Inherited Inequality: A General Framework and a 'Beyond-Averages' Application to South Africa

Brunori, P.[a,b], Ferreira, F.[a], Salas-Rojo, P.[a]

[a]III London School of Economics, [b]University of Florence

December, 2024

I'd like to start with a very bad joke...

... so the presentation can only get better.

[...] the pursuit of equality itself is a mirage. What's more desirable and more practicable than the pursuit of equality is the pursuit of equality of opportunity."" Margaret Thatcher, 1975.

"Socialismo significa igualdad de oportunidades, no de ingresos." Raúl Castro, 2008.

(In)equality of opportunity looks like an empty-meaning concept. Let's try to add something.

In the next minutes...

Inherited inequalities are studied in the mobility and Inequality of Opportunity (IOp) literatures.

There is a strong theoretical ground behind IOp, but it is not easy to put theory into practice.

Transformation Trees help us obtaining IOp estimates from sample data and overcome some (many) empirical limitations. We apply this algorithm to South Africa...

.... and many more countries! I will abuse your patience and show results from our GEOM project.

Inherited Inequalities General framework

The intergenerational persistence of any factor y (income, education, wealth...) is typically measured as:

- Association between $F(y_{children}, y_{parents})$.
  - Transition matrix (Markov), rank correlation, regression coefficient, etc.
  - We call it "Intergenerational Mobility".
- Inequality in $y_{children}$ predicted by "circumstances".
  - Circumstances are factors beyond individual's control.
  - We call it "Inequality of Opportunity".

Why care about IOp

If you are fully egalitarian, in principle you should want to erase **all** inequalities from your outcome of interest (like income).

Rawls (1971), Sen (1980), Dworkin (1987), Roemer (1993, 1998), Fleurbaey (1994) provide a philosophical framework to claim that all inequalities are not the same.

There is a well established consensus on Inequality of Opportunities being "unfair".

Higher IOp values are harmful for economic growth (Ferreira, 2007; Marrero and Rodriguez (2023) due to a misallocation of talent.

Two approaches to IOp

Take an outcome y and a vector c denoting *circumstances*. Define *types* as groups of individuals sharing the same circumstances. Define two approaches to Equality of Opportunity (EOp).

- Ex-Ante EOp: same average outcome across types (Van de Gaer, 1993).

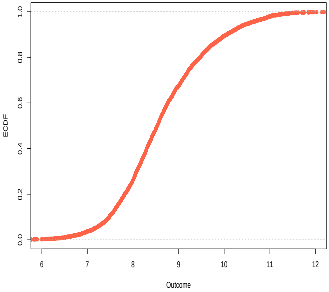
$$\text{EOp} \iff \forall(c, c'), \bar{y}|c = \bar{y}|c' \tag{1}$$

- Ex-Post EOp: same ECDFs across types (Roemer: 1993, 1998). Consider $F(.)$ as CDF of y. Then,

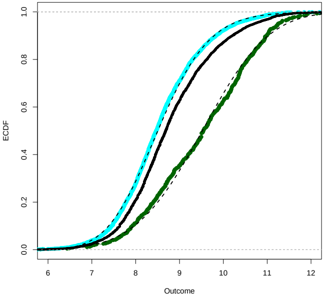
$$\text{EOp} \iff \forall(c, c'), F(y|c) = F(y|c') \tag{2}$$

Deviations from EOp are IOp.

Ex-Post IOp



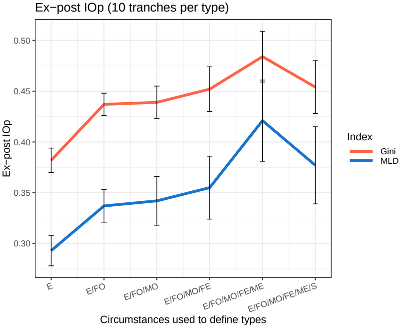
(a) Equality of Opportunity

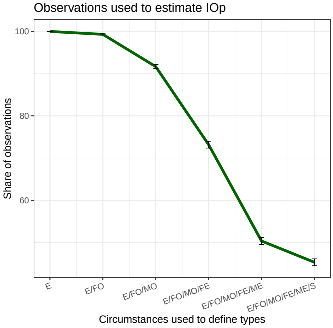


(b) Inequality of Opportunity

IOp estimation in survey data: Limits on empirics

- Partial observability of circumstances (Luongo, 2011).
- Uncertainty on how circumstances interact.
- In small types it is not always possible to get a distribution.

