Working Paper Series

On the link between the coefficient of determination and polarization

Federico Palacios-González
Rosa María García-Fernández

ECINEQ WP 2011 – 196

ECINEQ 2011-196
April 2011

www.ecineq.org

On the link between the coefficient of determination and polarization

Federico Palacios-González
Rosa María García-Fernández*
University of Granada

Abstract

In this paper, taking as starting point the link between polarization and dispersion, we reformulate the measure of polarization of Zhang and Kanbur (2001) using the decomposition of the variance instead of the decomposition of the Theil index. The proposed measure is equivalent to the coefficient of determination of an ANOVA Linear Model, that explains the income of the households as a function of any population characteristic e.g. education, gender, occupation etc. This result provides an alternative way to analyze polarization by household characteristics and at the same time allows us to compare sub-populations via the estimated coefficients of the ANOVA model.

Keywords. Polarization, coefficient of determination, ANOVA model
JEL Classification: D31,D63.

* **Addresses of correspondence**: Department of Quantitative Methods for Economics and Business. University of Granada, Granada, Spain. Tel:+34 958241956; Fax:+34 958 240620. rosamgf@ugr.es

1. Introduction

The concept of income polarization was introduced by Wolfson (1994) and Esteban and Ray (1994) independently with the aim of explaining distributional changes, such as the clustering around local means, that are not explained by the standard measures of inequality which refer to convergence to global mean. Since then until now different approaches to measure income polarization have been provided [see among others Tsui and Wang (2000), Zhang and Kanbur (2001), Duclos et al. (2004), Silber et al. (2007), Zelli and Pittau (2007), Gasparini et al. (2008) and Hussain (2009)].

Throughout this paper we focus on the approach to polarizarion of Zhang and Kanbur (2001). Quoting these authors, for a k exogenously given groups, " as income differences within group diminish, that is as the groups become more homogeneous internally, differences across groups are, relatively speaking, magnified and polarization is higher. Similarly, for given within group differences, as the groups means drift apart, polarization increases". Zhang and Kanbur quantified polarization by using the ratio of the between groups inequality to the within group inequality.

The aim of this work is to reformulate, by the reasons explained in the second section, the measure of polarization of Zhang and Kanbur (2001), using the decomposition of the variance instead of the property of decomposition of the inequality indices. Additionally, it is demonstrated that the new measure is equivalent to the coefficient of determination of an ANOVA Linear Model, that explains the household income as a function of any characteristic of the population as for instance education, gender, occupation etc. This result provides an alternative to the proposed measures in the specialized literature to analyze polarization by household characteristics (see among others Zhang and Kanbur, 2001 and Gradín, 2000). The

new approach allows us to compare sub-populations via the estimated coefficients of the ANOVA model. In this way, we can determine the most relevant variables to characterize the population into homogeneous groups that are antagonistic in terms of income. The proposed approach is applied to Spanish household data for the years 2006 and 2008. We utilize the information provided by the Household Budget Survey conducted by the National Statistics Institute. We have focused on the monthly equivalized net income and on the following economic and demographic characteristics of the main breadwinner of the household: gender, education, work status, branch of activity, town size and geografical area. In addition, the measures of Zhang and Kanbur (2001) and Esteban et all (see Gradín,2000), which are frequently utilized to analyze polarization by sub-populations, are obtained to establish comparisons.

The paper is organized as follows. Section two defines the measure of polarization and its links with the coefficient of determination of an ANOVA linear model. Section three contains the empirical applications and section four contains the conclusions.

2. Polarization and the coefficient of determination of an ANOVA linear model

In this section, we begin by developing a variance based measure of polarization focusing on the approach to polarization of Zhang and Kanbur (2001). Secondly, we show the link between the proposed measure and the coefficient of determination (R^2) of an ANOVA Linear Model that explains the household income as a function of any population characteristic e.g. education, gender, branch of activity etc.

2.1 Polarization measurement

Zhang and Kanbur (2001, ZK henceforth) considered that polarization was generated by two tendencies. According to these authors, for a k exogenously given groups, as income difference within the group decreases, that is the groups become more homogeneous

internally, differences across groups are magnified and polarization is higher. In a similar way, given the intra-group differences, the further apart the mean incomes of the groups are, the higher the polarization. These tendencies of polarization agree with the concepts of identification and alienation defined by Esteban and Ray (1994) to measure polarization. Indeed, they can be considered as an alternative way to quantify identification and alienation respectively. ZK quantified these tendencies by the ratio of the between groups inequality to the within group inequality.

