Regression-based approaches for the decomposition of income inequality in Italy, 1998-2008

Rosalba Manna[1], *Andrea Regoli*[2]

Abstract

Decompositions by population subgroups and by income sources represent the traditional techniques for decomposing income inequality. Compared with the classical methodologies, the regression-based method gives the opportunity of quantifying the contribution to the inequality of a set of factors, while taking the correlations among them into account. In this framework, two regression-based decomposition methodologies are used: the Fields method and the Shapley value approach, with the aim of measuring the relative contributions of individual as well as household factors to inequality in individual disposable incomes. The factors are introduced as explanatory variables in an income generating model that is estimated through a panel data regression model with time-invariant unobserved random effects. The results suggest that the most relevant factors in explaining the observed income inequality are gender, human capital as well as non-human capital whereas the work status and the area of residence only affect income differentials in a marginal way.

Keywords: Inequality decomposition; regression-based methods; Shapley value; panel data models.

1. Introduction

This work addresses a relevant and topical issue: inequality in the Italian income distribution. Recently, inequality has raised growing concerns at the global level as well as in the Italian society, where income differences are widening against the background of a deep macroeconomic recession and the entailed negative perception of the economic and financial situation at the household level. Moreover, the family background in Italy can powerfully limit the chances of moving up the social ladder, with the consequence that those who lack resources are likely to be disadvantaged in terms of opportunities too.

A number of recent studies have analyzed the evolution of household income inequality in Italy. Boeri and Brandolini (2004) found evidence of significant distributive changes across socio-economic groups despite inequality at the aggregate level was stable between 1993 and 2002. Through a decomposition by population subgroups defined by the

[1] Post Ph.D student (University of Naples “Parthenope”, Department of Statistics and Mathematics for Economic Research), e-mail: rosalba.manna@uniparthenope.it.

[2] Associate Professor (University of Naples “Parthenope”, Department of Statistics and Mathematics for Economic Research), e-mail: andrea.regoli@uniparthenope.it.

occupational status of the household head, the authors concluded that the income distribution shifted to the advantage of the self-employed and managers and to the disadvantage of the employees.

For the time period 1991-2004, Quintano *et al.* (2009) confirmed that, from 2000 on, a marked segmentation of households emerged, with widening gaps in the average incomes between the group of managers and self-employed and the group of employees. Moreover, a decomposition by income sources added new evidence upon the role played by the different sources in accounting for both the level of inequality and its trend. The increase in wage differentials was deemed to be the main driving force behind the dramatic rise in inequality in disposable household incomes between 1991 and 1993. The peak in inequality observed in 1998 was driven by the income from financial assets, whereas in more recent years the income from self-employment was found to be the main disequalizing factor.

The disappearance of the middle class was the main implication of Massari *et al.* (2009) work, which investigated the differences across the whole income scale by comparing the household income distributions in 2002 and in 2004 through a nonparametric approach, based on the relative distribution density function. The authors found an increased income polarization, due to downgrading of the incomes earned by the households headed by employees or by the self-employed.

Unlike the above mentioned studies, the present paper focuses upon the determinants of the observed income differentials. More precisely, the aim of this study is mainly empirical and it has to do with the assessment of the contribution of several individual and household factors to income inequality among individuals through a regression-based decomposition strategy.

A wide literature exists on the decomposition of inequality measures. The traditional methods include the decomposition by income sources (Shorrocks, 1982) and by population subgroups (Shorrocks, 1984). The former method estimates the contribution of individual income components to the observed inequality, whereas the latter allows to measure inequality both within and between subgroups of the population. Both of them are typically descriptive methods that tell us what sources of incomes or subgroups account for inequality but they fail to detect and measure the contributions of individual determinants to income inequality. For this reason, the information provided by those methods is of limited usefulness for policy-makers seeking to address income inequality problems.

Unlike the traditional methods, the regression-based approaches followed in this work have the advantage of going beyond decomposing inequality simply in terms of income components or discrete population categories. Indeed, they enable to include any factor that may drive the observed inequality, such as economic, social, demographic and policy variables, both discrete and continuous. Moreover the regression-based methods can manage problems of endogeneity due to reverse causality.

