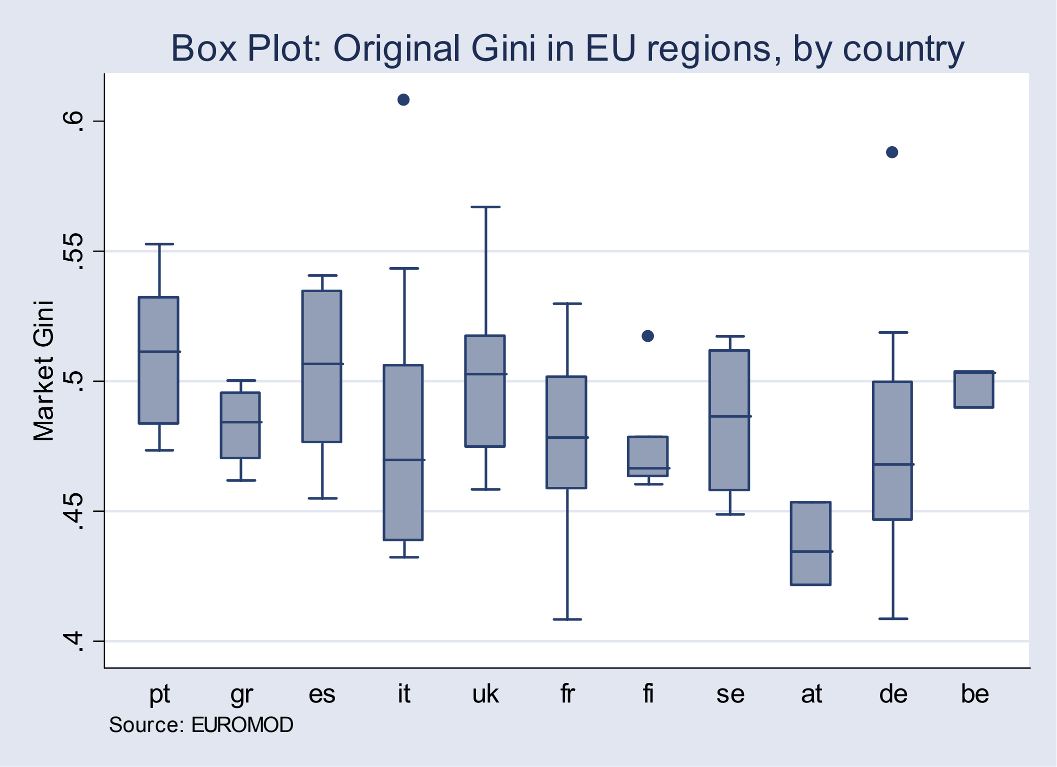
Mercader and Levy (2004) study the role of the country in explaining the variability of regional Ginis in Europe before and after the application of national taxes and transfers systems. They use the same data we use here. One of their main finding is that distribution of regional Ginis indices in Europe using the before taxes and transfers income distribution is not explained by the country to which the region belongs, i.e. the dispersion of the Ginis is not significantly reduced when we control for the country variable. On the contrary, there is a clear dependency between the regional Ginis distribution and the country when the income distribution after taxes and transfers is considered. This is illustrated in Figures 1 and 2 showing a box-plot of regional Gini indices by country before and after redistribution takes place.[9]

[7] The Danish, Swedish and UK currencies were converted into Euro using the exchange rate of December 31st, 1998.

[8] NUTS2 is a more disaggregated classification than NUTS1.

[9] The internal box line represents the median regional Gini and while the box upper and lower bands indicate the regional gini Gini at the 25th and 75th percentiles. The lines extending above and below the box report the minimum and the maximum ginis Ginis in each country and the symbol ° the outliers.

Figure 1.

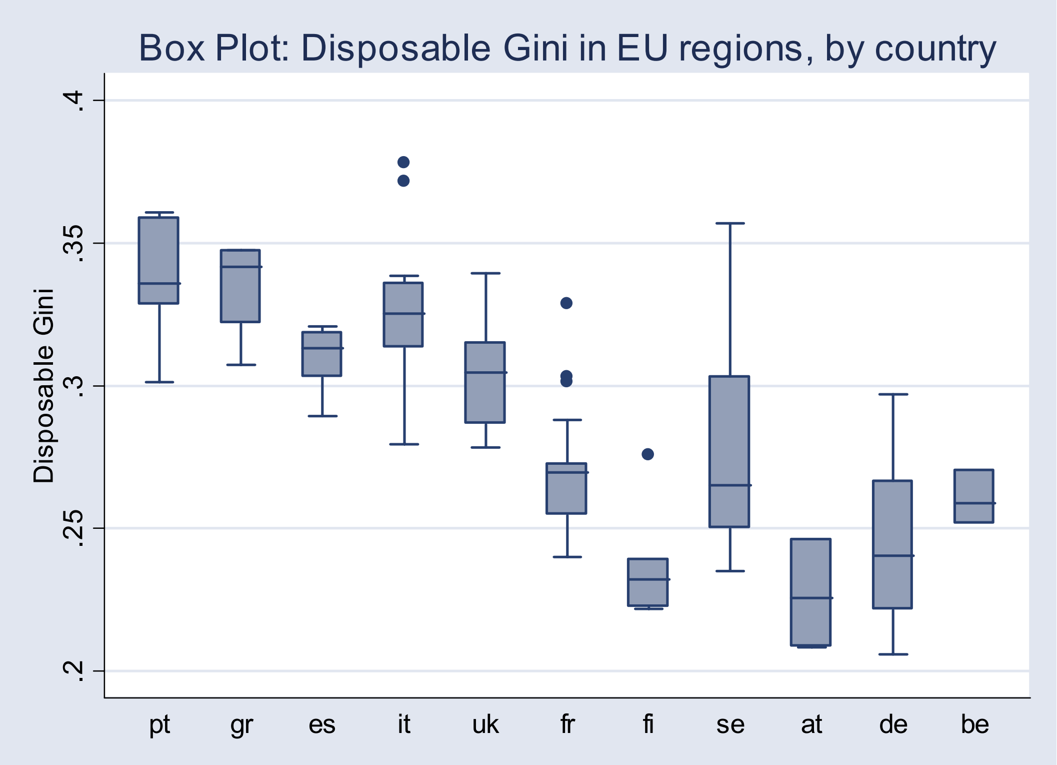


Source: EUROMOD

Note: Market Gini is calculated using the household income before taxes and transfers. The distribution is equivalised using the modified OECD scale. Each household is weighted according to the number of members. Taken from Mercader and Levy (2004).

. T

Figure 2.