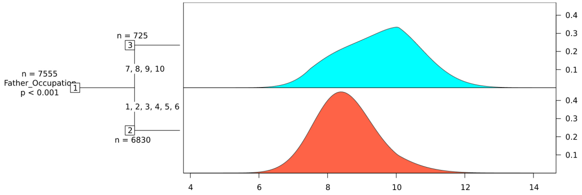




Transformation Trees: main idea

Transformation Trees (TrT, Hothorn and Zeileis, 2021) use the distribution of an outcome to partition the regressors space.

- First, the unconditional distribution is estimated.
- Second, the algorithm searches for binary splitting variables. A splitting is allowed if the shape of two resulting conditional distributions is sufficiently different.

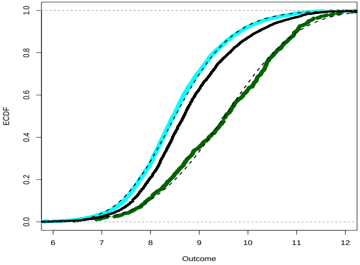


Transformation Trees: (very basic) outline

1. tune basic parameters (α, *m*,...); ▸ Go to tuning
2. approximate the unconditional distribution.
3. Run two statistical tests:
   - non-parametric test to select the most correlated regressor x. Adjust the p-value with Bonferroni. If $p - value > \alpha$, stop the algorithm. If not;
   - test the hypothesis of binary partitions derived from x being the same.
4. choose the variable and the splitting value producing the smallest *p-value* to obtain two subgroups,
5. repeat step 3:5 for the resulting subgroups.

Transformation Trees produces CDF to estimate IOp

Once TrT splits, interpolate the shape of the distributions.



$$\forall i, p, t, \;\; \tilde{y}_i^{p,t} = \frac{\mu^{p,t}}{\mu^{p}} \tag{3}$$

IOp is estimated with any inequality measure applied over $\tilde{y}_i^{p,t}$.

The Data: South Africa, 2017

National Income Dynamics Study (NIDS, 5^{th} wave) includes rich oversampling. Sample size: 7,297 observations.

- Outcome: age-adjusted equalized disposable household income ($2017 PPP).
- Circumstances: sex (2 categories), ethnicity (4), parents' education (13 + 13) and parents' occupation (11 + 11).