The regression-based decomposition methodology was proposed in the early 1970s (Blinder, 1973; Oaxaca, 1973), but failed to arouse much interest until Morduch and Sicular (2002) and Fields (2003) devised a regression-based decomposition by income determinants through the extension of the decomposition by income sources. Regression-based decompositions start with the estimation of an income-generating function, and then use the estimated coefficients to derive the inequality weight of every explanatory variable.

In the context of regression-based decomposition, many recent studies proposed the application of either the Fields method or the Shapley value approach, a concept taken from cooperative game theory.

Sastre and Trannoy (2002) measured the impact of different income sources on income inequality for UK and USA household income data, focussing on some methodological issues regarding the Shapley decomposition of Gini index.

The studies by Wan (2004), Wan and Zhou (2005) and Wan *et al.* (2007) combined the Shapley value approach and the regression-based decomposition technique in order to disentangle the contribution of different factors to household income inequality in China by using several inequality indices.

Israeli (2007) suggested a method for decomposing the R-Square of a linear regression that combines the Shapley approach with the Fields method and added an empirical illustration of this methodology on Israeli earnings data.

Guanatilaka and Chotikapanich (2009) investigated the evolution of Sri Lanka's expenditure inequality as well as its underlying causes by using three regression-based methodologies of decomposition: the Fields approach, the Shapley value decomposition and the Yun method.

Devicienti (2010) applied a Shapley value-based methodology for decomposing changes in the Italian wage distribution by using WHIP (Worker History Italian Panel) data on employees in private firms for the years between 1985 and 1999. The only other decomposition analysis that employed the Shapley value approach on Italian data is the recent study by Celidoni *et al.* (2011) who investigated the determinants of expenditure inequality on a pseudo panel based on Istat Household Budget Survey data for the years from 1997 to 2004.

In the wake of these studies, the present paper intends to contribute to the identification of the main driving factors for the inequality levels through the application of regression-based decomposition approaches. Unlike the above mentioned empirical studies for Italy, however, in this contribution the inequality is measured on individual incomes. Heterogeneity across individuals and across time is accounted for by using the longitudinal information from the Historical Archives of Bank of Italy's Survey of Household Income and Wealth.

The comparative discussion of the results derived from the application of Fields and Shapley approaches is also a key contribution of this paper.

This paper is organized as follows. In Section 2 the theoretical background of the regression-based methods is presented. Section 3 deals with model selection and specification issues, whereas the empirical data from the Survey of Household Income and Wealth (SHIW) are illustrated in Section 4. Section 5 shows the model estimates and the decomposition results, while conclusions are drawn in the final Section 6.

2. The regression-based decomposition according to the Fields method and the Shapley approach

Generally speaking, the regression-based inequality decomposition methods allow quantifying the impact of the determinants of inequality. Both the number and the kind of the explanatory factors are arbitrary, introducing some flexibility in the analysis that is not granted by the traditional decomposition methods.

Let us consider an income generating function such as:

$$\ln y = \sum_{j=1}^{k} b_j X_j + \varepsilon \tag{1}$$

where y denotes income, X_j the *j*-th explanatory variable, b_j its coefficient and ε the error term. The Fields method (Fields, 2003) estimates the share of the log-variance of income that is attributable to the *j*-th explanatory factor (the relative factor inequality weight) as:

$$s_{j,FIELDS} = \frac{\hat{b}_j \cdot \operatorname{cov}(X_j, \ln y)}{\sigma^2(\ln y)} \tag{2}$$

where $\hat{b}_j$ is the coefficient of the *j*-th explanatory factor estimated from an OLS multiple regression, $\sigma^2(\ln y)$ is the variance of the dependent variable and $\operatorname{cov}(X_j, \ln y)$ is the covariance between the *j*-th factor and the dependent variable.

