STATISTICS IN TRANSITION new series, June 2016
Vol. 17, No. 2, pp. 221–236

ON MEASURING INCOME POLARIZATION: AN APPROACH BASED ON REGRESSION TREES

Mauro Mussini[1]

ABSTRACT

This article proposes the application of regression trees for analysing income polarization. Using an approach to polarization based on the analysis of variance, we show that regression trees can uncover groups of homogeneous income receivers in a data-driven way. The regression tree can deal with nonlinear relationships between income and the characteristics of income receivers, and it can detect which characteristics and their interactions actually play a role in explaining income polarization. For these features, the regression tree is a flexible statistical tool to explore whether income receivers concentrate around local poles. An application to Italian individual income data shows an interesting partition of income receivers.

Key words: polarization, regression trees, recursive partitioning, ANOVA, JEL D31, D63, C14.

1. Introduction

The measurement of income polarization has developed by following two distinct approaches. One approach focuses on the concept of bipolarization that considers the extent to which incomes spread from the middle to the tails of the distribution, implying the disappearance of the middle class (Wang and Tsui, 2000; Wolfson, 1994). The other approach relies on the concept of identification-alienation: individuals identify themselves with those having similar income levels, whereas they feel alienated from individuals with different income levels (Deutsch *et al*. 2013; Duclos *et al*., 2004; Esteban and Ray, 1994; Poggi and Silber, 2010); therefore, polarization is investigated from the perspective of grouping of individuals around local poles and within-group identification. Following the second approach, we show that the regression tree is a useful statistical tool for measuring polarization in income distribution.

[1] Department of Economics, University of Verona, Via dell'Artigliere 8, Verona (Italy).
E-mail: mauro.mussini@univr.it.

Recently, Palacios-González and García-Fernández (2012) have pointed out that the coefficient of determination (R^2) of an ANOVA linear model can be interpreted as a measure of polarization. Since R^2 increases as within-group variance decreases (i.e. groups are internally more homogeneous), Palacios-González and García-Fernández state that R^2 can be seen as a (normalised) measure of polarization. Moreover, linking the ANOVA coefficient of determination with polarization enables one to analyse polarization by the characteristics of income receivers when groups are defined by such characteristics (Palacios-González and García-Fernández, 2012).

The variance decomposition approach proposed by Palacios-González and García-Fernández is analogous in the spirit to the Zhang and Kanbur (2001) approach to polarization measurement, since the latter is based on the income inequality decomposition by groups. Both the Palacios-González and García-Fernández approach and the Zhang and Kanbur one assume that groups are pre-established, and then measure polarization for that population partition; therefore, both approaches tell us whether polarization is high or low for the population partition defined *a priori*. Duclos *et al*. (2004) suggested letting the population partition arise in a data-driven way rather than taking the population partition as exogenous. In our approach to polarization analysis, we initially face the issue of identifying the most homogeneous groups in a data-driven way and then we measure the degree of income polarization for the population partition showing maximal within-group identification.

We show that groups can be naturally formed from the data exploration by using regression trees to recursively partition the population. We assume that income is the response variable and income receiver's characteristics are the explanatory variables; then, the population is recursively partitioned to maximally reduce the within-group variance, which is maximizing the gain in homogeneity within groups. Once groups clustering income receivers with similar income levels have been detected, R^2 is used to measure the extent to which incomes are polarized.

In our empirical analysis, regression trees are applied to Italian individual income data in order to detect the characteristics relevant for polarization. Our findings show that the interactions among employment status, education and age form well-identified groups of income receivers.

The article is organised as follows. Section 2 briefly reviews the Palacios-González and García-Fernández approach to polarization measurement. Section 3 introduces regression trees and shows how this technique is suitable for analysing income polarization. In Section 4, the regression tree approach is applied to Italian income data from the Survey on Household Income and Wealth (SHIW) conducted by the Bank of Italy in 2010 (Banca d'Italia, 2012).