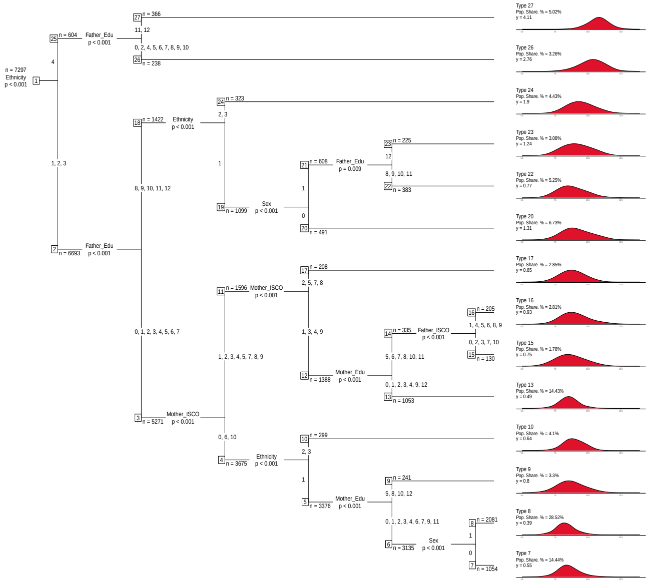
Main results: IOp estimates

Table: Main Results for South Africa

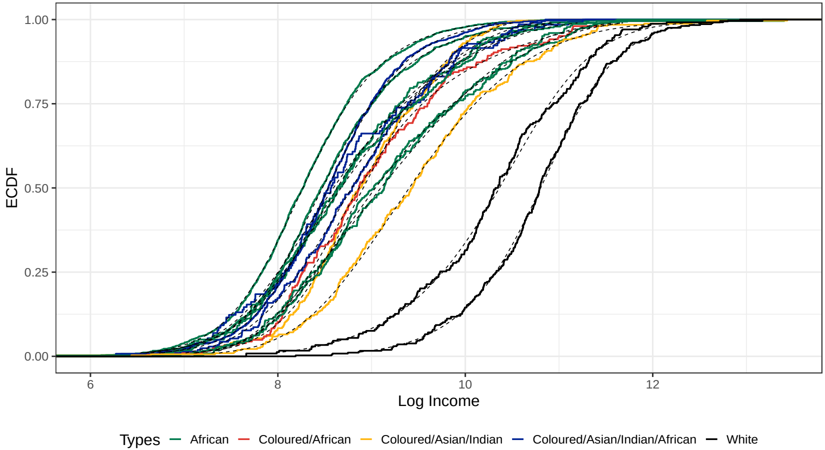
Approach	Types	Gini	Gini Abs	Gini Rel (%)	MLD	MLD Abs	MLD Rel (%)
Ex-post (Tree)	14	0.61	0.44	72.82	0.68	0.32	47.79
		(0.03)	(0.02)	(0.34)	(0.02)	(0.01)	(0.21)
Ex-ante (Tree)	12	0.61	0.41	67.44	0.68	0.27	40.39
		(0.03)	(0.02)	(0.31)	(0.02)	(0.01)	(0.18)
Ex-ante (Forest)		0.61	0.43	71.05	0.68	0.30	44.12
		(0.03)	(0.02)	(0.33)	(0.02)	(0.01)	(0.19)

Source: 2017 NIDS. N = 7,297. Note: Random Forests is grown with 200 trees. Bootstrapped standard errors in parenthesis.

Transformation Tree



CDF plot



ECDF
1.00
0.75
0.50
0.25
0.00
6
8
10
12
Log Income
Types
African
Coloured/African
Coloured/Asian/Indian
Coloured/Asian/Indian/African
White

Shapley-value decomposition

Values are percentages of absolute IOp (expost 0.44 Gini points, exante 0.41 Gini points)

Table: Share of Gini inequality (%) explained by circumstances

Method	Ethnicity	Father Occ.	Mother Occ.	Father Edu.	Mother Edu.	Sex
Ex-post						
	30.59	14.16	16.07	17.23	17.62	4.33
Ex-ante						
	19.34	15.04	13.55	23.5	23.41	5.15

Source: 2017 NIDS.

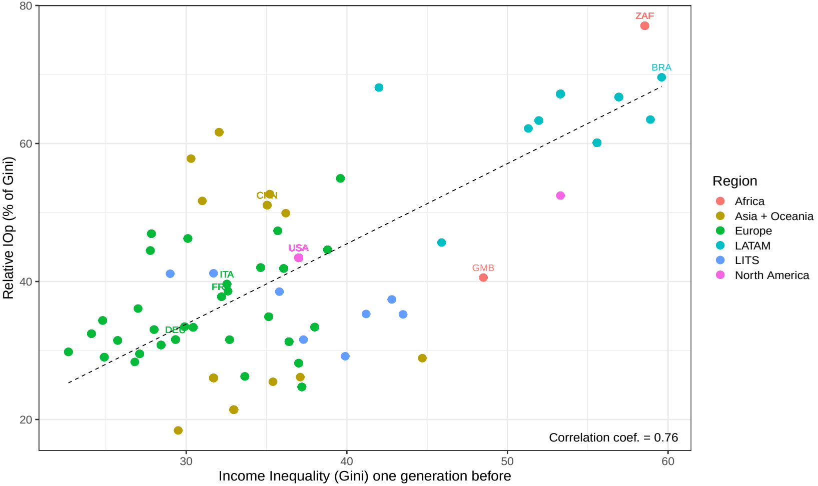
Abusing your patience 1/2: The GEOM Project

The Global Estimates of Opportunity and Mobility (GEOM) Project aims to explore inherited inequalities for as many countries and time periods as feasible, as comparably as possible.