The sign of s_j indicates whether the contribution of factor x_j is inequality-increasing $(s_j > 0)$ or decreasing $(s_j < 0)$. It holds that

$$\sum_{j=1}^{k} s_{j,FIELDS} = \frac{\sum_{j=1}^{k} \hat{b}_j \cdot \operatorname{cov}(X_j, \ln y)}{\sigma^2(\ln y)} = \frac{\sigma^2(\ln \hat{y})}{\sigma^2(\ln y)} = R^2 \tag{3}.$$

When the error term ε of the regression is considered, its inequality contribution is given by the proportion of inequality unexplained by the explanatory variables included in the income regression, that is:

$$s_\varepsilon = 1 - R^2 \tag{4}$$

Under some assumptions, Fields extended this result to any inequality index with certain properties, including the most common measures such as the Gini index and the indexes belonging to the generalized entropy family. One limitation of the Fields method is that the functional form for the income generating function must be log-linear.

Unlike the Fields method, the Shapley value approach, as introduced by Shorrocks (1999), yields an exact additive decomposition of any inequality measure into its contributory factors. Indeed, the decomposition of a given inequality measure through a regression-based method combined with the Shapley value approach aims at assessing the contributions of a set of factors (the explanatory variables in the income regression model 1) whose sum accounts for the inequality indicator. Moreover, the income generating model can have any functional form (including linear, logarithmic and semi-logarithmic functions).

As in the framework of a general decomposition problem, the inequality measure calculated on the predicted income values $I(\hat{y} \mid X_1, X_2, ..., X_k)$ is expressed as the sum of the contributory factors:

$$I(\hat{y} \mid X_1, X_2, \ldots, X_k) = \Phi(X_1, I) + \Phi(X_2, I) + \ldots + \Phi(X_k, I) \qquad (5).$$

The rationale behind the Shapley approach is that the contribution of a single factor can be assessed as the difference between the overall income inequality and the inequality that would be observed should that factor be removed from the set of income determinants. As a consequence, the marginal impact of each factor $\Phi(X_j, I)$ $j = 1,2,\ldots,k$ is calculated through the estimation of a sequence of regression models starting from the specification which includes all the regressors, and then successively eliminating each of them. The overall marginal contribution of each variable is then obtained as the average of its marginal effects: since the contribution of any factor depends on the order in which the factors appear in the elimination sequence, this average is calculated over all the possible elimination sequences.

The contribution $\Phi(X_j, I)$ of the factor X_j to the explanation of the inequality measure I is given by the following formula:

$$\Phi\left(X_j, I\right) = \frac{1}{k!} \sum_{\pi \in \Pi_k} \left[I\left(\hat{y} \mid B\left(\pi, X_j\right) \cup \left\{X_j\right\}\right) - I\left(\hat{y} \mid B\left(\pi, X_j\right)\right)\right] \qquad (6)$$

where $I(\hat{y} \mid X)$ is the inequality indicator calculated on the predicted income values from the regression on the vector of explanatory variables X ;
Π_k is the set of all the possible orderings (permutations) of the k variables;
$B(\pi, X_j)$ is the set of the variables preceding X_j in the given ordering π .

The calculation of each factor's contribution requires the estimation of $2^k - 1$ income generating models, and then the derivation of the inequality indicator I using the income predicted values for every model.

Finally, the proportion of unexplained inequality $I_R(y)$ is obtained as the difference between the inequality measure calculated on the observed income values $I(y)$ and the same measure calculated on the predicted income values, as follows:

$$I_R(y) = I(y) - I(\hat{y} \mid X_1, X_2, \ldots, X_K) \qquad (7).$$

The relative inequality weight of the factor X_j may be written as:

$$s_{j,SHAPLEY} = \frac{\Phi(X_j, I)}{I(y)} \qquad (8)$$

such that

$$\sum_{j=1}^{k} s_{j,SHAPLEY} = \frac{I(\hat{y} \mid X_1, X_2, \ldots, X_k)}{I(y)} \qquad (9).$$

As pointed out by Israeli (2007), when income is expressed in log terms and the variance is used as inequality index, the Shapley decomposition according to formula (9) matches the Fields decomposition of the R-square according to formula (3) since

$$\sum_{j=1}^{k} s_{j,SHAPLEY} = \frac{I(\hat{y} \mid X_1, X_2, ..., X_k)}{I(y)} = \frac{\sigma^2(\ln \hat{y})}{\sigma^2(\ln y)} = R^2 \qquad (10).$$

This does not mean that the factor contributions evaluated through the two approaches coincide. They do so only in the absence of correlation between the explanatory variables.