As can be observed by the reader, the approach of ZK links the concept of polarization with dispersion. Based on this relationship we are going to reformulate the measure of ZK using the intra-group and the inter-groups variance instead of the intra-group and inter-groups inequality. To justify why we utilize the decomposition of the variance, we refer to Fisher (1958, among others) who pointed out that when the representative magnitude of each group is the mean of the variable of interest, in our case the mean income, the intra-group variance and the inter-groups variance are the most appropriate approaches to evaluate the homogeneity within a group, and the heterogeneity across groups respectively. Moreover, we would like to stress that concentration and homogeneity are not equivalent concepts (see for instance Hermoso and Bastida, 2000). The former refers to the way in which total income is distributed among individuals, and the latter is related to the degree of homogeneity of the values of the statistical variable (in our case income). Although there is a correspondence between equidistribution or null concentration and null dispersion, maximum concentration is not associated with either maximum dispersion or with null dispersion. As a consequence, the measures of concentration should not be used to compute dispersion and vice versa.

To reformulate the measure of Zhang and Kanbur in terms of dispersion we proceed as follows. Given a number of groups determinated exogenously, we assume that the income difference within the group decreases when the income of the individuals are closer to the

average income of the group to which they belong. The smaller the distance the higher the homogeneity within the group. We presume that heterogeneity is linked to the distance between the mean incomes of the groups. The larger the distance, the higher the heterogeneity between groups. In line with the previous arguments we consider on the one hand, that a global measure of income homogeneity within a group should be inversely proportional to the intra-group variance (V_W). On the other hand, a global measure of income heterogeneity between individuals belonging to different groups, should be proportional to the variance between groups (V_B). Hence, following Zhang and Kanbur (2001) we can compute polarization by the expresssion

$$P^* = \frac{Inter\ groups\ variance}{Intra\ group\ variance} = \frac{V_B}{V_W} \in [0, +\infty]$$

Observe that the above measure, as well as the expression of Zhang and Kanbur is not normalized, making its interpretacion difficult. However, taking into consideration that the variance of the overall population (*V*) can be broken down into the intra-group variance plus the inter-groups variance, that is

$$V = V_W + V_B$$

we can normalize P^* obtaining

$$P = \frac{V_B}{V} = 1 - \frac{V_W}{V} \in [0,1]$$

Multiplied by 100 the result of the proposed measure can be interpreted as a percentage of polarization.

As we can note, P^* is closer to the measure of polarization of Zhang and Kanbur (2001). Indeed, if we normalize the latter measure using the decomposition property of the Theil[1]

[1]The index of Theil can be broken down in a similar way as the variance. That is, the overall inequality is equal to the inter-groups inequality plus the intra-group inequality.

index, we obtain an expression that resembles our index. The main modification concerns the way in which we compute homogeneity and heterogeneity, since we focus on dispersion instead of concentration indices.

2.2 Link between polarization and the R^2 of an ANOVA linear model

In this Section it is demonstrated that the defined measure is equivalent to the coefficient of determination of an ANOVA linear model. Specifically, the measure P is the coefficient of determination (R^2) of a general linear model in which the income is explained by the dummy variables that assign each individual to a group. This relation between P and R^2 provides an alternative way of looking at polarization by features of the households.

For a more detailed explanation, let us consider any individual characteristic, for instance education, gender, occupation etc., that provides an exhaustive partition of the whole population into k groups or sub-populations $G_1, G_2, \dots, G_k$.

Let us consider a sample of n individuals randomly selected, over which is observed a variable Y, that in our case will be the income of the household. Each sample will provide the following table of data

$D_1(j)$	$D_2(j)$	....	$D_k(j)$	Y_j
$D_1(1)$	$D_2(1)$	....	$D_k(1)$	Y_1
$D_1(2)$	$D_2(2)$	....	$D_k(2)$	Y_2
....		....		
$D_1(n)$	$D_2(n)$		$D_k(n)$	Y_n

where D_i , $\forall i = 1,2, \dots, k$, is a dummy variable that takes the value one if and only if the individual j of the sample belongs to the groups G_i, and zero in other cases.