2. Measuring polarization VIA ANOVA

The link between polarization and ANOVA is outlined by Palacios-González and García-Fernández (2012) in the generalized linear model framework. Palacios-González and García-Fernández follow the Zhang and Kanbur (2001) approach to polarization, which assumes that for *k* predetermined groups of income receivers the larger the ratio of between-group income inequality to within-group income inequality, the larger the polarization. Similarly to the Zhang and Kanbur approach, Palacios-González and García-Fernández assume the mean income of a group as the representative income for the income receivers within that group; moreover, they observe that the larger the disparities among the mean income of a group and the mean incomes of the other groups, the more the income receivers belonging to that group feel alienated from income receivers included in the other groups. However, the Palacios-González and García-Fernández approach differs from the Zhang and Kanbur one since the former is based on variance decomposition by group. Indeed, Palacios-González and García-Fernández propose to measure polarization using the ratio between the variance between groups and the variance within groups

$$P = \frac{\sigma_b^2}{\sigma_w^2}, \tag{1}$$

where σ_w^2 denotes the within-group variance and σ_b^2 is the between-group variance. Then, Palacios-González and García-Fernández suggest to normalise expression in (1) by replacing σ_w^2 with the overall variance $\sigma^2 = \sigma_w^2 + \sigma_b^2$,

$$P^* = \frac{\sigma_b^2}{\sigma^2} = 1 - \frac{\sigma_w^2}{\sigma^2}. \tag{2}$$

The polarization measure in (2) is equivalent to the (unadjusted) R^2 used in ANOVA when one investigates the effect of grouping on income. Palacios-González and García-Fernández formulate a fixed-effects ANOVA model in the framework of generalized linear models, where *n* income receivers are partitioned into *k* groups on the basis of the *k* different values (levels) taken by one of the characteristics of income receivers (e.g. gender, age, employment status, etc.). Let Y_i denote the income receiver *i*'s income and D_{ih} be the dummy variable that equals 1 if the income receiver *i* belongs to group *h* and 0 otherwise. In matrix notation, the model is expressed as

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} d_{11} & \cdots & d_{1h} & \cdots & d_{1k} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ d_{i1} & \cdots & d_{ih} & \cdots & d_{ik} \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ d_{n1} & \cdots & d_{nh} & \cdots & d_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_h \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix} \tag{3}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

where $\mathbf{X}$ is the $n \times k$ matrix with the known constants d_{ih}, $\boldsymbol{\beta}$ is the $k \times 1$ vector of unknown parameters, $\mathbf{u}$ is the $n \times 1$ vector of unobservable errors. Given the model specification in (3), it can be immediately verified that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & \vdots \\ 0 & & n_h & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & n_k \end{bmatrix} \tag{4}$$

and

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^{n_1} y_{i1} \\ \vdots \\ \sum_{i=1}^{n_h} y_{ih} \\ \vdots \\ \sum_{i=1}^{n_k} y_{ik} \end{bmatrix}, \tag{5}$$

where n_h is the size of group *h*. Therefore, the elements of the least squares estimator $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Y}$ are the group mean incomes $\bar{y}_h$,

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_h \\ \vdots \\ \bar{y}_k \end{pmatrix}. \tag{6}$$

As shown in Palacios-González and García-Fernández (2012, p.1546), even though the model in (3) does not include the intercept, the decomposition of the total sum of squares (*TSS*) into the explained sum of squares (*ESS*) and the residual sum of squares (*RSS*) is valid. Then, the coefficient of determination of the model in (3)

$$\begin{aligned} R^2 &= 1 - \frac{RSS}{TSS} \\ &= 1 - \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\left(\mathbf{X}'\mathbf{X}\right)\hat{\boldsymbol{\beta}}}{\left(\mathbf{y} - \mathbf{1}\bar{y}\right)'\left(\mathbf{y} - \mathbf{1}\bar{y}\right)} \end{aligned} \tag{7}$$

is equivalent to P^* in (2).[2] Using (7) the link between the income polarization and the levels of one of the characteristics of income receivers can be investigated: values of R^2 close to 1 suggest that grouping income receivers by the levels of one of their characteristics creates groups which are internally homogenous; on the contrary, low values of R^2 indicate that an income receiver does not identify himself much with the other members of his group (i.e. those sharing the same level of the characteristic under consideration).