3. Model selection

The first step in the regression-based decomposition of income inequality requires the specification and the estimation of an income generating function, that is a model where income is regressed on some explanatory variables accounting for individual and household characteristics. For the estimation of the income generating function, we decided to exploit the potential of panel data by pooling the observations on a cross-section of individuals over several time periods.

We specified a panel data regression model with time-invariant unobserved effects (Wooldridge, 2002), which can be written as:

$$\ln y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it} \quad t = 1,2,...,T \quad i = 1,2,...,N \qquad (11)$$

where $\mathbf{x}_{it}$ is a $1 \times K$ vector of regressors, c_i is the time-constant, individual-specific effect and u_{it} is the disturbance term for which the strict exogeneity condition is assumed to hold, that is

$$E(u_{it} \mid \mathbf{x}_{i1}, ..., \mathbf{x}_{iT}, c_i) = 0 \quad t = 1,2,...,T \qquad (12).$$

This assumption implies that each error term u_i is uncorrelated with the regressors at all time periods, namely

$$E(\mathbf{x}'_{is} u_{it}) = 0 \quad s,t = 1,2,...,T \qquad (13).$$

The two core specifications of such models are known as Random Effects (RE) and Fixed Effects (FE) models. In particular, we have specified a RE model, where the individual effect c_i is treated as a random variable that adds to the error term u_{it}. This choice is justified primarily by the RE model using both the “between variation” (the variability across individuals) and the “within variation” (the variability over time). For this reason, unlike the FE model, it allows both to estimate the coefficients of the regressors that do not vary at all over time (with null within variation) and to measure with no efficiency loss the effects of regressors that display a small within variation. In this study, the dependent variable is represented by the (log of) individual net disposable income, whereas the regressors include, among others, gender (that is invariant over time) and the years of completed studies (that exhibit a little variation over time).

Our preference for the RE model is also explained by the fact that we are not interested in estimating the values of the unobserved term for some specific individuals, but instead we concentrate on the influence of individual and household factors on the disposable income of hypothetical individuals with given characteristics. In the situation where the individuals are drawn randomly from a large population, as is usually the case for household panel studies, the RE model is an appropriate specification (Baltagi, 2008).

The RE estimator is derived under the further assumption of uncorrelation (orthogonality) between the individual effect c_i and the observed explanatory variables $\mathbf{x}_{it}$:

$$E(\mathbf{x}'_{it}c_i) = 0 \qquad t = 1,2,...T \tag{14}.$$

This means that all the regressors $\mathbf{x}_{it}$ are considered to be exogenous. Many applications of the income generating function in a longitudinal framework have studied the effects of human capital accumulation on individual wages through the specification of a panel data model where the unobserved, individual term was intended to capture such features as individual ability. According to those studies, unobserved heterogeneity among individuals is likely to be correlated with some observed explanatory variables, which can generate potential endogeneity problems. Our study differs in that it is not restricted to the analysis of the returns to schooling and/or work experience on a sample of employees but it focuses on the estimation of the impact of several factors (that include both human and non-human capital as well as individual attributes) on the inequality in a measure of living conditions.

The above statement drove the choice of defining this measure as the individual disposable income, made up of wages, income from self-employment, transfers and income from capital. Traditionally the economic inequality is evaluated on household equivalent income or consumption, if one is willing to assume that the well-being of an individual depends on the combined resources of all the household members. The choice of measuring the inequality on individual income and, consequently, defining the individual as the unit of analysis is motivated here by the interest in explaining the determinants of inequality in the individual capacity to earn income, regardless of how the individual resources may be pooled together and then shared within the household.

4. Data, variables and summary statistics

The data used in this work are drawn from the Survey of Household Income and Wealth (SHIW) conducted every two years by the Bank of Italy on a sample of about 8,000 Italian households.

For every survey, the sample is composed of both households that have been already interviewed in previous years (panel households) and fresh households.

