STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 277–296, DOI 10.21307/stattrans-2018-016

ON MEASURING POLARIZATION FOR ORDINAL DATA: AN APPROACH BASED ON THE DECOMPOSITION OF THE LETI INDEX

Mauro Mussini[1]

ABSTRACT

This paper deals with the measurement of polarization for ordinal data. Polarization in the distribution of an ordinal variable is measured by using the decomposition of the Leti heterogeneity index. The ratio of the between-group component of the index to the within-group component is used to measure the degree of polarization for an ordinal variable. This polarization measure does not require imposing cardinality on ordered categories to quantify the degree of polarization in the distribution of an ordinal variable. We address the practical issue of identifying groups by using classification trees for ordinal variables. This tree-based approach uncovers the most homogeneous groups from observed data, discovering the patterns of polarization in a data-driven way. An application to Italian survey data on self-reported health status is shown.

Key words: polarization, ordinal data, Leti index, classification trees.

JEL: D31, C40, C46

Introduction

Surveys frequently comprise one or more questions asking a respondent to self-assess his status (e.g., health, well-being, satisfaction) by choosing a response category from a set of ordered categories. When analyzing polarization in the distribution of an ordinal variable, one approach consists in imposing cardinality on ordinal categories to calculate conventional polarization measures. However, Apouey (2007) argued that transforming ordinal data into cardinal data is a supra-ordinal assumption, and she proposed bi-polarization indices which do not require supra-ordinal assumptions. Apouey's indices measure bi-polarization in the distribution (Wolfson, 1994); that is, the disappearing of the central class induced by the distribution of the observations towards the lower and upper categories rather than around the central categories. The concept of bi-polarization differs from that of polarization, since the latter is the tendency of grouping around local poles (Deutsch et al., 2013), which can be more than two and different from the extreme categories. In this paper, we use classification and regression trees (CART) (Breiman et al., 1984) for uncovering polarization

[1] Department of Economics, University of Verona. E-mail: mauro.mussini@univr.it.

patterns when dealing with ordinal data. The use of regression trees to explore polarization in income distribution has been recently investigated by Mussini (2016). Classification trees for ordinal variables (Piccarreta, 2008) are used to handle ordinal data. To quantify the polarization uncovered from ordinal data exploration, a measure based on the decomposition of the Leti heterogeneity index by group is applied. We show that this polarization measure is coherent with a criterion for identifying groups of observations in classification trees for ordinal variables.

Polarization is a relevant topic in studies on income distribution (Esteban and Ray, 1994; Duclos et al., 2004; Palacios-González and García-Fernández, 2012; Jenkins, 1995; Bossert and Schworm, 2008; Wang and Tsui, 2000; Yitzhaki, 2010; Gigliarano and Mosler, 2009; Foster and Wolfson, 2010) and its original notion is based on the concept of identification-alienation: individuals identify themselves with those having similar income levels, whereas they feel alienated from those with different income levels. When measuring polarization for an ordinal variable whose categories describe the status of an individual, there is polarization when groups of individuals characterized by within-group homogeneity (identification) and between-group heterogeneity (alienation) are observable. A similar approach was suggested by Fusco and Silber (2014), who defined the situations with the lowest and highest levels of polarization under the assumption that groups are defined a priori. According to Fusco and Silber (2014), polarization is lowest if each group shows the same relative frequency distribution of individuals between the various ordered categories; that is, if an individual cannot identify himself with the members of his group or distinguish himself from those of the other groups. Polarization is highest if all the individuals within a group belong to one category and such category varies according to the group considered; that is, if an individual can fully identify himself with the members of his group and feel alienated from those of the other groups. This approach based on within-group homogeneity and between-group heterogeneity is in line with that suggested by Zhang and Kanbur (2001) for measuring income polarization[2], however it suffers from the practical limitation that groups must be defined a priori (Duclos et al., 2004). We overcome this limitation by identifying groups through data exploration. We show that groups can naturally emerge from data by using classification trees to recursively partition individuals into groups. We assume that the ordinal variable is the response variable and some variables describing respondents (e.g., earned income, age, gender, education) are the explanatory variables. The population is recursively partitioned to maximize the between-group heterogeneity, which is equivalent to searching for the partition maximizing the gain in homogeneity within groups. A classification tree can uncover groups of homogeneous respondents in a data-driven way by selecting the explanatory variables which play a role in the polarization of the distribution of the response variable. Thus, polarization is examined on the basis not only of the response variable distribution but also of the socio-demographic characteristics of individuals, as suggested by Permanyer and D'Ambrosio (2015).

The classification tree is obtained by applying the ordinal Gini-Simpson criterion proposed by Piccarreta (2008), which is based on a measure of

[2] Given an inequality index (e.g. the Theil index), Zhang and Kanbur (2001) suggested measuring polarization by the ratio of the between-group component of the index to the within-group component.

heterogeneity for ordinal variables that can be expressed as a function of the between-group component of the Leti index of heterogeneity for ordinal variables (Leti, 1983). Grilli and Rampichini (2002) decomposed the Leti index of heterogeneity into two components: a within-group component measuring heterogeneity within groups, and a between-group component measuring heterogeneity between groups.[3] Building on the Zhang and Kanbur approach to the measurement of polarization for numerical variables, polarization in the distribution of an ordinal variable is measured by the ratio of the between-group component of the Leti index to the within-group component. Since both the recursive partition and the polarization measure depend on the between-group component of the Leti index, this link is used to define a procedure for measuring polarization which consists of two phases. First, the most homogeneous groups are identified by using classification trees for ordinal variables. Second, polarization is measured by breaking down the Leti index into between-group and within-group components.

We measure polarization in self-reported health data for a sample of Italian householders interviewed in the Survey on Household Income and Wealth in 2010 (Banca d'Italia, 2012). Our findings show that polarization is low and that the interaction effect of income and age contributes to explaining the polarization pattern.

The paper is organized as follows. Section 2 introduces the measure of polarization for ordinal variables. Section 3 outlines the procedure to recursively partition individuals into homogeneous groups. In section 4, an application to Italian household data on self-reported health status is shown. Section 5 concludes.

2. Measuring Polarization for Ordinal Variables

We briefly review the Leti heterogeneity index and its decomposition by group (subsection 2.1); we then introduce the measure of polarization based on the decomposition of the Leti index (subsection 2.2).