Note: Disposable Gini is calculated using the household income after taxes and transfers. The distribution is equivalised using the modified OECD scale. Each household is weighted according to the number of members. Taken from Mercader and Levy (2004).

It is interesting to notice that while regional gross income inequality levels appear to be rather independent from the country in which regions belong, the level of disposable income inequality of a given region is to be observed as much more dependent on the country in which this region belongs. Most regions in Southern Europe (Portugal, Italy, Greece and Spain) show a Gini coefficient above .3. The UK regions follow with Gini coefficients around 0.3. Most French and Belgian regions are in the interval .25-.3. The lowest Gini are found in the Finnish regions and regions from Austria (values around .2-.25). There are significant differences in disposable inequality levels in regions in Sweden (between .25 and .35) and Germany (from around .2 to .3).

The results of fitting a simple ANOVA model to the regional Ginis before and after the action of the tax-benefit allow to confirm these findings. The null hypothesis that average regional Ginis among countries are equal cannot be rejected on the Ginis before redistribution. In fact the country factor explains only 16% of the variance observed. On the other hand, the null hypothesis that the average Ginis among countries are the same is rejected on the after taxes and transfers case. In this second case, the country factor explains 69% of the variance of the Ginis. Thus, the country is, through the national tax-benefit system, a decisive factor on regions' final income inequality levels while this is not so in the case of market income inequality (See Mercader and Levy, 2004).

4. Funtional Data Analysis and functional ANOVA

The first steps of Statistics were characterized by the analysis of random univariate data. Even in regression models, the response variable and the regressor were originally one dimensional objects. Ramsay and Silverman (1997) express it saying that at the beginning the statistical atom was a number (an element of R). The collection of statistical techniques that form introductory courses in Statistics are designed to analyzed this kind of data. In the first half of the twenty century Statistics reached maturity and started to deal with multivariate data: a collection of p random numbers observed simultaneously (in the same subject) was treated as a statistical atom (an element of R^p). Principal component analysis, discriminant analysis, and cluster analysis, among others, are part of what we know today as Multivariate Data Analysis.

At the end of the last century a new statistical object has started to be considered. In the last years, the joint development of real-time measurement instruments and data storage computer resources has made possible to observe and save complete functions as results of random experiments. For instance, continuous-time monitoring clinical diagnostics or stock market information are common nowadays. From this perspective, random functions are now the statistical atoms. Functional Data Analysis (FDA) deals with the statistical description and modelization of samples of random functions. It's well worthwhile noting that random functions can also be obtained from standard random samples, by the application of nonparametric curve estimation methods. For instance, Kneip and Utikal (2001) study the temporal evolution of density function of incomes in United Kingdom from 1968 to 1988. They work with yearly cross-sectional sample of households, and use nonparametric density estimation methods (kernel methods, to be specific) to obtain 21 income density functions over time, one corresponding to each year. Other kind of data can be transformed into functions using nonparametric regression (also known as smoothing methods).

A broad outlook to Functional Data Analysis is given in the books by Ramsay and Silverman (1997, 2002). They show interesting applications and case studies and include many of the available techniques. Most of them are versions of standard statistical methods adapted to functional data (for instance functional descriptive statistics, functional linear models or functional principal component analysis; see also Cuevas et al, 2004, for more references on these topics) but others are specific for this kind of data because they exploit the nature of functions: for example, principal differential analysis is a kind of principal component analysis made on the derivatives of the observed functions, and regularization is a pre-process step where a change of variable is done in each observed function in order to made them as similar as possible.

One of the standard methods that have been generalized to be used in Functional Data Analysis is the one-way analysis of variance (*functional ANOVA*). In Ramsay and Silverman (1997) it is included in what they call functional linear models (a generalization of linear regression models). In addition to this approach, there are works devoted specifically to this topic, among the ones we highlight Cuevas et al. (2004) because it is the base of our work.

In functional ANOVA it is assumed that the n observed functions $f_{ij}(x)$ follow the model

$$f_{ij}(x) = m_i(x) + e_{ij}(x),\ \ j = 1,\dots,n_i, i = 1,\dots,k, x \in [a,b],\ \sum_i n_i = n,$$

where $m_i(x)$ are k unknown mean functions and $e_{ij}(x)$ are independent trajectories drawn from a L^2-process with zero mean and covariance function $Cov(e_{ij}(x), e_{ij}(y)) = K(x,y)$. So the observations $f_{ij}(x)$ constitute k independent samples of random functions, each with a specific function mean and a common covariance structure. This is the homocedastic case. The heteroscedastic version allows for a different covariance function in each sample: $Cov(e_{ij}(x), e_{ij}(y)) = K_i(x,y)$. The hypothesis to be tested in both cases is

$$H_0 : m_1(x) = \cdots = m_k(x).$$

As stated, in the context of the present paper, $f_{ij}(x)$ is the relative equivalent income (before or after taxes and transfers) estimated density function of region j in country i, one of the $k = 15$ countries forming the European Union before May 2004. So the data x corresponding to a person represents that its household has a relative income equal to x times the median regional equivalent income. The interval $[a,b]$, where values x belongs to, is $[-1,7]$ when incomes before taxes are transfers are used, and it is $[-1,5]$ for data after taxes and transfers. Section 5 deals with the estimation of density functions. The null hypothesis establishes that regional income densities have the same mean value in each country. Another way of wording it is to say that under the null hypothesis there is no country effect in the observed variability of regional income densities.

The classical F statistic for the univariate one-way ANOVA computes the ratio of variabilities *between* samples and *intra* sample. The functional version, as stated in Cuevas et al. (2004), would be

$$F_n = \frac{\sum_{i=1}^{k} n_i \left\| \bar{f}_{i\bullet} - \bar{f}_{\bullet\bullet} \right\|^2 /(k-1)}{\sum_{i,j} \left\| f_{ij} - \bar{f}_{i\bullet} \right\|^2 /(n-k)},$$

where $\bar{f}_{\bullet\bullet} = \bar{f}_{\bullet\bullet}(t)$ is the global mean function, $\bar{f}_{i\bullet} = \bar{f}_{i\bullet}(t)$ is the mean function in the i-th sample, and $\|f\| = \left(\int_a^b f^2(x)dx \right)^{1/2}$ is the usual L^2 norm. The null hypothesis H_0 should be rejected if the numerator of F_n (a measure of the differences between groups) is to big, compared with the denominator of F_n (a measure of the variability of the noise process generating $e_{ij}(x)$).