This survey is the only relevant source at the national level for household and individual longitudinal income data over a relatively large time interval. In particular, we referred to the Historical Database of the survey (Banca d'Italia, 2010) from which we selected information on the income earners born between 1938 and 1980 who have been successfully interviewed from 1998 to 2008 (they were between 18 and 60 years old in 1998). Such information took the form of a balanced micro panel where a large number of individuals N (N=1226) have been observed over a short time period T (T=6 years covering on the whole a time span of 10 years).

The selection of the individual longitudinal data was not an easy task. In the SHIW database every household is assigned a fixed identification number across waves whereas a fixed personal identification number for the individuals is missing. At every wave an individual is identified by both the household number and the order number of the individual within the household. The possibility of linking longitudinally the personal information exists however for couples of subsequent waves. Therefore the construction of the longitudinal dataset was achieved through the merging of individual data collected for couples of subsequent waves (1998-2000, 2000-2002, 2002-2004,...) followed by a further check for basic personal attributes such as gender and birth year.

In a longitudinal framework across waves, the situations of exits from the sample and/or later entries in the sample are not easily manageable since it cannot be ruled out that different individuals not constantly present in a household may be given the same order number within the household.

This remark drove the selection of a balanced panel (where only the income earners who participated continuously in the survey were included) though this choice may raise some theoretical issues relating to the attrition bias that could be addressed more deeply.

The choice of the semi-log functional form along with the selection of the explanatory variables were informed by the human capital theory suggesting that the ability to earn income is influenced by educational level and age. Gender is expected to play a special role in Italian income inequality due to the large gaps between men and women in the economic participation and opportunity, especially with reference to wage equality for similar work (Hausmann *et al.*, 2010). The remaining individual factors introduced as explanatory variables in the income generating model are work status (separating those who are employed and presumably receive an income from work from those who are not employed and whose income comes from other sources) and position in the household (accounting for whether or not the individual is the head of the household, according to his/her declaration). A measure of household wealth was also included accounting for the stock of non-human capital that is supposed to generate flows of income in the form of interests or rents. The net household's wealth is defined as the sum of real assets (property, businesses and valuables) and financial assets (deposits, securities, shares, etc) net of financial liabilities (such as mortgage loans and other debts). Then the geographical area of residence has been considered.

Table 1 - Descriptive statistics

VARIABLE	Definition	Obs	Mean	Std. dev.
logY	(Log of) Net disposable income	7356	9.709	0.763
Gender	=1 for male; =0 for female	7356	0.622	0.485
Education	Years of completed study	7356	10.259	3.840
Age	Age (in years)	7356	49.781	10.360
Household head	=1 for head of household =0 for other household member	7356	0.618	0.486
Work status	=1 for employed; =0 for not employed	7356	0.694	0.461
Geographical area	=1 for North and Centre; =0 for South and Islands	7356	0.725	0.4462
Household wealth	Real and financial wealth (in thousands of euro)	7356	268.330	369.917

Conditional on the information provided by the SHIW Historical Database, both the number and the kind of explanatory variables included as determinants seem to be broad enough to account for the main factors that are likely to explain income inequality.

Descriptive statistics for the variables introduced in the model are presented in Table 1.

The net disposable income is defined as the sum of individual income from wages, self-employment, pensions and other transfers, and property income, from both real and financial assets. Every income item is reported after tax and social security contributions. Negative or null income values were given null log (income) values.

5. Results

5.1 Model estimation

The Random Effects model in Table 2 shows the log of individual net disposable income as a function of demographic, human capital, work status, location and household wealth variables. The reported regression coefficients come from the estimation of the saturated model, that is the model including all the explanatory variables. Robust standard errors are computed in order to correct for potential heteroscedasticity.

Table 2 - Random effects model estimation

EXPLANATORY VARIABLE	Coefficient (robust std error)
Gender	0.3803*** (0.0312)
Education	0.0459*** (0.0032)
Age	0.0240*** (0.0018)
Head of household	0.3325*** (0.0228)
Work status	0.3743*** (0.0407)
Geographical area	0.2174*** (0.0305)
Household wealth	0.0004*** (0.0000)
Constant	7.0676*** (0.1318)

R^2 overall = 0.3534
N=1226; T=6
Wald chi-squared(7)=1323.9; *p-value*=0.00
***: significant at the 1% level

The signs of the estimated coefficients are in line with the theoretical expectations. Significant income gaps are due to gender, level of education, age, position in the household, work status and area of residence: ceteris paribus, on average the males, the more educated, the oldest, the heads of household, the employed and those who live in northern or central regions enjoy higher income levels. Larger income flows are also associated with larger stocks of wealth.