2.1 The Leti Index and Its Decomposition

Suppose that Y is an ordinal variable with k ordered categories $y_1, \cdots, y_j, \cdots, y_k$. Let n be the number of individuals and $n_1, \cdots, n_j, \cdots, n_k$ be the frequencies observed for the k ordered categories of Y. Let $F(y_j)$ be the cumulative relative frequency of y_j:

$$F(y_j) = \frac{\sum_{i=1}^{j} n_i}{n}. \tag{1}$$

The Leti index (Leti, 1983, pp. 290-297) is

$$L = 2\sum_{j=1}^{k-1} F(y_j)\left[1 - F(y_j)\right], \tag{2}$$

[3] Shorrocks (1980) defined a class of decomposable inequality measures for the measurement of inequality in the distribution of a numerical variable. Shorrocks (1984) also studied the properties of the inequality measures which can be decomposed by population subgroups.

and measures the degree of heterogeneity in the distribution of Y. The Leti index equals 0 if frequencies are concentrated in one category. The Leti index equals $(k-1)/2$ if heterogeneity is highest; that is, when frequencies are equally split between the lowest category y_1 and the highest category y_k. The Leti index can be normalized by dividing L by $(k-1)/2$.[4] Building on the conceptualization of maximum heterogeneity for an ordinal variable suggested by Leik (1966), Blair and Lacy (1996, 2000) developed a measure of heterogeneity for ordinal variables, which is equivalent to the normalized version of the Leti index. This index was used by Reardon (2009) to measure segregation in the case of an ordinal variable. In addition, the index is a member of a class of inequality measures for ordinal data that was axiomatically derived by Lv et al. (2015).

Grilli and Rampichini (2002) showed that the Leti index is decomposable by groups. Suppose the n individuals are split into h groups. Let $n_{j,g}$ be the frequency observed for category y_j within group g (with $g = 1, \cdots, h$) and n_g be the size of group g. Let $F(y_j|g)$ be the cumulative relative frequency of y_j within group g:

$$F(y_j|g) = \frac{\sum_{i=1}^{j} n_{i,g}}{n_g}. \tag{3}$$

The heterogeneity within group g can be measured by using the Leti index:

$$L_g = 2\sum_{j=1}^{k-1} F(y_j|g)[1 - F(y_j|g)]. \tag{4}$$

$p_g = n_g/n$ being the population share of group g, the within-group component of the Leti index is

$$L^W = \sum_{g=1}^{h} p_g L_g. \tag{5}$$

The between-group component of the Leti index is

$$L^B = 2\sum_{g=1}^{h} p_g \sum_{j=1}^{k-1} F(y_j|g)[F(y_j|g) - F(y_j)]. \tag{6}$$

L^B in eq. (6) measures the heterogeneity between the cumulative relative frequency distribution in the population and the cumulative relative frequency distributions in the various groups.

Since $F(y_j) = \sum_{g=1}^{h} p_g F(y_j|g)$, L^B can be rewritten as

$$L^B = 2\sum_{g=1}^{h}\sum_{j=1}^{k-1} p_g F(y_j|g)[\sum_{i\neq g} p_i\, F(y_j|g) - \sum_{i\neq g} p_i F(y_j|i)]. \tag{7}$$

Hence, after simple manipulations, an alternative expression for L^B is obtained:

$$L^B = 2\sum_{g=1}^{h}\sum_{j=1}^{k-1} p_g F(y_j|g)\left\{\sum_{i\neq g} p_i\,[F(y_j|g) - F(y_j|i)]\right\}$$

[4] When n is odd, the maximum value of the Leti index is $\frac{k-1}{2}\left(1-\frac{1}{n^2}\right)$ instead of $\frac{k-1}{2}$. However, this difference is negligible when n is sufficiently large.

$$L^B = 2\sum_{g=1}^{h}\sum_{j=1}^{k-1}\left\{\sum_{i\neq g} p_i\, p_g F(y_j|g)[F(y_j|g) - F(y_j|i)]\right\}$$

$$L^B = 2\sum_{g=1}^{h}\sum_{i\neq g} p_g p_i \sum_{j=1}^{k-1} F(y_j|g)[F(y_j|g) - F(y_j|i)]$$

$$L^B = 2\sum_{g=1}^{h}\sum_{i=g+1}^{h} p_g p_i \sum_{j=1}^{k-1}[F(y_j|g) - F(y_j|i)]^2$$

$$L^B = 2\textstyle\sum_{g=1}^{h}\sum_{i=g+1}^{h} p_g p_i D_{gi}. \qquad (8)$$

In eq. (8), D_{gi} measures the heterogeneity between the cumulative relative frequency distributions of groups g and i. If the two groups have the same cumulative relative frequency distribution, then $D_{gi} = 0$. L^B in eq. (8) is expressed as a function of the pairwise differences between the within-group cumulative relative frequency distributions. In this respect, there is a similarity between L^B and an index of inequality in life chances suggested by Silber and Yalonetzky (2011).[5] When all groups have the same cumulative relative frequency distribution, D_{gi} is 0 for every $g, i = 1, \cdots, h$ (with $g \neq i$) and L^B equals 0 since there is no heterogeneity between the cumulative relative frequency distributions of different groups. L^B coincides with L if the frequencies are concentrated in one category within every group; that is, when heterogeneity is fully explained by the between-group heterogeneity.

Originally, Grilli and Rampichini (2002) interpreted the ratio of L^B to L as the share of heterogeneity explained by a generic variable X used to form groups (Grilli and Rampichini, 2002, pp. 114). In the next section, we show that the ratio of the between-group component to the within-group component can be seen as a measure of polarization for ordinal variables.