Let us focus on the General Linear Model (GLM)

$$Y_j = \sum_{i=1}^{k} \delta_i D_i(j) + u_j \qquad (1)$$

where $\delta_i \; i = 1,2, \dots k$ are the regression parameters and u_j is the error term which verifies that $E[u_j] = 0, V(u_j) = \sigma^2$and $Corr(u_i, u_j) = 0$ for all $i \neq j$.

Observe that Model (1) explains the variable Y in relation to the group to which each individual belongs. Given that $D_i(j)$ is a dicotomic (dummy) variable, model (1) can be interpreted as an ANOVA model (see for example Gujarati, 1997, p. 490).

Denoting by

$$\mathbf{X} = \begin{pmatrix} D_1(1) & \cdots & D_k(1) \\ \vdots & \ddots & \vdots \\ D_1(n) & \cdots & D_k(n) \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \boldsymbol{\delta} = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_n \end{pmatrix}$$

we have

$$\widehat{\boldsymbol{\delta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad (2)$$

Note that there is only one element equal to one in each row of matrix $\mathbf{X}$, the rest of the elements of the row are equal to zero. This is because the groups G_i constitute a partition of the population, and each individual belongs only to one group. As a consequence the columns of $\mathbf{X}$ are orthogonal which allows us to affirm that

$$(\mathbf{X}'\mathbf{X}) = \begin{pmatrix} n_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & n_k \end{pmatrix}$$

is a diagonal matrix in which, every $n_i \;\; i = 1,2, \dots, k$ is the number of individuals of the sample that belong to the groups $G_i \; i = 1,2, \dots . k$. That is

$$n_i = \sum_{j=1}^{n} D_i(j)^2 \; = \sum_{j=1}^{n} D_i(j) \;\; \forall i = 1,2, \dots, k$$

Moreover

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum_{j=1}^{n} D_1(j) Y_j \\ \vdots \\ \sum_{j=1}^{n} D_k(j) Y_j \end{pmatrix}$$

that is, the k components of the column vector $\mathbf{X}'\mathbf{Y}$ are the total of the variable Y in each group. Hence, it is evident that the components of the vector $\widehat{\boldsymbol{\delta}}$ given by (2) are the sample means of the variable Y in the different groups which are denoted by $\bar{Y}_i$.

We want to highlight two issues related to model (1). Firstly, note that the model does not present exact collinearity since it does not have constant term. An alternative way to eliminate exact multicollinearity is to retain the constant and remove one dummy variable of the model. Nonetheless, there is no difference to the fitted values for Y and the residual of the regression if the constant, or any of the dummy variables, is dropped from the model[2] and hence the coefficient of determination is always the same (see for instance Russell and MacKinnon, 2004, pp. 70).

Secondly, observe that the sum of the residuals of model (1) is zero despite the fact that the model does not have intercept since the elements of the rows of matrix $\mathbf{X}$ add up to one, that is

$$\sum_{j=1}^{k} \mathbf{X}_j = \mathbf{1}_n$$

[2] In the empirical applications is more interesting to eliminate a dummy variable instead of the constant because the coefficients of the model show the difference between the expected income of the groups included in the model and the omitted group.

where $\mathbf{X}_j$ is the *j- th* column of matrix $\mathbf{X}$ and $\mathbf{1}_n$ is a column vector whose n components are equal to one. Thus, it is verified[3]

$$\sum_{i=1}^{n} e_i = \mathbf{1'}_n\mathbf{e} = \left(\sum_{j=1}^{k} \mathbf{X'}_j\right)\mathbf{e} = \sum_{j=1}^{k}\left(\mathbf{X'}_j\mathbf{e}\right) = 0$$

where $\mathbf{e}$ is the residual vector.

Hence, although model (1) does not have constant term the decomposition of the total sum of squares into the explained sum of squares plus the residual sum of squares is valid, and the coefficient of determination takes values between $[0,1]$.