Cuevas et al. (2004) argue that it is enough to consider the numerator of F_n when you are comparing values of the statistic coming from functional ANOVA models with noise processes having the same variability (all the denominators are estimating the same quantity). This is the case when an observed F_n value is compared with Monte Carlo simulated values and the simulation is done to produce data according to the null hypothesis and having the same noise variability as the observed data. Technical reasons lead Cuevas et al. (2004) to measure differences between groups using the statistic

$$V_n = \sum_{i<j} n_i \left\| \bar{f}_{i\bullet} - \bar{f}_{j\bullet} \right\|^2,$$

that is equivalent to use the numerator of F_n. Their Theorem 1 establishes that the asymptotic distribution of V_n under H_0 coincides with that of the statistic

$$V = \sum_{i<j} \left\| Z_i - C_{ij} Z_j \right\|^2,$$

where $C_{ij} = (p_i / p_j)^{1/2}$, $(n_i / n) \to p_i$ as $n \to \infty$, and $Z_i = Z_i(x), i = 1, \ldots, k$, are independent Gaussian processes with 0 mean and covariance function $K_i(x, y)$, that can be consistently estimated by

$$\hat{K}_i(x,y) = \sum_{j=1}^{n_i} \frac{1}{n_i - 1} \left(X_{ij}(x) - \bar{X}_{i\bullet}(x) \right) \left(X_{ij}(y) - \bar{X}_{i\bullet}(y) \right).$$

In the homoscedastic case the natural estimator of the common covariance function is

$$\hat{K}(x,y) = \frac{1}{n-1} \sum_{j=1}^{n_i} (n_i - 1) \hat{K}_i(x,y).$$

This theoretical result offers an asymptotic Monte Carlo procedure to tabulate the null distribution of V_n: a large number N of values of statistic V are simulated (say $V_l^*, l = 1, \ldots, N$) and the p-value corresponding to the observed V_n is computed as the proportion of simulated values V_l^* greater than V_n.

Assuming homoscedasticity, an alternative way to obtain valid approximate the null distribution of V_n is based on permutations. The procedure to obtain pseudo functional data sets according to the null hypothesis consists in randomly permute the group label of the observed functions. Under the null hypothesis the original sample and the permutated sample are interchangeable, and so they are the observed V_n and the value

V^p computed in the permutated sample. N values of statistic V^p are obtained by permutation and the *p*-value corresponding to the observed V_n is computed as the proportion of times the permuted statistics V^p is greater than V_n. Also in the context of functional ANOVA, Muñoz-Maldonado et al. (2002) use a permutation test based on a different statistic.

The permutation test presents a drawback under the alternative hypothesis. The between group variability is translated by the permutation procedure to noise variability in the permuted samples. Therefore the artificial samples verify the null hypothesis of equal mean in different groups, but the noise variability is greater than the corresponding to the original data. The main consequence of the increment in noise variability is the reduction of the test power: small deviation from the null hypothesis would not be detected because of the precision loss.

The preceding procedure can be modified to obtain a more powerful permutation test. Instead of permuting the observed function, we propose to permute the estimated residuals and define the artificial functions as the sum of the global mean plus a permutated estimated residual,

$$f_{ij}^{p}(x) = \bar{f}_{\bullet\bullet}(x) + \hat{e}_{ij}^{p}(x),$$

and $\hat{e}_{ij}^{p}(x)$ is selected from the estimated residuals, $\hat{e}_{rl}(x) = f_{rl}(x) - \bar{f}_{r\bullet}$, by random permutation. When the null hypothesis is false this modified permutation test guarantees that the artificial data verifies the null hypothesis and have approximately the same noise variability as the original sample. Therefore this modified test would detect deviations from the null hypothesis that would be overlooked by the standard permutation test. From now on we use this modified permutation test when we refer to permutation test in text.

Let us now discuss an important question arising in economic and social micro data bases: in theses contexts it is usual each observation has a different weight, that is proportional to the amount of people in the population it is representing. The functional ANOVA test based on the statistic V_n, as in Cuevas et al. (2004), is not directly applicable to such weighted samples. Delicado (2006) establishes for weighted samples the result analogue to Theorem 1 in Cuevas et al. (2004). A new statistic V_n^w, the version of V_n appropriate for weighted samples, is introduced. Its null distribution can be approximated by Monte Carlo simulation or by permutation methods.

The procedures described here can be used to test the null hypothesis of no factor effects in the functional ANOVA model for any k samples of functions in L^2. If we are comparing samples of density functions (as it is the case in this work), that are always in L^1, we can assume that they also belong to L^2 and use the above methods. Moreover a L^1 version of the F_n statistic can be defined as

$$F_n^{L_1} = \frac{\sum_{i=1}^{k} n_i \left\| \bar{f}_{i\bullet} - \bar{f}_{\bullet\bullet} \right\|_1 /(k-1)}{\sum_{i,j} \left\| f_{ij} - \bar{f}_{i\bullet} \right\|_1 /(n-k)}, \qquad (1)$$

where the norm L^1 of a function f is defined as $\|f\|_1 = \int_a^b |f(x)|dx$.

5. Density estimation in EU regions

The information about the income distribution in European countries and regions was extracted from household budget surveys as described in Section 2. This information is summarized by the estimation of probability density functions representing income distributions. In this section we detail how we have done this estimation, and we show the estimated income density functions for European regions. These functions are the basic inputs to the functional ANOVA models we are fitting in Section 6.

We use nonparametric density estimation techniques to estimate income density functions. Specifically our estimations are based on kernel methods (see Silverman 1986, or Bowman and Azzalini 2001, for instance). Let $x_1,\ldots,x_n$ be a sample from a random variable with density function f, and let $w_1,\ldots,w_n$ be their respective weights. The weighted kernel estimator for the value of f at the point x is

$$\hat{f}_w(x)=\sum_{i=1}^{n}\frac{w_i}{\sum_j w_j}\frac{1}{h}K\left(\frac{x-x_i}{h}\right),$$

where K is a unimodal density function symmetric around 0 (a standard normal density, for instance), and h is to be known as *bandwidth* or *smoothing parameter*. Observe then that $K((x-x_i)/h)/h$ is a unimodal density function symmetric around x_i and rescaled in such way that it reproduces in the interval $[x_i-h,x_i+h]$ the shape of K in $[-1,1]$. Therefore $\hat{f}_w(x)$ can be interpreted as the mixture of densities $K((x-x_i)/h)/h$, each with weight w_i, that is the density corresponding to a random variable that randomly observe one of the original data $x_1,\ldots,x_n$ with probabilities $w_1,\ldots,w_n$, plus a noise e with density $K(e/h)/h$. Big values of h produce an estimator $\hat{f}_w(x)$ very smooth as a function of x, whereas small values of h lead to bumpy functions.

