Estimating Inequality of Opportunity in Ghana Across Cohorts: A Machine Learning Approach

Vito De Sandi

May 2023

Abstract

In the last decades, the inequality of opportunity issue is capturing more and more the attention of researchers and policymakers. While extensive theoretical and empirical work has been carried out in this field, only a small fraction has been conducted in developing countries, particularly in Africa, due to a lack of data. The paper presents new estimates of inequality of opportunity in Ghana across cohorts and implementing a machine learning approach to measure and analyze IOp. The study finds that absolute IOp follows a decreasing pattern over generations, with more recent cohorts experiencing more equality in terms of opportunity than older ones. The relative contribution of each circumstance in every cohort is discussed, birth place and ethnicity seem to matter considerably, particularly in the more recent cohorts. Overall, Ghana is shown to be a complex society with a complex interaction of circumstances generating opportunity.

1 Introduction

In the last decades, the inequality of opportunity issue is capturing more and more the attention of researchers and policymakers. Equality of opportunity can be described as the social ethics which seeks to level differences in outcomes attributable to luck, but not those for which individuals are responsible. The concept of equality of opportunity, as discussed by Dworkin (1981), Arneson (1989) and Cohen (1989), emphasizes the individual responsibility in determining economic advantage and disadvantage. However there is no a general consensus in the literature on what luck and responsibility really are. Following Roemer (1998), we distinguish between inequalities due to factors beyond the individual control such as socioeconomic status, race, gender, birth place and factors within their control, and we consider the former as

unfair, because they represent forms of inequality that hinders social mobility and perpetuates systemic disadvantage for certain groups (Ramos and Gaer, 2012; Fleurbaey and Maniquet, 2011). Moreover, IOp ideal can be broken down into two ethical principles. The first principle of compensation asserts that differences in individual accomplishments that can be clearly linked to factors beyond individual responsibility (circumstances) should be compensated by the society. The second principle, natural reward principle, requires to not compensate differences in achievements due to factor within the individual's responsibility (effort) (Checchi and Peragine, 2010). Hence, we can define an IOp measure as an estimate of how far a given distribution of outcomes is from equal opportunity counterfactual distribution, in which differences in outcomes caused by circumstances have been compensated. Recent papers highlight the extensive theoretical and empirical work that has been carried out in this field (Peragine and Ferreira, 2015; Ramos and Gaer, 2012; Roemer and Trannoy, 2016) Nevertheless only a small fraction of this studies have been conducted on the developing countries and in particular in Africa. Even though the overall inequality in African states is well documented (Thorbecke, 2013; Easterly and Levine, 1997), the sources of such inequality are yet unknown. Very few studies have analyzed the phenomenon with cross country comparisons.(Brunori, Palmisano, and Peragine, 2019; Roemer and Trannoy, 2016). This is due to the fact that in studying inequality of opportunity, an high level of information for both circumstances and outcome variables is required and the developing Countries often lack of this kind of data. However, using 4 large surveys conducted in Ghana (GLSS), in this work I'll present the IOp measurements in this State and discuss the pattern of such measures over time and generations.

A relevant challenge for the IOp analysis is the robustness of the estimation. The current approaches to estimating inequality of opportunity are often hindered by ad-hoc model selection, which can lead researchers to either overestimate or underestimate the real amount of inequality of opportunity. For such reason we implement a machine learning approach (regression trees and forest) for measuring and analyzing inequality of opportunity (Brunori, Hufe, and Mahler, 2023).

2 Objective

The present work will expand the inequality of opportunity literature in a double direction. First, we contribute to the measurement on inequality of opportunity in developing countries, in particular in Ghana. Differently

form other works the analysis will be conducted by cohorts to capture the evolution of the IOp for individuals born in different political and economical conditions. Four large Cross-section datasets (1998, 2005, 2013, 2017) have allow as to built the following 4 cohorts:

1 Born before 1957 (Colonial period),
2 Born between 1957-1968 (Independence period),
3 Born between 1969-1980 (Second Republic period)
4 Born between 1981-1991 (Rowling period)

Previous researches on inequality have shown a varied treatment of age when building cohorts. Some studies either overlook it or treat it as a circumstance. By contrast,we run a regression analysis using the age variable as a predictor to define cohort-level measure and eliminate the age effect (Aaberge, 2011; Andreoli et al., 2021).