An overall R^2 equal to 0.35 indicates a satisfactory fit of the income regression model, when compared with other studies on the same phenomenon. We might have improved the fit by including interaction terms, but this would have created some problems in correctly assigning the resulting effect to the variables included in the interaction term.

5.2 Decomposition results through Fields method

Since the RE estimator is equivalent to an OLS estimator applied to conveniently transformed variables (Wooldridge, 2002), the results of the decomposition analysis according to the Fields method have been obtained from the OLS regression of the transformed variables: the new variables are obtained by removing from the original ones a fraction θ of their average over time, where θ is estimated as a function of the variances of both the error and the individual effects term.

The inequality weight of each factor (column 1 of table 3) was calculated through the formula (2) as a function of the corresponding OLS coefficient, the covariance between the log income and the factor, and the variance of log income. The inequality weights associated sum up to 36.3, which is the value of R^2 from the above regression. The remaining proportion (63.7) is attributed to the residual term, which means that a large portion of inequality is not explained by the variables included among the income determinants. Column 2 reports the percentage contributions of each factor to the explained inequality level. The most important variables in determining the explained income inequality are gender (21.3%) and household wealth (21.1%), followed by educational level (19.4%) and household head (16.5%). Smaller weights are attached to working status (9.4%), age (8.8%) and geographical area (a bare 3.5%).

Table 3 - Factor contributions to inequality using the Fields method

FACTOR X_j	Factor inequality weight s_jx100	Percentage contribution net of residual
Gender	7.7	21.3
Education	7.0	19.4
Age	3.2	8.8
Head of household	6.0	16.5
Work status	3.4	9.4
Geographical area	1.3	3.5
Household wealth	7.7	21.1
Residual	63.7	
Total	**100.0**	**100.00**

5.3 Decomposition results through Shapley approach

The results from the inequality decomposition using the Shapley value approach are reported in Table 4.

Since the decomposition results are influenced by the choice of the inequality index, the estimates are presented for four inequality measures: Gini index, Theil index, the mean logarithmic deviation, and the variance of logarithms.

The table shows the contributions to the income inequality in absolute terms (first column) and in percent of both the observed inequality (second column) and the explained inequality (third column).

When using either Gini or Theil index, the contributions of individual and household factors altogether account for more than 80% of the observed inequality. The explained inequality is smaller both for the mean log deviation (57.8%) and especially for the variance of logarithms (36%). In the latter case, as expected, the percentage of unexplained inequality is very similar to that resulting from the Fields method. Indeed, when applied to the variance of log income, the Shapley value approach is equivalent to the Fields decomposition of the R-square. The actual differences in the factor contributions are due to the presence of correlation among the regressors, which is not accounted for by the Fields method.

Table 4 - Factor contributions to inequality using the Shapley method

FACTOR X_i	Inequality measure					
	Gini			Theil		
	s_jx100	In % of (2)	In % of (1)	s_jx100	In % of (2)	In % of (1)
Gender	5.5	16.8	20.4	2.9	14.8	18.4
Education	4.1	12.5	15.3	1.8	9.1	11.3
Age	5.3	16.4	20.0	0.8	3.9	4.9
Head of household	4.2	12.7	15.5	2.2	11.1	13.8
Work status	1.4	4.3	5.2	0.9	4.7	5.8
Geographical area	1.3	3.9	4.8	0.4	1.8	2.2
Household wealth	5.0	15.4	18.8	6.9	35.2	43.6
(1) Total Explained Inequality	26.7	82.0	100.0	15.8	80.7	100.0
Unexplained Inequality	5.9	18.0		3.8	19.3	
(2) Observed Inequality	32.6	100.0		19.5	100.0	