3. Using regression trees for detecting homogenous groups

The regression tree is a nonparametric method for finding patterns or predicting new observations in data mining (Hsiao and Shih, 2006). Regression trees are able to capture nonlinear relationship between the response variable and explanatory variables, and to summarize results with an intuitive graphic. In addition, unlike other statistical methods (e.g. linear regression, ANOVA) regression trees do not require specific distribution assumptions. For these reasons, the regression tree method is a flexible statistical tool which has been applied in various research fields, such as ecology (De'ath and Fabricius, 2000), finance (Campanella, 2014) and epidemiology (Gass *et al.*, 2014). Here we present the regression trees as an explorative statistical tool for uncovering the relationships between income and the characteristics of income receivers. Let $(Y,\mathbf{X}):\Omega \to \left(S_Y \times S_{X_1} \times \cdots \times S_{X_p}\right) \equiv S$ be a vector random variable defined on the probability space $\left(\Omega, F, P\right)$, where *Y* is a numerical response variable and $\mathbf{X} = \left\{X_1,\dots X_j,\dots,X_p\right\}$ is a set of *p* explanatory variables. Assume that *Y* is income and $\mathbf{X}$ is the vector collecting *p* income receiver's characteristics. The regression tree is built by recursively partitioning the space *S* into disjoint subsets, such that each subset comprises income receivers who are as homogenous as possible with respect to *Y*. The income receivers comprised in a subset constitute a group which is characterized by the group mean income and the combination of the levels of the characteristics that defines the group. From this standpoint, maximizing within-group homogeneity is equivalent to minimizing variance within groups. Therefore, a rule based on ANOVA is used to repeatedly split income receivers into more homogeneous groups.

Define the variance of the values of *Y* within subset *t* as follows:

$$\sigma_t^2\left(Y\right) = n_t^{-1} \sum_{\mathbf{X}_i \in t} \left(y_{it} - \bar{y}_t\right)^2, \tag{8}$$

[2] A model with more explanatory variables can produce a higher R^2, but this result may be caused by overfitting. To avoid this problem, the adjusted R^2 can also be used as a measure of polarization, as noted by an anonymous referee.

where $\bar{y}_t$ is the mean income within subset *t* and n_t is the number of income receivers in subset *t*. Let $c \in S_{X_j} \mid t$ stand for a value of *X*$_j$ within the domain of *X*$_j$ restricted to subset *t*. The variance reduction due to splitting *t* into two parts, t_L and t_R, at the threshold *c* is

$$\Delta_t(Y,c) = \sigma_t^2(Y) - \left[\frac{n_{t_L}}{n_t}\sigma_t^2(Y \mid X_j \leq c) + \frac{n_{t_R}}{n_t}\sigma_t^2(Y \mid X_j > c)\right], \tag{9}$$

where $n_{t_L} = \sum_{i=1}^{n_t} I_{\{X_{ij} \leq c\}}$ and $n_{t_R} = \sum_{i=1}^{n_t} I_{\{X_{ij} > c\}}$ are the numbers of income receivers in subsets t_L and t_R, respectively. For subset *t*, the splitting variable and the variable split *c* are selected from all possible splits of the explanatory variables in order to maximize the variance reduction in (9). We note that maximizing (9) is equivalent to maximizing $\Phi_t(Y,c) = n_t\Delta_t(Y,c)$; that is, one searches for the split that minimizes the residual sum of squares

$$RSS_t(Y,c) = \sum_{\mathbf{X}_i \in t_L}\left(y_{it_L} - \bar{y}_{t_L}\right)^2 + \sum_{\mathbf{X}_i \in t_R}\left(y_{it_R} - \bar{y}_{t_R}\right)^2. \tag{10}$$

It follows that a subset is formed in *S* by splitting a parent subset into two parts through a binary split of the support of an explanatory variable *X*$_j$; therefore, a subset is characterized by the explanatory variables and variable splits which define it.