A well known characteristic of income distributions is their marked right asymmetry, that classically lead to model them as log-normal random variables. From a nonparametric point of view asymmetry implies that different degree of smoothness should be used in different levels of income (typically, more smoothness is needed in the right tail of the distribution, whit low density, corresponding to high incomes). In kernel estimation there are two main ways to apply different degree of smoothness to different zones. One of them consists in using variable bandwidths (that can depend on the point x where the density is estimated, or on each observation x_i). The other way, that we follows, is based on transformations. Data $x_1,\ldots,x_n$ are transformed by a known function g (we use here the logarithm function) achieving that the transformed data have almost symmetric distribution. Then constant bandwidth kernel estimation is done in the transformed scale, and a change-of-variable formula is used to have a density estimation in the original scale.

An additional problem should be noted here. The logarithm transformation has to be applied to positive data, but the income data we are analyzing are not all positive. Then a positive constant c has to be added to each observation before taking logs.

Finally, the nonparametric density estimator we are using is as follows:

$$\hat{f}(x) = \hat{f}_w^{\lg}(\log(x+c))\frac{1}{x+c}, \text{ for incomes } x > -c,$$

where $\hat{f}_w^{\lg}$ is the weighted kernel density estimator derived from $\log(x_i + c)$, $i = 1,\dots,n$. Observe that two parameters (bandwidth h and c) have to be chosen.

The bandwidth choice is the cornerstone of nonparametric estimation, because the aspect and properties of the estimation strongly depend on it. This choice has received much attention in the last years (see for instance Wand and Jones 1995). In this paper we use the *normal reference rule*, that takes the theoretical expression of optimal bandwidth h and replace there the unknown terms (because they depend on the density that is being estimated) by the values they would take if data were normally distributed, with the same mean and variance. It is well known that this rule is only appropriate when data are near normality (that is the case for $\log(x_i + c)$) and that it tends to oversmooth (to produce too high values for h). In order to correct the oversmoothing, a common practice is to multiply the proposed values by a positive constant lower than 1. In our case, we always take 2/3 of the values provided by the normal reference rule. The constant 2/3 was chosen by visual inspection.

An alternative bandwidth choice rule, nowadays accepted as much satisfactory the method, is the *plug-in* method (Sheather and Jones 1991) that consists in replacing in the theoretical expression of optimal bandwidth h the unknown terms (because they depend on derivatives of the unknown density that is being estimated) by estimations (based in kernel estimation of the derivatives). The involved computations are far from being trivial, and in fact there are not available implementations covering the possibility of weighted samples. This is the main reason that leads us to use the normal reference rule, jointly with the fact that data $\log(x_i + c)$ are almost normally distributed.

The effect of the choice of constant c on the density estimator has deserved little attention in the literature (compared to the choice of bandwidth h), but it is not at all negligible from a practical point of view, as our experience clearly showed. On the one hand, c must be greater than the absolute value of the lowest observation, if it is negative. On the other hand, it is sensible to take c a little bit greater than this quantity, in order to not having log-scale data too negative (almost minus the machine infinity). Taking that into account, we have implemented the following rule to determine the constant c. Let c_0 be the absolute value of the minimum of $x_1,\dots,x_n$ if this minimum is negative and 0 otherwise. Let c_1 be the percentile 10% of the subset of positive values in $x_1,\dots,x_n$. Then we define $c = c_0 + c_1$.

Lets now go to the last point to concrete before having estimated density functions: we should state the set of points where the density function will be estimated, namely $t_1,\dots,t_m$. The specific choice of this set of evaluation points has implications in the final aspect of the estimated densities: a very low m or a very bad distributed set of points could produce unfair density functions. Moreover, it should be desirable to have a unique set of evaluation points, common to all the regional densities we are estimating in this paper, because all computations between functions (sums, differences, products) are much easier when they are evaluated in the same set of points.

The way we are chosen the set of evaluation points is as follows. Let us first talk about density estimation of equivalent income distributions before taxes and benefits, denoted by x. We take the whole European sample (99120 data) and compute the quantiles of order $j/1000$, $j=1,\ldots,999$. We drop out repeated quantile values and those being lower than -1000 € or greater than 7000 €, because observations outside these limits (less than 0.5% of the data) can be consider outliers and all the regional densities are almost constant outside this interval. In this way we obtain 926 evaluation points, ranging from 0 € to 6811 €. Then we add 10 equispaced points between -1000 € and 0, and 10 more between 6811 € and 7000 €. In this way we have 946 evaluation points between -1000 € and 7000 €. We are interested in the regional relative equivalent income before taxes and benefits, computed dividing x by the regional median of x. The evaluation points we use to estimate the corresponding density functions are the previous evaluation points divided by the global European median of x, because this way we still have a common evaluation points set for all the regional estimated densities.

When working with density estimation of equivalent income distributions after taxes and benefits, denoted by y, the way we select the evaluation points is analogous. In this case the interval where density estimators are evaluated is $[-1000€,5000€]$, which include more than 99.5% of the data). There are 1015 evaluations points now.

The density estimation has been done in the package R (R Development Core Team, 2005) using the library *sm* (Bowman and Azzalini 2001), that implements the normal reference bandwidth choice rule and kernel estimation for weighted data. The estimated regional densities are shown in Figure 3 (before taxes and benefits) and Figure 4 (after taxes and benefits). They are grouped by countries.

Relative income density functions before tax and transfers in all regions show a bimodal shape, with one mode (of maximum height) at zero incomes and a second one located in incomes between 1 and 2 times the regional median. The height of the first mode is substantially higher than the one of the second mode as well as more sharp. There is a significant variation in heights among regions of different countries (See the case of Italy with similar heights for the two modes). The second mode shows also important differences among regions of a given country (See for instance Germany or Finland).

Relative income density functions after tax and transfers look rather different. The concentration around zero income disappears, as well as the bi-modal shape. In almost all cases now the density is unimodal with a mode close to median regional income. Regional variation in densities within a country seem to have reduced.