2.2 A Measure of Polarization for Ordinal Variables

Polarization is the tendency of individuals to concentrate around local poles, forming groups of reasonable size in which every individual can identify himself with the members of his group and feel alienated from those of the other groups (Esteban and Ray, 1994; Duclos et al., 2004). The concept of polarization has been applied to studies on income distribution in which the original notion of identification-alienation has been adapted to the topic: individuals identify themselves with those having similar income levels, whereas they feel alienated from those with different income levels. This idea of polarization can be extended to the distribution of an ordinal variable by observing that there is polarization if groups have different relative frequency distributions and the relative frequency distribution within each group tends to converge towards a single category; that

[5] Silber and Yalonetzky (2011) proposed a set of new indices for measuring inequality in life chances in the case of an ordinal variable. One of these indices is based on pairwise comparisons between the within-group cumulative relative frequency distributions of the ordinal variable.

is, polarization occurs if groups are characterized by within-group homogeneity (identification) and between-group heterogeneity (alienation). For example, Figure 1 shows the relative frequency distribution of an ordinal variable with five response categories (ranging from “Very Poor” to “Excellent”). If we suppose that the individuals belonging to the same response category form a group of respondents with the same characteristics, we can say that the population is “polarized” in line with the Esteban and Ray general idea of polarization (Esteban and Ray, 1994). Fusco and Silber (2014) defined the situations with the lowest and highest levels of polarization for an ordinal variable, under the assumption that groups are pre-established. Polarization is lowest if each group has the same relative distribution of individuals between the various ordered categories; that is, if an individual cannot identify himself with the members of his group and distinguish himself from those of the other groups. Polarization is highest if the individuals within a group belong to a single category, and this category varies according to the group considered; that is, if an individual can fully identify himself with the members of his group and feel alienated from those of the other groups.

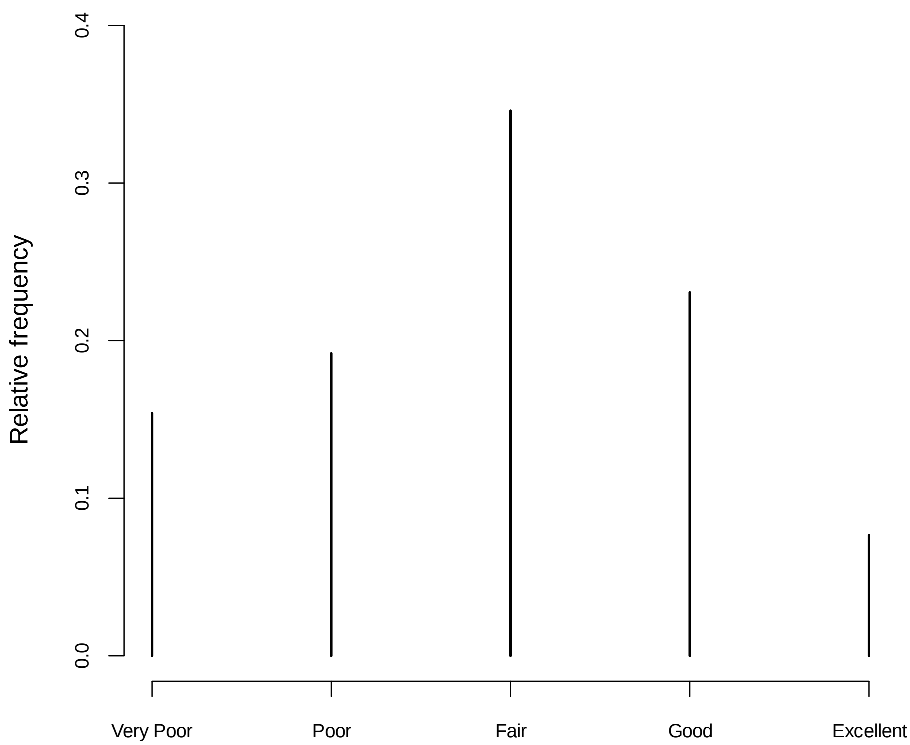

Figure 1. Relative frequency distribution of an ordinal variable

In this framework, we establish a link between the measurement of polarization and the decomposition of the Leti index. Since the between-group component measures between-group heterogeneity and the within-group component measures within-group heterogeneity, we note that polarization increases as the share of the Leti index attributable to the between-group component increases. The higher the between-group heterogeneity, the lower the within-group heterogeneity. In line with the Zhang and Kanbur approach (2001), the ratio of the between-group component to the within-group component can be interpreted as a measure of polarization:

$$PO = \frac{L^B}{L^W} = \frac{2\sum_{g=1}^{h}\sum_{i=g+1}^{h} p_g p_i D_{gi}}{\sum_{g=1}^{h} p_g L_g}. \tag{9}$$

PO equals 0 if the cumulative relative frequency distribution within each group is the same; that is, the cumulative relative frequency distribution within each group is equal to that of the whole population. In this case, polarization is lowest since there is no between-group heterogeneity. PO increases as the share of overall heterogeneity due to the between-group heterogeneity increases. While the index equals 0 in the case of minimum polarization, there is no upper limit for the index. In this respect, PO differs from conventional inequality indices, which usually range from 0 (perfect equality) to 1 (maximum inequality). The polarization index satisfies the principle of population size invariance, which is a desirable property for inequality indices. This property states that the value of the index does not change if every individual is replicated m times.[6]

The formulation of PO takes the between-group heterogeneity, within-group homogeneity and group population shares into account; that is, the three main features of polarization (Esteban and Ray, 1994, p. 824) are included in the polarization measure. While the role of between-group heterogeneity is clear, those of the other two features deserve some additional explanations. The role of within-group homogeneity is considered by the within-group heterogeneity component in the denominator of the ratio in eq. (9). The higher the within-group homogeneity, the lower the denominator. Therefore, a gain in within-group homogeneity increases polarization, all other things being equal. From eq. (9), we see that smaller groups carry less weight in the measurement of polarization than larger groups. In addition, considering groups g and i and holding the sum of their population shares constant, the more similar their population shares, the greater the weight assigned to the heterogeneity between their cumulative relative frequency distributions. In eq. (9), D_{gi} is weighted by the product $p_g p_i$, which increases as p_g and p_i become closer, holding the sum of the population shares of the two groups constant.

[6] Silber and Yalonetzky (2011) introduced an alternative property linked to population replication for indices measuring inequality in the case of ordinal data, named population composition invariance. The property of population composition invariance states that the value of the index is unchanged if every individual within a group g is replicated m times. This property is not satisfied by PO since the population share of group g and those of other groups would change if the population of group g were replicated a certain number of times.

To apply the Leti-based measure of polarization, the partition of individuals into groups is required. However, assuming that groups are pre-established does not necessarily reflect the actual polarization in the distribution of an ordinal variable. Moreover, the choice of the criterion to form groups is a practical issue to be addressed (Duclos et al., 2004). We overcome these issues by letting homogeneous groups be formed in a data driven way. To uncover the most homogeneous groups, we use classification trees for ordinal variables (Piccarreta, 2008), in which the recursive partition relies on a heterogeneity measure that can be expressed as a function of the between-group component of the Leti index. In the next section, we show that classification trees are useful to detect the most homogenous groups, since each group is composed of individuals who have the same characteristics (e.g. age, gender, occupational attainment, education) and are similar in terms of ordinal response categories. In fact, the classification tree procedure includes some individuals in the same group if they are similar in terms of a set of variables and the variable values characterizing that group differ from those characterizing the other groups, in line with the original idea of polarization proposed by Esteban and Ray (1994).