Hence we have been capable to catch the evolution of the IOp across generations and understand the role of each circumstances in this results. As far as I know, There are no other works which attempt this measurement in an African state. This is not only a empirical exercise. The results can help to understand "the social and economic mechanisms that generate inequalities and can help in identifying priorities in anti-poverty policies..."(Brunori, Palmisano, and Peragine, 2019). The second contribution is merely related to the methodology for the estimations. IOp results biases from critical choices made during model selection. On one hand, excluding relevant circumstances from the model causes underestimation of inequality of opportunity (Ferreira and Gignoux, 2011). Conversely, including too many circumstances leads to overfitting of the results. Following the novel literature in this field we employ a supervised machine learning approach that through the implementation of regression trees and forests is able to overcome the discretionary in the model and circumstances selection.

3 Methodology

Consider a population $N \in N_+$ and a vector of non negative incomes $x_i \forall i \in 1, ..., N$. We assume that x_i is a function of effort $e_i \in \eta$ with $i \in 1, ..., N$ and circumstances $c_i \in \theta$ with $i \in 1, .., N$. Hence each income will be $x = f(c, e)$. The population can be sub grouped in type $\tau = t_1 ..., t_M$ two individuals belong to the same type if they share the same circumstances. A key aspect of the Iop measurement is the selection of relevant circumstances and types.

In this work we focus on ex-ante utilitarian measures of IOp. This approach requires first to built a counterfactual distribution assigning to individuals with the same circumstances the mean level of outcome of their type, then it assess the inequality between types by means of an inequality index. This will be the absolute amount of IOp. However, given the shortcomings earlier expressed we follow the proposal by Brunori, Hufe, and Mahler, 2023 for the ex-ante measurement through regression trees and forest.

3.1 Regression Trees

Tree algorithm employs all the circumstances and through sequential and hierarchical decisions based on statistical criteria it divides a data set into exhaustive and exclusive groups of observations. The algorithm assigns each observation to a unique terminal node. From this partitioning we get types t_m. To each individual in these types is assigned the average value of the depended variable of interest $\hat{x_m}$. Hence we obtain a counterfactual distribution of outcome $\hat{x} = f(\hat{x_i}) \forall i \in N$ that resembles the non parametric ex ante counterfactual distribution.

More concretely a conditional inference regression tree predicts a dependent variable X based on a set of j regressors C (Hothorn, Hornik, and Zeileis, 2006). The conditional distribution of X is assumed to depend on a function f of the regressors: $D(x|C = D(x|f(C_1, ..., C_j))$. The regression tree generate M types by recursive binary splittings, described by the following steps:

1. Select an appropriate alpha-value

2. The algorithm conducts a correlation test on the null hypothesis of independence for every circumstance being examined, using a predefined level of significance (alpha-value)

   $$H_0^c : D(x|C) = D(x)$$

   for all circumstances. We get a p-value for each test. (p-value are adjusted by Bonferroni correction)

3. Take the circumstance with the lowest p-value C^* (the one with the stronger association with the outcome) if the adjusted P-value $> \alpha$ the algorithm stops, if adjusted P-value $< \alpha$ it selects C as splitting circumstance;

4. If a circumstance is selected as splitting point, conditional inference trees decide where to operate the partition. In case of a continuous

variable all possible value are checked applying a difference-in-means t-test and obtaining an associated p-value. The value with lowest p-value will be selected for the partition.

5. For all splitting points, the algorithm tests the null hypothesis of point two for the conditional expectation and stores the p-value associated with each test. Again the circumstance with the lowest p-value is chosen, and if the adjusted p-value $< \alpha$, C is selected as splitting circumstance.

6. Repeat step 2-5 for all sub samples.

The process avoids the arbitrariness in either the selection of circumstances and the selection of a model: The algorithm selects only those features that exhibit the most significant association with the outcome variable. The interactions between circumstances in determining the output is entirely due to the algorithm.

3.2 Forests

At least two issues are involved in the implementation of regression trees. First, the structure of the tree is highly dependent on the sample size. Indeed, regression trees have a tendency to exhibit low bias but high variance. To mitigate these issues, the algorithm employs bootstrapped subsamples of the original data to construct trees for each subsample. The procedure is called Random Forest. Each tree within the Random Forest follows the same procedure previously explained. In this work, we obtained the forest using three-fold cross-validation and by running the algorithm with 200 trees, allowing the forest to growth as much as they can.

4 Data

We make use of the Ghana Living Standard Survey (GLSS) a nation-wide household survey conducted in 1998, 2005, 2013, 2017. It collects detailed information on demographic characteristics of the population. The sample considered in each wave is restricted to adult individuals above 15 (working age) and younger than 70 (still economical active according to Ghana statistical services). The outcome is the total consumption per household calculated as follows: total consumption per household is divided by equivalence scale provided by the data set and then it is adjusted for the CPI PPP value considering the PPP in 2010 and the CPI in wave's year. As

expected has been impossible the use of income data due to the large number of very low family annual income (near to 0) that overemphasizes the true inequality making this variable unreliable. In line the relevant literature we took information about sex, ethnicity, birth place, father occupation, father education, mother occupation and mother education as circumstances (Brunori, Palmisano, and Peragine, 2019; Bourguignon, Ferreira, and Menéndez, 2007). All this variables were available in the considered waves, table 1 present for each cohort the sample and all the circumstances in detail.