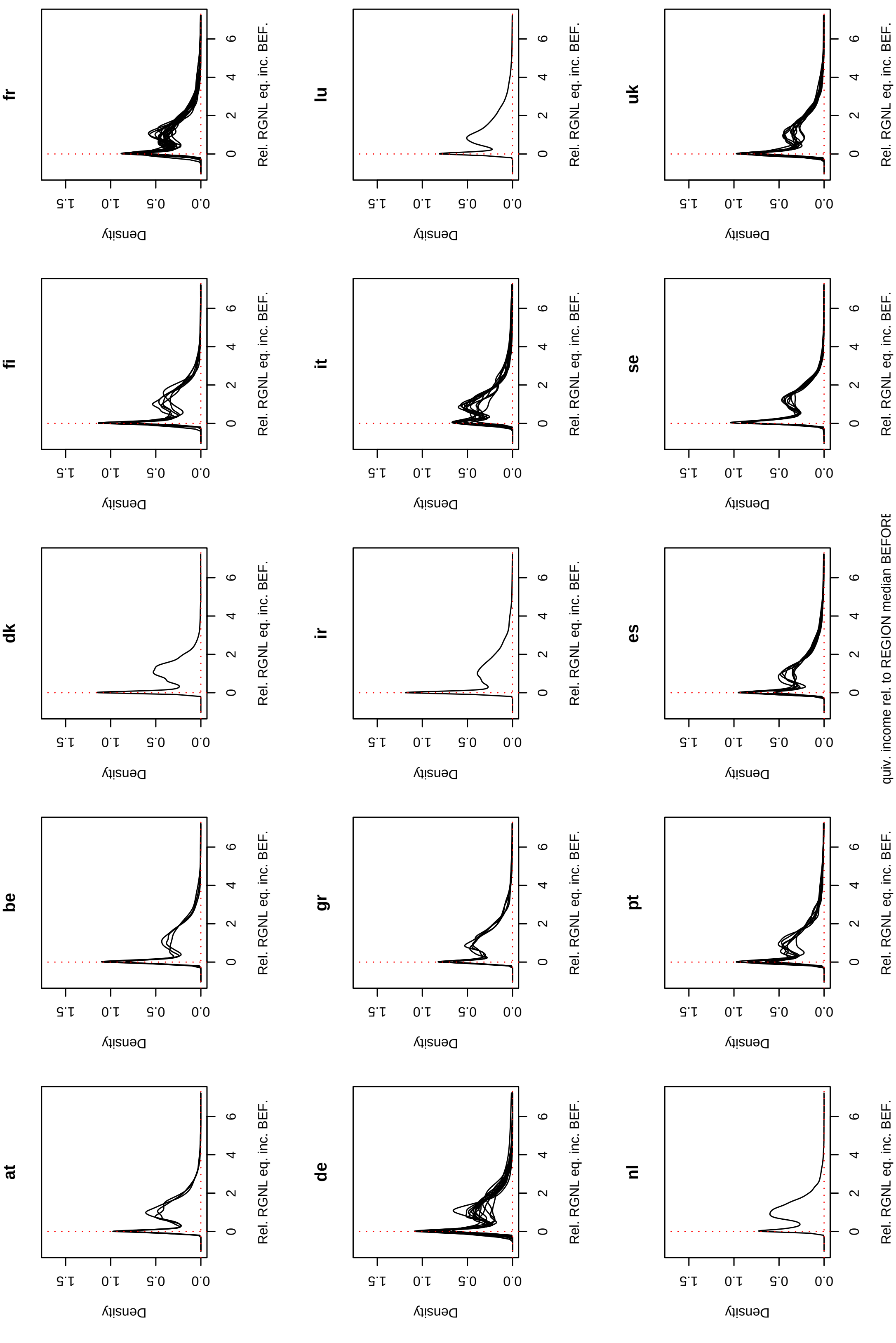


Figure 3. Density estimation for regional equivalent income before taxes and benefits, grouped by countries.

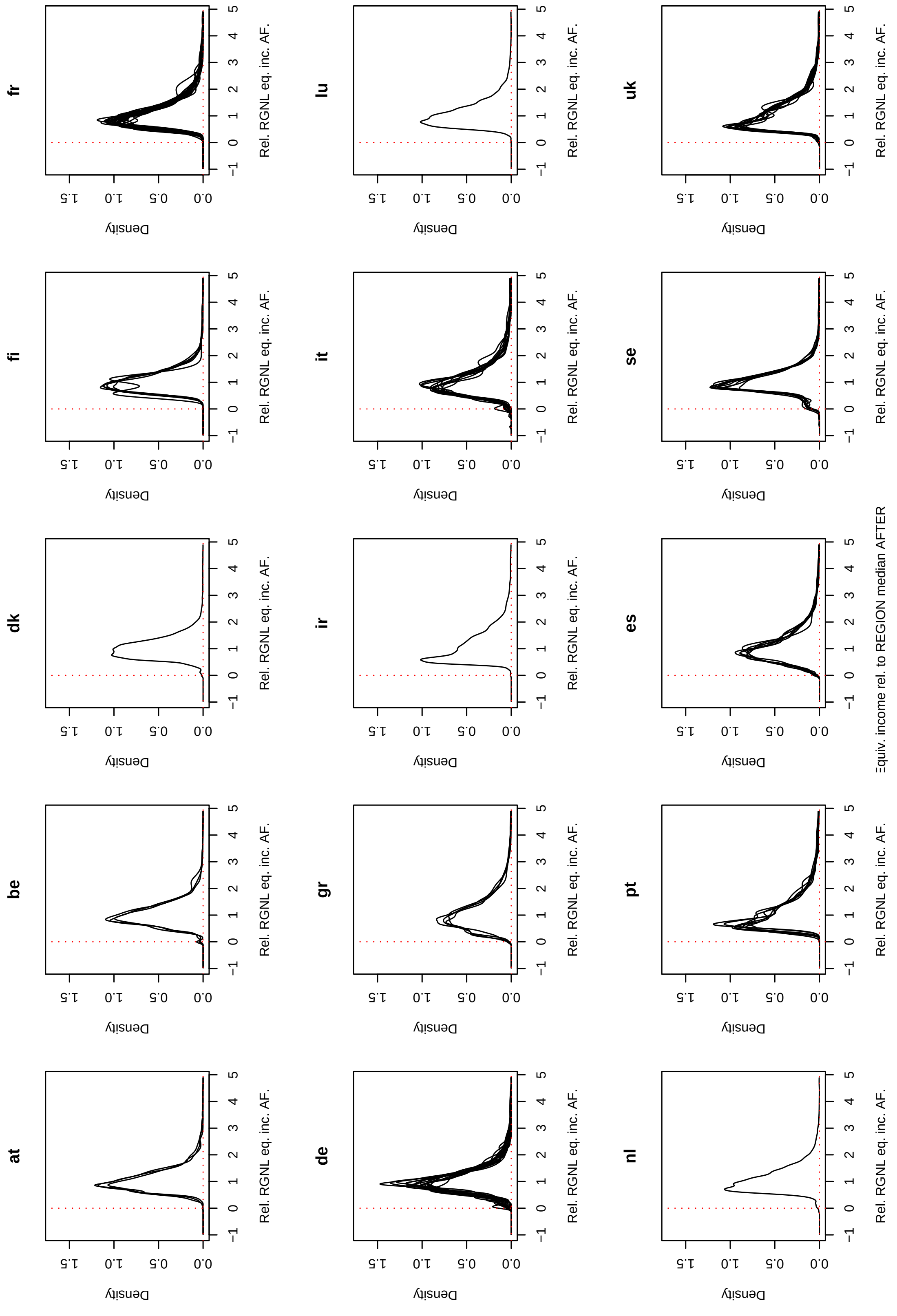


Figure 4. Density estimation for regional equivalent income after taxes and benefits, grouped by countries.