3. Using Classification Trees for Detecting Homogenous Groups

Classification and regression trees (Breiman et al., 1984) are nonparametric methods for exploring data or predicting new observations. If the response variable is categorical (numerical), a classification (regression) tree is produced. In a classification tree, the variation of a response categorical variable is explained by a set of explanatory variables. The classification tree is produced by recursively partitioning individuals into more homogeneous groups, each of which is characterized by both the within-group distribution of the response variable and the values of explanatory variables describing the members of the group. When dealing with an ordinal response variable, the conventional criteria for partitioning individuals into groups may not lead to the best partition (Piccarreta, 2008). Piccarreta (2008) extended the classification tree method and introduced splitting criteria to deal with an ordinal response variable. Here, we use ordinal classification trees as an explorative statistical tool for uncovering the relationships between an ordinal response variable and a set of individual's characteristics.

3.1 Classification Trees for Ordinal Variables

Let $(Y, \boldsymbol{X}): \Omega \rightarrow \left(S_Y \times S_{X_1} \times \cdots \times S_{X_p}\right) \equiv S$ be a vector random variable on the probability space (Ω, F, P), where Y is an ordinal variable and $\boldsymbol{X} = \{X_1, \cdots, X_m, \cdots, X_p\}$ are p explanatory variables. Assume that Y is the response variable, with k ordered categories $\left(y_1, \cdots, y_j, \cdots, y_k\right)$, and $\boldsymbol{X}$ is the vector collecting p individual's characteristics. The classification tree is built by recursively partitioning the space S into disjoint subsets, such that each subset includes individuals who are as homogeneous as possible in terms of Y. Initially, all individuals are included in one set, called the root node, and then are split into subsets, called nodes. The degree of heterogeneity of the response variable within a node is measured by defining an impurity measure. In the case of an ordinal response variable, impurity can be measured by using the Gini index of heterogeneity of an ordinal variable (Gini, 1954):

$$I_t(Y) = \sum_{j=1}^{k} F(y_j|t)[1 - F(y_j|t)], \tag{10}$$

where t is a generic node, which coincides with the root node at the beginning of the recursive partitioning procedure. The partitioning procedure starts by splitting a parent node (the root node) into two descendent nodes according to a cut-off value chosen among all the observed values of the explanatory variables $\boldsymbol{X}$. Such a cut-off value is selected to maximize the decrease in the impurity measure in eq. (10). In the next step, each descendent node is split into two further subsets according to the partition maximizing the decrease in impurity. In each step of the splitting procedure, the decrease in impurity is measured by subtracting the impurity within the descendent nodes from the impurity of the parent node. To explain the criterion for partitioning a parent node into two descendent nodes, consider a generic node t with n_t individuals. Without loss of generality, we may assume that X_m is a numerical explanatory variable. Let $c \in S_{X_m}\big|t$ stand for a value of X_m, with the domain of X_m restricted to node t. Let t_l and t_r be the descendent nodes obtained by splitting t at the cut-off c. Let $n_{t_l} = \sum_{i=1}^{n_t} I_{\{X_{m,i} \le c\}}$ and $n_{t_r} = \sum_{i=1}^{n_t} I_{\{X_{m,i} > c\}}$ be the numbers of individuals in nodes t_l and t_r, respectively. The decrease in impurity obtained by splitting t into two nodes, t_l and t_r, at c is

$$\Delta_t(Y, c) = I_t(Y) - \frac{n_{t_l}}{n_t} I_{t_l}(Y) - \frac{n_{t_r}}{n_t} I_{t_r}(Y), \tag{11}$$

where $I_{t_l}(Y) = \sum_{j=1}^{k} F(y_j|t_l)[1 - F(y_j|t_l)]$ and $I_{t_r} = \sum_{j=1}^{k} F(y_j|t_r)[1 - F(y_j|t_r)]$ are the impurity measures calculated for nodes t_l and t_r, respectively. After simple manipulations, eq. (11) can be rewritten as

$$\Delta_t(Y, c) = \frac{n_{t_l} n_{t_r}}{n_t^2} \sum_{j=1}^{k} \left[F(y_j|t_l) - F(y_j|t_r)\right]^2. \tag{12}$$

Piccarreta (2008) suggested the use of the expression in eq. (12) for measuring the decrease in impurity due to splitting t into t_l and t_r, with the exclusion of the comparison between $F(y_k|t_l)$ and $F(y_k|t_r)$:

$$\Delta_t^*(Y, c) = \frac{n_{t_l} n_{t_r}}{n_t^2} \sum_{j=1}^{k-1} \left[F(y_j|t_l) - F(y_j|t_r)\right]^2. \tag{13}$$

For node t, the splitting variable and the variable threshold c are selected from all the observed values of the explanatory variables to maximize the impurity reduction in eq. (13). This splitting procedure recursively runs until a stopping rule establishes that no further partition is useful since it does not produce any important gain in terms of within-group homogeneity and between-group heterogeneity. At the end of the procedure, the individuals in a subset (terminal node) constitute a group characterized by the distribution of Y within the group and the combination of the values of the explanatory variables which identifies that group.

3.2 Linking the Decomposition of the Leti Index with the Splitting Criteria for a Classification Tree

We show that maximizing $\Delta_t^*(Y,c)$ is equivalent to searching for the breakdown maximizing the between-group component of the Leti index calculated for node t. The Leti heterogeneity index for t is

$$L_t = 2\sum_{j=1}^{k-1} F(y_j|t)[1-F(y_j|t)]. \tag{14}$$

Supposing that t is split into t_l and t_r, the decomposition of L_t is

$$L_t = L_t^W + L_t^B, \tag{15}$$

where the within-group component is

$$L_t^W = \frac{n_{t_l}}{n_t} 2\sum_{j=1}^{k-1} F(y_j|t_l)[1-F(y_j|t_l)] + \frac{n_{t_r}}{n_t} 2\sum_{j=1}^{k-1} F(y_j|t_r)[1-F(y_j|t_r)] = p_{t_l} L_{t_l} + p_{t_r} L_{t_r} \tag{16}$$

and the between-group component is

$$L_t^B = 2\left\{\frac{n_{t_l}}{n_t}\sum_{j=1}^{k-1} F(y_j|t_l)[F(y_j|t_l) - F(y_j|t)] + \frac{n_{t_r}}{n_t}\sum_{j=1}^{k-1} F(y_j|t_r)[F(y_j|t_r) - F(y_j|t)]\right\}$$

$$L_t^B = 2p_{t_l}p_{t_r}\sum_{j=1}^{k-1}[F(y_j|t_l) - F(y_j|t_r)]^2. \tag{17}$$

Irrespective of the multiplicative factor 2 in eq. (17), the comparison of eq. (17) and (13) leads to the conclusion that the decrease in heterogeneity produced by splitting node t is measured by the between-group component of the Leti index calculated for that subset. The partitioning procedure iteratively searches for the breakdown maximizing the between-group component of the Leti index.