https://geom.ecineq.org/

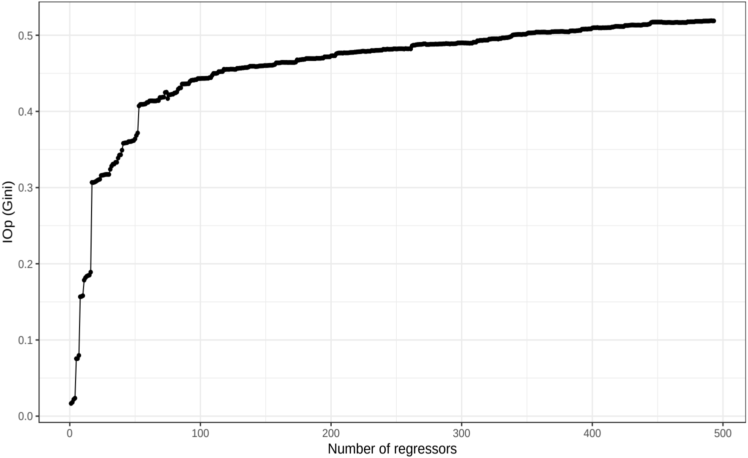
At the moment we have results for more than 80 countries, from more than 200 surveys around the world.

Abusing your patience 2/2: Great Gatsby Curve



Take-home idea. Type choices matter!

Unless you have really good admin/complete data or a solid theory, IOp values are highly volatile.



Q&A

Thank you!

All remaining questions - p.salas-rojo@lse.ac.uk

Appendix I: The Bernstein polynomial

The algorithm approximates the shape of the distribution with a linear combination of Bernstein basis polynomial that, for an order m and some positive continuous variable $y \in [a, z]$, is defined as a set of polynomials:

$$\left\{ b_{j,m}(y, a, z) = \frac{1}{(z-a)^m} \binom{m}{y} (y-a)^j (z-y)^{m-j} . \forall j = 1, ..., m \right\} \tag{4}$$

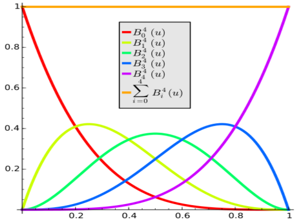
A linear combination of Bernstein basis polynomial has the form:

$$B_m(y, a, z) = \sum_{i=0}^{m} \beta_i b_{j,m}(y, a, z) \tag{5}$$

So, for a distribution approximated with a Bernstein polynomial of order m, we get m+1 parameters defining the shape of the objective distribution. Go to Bernstein Plot

Appendix I: The Bernstein polynomial

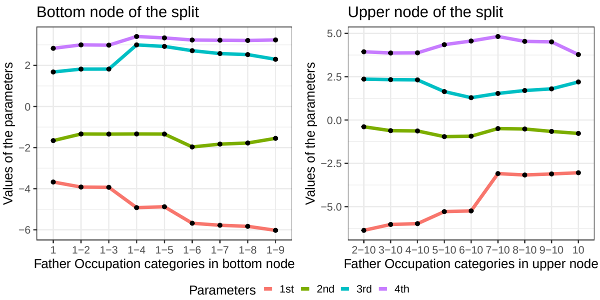
The shape of the distributions is approximated with a linear combination of Bernstein basis polynomials.

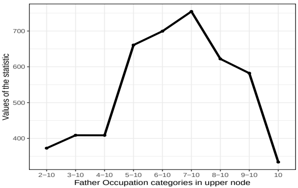


$$B_m(y, a, z) = \sum_{i=0}^{m} \beta_i b_{j,m}(y, a, z) \tag{6}$$

For a distribution approximated with a Bernstein polynomial of order m, we get $m+1$ parameters. Go to Bernstein

Appendix I: Parameters and splitting criteria





. Log likelihood function and parameters:

$$l_i(\theta) = log[f_z(a(y)^T\theta)] + log(a(y)^T\theta) \tag{7}$$

$$\hat{\theta}^N_{ML}(c) = argmax_{\theta\in\Theta}\sum_{i=1}^{N} l_i(\theta) \tag{8}$$

Null hypothesis.

$$H_0 : s(\hat{\theta}^N_{ML}|y) \perp C \tag{9}$$

Obtain the score.