6. Functional ANOVA on regional EU density functions

In this section we deal with the main contribution of the paper: to validate our thesis that the country to the regions belong is a much more powerful explanatory factor for income distribution after taxes and benefits than before. We approach the problem fitting two functional ANOVA models to the regional income estimated density functions, before and after taxes and benefits, where the null hypothesis of no country effect is tested. Then we compare the evidence against the null hypothesis before and after taxes and benefits. Our thesis will be validated if the evidence corresponding to data after taxes and benefits is much grater than the obtained with the data before taxes and benefits.

At this point we should specify how the evidence against the null hypothesis can be measured. Any distance between the observed value T_{Obs} of the test statistic T and their null distribution is a valid such measure, that is, any distance between the observed data and the null hypothesis could be used. A first attempt is to use 1 minus the test p-value as a distance:

$$d_{T,1} = 1 - p \cdot value = P_{H0}(T < T_{Obs}) .$$

This quantity is always between 0 and 1, and values near to 1 (p-values lower than 0.05, typically) lead to reject the null hypothesis. Other distance that we are using in this paper is the standardized value (using the null distribution) of the observed test statistic,

$$d_{T,2} = \frac{T_{Obs} - m_T}{S_T},$$

where m_T is the mean of the statistic T under the null distribution, and S_T is its standard deviation. (both quantities estimated from simulated or permuted samples).

In order to make conclusions robust against the choice of the specific functional ANOVA test, we have used two different statistics (introduced in Section 4) to test the null hypothesis: the statistic V_n^w defined in equation (1) and based in the norm L_2, and the statistic $F_n^{L_1}$ defined in equation (1) and based in the norm L_1. In Section 4 was indicated that the null distribution of V_n^w and $F_n^{L_1}$ must be approximated by simulation. There were presented three alternative methods to approach the null distribution of V_n^w: the heteroscedastic and homoscedastic asymptotic approximation (both based in Proposition 1), and the permutation test approximation (that implicitly assumes homoscedasticity). The last approach is the only available for approximating the null distribution of $F_n^{L_1}$. In our implementation the number of simulated or permuted samples of functions, N, has been taken equal to 200.

Let us now describe the results obtained on the density functions estimated in Section 5. Figure 5 shows the observed values of statistics V_n^w and $F_n^{L_1}$ as big solid circles, in panels referred as L2 and L1, respectively. The approximated null distributions of these statistics are represented by the box-plots of the N simulated (or permuted) values. These graphics reflect well the distances between observed statistics and null distributions. It is visually clear that these distances corresponding to income densities after the effect of countries tax-benefit systems are much grater than those corresponding to the net income densities, what corroborates our thesis on country effects. More formal support can be obtained computing distances of $d_{T,1}$ and $d_{T,2}$ kind.

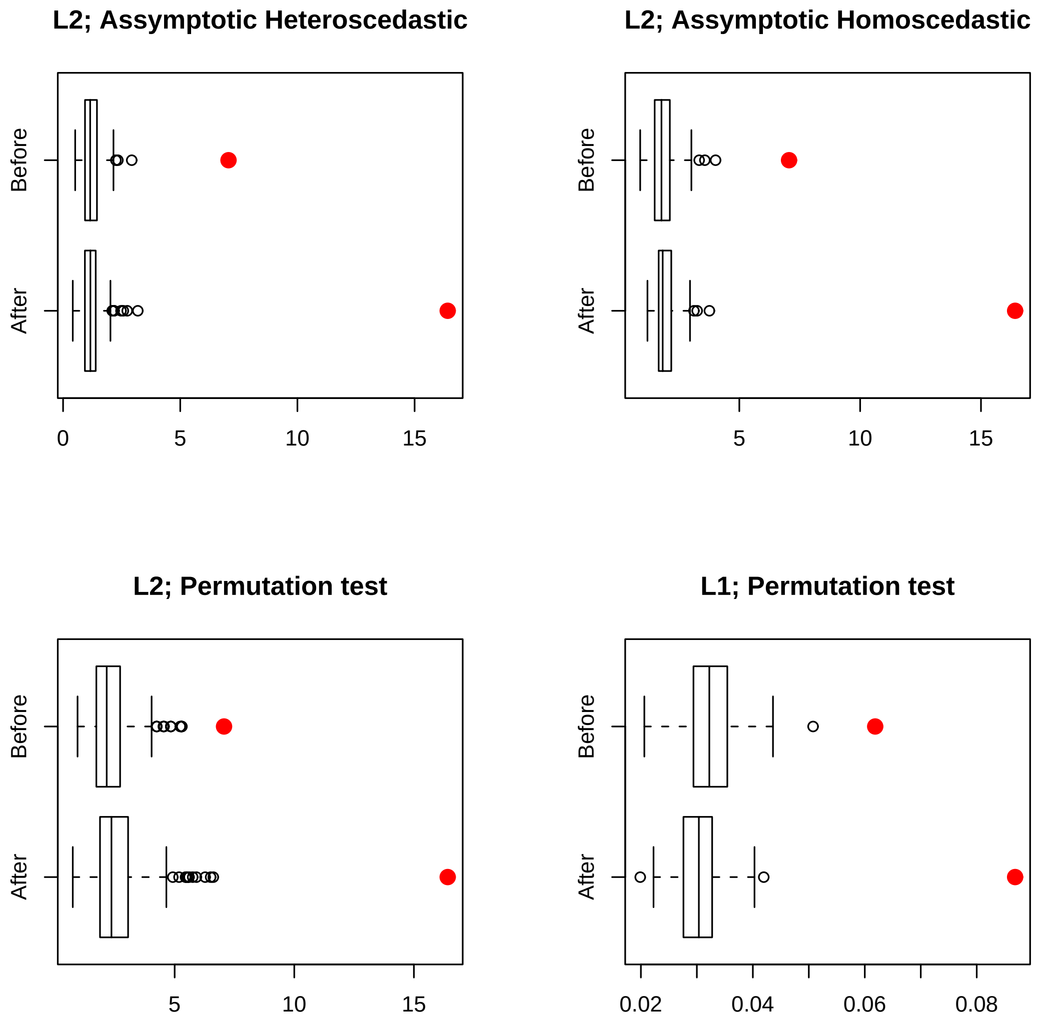


Figure 5. Box-plots of the N simulated (or permuted) values of distributions of statistics V_n^w and $F_n^{L_1}$. Observed values of statistics V_n^w and $F_n^{L_1}$ as big solid circles.