The splitting procedure can be repeated until the terminal nodes are very small, resulting in an overlarge tree that could be difficult to interpret. Therefore, a stopping rule is needed to select the optimal tree size. A tree pruning procedure (Breiman et al., 1984) is used to find the best tree. Pruning can be performed by minimizing the following cost-complexity function for a tree T:

$$R_\alpha(T) = R(T) + \alpha \cdot |T|. \tag{18}$$

In eq. (18), $|T|$ is the tree size (i.e. the number of terminal nodes), α is a complexity parameter ranging within the interval $(0,\infty)$, and $R(T)$ is the resubstitution error. The functional form of $R(T)$ depends on the nature of the response variable Y. If Y is ordinal, $R(T)$ may coincide with either the total misclassification rate or the total misclassification cost. A misclassification occurs when the true response category of an individual is different from that assigned to him by the tree. Following Galimberti et al. (2012), the response category assigned to an individual is equal to the median category of the terminal node in which the individual is included. The total misclassification rate is equal to ratio of the number of misclassified individuals to the total number of individuals. The total misclassification rate is commonly used when dealing with a nominal variable. Piccarreta (2008) suggested assigning a cost to each misclassification given that the response variable is ordinal instead of nominal. The misclassification cost is set equal to the number of categories separating the true response category of an individual from the response category assigned to him by the tree: for example, if the two categories are adjacent, the misclassification cost equals 1; if the true response category of an individual is y_j and the response category assigned to him is y_{j-2}, then the misclassification cost equals 2. The total misclassification cost is equal to the sum of misclassification costs.

As shown in Breiman et al. (1984), for any α there is a unique smallest tree minimizing eq. (18), therefore, finding the best tree reduces to selecting the optimal tree size. Since $R(T)$ in eq. (18) is always minimized by the largest tree, Breiman et al. (1984) suggested using V-fold cross-validation to improve the reliability of misclassification error estimates. V-fold cross-validation is performed in various steps: (i) individuals are divided into V (usually V is set equal to 10) subsets of approximately equal size; (ii) each subset in turn is left out, a tree of size $|T|$ is built by using the remaining subsets and this tree is used to predict the response categories for the members of the omitted subset; (iii) the misclassification costs are calculated for each omitted subset; (iv) the misclassification costs calculated for the V subsets are added up and the cross-validated total misclassification cost is obtained, $R^{CV}(T)$; (v) steps (i)-(iv) are repeated for every tree size. Then, $R(T)$ is replaced with $R^{CV}(T)$ in eq. (18) to select the optimal tree size.

After pruning the classification tree, the terminal nodes identify groups characterized by within-group homogeneity and between-group heterogeneity in terms of a set of variables comprising the response ordinal variable and the explanatory variables used to produce the tree. Different from the Silber and Fusco (2014) approach, groups are directly identified through data exploration by clustering individuals who are similar. Therefore, using classification trees, polarization patterns can be naturally uncovered in a data driven way. A further advantage of the tree-based approach to the identification of groups is the selection of the most important explanatory variables in determining between-group heterogeneity, since only the explanatory variables producing an appreciable decrease in impurity are shown in the classification tree.

4. Application to Data on Self-Reported Health Status

We measure the polarization in the distribution of data on self-reported health status (hereafter, SRHS) collected by the Survey on Household Income and Wealth (henceforth, SHIW) carried out by the Bank of Italy in 2010 (Banca d’Italia, 2012). SRHS data include respondents’ perceptions of their general health condition, with the response categories ranging from “Very Poor” to “Excellent”. The use of SRHS is very common in epidemiological surveys since it is a good predictor of mortality (Allison and Foster, 2004); moreover, socio-economic surveys frequently ask SRHS to investigate the relationship between health status and socio-economic status (Kakwani et al., 1997; Idler and Benyamini, 1997). In our analysis, polarization in SRHS is measured by exploring the relationship between SRHS and a set of explanatory socio-economic variables collected in the 2010 SHIW. First, we run the classification tree procedure to partition respondents into homogeneous groups. Second, we measure polarization in the SRHS distribution by using PO.

The 2010 SHIW collected information on income, wealth and socio-economic variables for a sample of 7,951 households. In addition, the survey asked each householder to assess his health status and that of each household member. We focus our attention on the householder SRHS and 7,950 householders are considered[7]. Table 1 shows the description and coding for the ordinal response variable and explanatory variables. SRHS is measured with an ordinal variable having five response categories: “Very Poor”, “Poor”, “Fair”, “Good”, “Excellent”.

Table 1. Variable description and coding.

Response variable			
name	description	type	ordered categories
SRHS	self-reported health status	ordinal	"Very Poor", "Poor", "Fair", "Good", "Excellent";
Explanatory variables			
name	description	type	categories coding (for categorical variables) or range (for numerical variables)
AGE_CLASS	age class	ordinal	up to 34 years, 35-44, 45-54, 55-64, more than 64 years
AREA	geographical area of residence	nominal	N="North", C="Centre", S="South and Islands"
INCOME	household income	numerical	$(0,\infty)$
EMPLOYMENT	employment status	nominal	(BC="blue-collar worker", OW="office worker or school teacher", M="cadre or manager", P="sole proprietor/member of the arts or professions", SE="other self-employed", R="retired", NE="other not-employed")
EDUCATION	educational qualification	ordinal	N="none", P="primary school certificate", LS="lower secondary school certificate", VS="vocational secondary school diploma", US="upper secondary school diploma", B="3-year university degree", G="5-year university degree", PG="postgraduate qualification"
ACTIVITY	sector of activity	nominal	A="agriculture, fishing", I="industry", G="general government", O="other", NA="do not know"
GENDER	gender	dichotomous	F="Female"
SIZE_TOWN	size of the town of residence	ordinal	ST="0-20,000 inhabitants", MT="20,000-40,000", LT="40,000-500,000", C="more than 500,000 inhabitants"