Taking into consideration that the explained sum of squares is given by

$$SSE = \widehat{\boldsymbol{\delta}}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2 = \sum_{i=1}^{k} n_i\bar{Y}_i^2 - n\bar{Y}^2 \qquad (3)$$

it can be affirmed that

$$\frac{SSE}{n} = \frac{1}{n}\sum_{i=1}^{k} n_i\bar{Y}_i^2 - \bar{Y}^2$$

is the inter groups variance.

Observe that if the variable Y is the household income, the polarization measure P is the coefficient of the determination of the model that explains the household income in relation to the group to which each individual belongs. That is

$$P = \frac{Inter\ groups\ variance}{Total\ variance} = \frac{SSE}{SST} = R^2$$

[3] Remember that the normal equations of a GLM can be written as $\mathbf{X}'\mathbf{e} = \mathbf{0_k}$ and hence $\mathbf{X'}_j\mathbf{e} = 0\ \forall j = 1, \dots, k$.

3. An empirical illustration

In this Section, the proposed approach is applied to Spanish households' data. We utilize the information provided by the Household Budget Survey conducted by the Spanish National Statistics Institute which started in Jannuary 2006 and replaces the Household Continuous Survey. We use the two avalaible years which are, 2006 and 2008. This survey provides information on monthly households income and different features related to the main breadwinner of the households. To study income polarization by sub-populations we focus on the net monthly income and on the main economic and demographic characteristics of the households, which are shown in Table 2. The monthly incomes are normalized using the modified OECD equivalence scale. To make comparisons we express the equivalized net household income in constant euros at 2006 prices.

During the period considered and broadly speaking, since it is not the purpose of this paper, the Spanish real GDP increased, although at a decreasing rate. The was a rise of 3.7 % in the real GDP compared to 2006. However, the number of jobs created in 2007 decreased from the previous year, and towards the end of the year there were signs that the economy was slowing down. In 2008, the growth rate of real GDP was 1.2 % which implied a reduction of 2.5 percetage points with respect to the previous year. The effect of the crisis became more visible in 2008 when job creation declined and the unemployment rate increased (see for instance http://www.imf.org/external/data.htm).

Consistent with the previous comments, and considering that the income data refers to the year before carrying out the survey that is, 2005 and 2007, it can be observed that the mean net monthly equivalized income, as well as the median, increased (see Table 1). Inequality experienced a slight decrease according to the Gini and the Theil indices from 2006 to 2008. Table 2 includes the proposed measure of polarization for the overall households, considering that they are bunched into two groups. The dividing line between two groups is the mean

income. Households with equivalized net income below the mean belong to the group of "poor" or less favoured in terms of income and, those above belong to the group of "rich" or more favoured in term of income. The measures of Esteban Gradín and Ray (see Gradín, 2000; EGR henceforth), Zhang and Kanbur (2001) and a normalized expression of the latter (ZKN), whose formulas are referred to the Appendix 2, have been calculated to establish comparisons.

Table 1. Summary measure of montly equivalized net income in Spain

	Mean	Median	Gini index	Theil index	Sample Size
2006	983.741	850	0.301	0.151	19435
2008	1140.155	1000	0.296	0.149	22077

Source: own calculation on the Household Budget Survey

Table 2. Polarization measures

	P	ZKN	ZK	EGR
2006	0.52089257	0.63879318	1.76849702	0.21411059
2008	0.48785465	0.61551544	1.60088469	0.20941021

Source: own calculation on the Household Budget Survey

The proposed measure of polarization diminished during the considered period which implies on the one hand, that the means incomes of the two groups are closer and hence the heterogeneity among groups is smaller. On the other hand, the groups are less homogeneous and their contribution to polarization was minor. As it is observed in Table 2, the measures of ZK and EGR changed in the same direction as the proposed measure.

In this context, in which inequality as well as polarization decreased, we are going to analyze the polarization by sub-populations, with the aim of identifying the characteristics of the households that have contributed to the reduction in polarization. For this, firstly, the

households are classified into groups according to the categories of the qualitative variables considered. To do this, the economic and demographic variables are transformed into dummy variables. In a model with constant term, for each qualitative variable, the number of dummy variables is always one less than the number of categories to avoid exact multicollinearity. Therefore, a reference category, which appears in italics in Table 3 in the Appendix 1, is chosen for each variable. Secondly, an ANOVA linear intercept model equivalent to the non-intercept model given by (1) is estimated for each classification of the households. Thirdly, the coefficient of determination, which is equal to the measure of polarization proposed, of each fitted model is obtained. The estimated coefficients of each model are contained in Table 4 and the coefficients of determination are shown in Table 5.