The null hypothesis of no country effect is rejected in all cases: equivalent income before or after taxes and benefits, V_n^w or $F_n^{L_1}$, and the three ways to approximate the null distribution of V_n^w. In fact, the 8 resulting p-values have been estimated as 0: no simulated or permuted sample of functions presents statistics V_n^w or $F_n^{L_1}$ greater than those observed in our original functional data set, implying that distances $d_{V_n^w,1}$ and $d_{F_n^{L_1},1}$, based on p-values, are always equal to 1, for both sets of income densities (before and after taxes and benefits). Therefore this kind of distances between data and null hypothesis does not help to validate our thesis that country effect is more important after taxes and benefits.

When computing type $d_{T,2}$ measures, based on the standardized observed statistic value, things are different and it is possible to calibrate the dissimilar behaviour of regional densities before and after the action of tax-benefit systems. Mean and standard deviations under the null hypothesis are estimated from simulated or permuted samples. Table 4 shows distances $d_{V_n^w,2}$ (columns 1 to 3) and $d_{F_n^{L_1},2}$ (column 4) for situations before (first row) and after (second row) taxes and benefits. All those numbers indicate that the evidence against the null hypothesis of no country effects after is more than twice than before taxes and benefits (see the third row, what shows the quotient of rows 1 and 2).

Results corresponding to asymptotic approximations (columns 1 and 2) are consistent, as well as they are those based on permutation tests (columns 3 and 4). The similarity between the two first columns indicates that homoscedasticity is a valid assumption. Discrepancies between these columns and the last two lead to question the soundness of asymptotic approximations: in fact asymptotic approximations need a large number of functions per class (large number of regions per country, here) what does no happen in our case. Despite those discrepancies, values in the third row are similar for the four ways to carry on the functional ANOVA tests, what points out the robustness of our conclusion on the used method.

Equivalent income density...	$T = V_n^w$ Heteroscedastic asymptotic approximation	$T = V_n^w$ Homoscedastic asymptotic approximation	$T = V_n^w$ Permutation test approximation	$T = F_n^{L_1}$ Permutation test approximation
... before taxes and benefits	15.03	10.98	5.20	6.20
... after taxes and benefits	36.34	34.22	12.84	14.55
Quotient of the previous rows	2.42	3.12	2.46	2.35

Table 4. List of distances $d_{T,2}$ for four versions of the functional ANOVA test, and two scenarios.

7. Concluding comments

In this paper we have studied the role of the country factor in explaining the variability of regional income distributions before and after national tax-benefit systems operate in Europe-15. Our main aim has been to disentangle the relative contribution of countries' tax-benefit systems on income distribution shapes, in relation to other country-specific institutional features. Functional Data Analysis techniques have allowed to take into account of the full income distribution, instead of a single univariate index. Moreover, the Functional ANOVA model has permitted the analysis of the country explicative power on the dispersion of regional income density functions (estimated non-parametrically) before and after taxes and transfers.

Our statistical evidence shows that regional income distributions of different countries are statistically different, both before and after redistribution takes place –we reject in both cases the null hypothesis of equal average regional distributions among countries.

The analysis of the full distribution leaves therefore to different conclusions than the analysis based on a single inequality index –remember that using the Gini we did not reject the null hypothesis of equality of the Ginis on the distribution before. However, the null assumption is rejected more strongly in the after redistribution case. If we consider the value of the statistic used in the contrast as a distance between the observed data and the model under the null hypothesis, the distance separating (from equality) the regional distributions after is more than the double (and smaller than tree times) the distance before.

From a European perspective, our results suggest a different reading. The variability between countries in the shape of their regional income distributions is clearly increased by national tax-benefit systems. Our analysis suggests that a European taxes and benefits would contribute reducing disparities in disposable income distribution among countries.

Bibliography

Atkinson, A.B. (2000) "Increased Income Inequality in OECD Countries and the Redistributive Impact of the Government Budget", WIDER, World Institute for Development Economics Research, Working Papers No. 202, October.

Atkinson, *A. B.* Bourguignon, *F.* and Chiappori, P.A. (1988) « Fiscalité et transferts: une comparaison franco-britannique »**.** *Annales d'Economie et de Statistique*, v. 0, iss. 11, 117-40.

Atkinson, A.B. Rainwater, L. and Smeeding, T.M. (1995), *Income Distribution in OECD countries. Evidence from the Luxembourg Income Study*, Social Policy Studies nº 18, OECD, Paris.

Bowman, A.W. and Azzalini, A. (2001) *Applied Smoothing Techniques for Data Analysis.* Oxford. Oxford University Press.

Cuevas, A., Febrero, M., and Fraiman, R. (2004) An anova test for functional data. *Computational Statistics & Data Analysis*, **47**, 111-122.

Delicado, P. (2006) Functional ANOVA when data are a weighted sample of density functions. Preprint.

Jenkins, S.P. and Van Kerm, P. (2005) Accounting for income distribution trends: a density function decomposition approach. *Journal of Economic Inequality*, 3, 43-61.

Kneip, A., and Utikal K.J. (2001) Inference for Density Families Using Functional Principal Component Analysis. *Journal of the American Statistical Association*, **96**, 519-542.

Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T., and Cohen, K.L. (1999) Robust principal component análisis for functional data. *TEST*, **8** (1), 1-73.

Mercader, M. and Levy, H. (2004) The role of tax and transfers in reducing personal Income Inequality in Europe's regions: Evidence from EUROMOD. *EUROMOD Working Paper No. EM9/04.*

Muñoz-Maldonado, Y., Staniswalis, J.G., Irwin, L.N., and Byers, D. (2002) A similarity analysis of curves. *Canadian Journal of Statistics*, **30**, 373-381.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

Ramsay, J.O., and Silverman, B. W. (1997) *Functional Data Analysis*. New York: Springer.

Ramsay, J.O., and Silverman, B. W. (2002) *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer-Verlag.

Sheather and Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 683-690.

Stewart, K. (2002) "Measuring Well-Being and Exclusion in Europe's Regions", CASEpaper 53, CASE Centre for Analysis of Social Exclusion, London School of Economics.

Sutherland, H. (Ed.) (2001) "Final report. EUROMOD: an integrated European benefit-tax model", EUROMOD Working Paper nº. EM9/01.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Wagstaff et 24 other authors (1999). "The redistributive effect, progressivity and differential tax treatment: Personal income taxes in twelve OECD countries", *Journal of Public Economics*, 72, 73-98.