Figure 2 shows the relative frequency distribution of SRHS data. We observe that the median category is “Good” and that the relative frequencies in the upper categories (“Good” and “Excellent”) are greater than those in the others. We initially run the recursive partitioning procedure by setting a small value of the

[7] SRHS is not available for one of the surveyed householders; therefore, he is excluded from the empirical analysis. In all calculations, we use the sample weights provided by the SHIW.

complexity parameter (CP=0.01) to produce a large tree.[8] An overlarge tree avoids that the interaction effects between explanatory variables are not discovered because none of the associated main effects produces a split with an appreciable decrease in terms of misclassification costs.[9]

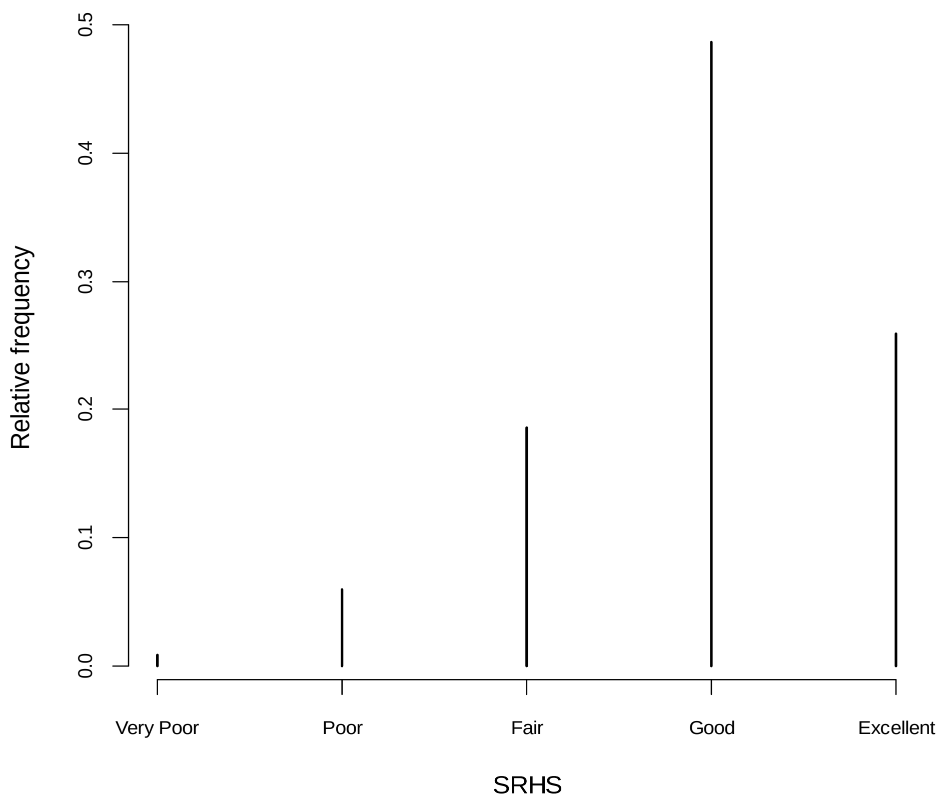

Figure 2. Relative frequency distribution of SRHS

8 We use the R package rpartScore (Galimberti et al., 2012) for recursive partitioning and we set the complexity parameter equal to the default value CP=0.01. The CP value in rpartScore is directly linked to α in eq. (18), since CP is equal to the ratio of α to the total misclassification cost calculated for the tree with no splits (i.e. the tree having no subsets). Therefore, α can be determined by setting CP.

9 Setting a large CP value serves the scope of excluding a split if it does not produce an appreciable reduction in total misclassification cost. However, if that split is made, one of the descendent subsets may be split in a way to produce an appreciable decrease in total misclassification cost. This can occur when a split based on the interaction between variables produces an appreciable decrease in total misclassification cost but none of the associated variable main effects produces an appreciable misclassification cost reduction.

Table 2 shows the tree size $|T|$ (column 1), the minimum CP value for a tree of size $|T|$ (column 2), the total misclassification cost (column 3), the 10-fold cross-validated total misclassification cost (column 4), and the standard error of the 10-fold cross-validated total misclassification cost (column 5).

Table 2. Tree size, total misclassification cost and 10-fold cross-validated total misclassification cost.

$\|T\|$	CP	$R(T)$	$R^{CV}(T)$	SE
1	0.0491	1.0000	1.0000	0.0172
3	0.0170	0.9018	0.9047	0.0174
6	0.0100	0.8507	0.8566	0.0220

Table 2 shows that the tree is not particularly successful in classifying individuals since the 10-fold cross-validated total misclassification cost is 0.8566. However, our aim is not finding a tree performing a good classification but exploring whether there are homogenous groups emerging from the data. From this standpoint, we need to handle the trade-off between the gain in within-group homogeneity and the tree size increase. We observe that passing from three to six terminal nodes does not imply a remarkable reduction of misclassification cost; that is, increasing the number of groups from three to six produces a small gain in terms of within-group homogeneity. Hence, we prune the tree by setting a complexity parameter greater than 0.01 to reduce the tree size. Figure 3 shows that the pruned tree has three terminal nodes in which the householders are split (groups 2, 6 and 7 in Figure 3). Figure 3 shows the size and the median category for each group. AGE_CLASS and INCOME are the explanatory variables playing a role in the partition of householders into groups. As expected, age has an effect on SRHS. SRHS of householders aged 65 years or older (group 3) is lower than SRHS of those younger than 65 years (group 2). Among householders aged 65 years or older (group 3), SRHS is better for householders with household income higher than 20,960.5 euros (group 7).