We proceed in building 4 cohorts according to the relevant Ghana historical events. First we gather the individuals born under the British empire government: Born before 1957 (colonial period). The independence in 1957 signs the begin of the second cohort: Born between 1957-1968. Following independence Ghana remained a constitutional monarchy and parliamentary democracy until 1960, between 1959 and 1964 the development secondary schools became the main priority. The Second Republic: Born between 1969-1980 (Second Republic period), despite its brief tenure, played a crucial role in highlighting the development challenges confronting the nation. The National Redemption Council (NRC) has controlled the Country until 1980. In the last cohort we include individuals born between 1981-1991 (Rowling period). Flight Lieutenant Jerry John Rawlings attempted two coups in 1979 and 1981. His second military coup established a Provisional National Defense Council as the supreme national government. (Gocking, 2005; Herbst, 1993). The table presents the years each cohort is composed and sample size without missing values.

Table 1: Cohorts Description

Cohort	**Year**	**Sample Size**
Colonial period	≤ 1957	9796
Independence period	1957-1968	15468
Second Republic period	1969-1980	18661
Rowling Period	1981-1991	14604

5 Results

Table 2 reports, for each cohort, the partitioned types,the estimates of total inequality through the gini index, the inequality of opportunity computed with regression trees and after the implementation of the forests, as abso-

lute value and as a percentage of the overall inequality. All the measures of IOp trees and forest are depurated of the effect due the age component as described in the section 2. The overall inequality (Gini) even though decreasing appears particularly high with a peak of 0.52 in the first cohort. The main findings are the IOp measurements. The opportunity Gini coefficient from the trees ranges from 0.1595 in for the individuals born during the Rowling period to 0.166 in the Colonial period. Overall the absolute values show a progressively decrease of the IOp index. The IOp computed with forest (lowering the sample size effect) exhibits a similar decreasing trend of absolute inequality of opportunity. The last cohort seems to be the one with highest opportunities.

Table 2: Inequality of Opportunity Estimations

Cohort	Types	Gini	IOp Tree	IOp Forest	R.IOp tree	R.IOp forest
Col.Period	25	0.52	0.166	0.187	31.92	35.96
Ind.Period	28	0.5	0.162	0.182	32.4	36.4
2nd Rep.	34	0.48	0.1592	0.178	31.6	37.08
Rowling.P.	42	0.496	0.1595	0.168	32.1	33.8

The shares of total inequality that are explained by Inequality of opportunity help us to catch the real dimension of IOp in Ghana. These shares are remarkably high, all around 32% for the IOp with trees. The results with forest are a bit higher: the 2^{nd} Republican period appears more unequal in relative terms. By contrast the relative level of inequality of opportunity for the generation born in ten years Rowing's government seem surprisingly low, this is due to the increase of total inequality (0.496). Overall the last generations appear less unequal in term of opportunity.

Table 3: Relative Shapely Value Decomposition

Circum.	**Col. Period**	**Ind. Period**	**2nd Rep.**	**Rowl. Period.**
Birth Place	28.08	21.24	20.55	33.28
Father Occ.	12.22	17.13	16.75	9.76
Father Edu.	12.24	13.87	16.69	12.85
Sex	5.13	1.144	1.65	1.98
Mother Occ.	13.91	19.71	16.27	8.48
Mother Edu.	4.54	9.87	12.80	11.97
Ethnicity	23.86	17.006	15.26	21.64

Another important aspect of the inequality of opportunity is: how im-

portant each circumstance is in accounting for the inequality of opportunity measure? Table 3 presents the estimates of the Shapley value decomposition of the between-type Gini coefficient (Shorrocks et al., 1999). The results show some interesting patterns. In all cohorts the birth place and the ethnicity are factors that influence IOp the most (33% and 21% respectively for the last cohort), these results were expected for both, in a country with a large ethnically diversified population and where most of the population gravitates towards the main cities, like Accra. Both circumstances reduce their shares during the Independence and 2^{nd} Republican period but then turn to grow. Parent's occupation gathers less and less shares, only 9% and 8% respectively for father's occupation and mother's occupation in the Rowling period. The last generation seems to the be influenced more by the parent's education. In literature (Brunori, Ferreira, and Neidhöfer, 2023), sex is often the variable that contributes less to the IOp. Our results are not different (this is due to the imputation of the same outcome level to all family members, making impossible to catch the inter family value of inequality), but worth of mention is the high contribution of sex for the oldest generation (4.6%) in comparison to the more recent one.