At the beginning of the recursive partitioning procedure ungrouped income receivers are considered, and then the whole space *S* is split into two parts by selecting the most effective variable (and variable split) in reducing the overall variance of *Y* by minimizing within-group variance. The binary splitting is repeated for each subset until the tree has grown large enough so that no further splitting yields a variance reduction, which overcomes a pre-established minimal threshold. As pointed out in Breiman *et al*. (1993), it is convenient to set a small value for the threshold,[3] growing an overlarge tree and then searching for the best tree. Tree pruning is used to find the best tree. Pruning can be performed by minimizing the following cost-complexity function for a tree *T*

$$R_\alpha(T) = R(T) + \alpha|T|, \tag{11}$$

[3] Setting a large threshold serves the scope of excluding a split if it does not produce an appreciable reduction in variance; however, if that split is made, one of the descendent subsets may be split in a way to yield an appreciable decrease in variance. This can occur when a split based on interactions among variables yields an appreciable decrease in variance, but none of the associated variable main effects produces an appreciable variance reduction (De'ath and Fabricius, 2000, pp. 3183).

where $|T|$ is the tree size, that is the number of terminal subsets, α is a complexity parameter ranging within the interval $[0,\infty)$, and $R(T)$ is the resubstitution estimate of error which coincides with the residual sum of squares of Y for a regression tree with size $|T|$ (De'ath and Fabricius, 2000).[4] As shown in Breiman *et al.* (1993), for any α there is a unique smallest tree which minimizes (11), therefore, finding the best tree reduces to choosing the best tree size. The strategy for selecting the optimal tree size is discussed in the empirical analysis shown in the next section (Section 4).

Once the regression tree has been pruned, $|T|$ homogenous groups are identified. The measure of polarization P^* (i.e. R^2) is calculated for this population partition. Unlike the Palacios-González and García-Fernández approach, where groups are pre-established, the identification of the $|T|$ groups arises from the structure of the data by clustering observations with similar income values. Therefore, using regression trees, polarization patterns can be naturally uncovered from data.

Another difference between the regression tree and the Palacios-González and García-Fernández ANOVA model is that the former can deal with high-order interaction effects among explanatory variables, whereas the latter can only capture the main effects of the variable used to define groups. It is worth mentioning that the Palacios-González and García-Fernández model could be extended to include interaction effects among explanatory variables; however, the interactions need to be specified *a priori*. Using regression tree, only the interactions which actually contribute to growing the tree are included in the fitting process; therefore, we can say that interactions are specified in a data-driven way, as noted in Strobl *et al.* (2009).

3.1. Comparison with other methods for measuring income polarization

Since other approaches for analysing income polarization have been proposed in the literature, it is worth underlining the differences between these approaches and the approach based on regression trees. Esteban and Ray (1994) define a class of indices to measure income polarization. The Esteban and Ray polarization index is based on the pairwise comparisons between groups, where each group is identified by its income level and size:

$$ER = \sum\nolimits_{i=1}^{k} \sum\nolimits_{j=1}^{k} n_i^{1+\alpha} n_j \left| y_i - y_j \right| \quad \text{with } \alpha \in [1;1.6], \tag{12}$$

where k is the number of groups, y_i is the income level of group i and n_i is the size of group i. The ER index depends on the choice of parameter α. To apply

[4] The introduction of α which handles the trade-off between $R(T)$ and the tree size is necessary since the residual sum of squares will always be minimized by the largest tree (Sutton, 2005, p. 311); however, the larger the tree, the lower its interpretability. The use of a cost-complexity measure avoids choosing trees with very small $R(T)$, but too large to be interpreted clearly.

the *ER* index, the choice of the criterion to form k groups is required. In doing so, the groups may be formed by partitioning the income distribution into k non-overlapping income ranges or by establishing an external criterion (e.g. age, occupation, geographical area, education level) which a priori splits the population into k groups. Unlike the Esteban and Ray approach to polarization, the approach based on regression trees finds groups in a data-driven way by searching for the partition maximizing within-group homogeneity. Using the tree-based approach, R^2 is used to measure polarization whereas the index of polarization in (12) is used by applying the Esteban and Ray approach.