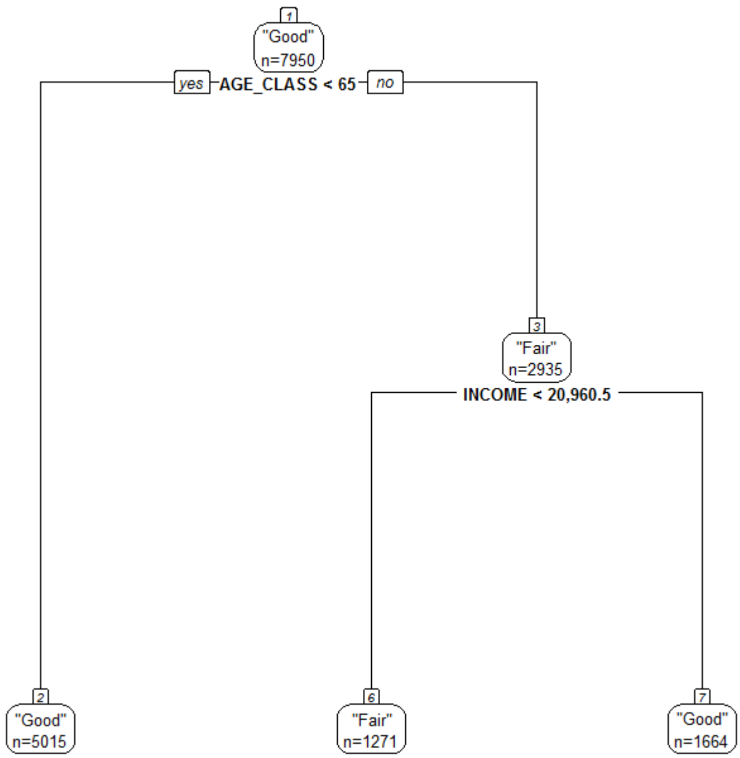

Figure 3. Classification tree for SRHS

Figure 4 shows the relative frequency distribution within each of the three groups. Although the median category of groups 2 and 7 is the same, the relative frequencies are concentrated in the upper two categories within group 2 whereas the relative frequencies spread towards the middle category within group 7. We observe that the SRHS distribution within group 2 is quite different from that within group 6, however group 6 is not very homogeneous in terms of SRHS. The normalized Leti index of the overall SRHS distribution equals 0.4542, indicating an intermediate level of heterogeneity. We break down the Leti index by group and we find that the within-group component is 0.3880 while the between-group component is 0.0662. The polarization measure PO is equal to 0.1706 and indicates that polarization is low. This means that the groups are not particularly characterized by within-group homogeneity and between-group heterogeneity with respect to SRHS and the socio-economic variables considered.

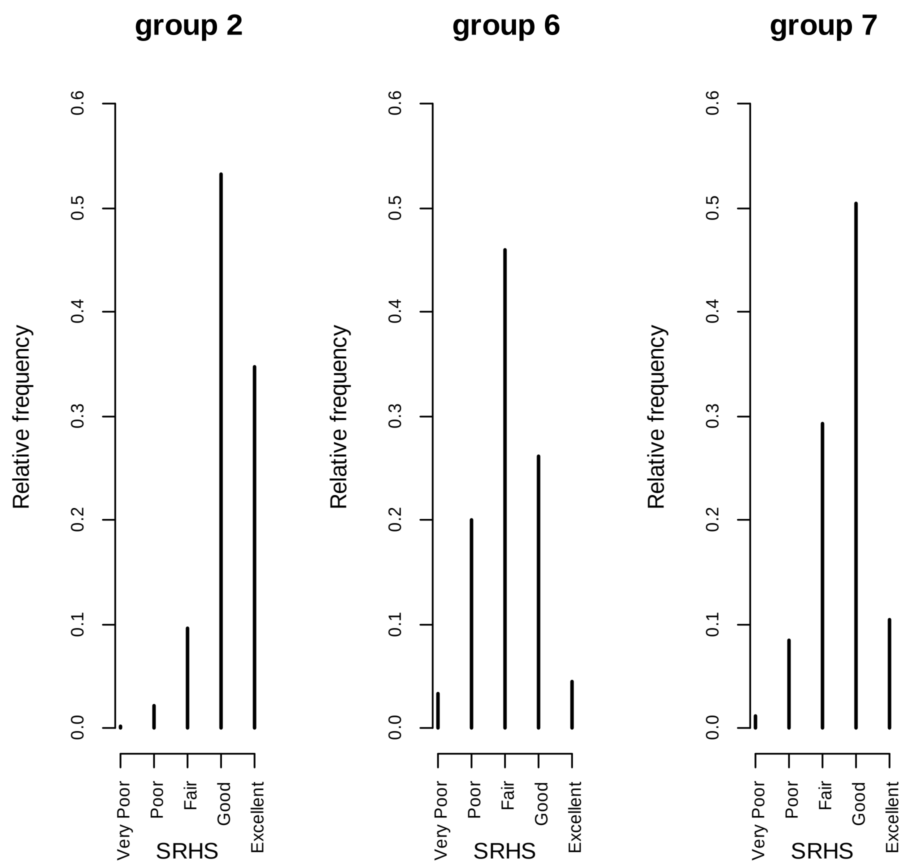

Figure 4. Relative frequency distributions of SRHS by group

5. Conclusion

This article deals with the measurement of polarization for ordinal variables. The contribution of the article is two-fold. First, we propose a synthetic measure of polarization based on the decomposition of the Leti heterogeneity index by group. Given a set of individuals split into groups by a certain criterion, the ratio of the between-group component of the Leti index to the within-group component indicates the extent to which the distribution of the ordinal variable is homogeneous within each group and heterogeneous between groups. If the within-group distributions are equal, the members of a group cannot distinguish themselves from those of the other groups. In this case, the measure of polarization equals 0, indicating that polarization is minimum. If the ordinal variable distribution within each group is mainly concentrated in a single category and this category varies according to the group considered, the within-group homogeneity is high. In this case, each member of a group can identify himself with the members of his group and feel alienated from those belonging to the other groups. The greater the within-group homogeneity, the greater the measure of polarization. An advantage of this polarization measure is that it does not require imposing cardinality on the ordered categories of an ordinal variable. Indeed, imposing cardinality is a supra-ordinal assumption altering the original variable type.

The second contribution of the article is the use of a tree-based approach to partition individuals into homogeneous groups when exploring polarization in the distribution of an ordinal variable. As noted by Duclos et al. (2004), a practical issue in polarization studies is finding groups characterized by within-group homogeneity and between-group heterogeneity in terms of a set of variables. We show that the between-group component of the Leti index is equivalent to the impurity measure used in the process generating a classification tree for an ordinal response variable. Using classification trees, we can uncover whether individuals are naturally split into homogeneous groups, each of which comprises individuals who are similar in terms of the ordinal response variable and a set of explanatory variables. In addition, this approach is useful for selecting the explanatory variables which play a role in the polarization of the ordinal variable. Since the recursive partitioning procedure also explores the interaction effects between the explanatory variables, analysts can discover polarization patterns which cannot be assumed a priori.