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

Table 1: Databases used in each country

Country	Base Dataset for EUROMOD	Date of collection	Reference time period for incomes
Austria	Austrian version of European Community Household Panel (W5)	1999	annual 1998
Belgium	Panel Survey on Belgian Households (W6)	1999	annual 1998
Denmark	European Community Household Panel (W2)	1995	annual 1994
Finland	Income distribution survey	1998	annual 1998
France	Budget de Famille	1994/5	annual 1993/4
Germany	German Socio-Economic Panel (W15)	1998	annual 1997
Greece	European Community Household Panel (W2)	1995	annual 1995
Ireland	Living in Ireland Survey (W1)	1994	month in 1994
Italy	Survey of Households Income and Wealth	1996	annual 1995
Luxembourg	PSELL-2 (W5)	1999	annual 1998
Netherlands	Sociaal-economisch panelonderzoek (W3)	1996	annual 1995
Portugal	European Community Household Panel (W3)	1996	annual 1995
Spain	European Community Household Panel (W3)	1996	annual 1995
Sweden	Income distribution survey	1997	annual 1997
UK	Family Expenditure Survey	1995/6	monthly in 1995/6

Table 2. Regional information in EUROMOD national databases

Country	Information used in our analysis	Average population per region	Number of regions
Austria	NUTS1	2,645,969	3
Belgium	NUTS1	3,299,186	3
Denmark	Not considered		
Finland	NUTS2	814,532	6
France	NUTS2	2,588,033	22
Germany	NUTS1	4,966,341	16
Greece	NUTS1	2,635,525	4
Ireland	Not considered		
Italy	NUTS1+ South Split	4,767,237	12
Luxembourg	Not considered		
Netherlands	Not considered		
Portugal	NUTS2	1,417,429	7
Spain	NUTS1	5,557,837	7
Sweden	NUTS2	1,123,557	8
UK	NUTS1+Greater London	4,786,980	12

Table 3: Regions in EUROMOD and sample sizes.

Country	Country Code	Region Name	NUTS abbreviation	Sample size
Portugal	12	Madeira	PT3	598
	12	Açores	PT2	599
	12	Algarve	PT15	637
	12	Centro	PT12	1,027
	12	Alentejo	PT14	514
	12	Norte	PT11	840
	12	Lisboa E Vale Do Tejo	PT13	591
Greece	7	Thraki, Makedonia, Thessalia	GR1	1,660
	7	Dellada, Sterea, Pelloponisos, Ionia Nisia Ipiros	GR2	1,251
	7	Notio Aigaio, Voreio Aigaio, Kriti	GR4	656
	7	Athina	GR3	1,601
Spain	13	Canarias	ES7	380
	13	Andalucia, Murcia	ES6	1,013
	13	Cast Leon, Cast Mancha, Extremadura	ES4	959
	13	Galicia, Asturias, Cantabria	ES1	895
	13	Catalunya, Valen, Baleares	ES5	1,375
	13	Euskadi, Navarra, Rioja, Aragón	ES2	960
	13	Madrid	ES3	537
Italy	9	Sicilia	ITA	559
	9	Basilicata/Calabria	IT92	389
	9	Campania	IT8	709
	9	Puglia	IT91	520
	9	Sardegna	ITB	295
	9	Abruzzo-Molise	IT7	396
	9	Lazio	IT6	411
	9	Center	IT5	1,250
	9	North-east	IT3	1,009
	9	North-west	IT1	1,048
	9	Lombardia	IT2	824
	9	Emilia	IT4	725
Belgium	2	Flandre	BE2	1,961
	2	Wallonie	BE3	1,300
	2	Bruxelles	BE1	393
Sweden	14	Smaland med oarna	SE03	1,607
	14	Norra Melansverige	SE06	1,235
	14	Mellersta Norrland	SE07	888
	14	Ovre Norrland	SE08	1,049
	14	Vastsverige	SE05	4,448
	14	Sydsverige	SE04	2,889
	14	Östra mellansverige	SE02	3,404
	14	Stockholm	SE01	4,114
Finland	4	Itä-suomi	FI13	1,394
	4	Väli-suomi	FI14	1,364
	4	Pohjois-suomi	FI15	762
	4	Etelä-suomi	FI12	3,276
	4	Ahvenanmaa/aland	FI2	44
	4	Uusimaa	FI11	2,152
UK	15	Northern Ireland	UKB	134
	15	North	UK1	405
	15	West Midlands	UK7	621

	15	Scotland	UKA	604
	15	Yorks & Humberside	UK2	594
	15	East Anglia	UK4	282
	15	North West	UK8	722
	15	Wales	UK9	339
	15	East Midlands	UK3	491
	15	South West	UK6	637
	15	South East	UK5	1,274
	15	Greater London	UK55	694
Germany	6	Thüringen	DEG	358
	6	Sachsen-Anhalt	DEE	356
	6	Sachsen	DED	594
	6	Mecklenburg-Vorpommern	DE8	216
	6	Brandenburg	DE4	318
	6	Rheinland Pfalz / Saarland	DEB	417
	6	Bremen	DE5	61
	6	Berlin-Ost	DE3	189
	6	Niedersachsen	DE9	626
	6	Baden-Würtemberg	DE1	934
	6	Hamburg	DE6	97
	6	Bayern	DE2	984
	6	Saarland	DEC	134
	6	Nordrhein Westfalen	DEA	1,508
	6	Schleswig-Holstein	DEF	190
	6	Hessen	DE7	498
France	5	Corse	FR83	38
	5	Nord-Pas De Calais	FR3	715
	5	Basse-Normandie	FR25	281
	5	Auvergne	FR72	234
	5	Champagne-Ardennes	FR21	305
	5	Poitou-Charentes	FR53	309
	5	Languedoc-Roussillon	FR81	437
	5	Bretagne	FR52	582
	5	Pays De La Loire	FR51	653
	5	Centre	FR24	434
	5	Limousin	FR63	153
	5	Franche Comte	FR43	256
	5	Bourgogne	FR26	320
	5	Aquitaine	FR61	585
	5	Midi-Pyrenees	FR62	523
	5	Haute-Normandie	FR23	355
	5	Picardie	FR22	305
	5	Rhone-Alpes	FR71	949
	5	Provence-Alpes-Cote Dazur	FR82	879
	5	Lorraine	FR41	472
	5	Alsace	FR42	367
	5	Ile De France	FR1	2,139
Austria	1	West: Oberösterreich, Salzburg, Tirol, Vorarlberg	AT3	879
	1	South: Kärnten, Steiermark	AT2	648
	1	East: Wien, Burgenland, Niederösterreich	AT1	1,145