Table 4. Estimated coefficients of the ANOVA models

Households Characteristics	2006 Coeff.	t-Statistics	2008 Coeff.	t-Statistics
Gender				
Female	-32.5773***	-3.2105	-32.3803***	-2.9689
Constant	991.33***	202.4013	1148.36***	209.2065
Education				
Middle school	182.589***	18.5575	197.208***	17.7306
High school	370.706***	31.2955	394.513***	29.3334
Tertiary	708.121***	70.1468	786.02***	67.7304
Constant	720.823***	115.4796	824.144***	105.2793
Work status				
Employee	406.025***	5.2912	443.771***	4.8302
Self-employed	289.106***	3.7378	348.29***	3.7613
Employer	634.09***	8.0661	645.006***	6.8731
Constant	598.515***	7.8159	713.463***	7.7791
Branch of Activity				
Agriculture	-168.065***	-7.3676	-192.101***	-7.5152
Manufacturing	133.508***	6.4428	163.192***	7.2445
Construction	-6.57227	-0.3010	-21.1038	-0.8793

Trade	106.192***	4.7178	99.2652***	4.0724
Hotel	44.6349*	1.6718	16.1062	0.5464
Transport	193.492***	7.8885	183.851***	6.8669
Financial	616.997***	19.4446	747.022***	21.0175
Real Estate	450.814***	21.6324	450.898***	20.0168
Constant	843.239***	44.6668	987.224***	48.4373
Town size				
More than 100,000 inhabitants	215.497***	19.6857	245.52***	20.2220
From 50,000 to 100,000	154.263***	10.3603	147.133***	8.9770
From 20,000 to 50,000	111.719***	8.1596	106.513***	6.9155
From 10,000 to 20,000	94.1498***	6.2155	86.0046***	5.0684
Constant	861.199***	103.5791	1006.69***	107.6677
Geographical area				
Northwest	18.5135	0.8119	95.8245***	3.7691
Northeast	124.853***	5.7645	295.336***	12.2944
Madrid	261.937***	10.0732	386.644***	13.5802
Centre	-77.629***	-3.5368	-5.58998	-0.2244
East	138.429***	6.4440	233.146***	9.5992
South	-95.5599***	-4.3187	-17.8057	-0.7164
Constant	939.054***	48.0978	989.411***	44.9741
***,**,* significant a1%, 5% and 10% respectively				

Source: own calculation on the Household Budget Survey

Table 5. Polarization measure-Coefficient of determination[4]

	GENDER	EDUCATION	WORK STATUS	B.ACTIVITY	TOWN SIZE	NUTS-1
2006	0.00053	0.209489	0.016199	0.115817	0.020184	0.033957
2008	0.000399	0.182176	0.008467	0.093001	0.019118	0.039843

Source: own calculation on the Household Budget Survey

[4] Although the R^2 are small, the ANOVA F-Test are significantly distinct from zero with p-values of order less than 0.005 in all cases. This is due to the sample size being extraordinarily big and the potency of the contrast is very high.

Focusing on economic household characteristics, which refers to the main breadwinner of the household, we observe that polarization is higher when the households are classified by level of education. Polarization changed from 21% in 2006 to 18 % in 2008. The estimated coefficients of the ANOVA model show a significant positive difference between the categories introduced in the model and the reference category (up to elementary school). This result confirms a positive relation between education of the main breadwinner and income of the household. If the households are clasiffied according to the branch of activity of the main breadwinner, polarization varied from 11.6 % in 2006 to 9.3% in 2008. The estimated coefficients show a significant negative difference between the agriculture branch of activity and the reference category (other activities and services). Those households whose income come from agriculture were worse in terms of income. The situation worsened from 2006 to 2008. In the other extreme are the households whose incomes came from the Financial sector. The income polarization was smaller if we consider the work status of the head of the household. It was equal to 1.6 % in 2006 and equal to 0.85 % in 2008. The estimated coefficients of the ANOVA model reveal a significant positive difference between the categories: employee, self-employed and employer and the reference category (other situation). The difference is higher if the head of the household is an employer. The worse position in terms of income corresponds to the households whose main breadwinner is self-employed. If the households are grouped by gender we notice that polarization slightly decreased from 0.05% in 2006 to 0.04% in 2008. This result shows that although there exist differences in average income of men and women, as is shown by the estimated coefficients of the model, the homogeneity within the groups is small. The small homogeneity within the group leads to a weakening of the identification among the households that belong to the same group that produces a diminishing of polarization.