In the income distribution literature, polarization has also developed by following an alternative approach focusing on the concept of bipolarization; that is, the extent to which incomes spread from the middle to the tails of the distribution, implying the disappearance of the middle class (Wolfson, 1994). Wolfson (1994) suggests an index to measure bipolarization in income distribution. Let $Me(\mathbf{y})$ stand for the median income. Let $\mathbf{y}^+$ be the vector with the incomes above the median income and $\mathbf{y}^-$ be the vector with the incomes below the median income. $\mu(\mathbf{y}^+)$ and $\mu(\mathbf{y}^-)$ being the mean incomes above and below the median respectively, the Wolfson index is

$$W = \frac{2\mu(\mathbf{y})}{Me(\mathbf{y})}\left[\frac{\mu(\mathbf{y}^+) - \mu(\mathbf{y}^-)}{\mu(\mathbf{y})} - G(\mathbf{y})\right], \tag{13}$$

where $\mu(\mathbf{y})$ is the overall mean and $G(\mathbf{y})$ is the Gini index of inequality. When measuring bipolarization, the median is considered as a threshold for partitioning the distribution into a lower portion and an upper portion; then, the concentration of incomes around two poles on opposite sides of the median is observed.

4. Application to income data

We apply regression trees to individual incomes collected by the Survey on Household Income and Wealth (SHIW) conducted by the Bank of Italy in 2010 (Banca d'Italia, 2012). The SHIW is carried out every two years, and each survey sample comprises households interviewed for the first time and households interviewed in previous surveys (panel households). The SHIW data is one of the most frequently used information source to investigate income inequality in Italy (Mussini, 2013; Zenga 2007), since the survey collects information on income and socioeconomic status for every household member. The sample size of the 2010 survey is 7,951 households, including 19,836 individuals. We perform the analysis on individual incomes, considering 13,733 income receivers. Table 1 shows some descriptive statistics for the subsample under consideration.

Table 1. Descriptive statistics for the income distribution

number of observations	minimum	first quartile	median	mean	third quartile	maximum
13,733	-7,345.2	10,131.7	16,073.0	19,155.3	23,711.4	573,383.9

Source: Calculations on SHIW 2010 data.

The set of characteristics of income receivers used as explanatory variables is shown in Table 2. Applying regression trees enables one to detect which characteristics play a role in explaining the income received by an individual. The combinations of the characteristics defining the $|T|$ terminal subsets identify $|T|$ exhaustive and mutually exclusive groups of income receivers.

Table 2. Explanatory variables description and coding

name	description	type	categories coding (for categorical variables) or range (for numeric variables)
age	age	numerical	(0, 102] years;
area	geographical area of residence	nominal	N="North", C="Centre", S="South and Islands";
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";
status	marital status	nominal	M="married", S="single", D="separated or divorced", W="widowed"
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", M="male";
size	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";
Italian	citizenship	dichotomous	I="Italian", F="not Italian";
health	state of health	ordinal	VP="very poor", P="poor", F="fair", VG="good", E="excellent";
home	individual's home status	nominal	O="owned", R="rented or sublet", UR="under redemption agreement", U="occupied in usufruct";

*Source: SHIW 2010. *Ordinal variable categories are listed in ascending order.*

The tree grows large by setting a small value of the complexity parameter (*cp*) to avoid that interaction effects among explanatory variables are not discovered because none of the associated main effects produces a split with an appreciable decrease in variance.[5] Table 3 shows the resubstitution relative error (RE=1-R^2), the 10-fold cross-validation relative error (RE^{CV}), and the standard error of the 10-fold cross-validation relative error (*SE*) for different tree sizes. From Table 3, we observe that the pre-pruning tree has twenty six terminal subsets. Cross-validation is used to obtain more accurate estimates of (prediction) relative error for trees of a given size (see Breiman *et al*., 1993, pp. 234-237).[6] Cross-validation estimates of relative error can be used to select the optimal tree size by choosing the size with minimum cross-validation relative error. However, to select the optimal tree size we follow the 1-*SE* rule proposed by Breiman *et al*. (1993). The 1-*SE* rule suggests choosing the smallest tree *T* such that

$$RE^{CV}\left(T\right) \leq RE^{CV}\left(T_{\min}\right) + SE\ , \tag{14}$$

where $T_{\min}$ is the tree with minimum cross-validation relative error and *SE* is the associated standard error estimate. The rationale for the use of the 1-*SE* rule is that it usually selects a much smaller (and more interpretable) tree than that suggested by the minimum cross-validation relative error, entailing a minimal increase in the cross-validation relative error (less than *SE*).