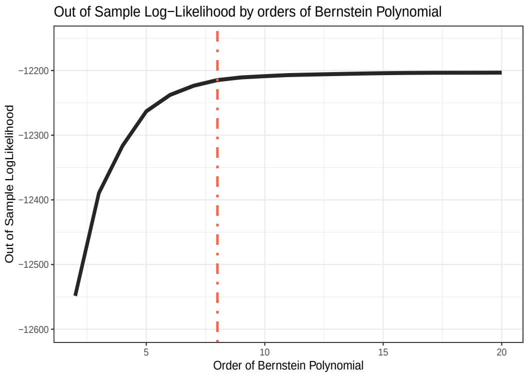
$$s(\hat{\theta}|y) = a(y)\frac{f_z'(a(y)^T\theta)}{f_z(a(y)^T\theta)} + \frac{a'(y)^T\theta}{a(y)^T\theta} \tag{10}$$

Appendix II: Transformation Trees: some limitations

- There is no debiased estimator for the ECDF inequality, but we crossfit the prediction (currently working on this on a different project).
- One can rely on a set of robustness techniques (Appendix).
- **Ex-Ante** IOp has its own tree *and* random forest (Brunori, Hufe, Mahler; 2023). These estimates include the double debiased (Terschuur and Escanciano, 2022).

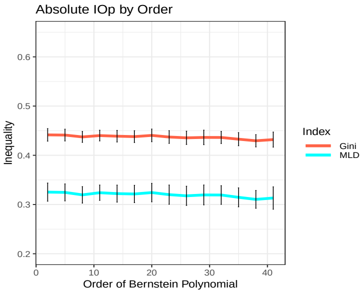
Appendix III: The Tuning

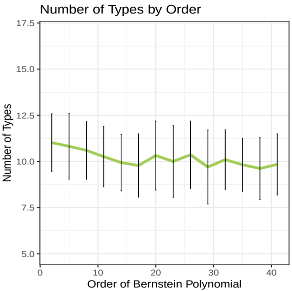
We reduce the out-of-sample loglikelihood approximation of the Bernstein Polynomial. We select the order where loglik improvement is smaller than 0.1%.



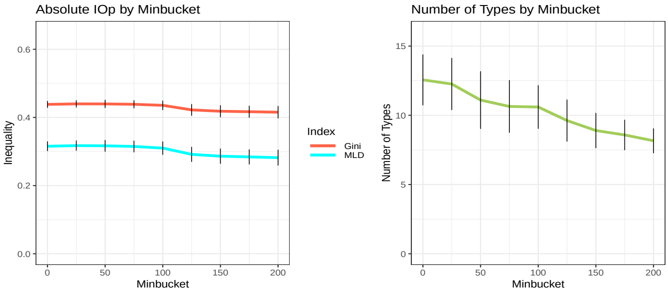
Return to algorithm outline

Appendix IV: Robustness (Order of Bernstein)

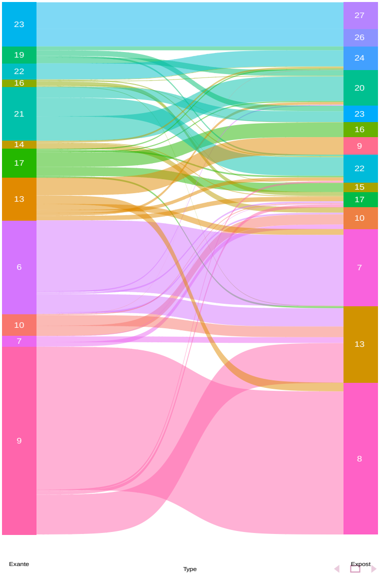




Appendix IV: Robustness (Minbucket)



Appendix V: Ex-ante vs Ex-post: Types comparison



Appendix VI: Previous IOp studies in South Africa

Piraino (2015) employs two different **ex-ante** approaches to IOp, creating 54 Romerian types. He focuses on individual labor income and a subsample of male (aged 20-44).

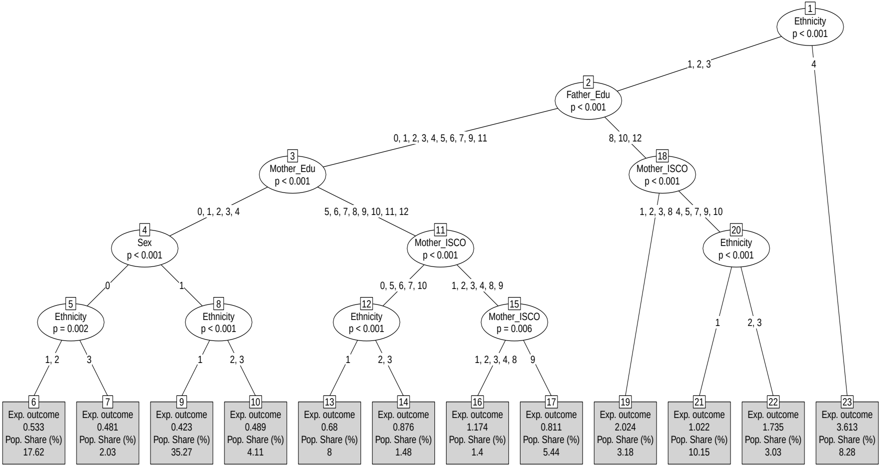
Table 4. *Index of inequality of opportunity*