6 Discussion and Conclusion

The 'novelty' of this work comes from the following elements. First, we have introduced new estimates of inequality of opportunity in Ghana focused on cohort-level inequality. We have shown that IOp follows a decreasing pattern; hence, the generations born in more recent cohorts experience more equality in terms of opportunity than the older ones. However, the reduction in total inequality has not determined a similar reduction in the relative IOp. Second, we set up the relative contribution of each circumstance in every cohort. Geography and ethnicity matter considerably, and they seem to matter more than in the past. Thus, for an individual born in the Rowling period, the main cause of inequality of opportunity is their birthplace and ethnicity, symptom of the polarization of the Country in the main cities. The socioeconomic background of parents weighs less especially in the last cohort. Overall, Ghana appears to be a much more complex society than at first sight, with a complex interaction of circumstances generating opportunity. About the methodological side, we have followed a data-driven approach: the regression trees and forest techniques, which are closely related to the ex-ante IOp definition, have allowed us to overcome the discretion in the selection of the model and in the construction of types.

References

Aaberge, Rolf et al. (2011). "Measuring long-term inequality of opportunity". In: *Journal of Public Economics* 95.3-4, pp. 193–204.

Andreoli, Francesco et al. (2021). "New estimates of inequality of opportunity across European cohorts (and some insights on the long-term impact of educational policy)". In.

Arneson, Richard J (1989). "Equality and equal opportunity for welfare". In: *Philosophical Studies: An international journal for philosophy in the analytic tradition* 56.1, pp. 77–93.

Bourguignon, François, Francisco HG Ferreira, and Marta Menéndez (2007). "Inequality of opportunity in Brazil". In: *Review of income and Wealth* 53.4, pp. 585–618.

Brunori, Paolo, Francisco HG Ferreira, and Guido Neidhöfer (2023). *Inequality of opportunity and intergenerational persistence in Latin America*. Tech. rep. World Institute for Development Economic Research (UNU-WIDER).

Brunori, Paolo, Paul Hufe, and Daniel Mahler (2023). "The roots of inequality: Estimating inequality of opportunity from regression trees and forests". In: *Scandinavian Journal of Economics*.

Brunori, Paolo, Flaviana Palmisano, and Vitorocco Peragine (2019). "Inequality of opportunity in sub-Saharan Africa". In: *Applied Economics* 51.60, pp. 6428–6458.

Checchi, Daniele and Peragine (2010). "Inequality of opportunity in Italy". In: *The Journal of Economic Inequality* 8, pp. 429–450.

Cohen, Gerald A (1989). "On the currency of egalitarian justice". In: *Ethics* 99.4, pp. 906–944.

Dworkin, Ronald (1981). "Part 1: Equality of Welfare". In: *Philosophy and Public Affairs* 10.3, pp. 185–246.

Easterly, William and Ross Levine (1997). "Africa's growth tragedy: policies and ethnic divisions". In: *The quarterly journal of economics*, pp. 1203–1250.

Ferreira, Francisco HG and Jérémie Gignoux (2011). "The measurement of inequality of opportunity: Theory and an application to Latin America". In: *Review of income and wealth* 57.4, pp. 622–657.

Fleurbaey, Marc and François Maniquet (2011). *A theory of fairness and social welfare*. Vol. 48. Cambridge University Press.

Gocking, Roger (2005). *The history of Ghana*. Greenwood publishing group.

Herbst, Jeffrey Ira (1993). *The politics of reform in Ghana, 1982-1991*. Univ of California Press.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis (2006). "Unbiased recursive partitioning: A conditional inference framework". In: *Journal of Computational and Graphical statistics* 15.3, pp. 651–674.

Peragine and Ferreira (2015). "Equality of opportunity: Theory and evidence". In: *World Bank Policy Research Paper* 7217.

Ramos, Xavier and Dirk Van de Gaer (2012). "Empirical approaches to inequality of opportunity: Principles, measures, and evidence". In.

Roemer, JE (1998). "Equality of opportunity, Harvard U". In: *Press, Cambridge*.

Roemer, John E and Alain Trannoy (2016). "Equality of opportunity: Theory and measurement". In: *Journal of Economic Literature* 54.4, pp. 1288–1332.

Shorrocks, Anthony F et al. (1999). *Decomposition procedures for distributional analysis: a unified framework based on the Shapley value*. Tech. rep. mimeo, University of Essex.

Thorbecke, Erik (2013). "The interrelationship linking growth, inequality and poverty in sub-Saharan Africa". In: *Journal of African economies* 22.suppl_1, pp. i15–i48.

Appendix A - Circumstances Description