[5] We use the R package rpart (Therneau *et al*., 2012) for recursive partitioning and we set the *cp* equal to 0.0025. The *cp* value in rpart has a meaningful interpretation since it is equal to the increase in R^2 that a split has to produce in order to be made. It immediately follows that the relationship between *cp* and α in equation (11) is $\alpha = TSS \cdot cp$, where *TSS* denotes the total sum of squares of *Y*. Therefore, when setting *cp*, one also defines α.

[6] 10-fold cross-validation is performed as follows: (*i*) observations are divided into ten subsets of approximately equal size; (*ii*) each subset in turn is left out, a tree of size $|T|$ is built using the remaining subsets, and this tree is used to predict the response variable values for the omitted subset; (*iii*) the prediction errors are calculated for each omitted subset by adding up the squared differences between the observed and predicted values; (*iv*) the sums of prediction errors calculated for the ten subsets are added up, and the total sum of prediction errors $R^{CV}(T)$ is divided by *TSS* to obtain the 10-fold cross-validation relative error $RE^{CV}(T)$ for a tree with size $|T|$; (*v*) steps (*i*)-(*iv*) are repeated for every tree size.

Table 3. Resubstitution relative error ($RE(T)$) and 10-fold cross-validation relative error ($RE^{CV}(T)$) by tree size

$\|T\|$	*cp*	Number of splits	$RE(T)$	$RE^{CV}(T)$	*SE*
1	0.106872	0	1.00000	1.00017	0.09527
2	0.049124	1	0.89313	0.89395	0.09285
3	0.035976	2	0.84400	0.84488	0.09240
4	0.029526	3	0.80803	0.80903	0.09083
5	0.022208	4	0.77850	0.78202	0.08846
6	0.013902	5	0.75629	0.76205	0.08779
7	0.013235	6	0.74239	0.75649	0.08743
8	0.011640	7	0.72916	0.73876	0.08732
9	0.010856	8	0.71752	0.72990	0.08707
10	0.007842	9	0.70666	0.71721	0.08689
11	0.007525	10	0.69882	0.71631	0.08696
12	0.007216	11	0.69129	0.71462	0.08692
13	0.004419	12	0.68408	0.70086	0.08669
14	0.004253	13	0.67966	0.70096	0.08675
15	0.003642	14	0.67541	0.69789	0.08681
16	0.003585	17	0.66438	0.70033	0.08680
17	0.003570	18	0.66079	0.69947	0.08679
18	0.003459	19	0.65722	0.69953	0.08679
19	0.003393	20	0.65376	0.69938	0.08679
20	0.003384	21	0.65037	0.69825	0.08678
21	0.002977	22	0.64699	0.69782	0.08679
22	0.002853	23	0.64401	0.69650	0.08676
23	0.002774	24	0.64116	0.69552	0.08663
24	0.002766	25	0.63838	0.69132	0.08666
25	0.002549	26	0.63562	0.68918	0.08665
26	0.002500	27	0.63307	0.68661	0.08665

Source: Calculations on SHIW 2010 data.

From Table 3, we see that the tree with six terminal subsets is the smallest tree which satisfies (14). Once the optimal tree size has been selected, the tree is pruned.[7] Figure 1 shows the pruned tree where six groups are detected by the terminal subsets 4, 6, 7, 10, 22, 23. Each terminal subset in Figure 1 shows its size and mean income.

[7] Practically speaking, pruning is performed through the R package rpart by replacing the *cp* value used to grow the overgrown tree with the *cp* value that generates a tree with six terminal subsets in Table 3 (i.e., *cp*=0.013902).