Circumstances	Non-parametric	Parametric
1. Father's education	0.179 (0.030)	0.171 (0.032)
2. Father's education & occupation	0.219 (0.031)	0.187 (0.040)
3. Father's education & province	0.229 (0.033)	0.189 (0.031)
4. Father's education & occupation & province	___	0.227 (0.042)
5. Father's education & own race	0.237 (0.032)	0.241 (0.035)

Notes: Author's estimation from NIDS (2008–12). Bootstrap standard errors in parentheses.

Figure: Source: Piraino (2015). IOp is measured as a fraction of total MLD.

Appendix VI (bis): Conditional Inference Tree (for Ex-Ante)



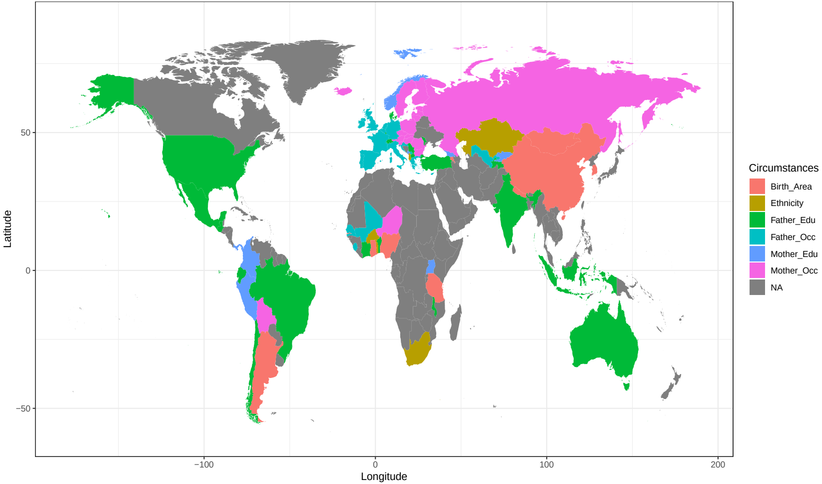
Appendix VII (bis): EOp (Lower Envelope)

From Roemer and Trannoy (2016)

" (...) We want to choose that policy which pushes up the lowest v^t function as much as possible – and as in Rawlsian maximin, (...). A natural approach is therefore to maximize the area below the lowest function v^t, or more precisely, to find that policy which maximizes the area under the lower envelope of the functions (...) "

In South Africa, around 28% of the population are in the worst-off type. The area below is around 2,203 (187 USD per month).

Appendix VII: Shapley Variable Importance in GEOM



Next Steps: Focus on GEOM (Short/medium term)

- In the short/medium term: work on GEOM specific papers, focused on big macro regions, to get insights on IOp dynamics.
  - Europe (now drafting the working paper, early 2025).
  - LATAM (for a Handbook on Inequality topics, to be published in the Chicago University Press, Editor: Steven Durlauf, mid 2025).
  - Asia and Oceania (Joint with Asian Development Bank, late 2025).
- And some other projects...
  - Wage inequality and technological change in Spain (Founded by La Caixa, mid 2025).
  - Project with the OECD on Wealth Inequality by regions (mid 2025).

Next Steps: Focus on Spain (medium term)

- Little evidence on inherited inequalities in the housing market/home ownershipx.
- Bauluz and Meyer (2024) show that the wealth accumulation patterns are changing for younger cohorts. Spain is a paradigm of this dynamic.
- So: I have tax records from the Panel de Hogares, including personal taxes, cadastral values, and so on. I am currently negotiating (advanced) to get the data used in The Opportunity Atlas.
- After matching both data sources I plan to explore the role of parental background on home ownership and wealth accumulation.