With respect to the demographic characteristics, we can observe that polarization slightly decreased if the households are classified according to the town size. The proposed measure

changed from 2 % in 2006 to 1.9% in 2008. The estimated coeffients show that the bigger the town the greater is the positive difference between the categories considered in the model and the reference category (less than 10,000 inhabitants). If the households are grouped by geographical areas (NUTS 1) we observe that polarization increased from 3.3 % in 2006 to 4% in 2008. This result shows that the income differences between geographical areas have increased. The estimated coefficients of the ANOVA model show a significant positive difference between the income of the households residing in Madrid, the Northeast and the East of Spain and those households residing in the reference category (Canary Islands).

To establish comparisons, the measure of ZK as well as the measure of EGR associated with each classification have been calculated (Table 6) for 2006 and 2008.

Table 6. Measures of ZK and EGR by households characteristics.

	ZKN-06	ZKN-8	EGR-06	EGR-08
GENDER	0.00066361	0.00053651	0.0059166	0.00537147
EDUCATION	0.25695068	0.23345056	0.08095845	0.07474023
WORK STATUS	0.00013272	0.00461653	0.0240841	0.01697729
B.ACTIVITY	0.13789863	0.11709413	0.03405072	0.03133461
TOWN SIZE	0.02475274	0.02447844	0.02424103	0.02449495
NUTS-1	0.04167485	0.05103587	0.02030265	0.02283306

Source: own calculation on the Household Budget Survey

Although the EGR measure is not normalized, it has varied as the defined measure. The only exception is the town size classification for which polarization slightly increased from 2006 to 2008. The proposed measure and the measure of ZKN changed in the same direction except for the work status clasification, for which P decreased and ZKN increased. This is due to the proportion of the inter-groups inequality over the total inequality increased from 2006 to 2008 whereas the between groups variance over the total variance diminished in the same period. In other words, the between groups income concentration is higher according to the Theil index, but the groups are more homogeneous. This result highlights the difference between concentration and lack of dispersion and reinforces the argument pointed out in section two.

4.Conclusions

In this paper, taking as starting point the link between polarization and dispersion, we reformulate the measure of polarization of Zhang and Kanbur (2001) using the decomposition of the variance. The proposed measure is equivalent to the coefficient of determination of an ANOVA Linear Model, that explains the income of the households as a function of any population characteristic e.g. education, gender, occupation etc. This result provides firstly an alternative way to analyze polarization by households characteristic and secondly the developed approach allows us at the same time to compare sub-populations via the estimated coefficients of the ANOVA model.

The proposed approach is applied to Spanish households' data for the years 2006 and 2008, using the information provided by the Household Budget Survey conducted by the National Statistics Institute. The results show a diminishing of income polarization and inequality from 2006 to 2008. Focusing on economic characteristics of the households, polarization is higher when they are classified by level of education followed by the branch of activity, work status and gender of the main breadwinner. With respect to the demographic characteristics, polarization slightly decreased when the households were classified according to the town size. In contrast, polarization increased when the households are grouped by geographical areas, meaning that the disparity between regions, in terms of income, augmented from 2006 to 2008.