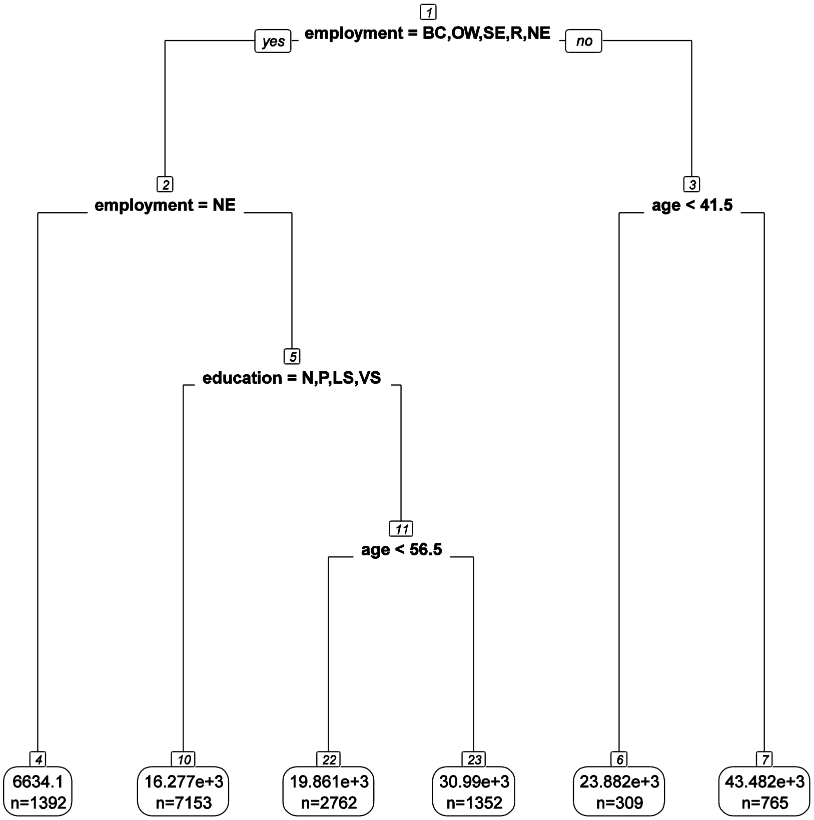

Figure 1. Regression tree analysis of income polarization

Only three variables (education, employment, age) from the set of explanatory variables in Table 2 are discriminating in recursive partitioning income receivers into subsets. The employment main effect distinguishes between individuals whose employment status is equal to M or P and the remaining individuals; that is, the employment status determines the initial partition between high-skilled workers or business owners (M or P) and the other workers (BC, OW, SE) or not working individuals (R or NE). This means that the main effect of the employment status is more important than those of the other variables in determining differences in income. Subset 4 comprises unemployed individuals and has the lowest mean income (6,634.1 EUR). The use of regression tree enables one to identify subsets 6 and 7, since the regression tree also accounts for interaction between employment and age: among the income receivers whose employment status is M or P, individuals younger than 41.5 years old (subset 6) receive much lower incomes than those older than 41.5 years old (subset 7). Education has an effect on income for individuals whose employment status is BC, OW, SE or R: individuals with educational qualifications lower than or equal to VS (subset 10) receive lower incomes than those with educational qualifications higher than VS (hereafter, high-educated workers); then, among high-educated workers, incomes are higher for individuals older than 56.5 years old (subset 23). Subset 10 is the largest subset, including more than half of the income receivers in the sample. It is worth mentioning that age does not play a role in determining income in subset 10 (low-educated workers), whereas age is discriminating in subset 11 (high-educated workers). This finding suggests that high-educated workers have chances of increasing their income during their career; this age effect is not present for low-educated workers. More specifically, the mean income of high-educated BC, OW and SE workers older than 56.5 years old (30,990 EUR) is almost twice the mean income of low-educated workers in the same occupations (16,277 EUR).

The above discussed partition is detected by discovering the different patterns of income existing in the income distribution: income receivers comprised in the same group share the same income pattern which differs from those of the other groups. Therefore, each income receiver identifies himself with those sharing the same income pattern and feels alienated from income receivers with different income patterns. The R^2 calculated for the partition detected by the regression tree is equal to 0.24371 and measures the polarization in the income distribution.