FACTOR X_i	Inequality measure					
	Mean log dev			Var log		
	s_jx100	In % of (2)	In % of (1)	s_jx100	In % of (2)	In % of (1)
Gender	2.6	12.2	21.3	5.0	8.6	24.0
Education	1.7	7.8	13.6	3.1	5.4	15.0
Age	1.9	8.8	15.3	4.7	8.1	22.4
Head of household	1.8	8.5	14.8	3.3	5.7	15.9
Work status	0.7	3.2	5.5	1.2	2.1	5.9
Geographical area	0.4	1.9	3.3	0.8	1.4	3.9
Household wealth	3.2	15.0	26.2	2.7	4.6	12.8
(1) Total Explained Inequality	12.2	57.3	100.0	20.9	36.0	100.0
Unexplained Inequality	9.1	42.7		37.3	64.0	
(2) Observed Inequality	21.3	100.0		58.2	100.0	

For Gini index as well as for the variance of log income, the factors that explain the largest part of income inequality are gender and age, whereas for the indexes belonging to the class of entropy measures (that is Theil index and mean log deviation) the main determinants are household wealth and gender. By comparing the weight of human capital and non-human capital factors, the human capital variables (age and education, jointly considered) show the highest contribution to the explained inequality but for the Theil index, for whom the relative contribution of household wealth is especially large (43.6%).

While apparently different, individual and household factors are strictly intertwined. Indeed the human capital endowments are quite strongly correlated with both real and financial assets of the family of origin.

For all the inequality measures, the contribution of position in the household is estimated between 14% and 16%, whereas the remaining variables - occupational status and geographical area - are much less important as determinants of the inequality. This seems to suggest that, once human capital, gender, wealth and position in the household are taken into account, whether an individual is unemployed or not, and whether he or she lives in the North or in the South, have only a minor impact on income differentials.

6. Conclusions

Unlike the traditional inequality decomposition methods, the regression-based approaches allow to measure the inequality contribution of any explanatory factor. For this reason, regression-based methods are able to highlight what factors are most important in determining the observed income differentials.

However, there is a portion of income inequality that is not captured by the explanatory factors. Whenever the R-square of the income regression model is not very high, the Fields method is expected to leave a large share of inequality unexplained. On the other hand the performance of the Shapley approach is not directly linked to the fit of the regression model, being evaluated through the marginal impact of each factor, which differs depending on the choice of the inequality index. In our analysis, what constitutes the first result, the unexplained percentage of inequality is much lower when the Shapley approach is used and the Gini or Theil measure is calculated (respectively 18% and 19.3%) than when the Fields method is used (about 63%).

Our results on the drivers of income inequality shed light on the dominant role of gender, human capital, and wealth. Whatever the decomposition method and the inequality measure, the gender is found to play a key role as a determinant of income inequality, its contribution being estimated between 18.4% and 24%. This is likely to be a distinctive feature of Italy, as pointed out by many comparative studies. In Italy, women are known to find difficulties in combining work and family duties, and for this reason their participation in the labour force is low. On the other hand, women who have a job earn on average lower salaries and have usually fewer opportunities to reach leadership positions than men with comparable skills. Further applications of inequality decomposition methods for cross-country comparisons would be needed in order to support this evidence.

Along with gender, the endowments of human capital (education and experience) as well as physical capital (household assets) are found to be crucial determinants of income differentials, too. The role played by the household wealth stock is of primary importance when the Shapley approach is applied to the generalized class of entropy measures and to Theil index in particular. This remark highlights another advantage of the Shapley value approach over the Fields method: the former procedure allows to evaluate whether the marginal impact of each factor is equally important for every inequality measure or else to stress the different sensitivity of every inequality index to the underlying factors; on the contrary, through the latter method, the decomposition results are the same for a large number of inequality measures. A further advantage in applying the Shapley approach is the fact that, in the presence of many explanatory variables that may be correlated (and this is the case for our analysis) the Shapley approach accounts for the correlation among the determinants whereas the Fields method does not.

In order to complete the comment on the role of the different determinants, we found that work status and geographical area play minor roles in explaining inequality. The small weight attached to the area of residence may be surprising since this result seems to controvert the ingrained belief that the North-South divide is a major driver of economic inequality in Italy.