5.References

Esteban, J.,M., and Ray, D.: On the Measurement of Polarization. Econometrica **62**, 819-851 (1994)

Duclos, J. Y., Esteban J.M., and Ray D.: Polarization: Concepts, Measurement, Estimation. Econometrica **74**, 1337-1772 (2004)

Fisher, W., D.: On Grouping for Maximum Homogeneity. Journal of the American Statistical Association **53 (284)**, 789-798 (1958)

Gasparini L., M. Horenstein, E. Molina and S. Olivieri: Polarización Económica, Instituciones y Conflicto. Ed. Uqbar (2008)

Gradín, C.: Polarization by sub-populations in Spain, 1973-91. Review of Income and Wealth **46 (4)**, 457-474 (2000)

Gujarati, D., N.: Econometria. McGraw-Hill Interamericana (1997)

Hermoso G., A.,A. and Hernández B., A.: Curso básico de estadística descriptiva y probabilidad, Ediciones Némesis (2000)

Hussain M.A.: The sensitivity of income polarization. Journal of Economic Inequality **7 (3)**, 207-223 (2009)

Russell D. and Mackinnon J.,G.: Econometric Theory And Methods . New York Oxford, Oxford University Press (2004)

Silber, J., J. Deutsch, and M. Hanoka: On the Link between the Concepts of Kurtosis and Bi-polarization. Economics Bulletin **4 (36)**, 1-5 (2007)

Tsui, K and Wang, Y.: Ordering and New Classes of Polarisation Indices. Journal of Public Economic Theory **2 (3)**, 349-363 (2000)

Wolfson, M., C.: When Inequalities Diverge?. American Economic Review **84**, 353-58 (1994)

Zhang, X. and R. Kanbur: What Difference Do Polarization Measures Make? An Application to China. Journal of Development Studies **37**, 85-98 (2001)

Zelli R. and Pittau M.G.:Exploring patterns of income polarization using siZer. Journal of Quantitative Economics **5**, 101-111 (2007)

Appendix 1

Table 3. Household groups by economic and demographic characteristic

Gender
Female
Male
Education
Up to elementary school
Middle school
High school
Tertiary
Work status
Employee
Self-employed
Employer
Other situation
Branch of Activity
Agriculture
Manufacturing
Construction
Trade
Hotel
Transport
Financial
Real Estate
Other activities and services
Town size
More than 100,000 inhabitants
From 50,000 to 100,000
From 20,000 to 50,000
From 10,000 to 20,000
Less than 10,000
Geographical area (NUTS1)

Northwest

Northeast

Madrid

Centre

East

South

Canary Islands

Appendix 2. Measures of Esteban, Gradín and Ray and Zhang and Kanbur

Esteban, Gradin and Ray (see Gradín 2000) defined the following measure

$$EGR(\alpha,\gamma)=\sum_{i=1}^{n}\sum_{j=1}^{n}{\pi_i}^{1+\alpha}\pi_j\left|\mu_i-\mu_j\right|-\gamma[G(f)-G(\rho)]\quad 1\leq\alpha\ \leq 1.6\,;\gamma\geq 1$$

where

$$\pi_i=\int_{y_{i-1}}^{y_i}f(y)dy$$

$$\mu_i=\frac{1}{\pi_i}\int_{y_{i-1}}^{y_i}yf(y)dy$$

represent the relative frequency and the conditional mean in group i for a density f of the logarithm of income respectively. The term in brackets is the Gini index of the original distribution, $G(f)$, minus the Gini coefficient of the distribution that gives each individual in a group their representative income, $G(\rho)$, and γ is a free sensitivity parameter that measures the sensitivity within group cohesion.

Zhang and Kanbur (2001) defined the following polarization index

$$ZK=\frac{between-group\ inequality}{within-group\ inequality}$$

For the Theil index the above expression can be written as follows

$$ZK = \frac{T_B}{T_W} = \frac{\sum_{j=1}^{K} \frac{n_j}{N}\frac{\mu_j}{\mu} ln\left(\frac{\mu_j}{\mu}\right)}{\sum_{j=1}^{K} \frac{n_j}{N}\frac{\mu_j}{\mu} T_j}$$

where

$$T_j = \frac{1}{n_j}\sum_{j=1}^{K} \frac{y_j}{\mu_j} ln\left(\frac{y_j}{\mu_j}\right)$$

K is the number of groups; N is the total population; n_j is the population of the jth group; μ is the total sample mean; μ_j is the mean of the jth group and y_j is the jth income.

Observe that the expression of ZK tends to infinite when the within-group inequality is equal to zero. This drawback can be corrected normalizing the measure taking into consideration that $T = T_W + T_B$. Proceeding in this way, we obtain the normalized index of Zhang and Kanbur which is given by

$$ZKN = 1 - \frac{T_W}{T}$$

where $T = T_W + T_B$.