5. Concluding remarks

The contribution of the article is two-fold. First, we show that the regression tree is a useful statistical tool to investigate whether incomes concentrate around local poles. The regression tree identifies groups which are internally homogeneous in a data-driven way: income receivers are recursively partitioned into groups by selecting the explanatory variables that actually contribute to defining groups of income receivers with similar income levels. Other distinguishing features of regression trees are the ability to capture nonlinear relationships between explanatory variables and income, and the intuitive graphic interpretation of results. Therefore, regression tree can be seen as a flexible and practical technique to explore income polarization.

Second, we extend the ANOVA-based approach to polarization measurement proposed by Palacios-González and García-Fernández (2012), since we point out that using regression trees instead of one-way ANOVA we are able to detect not only the main effects of explanatory variables but also their interaction effects. This enables analysts to discover polarization patterns that cannot be assumed *a priori*. For instance, our empirical analysis of Italian income data shows that the interactions among employment status, educational qualification and age form well-identified groups of income receivers, whereas the other characteristics do not play a clear role in explaining income polarization.

Further research will be devoted to extending the approach based on recursive partitioning to the analysis of polarization when the response variable is ordinal (e.g. level of satisfaction, health status) instead of numeric (e.g. income). In the first instance, this requires the definition of a proper polarization-sensitive impurity function that can be used for recursive partitioning, as the residual sum of squares is suited to the tree-based model for income polarization.

Acknowledgements

The author would like to thank two anonymous reviewers for helpful comments which helped to improve the manuscript.

REFERENCES

BANCA D'ITALIA, (2012). Survey on Household Income and Wealth 2010, Rome, Italy, 2012.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., STONE, C. J., (1993). Classification and regression trees, Chapman & Hall/CRC press, Boca Raton.

CAMPANELLA, F., (2014). Assess the Rating of SMEs by using Classification and Regression Trees (CART) with Qualitative Variables, Review of Economics & Finance, 4, 16–32.

DE'ATH, G., FABRICIUS, K. E., (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis, Ecology, 81, 3178–92.

DEUTSCH, J., FUSCO, A., SILBER, J., (2013). The BIP trilogy (Bipolarization, Inequality and Polarization): one saga but three different stories, Economics, Discussion Paper No. 2013–22.

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

ESTEBAN, J. M., RAY, D., (1994). On the measurement of polarization, Econometrica, 62, 819–51.

GASS, K., KLEIN, M., CHANG, H. H., FLANDERS, W. D., STRICKLAND, J., (2014). Classification and regression trees for epidemiological research: an air pollution example, Environmental Health, 13:17.

HSIAO, W. C., SHIH, Y. S., (2006). Splitting variable selection for multivariate regression trees, Statistics and Probability Letters, 77, 265–71.

MUSSINI, M., (2013). A matrix approach to the Gini index decomposition by subgroup and by income source, Applied Economics, 45, 2457–2468.

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, 1543–55.

POGGI, A., SILBER, J., (2010). On polarization and mobility: a look at polarization in the wage-career profile in Italy, Review of Income and Wealth, 56, 123–140.

STROBL, C., MALLEY, J., TUTZ, G., (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests, Psychological Methods, 14, 323–48.

SUTTON, C. D., (2005). Classification and regression trees, bagging, and boosting, in Handbook of statistics 24: data mining and data visualization (Eds.) C. R. Rao, E. J. Wegman and J. L. Solka, Elsevier, Amsterdam, pp. 303–29.

THERNEAU, T., ATKINSON, B., RIPLEY, B., (2012). Rpart: recursive partitioning and regression trees, R package version 3.1–55.

WANG, Y. Q., TSUI, K. Y., (2000). Polarization orderings and New Classes of Polarization Indices, Journal of Public Economic Theory, 2, 349–63.

WOLFSON, M. C., (1994). When inequalities diverge? American Economic Review, 84, 353–58.

ZENGA, M., (2007), Inequality curve and inequality index based on the ratios between lower and upper arithmetic means, Statistica & Applicazioni, 5, 3–27.

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