Finally, the regression-based decomposition methods employed in this paper may be further enhanced in order to provide policy insights. If among the explanatory factors some policy-relevant variables are introduced, e.g. labour market or redistributive intervention policies, the results of the decomposition may then be used in order to assess what decisions would be more effective in fighting the causes behind income inequality.

References

Baltagi B.H. (2008), *Econometric analysis of panel data*, Third edition, John Wiley & Sons Ltd, Chichester, England.

Banca d'Italia (2010), Historical Database of the Survey of Italian Household Budgets, 1977-2008, SHIW-HD, version 6.0, February 2010, On line at: http://www.bancaditalia.it/statistiche/indcamp/bilfait/docum/Shiw-Historical-Database.pdf

Blinder A.S. (1973), "Wage Discrimination: Reduced Form and Structural Estimates", *Journal of Human Resources*, 8, pp. 436-455.

Boeri T., Brandolini A. (2004), "The Age of Discontent: Italian Households at the Beginning of the Decade", *Giornale degli Economisti e Annali di Economia*, 63, pp. 449-487.

Celidoni M., Procidano I., Salmasi L. (2011), "Determinants of inequality in Italy: An approach based on the Shapley decomposition", *Review of Applied Socio-Economic Research*, vol.1, n.1, pp. 63-69.

Devicienti F. (2010), "Shapley-Value Decomposition of Changes in Wage Distribution: A note", *Journal of Economic Inequality*, 8 (1), pp.199-212.

Fields G. (2003), "Accounting for Income Inequality and Its Changes: A New Method with Application to the Distribution of Earnings in the United States", *Research in Labor Economics*, 22, pp. 1-38.

Guanatilaka R., Chotikapanich D. (2009), "Accounting for Sri Lanka's Expenditure Inequality 1980-2002: Regression-Based Decomposition Approaches", *Review of Income and Wealth*, 55 (4), pp. 882-906.

Hausmann R., Tyson L.D., Zahidi S. (2010), *The Global Gender Gap Report*, World Economic Forum, Geneva.

Israeli O. (2007), "A Shapley Based Decomposition of R-Squared of Linear Regression", *Journal of Economic Inequality*, 5 (2), pp.199-212.

Massari R., Pittau M.G., Zelli R. (2009), "A dwindling middle class? Italian evidence in the 2000s", *Journal of Economic Inequality*, 7 (4), pp. 333-350.

Morduch J., Sicular T. (2002), "Rethinking Inequality Decomposition, with Evidence from Rural China", *The Economic Journal*, 112, pp.93-106.

Oaxaca R. (1973), "Male-Female Wage Differences in Urban Labour Markets", *International Economic Review*, 14, pp.693-709.

Quintano C., Castellano R., Regoli A. (2009), "Evolution and Decomposition of Income Inequality in Italy, 1991-2004", *Statistical Methods and Applications*, 18 (3), pp. 419-443.

Sastre M., Trannoy A. (2002), "Shapley Inequality Decomposition by Factor Components: Some Methodological Issues", *Journal of Economics*, 9, pp. 51-89.

Shorrocks A. F. (1982), "Inequality Decomposition by Factor Components", *Econometrica*, 50, pp. 193-211.

Shorrocks A. F. (1984), "Inequality Decomposition by Population Subgroups", *Econometrica*, 52, pp. 1369-85.

Shorrocks A. F. (1999), "Decomposition Procedures for Distributional analysis: A Unified Framework Based on the Shapley Value", mimeo, University of Essex.

Wan G.H. (2004), "Accounting for Income Inequality in Rural China: a Regression-Based Approach", *Journal of Comparative Economics*, 32, pp.348-363.

Wan G. , Lu M., Chen Z. (2007), "Globalization and Regional Income Inequality: Empirical Evidence from Within China", *Review of Income and Wealth*, Vol. 53, No. 1, pp.35-59.

Wan G., Zhou Z. (2005), "Income Inequality in Rural China: Regression-Based Decomposition Using Household Data", *Review of Development Economics*, 9 (1), pp.107-120.

Wooldridge J.M. (2002), *Econometric Analysis of cross section and panel data*, The MIT Press, Cambridge, Massachusetts, London, England.