We measure the polarization of SRHS data for a sample of Italian householders interviewed in the 2010 SHIW. The polarization measure is equal to 0.1706, indicating that polarization is low. The classification tree for SRHS shows that the age and household income of respondents are the most important variables in the partition of householders in terms of SRHS. All other explanatory variables, like employment status, educational qualification or gender, do not play an important role in the polarization of SRHS.

Acknowledgements

The author thanks two anonymous reviewers for their valuable comments.

REFERENCES

ALLISON, R. A., FOSTER J., (2004). Measuring health inequality using qualitative data, Journal of Health Economics 23, pp. 505–524.

APOUEY, B., (2007). Measuring health polarization with self-assessed health data, Health Economics 16, pp. 875–894.

BANCA D'ITALIA, Survey on Household Income and Wealth 2010, Rome (2012). http://www.bancaditalia.it/statistiche/indcamp/bilfait/boll_stat;internal&action=_setlanguage.action?LANGUAGE=en.

BLAIR, J., LACY, M. G., (1996). Measures of Variation for Ordinal Data as Functions of the Cumulative Distribution, Perceptual and Motor Skills 82, pp. 411–418.

BLAIR, J., LACY, M. G., (2000). Statistics of Ordinal Variation, Sociological Methods & Research 28, pp. 251–280.

BOSSERT, W., SCHWORM, W., (2008). A Class of Two-Group Polarization Measures, Journal of Public Economic Theory 10, pp. 1169–1187.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. , STONE, C. J., (1984). Classification and regression trees. Chapman & Hall/CRC press, Boca Raton.

DEUTSCH, J., FUSCO, A., SILBER, J., (2013). The BIP trilogy (Bipolarization, Inequality and Polarization): one saga but three different stories. Economics: The Open-Access, Open-Assessment E-Journal 7, pp. 2013–2022. http://www.economics-ejournal.org/economics/journalarticles/2013-22.

DUCLOS, J. Y., ESTEBAN, J. M., RAY, D., (2004). Polarization: Concepts, measurement, estimation, Econometrica 72, pp. 1737–1772.

ESTEBAN, J. M., RAY, D., (1994). On the measurement of polarization, Econometrica 62, pp. 819–851.

FOSTER, J. E., WOLFSON, M. C., (2010). Polarization and the decline of the middle class: Canada and the U.S., Journal of Economic Inequality 8, pp. 247–273.

FUSCO, A., SILBER, J., (2014). On social polarization and ordinal variables: the case of self-assessed health, European Journal of Health Economics 15, pp. 841–851.

GALIMBERTI, G., SOFFRITTI, G., DI MASO, M., (2012). Classification Trees for Ordinal Responses in R: The rpartScore Package, Journal of Statistical Software 47, pp. 1–25.

GIGLIARANO, C., MOSLER, K., (2009). Constructing indices of multivariate polarization, Journal of Economic Inequality 7, pp. 435–460.

GINI, C., (1954). Variabilità e concentrazione, Veschi, Rome.

GRILLI, L., RAMPICHINI, C., (2002). Scomposizione della dispersione per variabili statistiche ordinali, Statistica 62, pp. 111–116.

IDLER, E., BENYAMINI, Y., (1997). Self-rated health and mortality: a review of twenty-seven community studies, Journal of Health and Social Behaviour 38, pp. 21–37.

KAKWANI, N., WAGSTAFF, A., VAN DOORSLAER, E., (1997). Socioeconomic inequalities in health: Measurement, computation, and statistical inference, Journal of Econometrics 77, pp. 87–103.

JENKINS, S. P., (1995). Did the Middle Class Shrink during the 1980s? UK Evidence from Kernel Density Estimates, Economics Letters 49, pp. 407–413.

LEIK, R. K., (1966). A Measure of Ordinal Consensus, The Pacific Sociological Review 9, pp. 85–90.

LETI, G., (1983). Statistica descrittiva, il Mulino, Bologna.

LV, G., WANG, Y., XU, Y., (2015). On a new class of measures for health inequality based on ordinal data, Journal of Economic Inequality 13, pp. 465–477.

MUSSINI, M., (2016). On measuring income polarization: an approach based on regression trees, Statistics in Transition – new series 17, pp. 221–236.

PALACIOS-GONZÁLEZ, F., GARCÍA-FERNÁNDEZ, R. M., (2012). Interpretation of the coefficient of determination of an ANOVA model as a measure of polarization, Journal of Applied Statistics 39, pp. 1543–1555.

PERMANYER, I., D'AMBROSIO, C., (2015). Measuring Social Polarization with Ordinal and Categorical Data, Journal of Public Economic Theory 17, pp. 311–327.

PICCARRETA, R., (2008). Classification trees for ordinal variables, Computational Statistics 23, pp. 407–427.

REARDON, S. F., (2009). Measures of ordinal segregation, in Y. Flückiger, S. F., Reardon and J., Silber (eds.), Occupational and Residential Segregation, Research on Economic Inequality 17, pp. 129–155, Emerald, Bingley.

SHORROCKS, A. F., (1980). The Class of Additively Decomposable Inequality Measures, Econometrica 48, pp. 613–625.

SHORROCKS, A. F., (1984). Inequality Decomposition by Population Subgroups, Econometrica 52, pp. 1369–1385.

SILBER, J., YALONETZKY, G., (2011). On Measuring Inequality in Life Chances when a Variable is Ordinal, in Juan Gabriel Rodríguez (ed.) Inequality of Opportunity: Theory and Measurement, Research on Economic Inequality 19, pp.77–98, Emerald, Bingley.

WANG, Y. Q., TSUI, K. Y., (2000). Polarization orderings and new classes of polarization indices, Journal of Public Economic Theory 2, pp. 349–363.

WOLFSON, M. C., (1994). When Inequalities Diverge?, American Economic Review 84, pp. 353–358.

YITZHAKI, S., (2010). Is there room for polarization?, The Review of Income and Wealth 56, pp. 7–22.

ZHANG, X., KANBUR, R., (2001). What difference do polarization measures make? An application to China, Journal of Development Studies 37, pp. 85–98.