

UNIVERSITY OF OXFORD

DOCTORAL THESIS

Essays in Spatial Economics and Inequality

Author:

Luke HEATH MILSOM

Supervisors:

Prof. Ferdinand RAUCH

Prof. Gabriel ULYSSEA

Dr. Niclas MONEKE

*A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Economics*

June 2023

Declaration of Authorship

I, Luke HEATH MILSOM, here declare the existence of joint work in this thesis. The first and second chapters of this thesis are solely my own work. The third chapter is co-written with Mr Shihang Hou. We certify that the work on both the conceptualization and execution of the work within this chapter was split evenly between the two of us.

Luke Heath Milsom, April 2023

UNIVERSITY OF OXFORD

Abstract

Doctor of Philosophy in Economics

Essays in Spatial Economics and Inequality

by Luke HEATH MILSOM

This thesis examines inequality both over space and in labor markets. The first chapter finds that birth location strongly affects educational outcomes in Benin, Cameroon, and Mali — opportunity is unevenly distributed over space. Place matters more for girls than boys, and local characteristics such as urbanization, access to electricity, and agricultural employment play a key role. Religion is also found to be correlated with gender differences in the effect of place on educational attainment. The second chapter uses historical road maps, a novel identification strategy, and a quantitative spatial economics framework, to show how road building alters the spatial distribution of opportunity. Some roads decrease inequality while others increase it. The chapter also highlights the equity-efficiency trade-off that policymakers face when choosing which roads to upgrade. Finally, the third chapter explores the role of occupations in explaining wage inequality using a two-way fixed effect model on UK administrative data. It finds that between-occupation heterogeneity accounts for a substantial proportion of log-wage variance, particularly in higher-wage occupations and larger labor markets. Overall, this thesis sheds light on the complex and interconnected factors shaping spatial inequality and suggests policy implications for promoting equitable access to opportunity.

Word Count

This thesis consists of approximately 88,500 in its entirety.

Acknowledgements

This P.h.D. would not have been possible without the kindness, generosity, patience, and help of many people.

I have had the privilege and good fortune to work under the supervision and benefited from letters from Ferdinand Rauch, Gabriel Ulyssea, Niclas Moneke, and Douglas Gollin. I am extremely grateful for their time, wisdom, and knowledge, and in their ability to make the whole process as painless as possible.

Doing a P.h.D., and indeed the masters before it is an expensive business, and not one I could have hoped to have entertained without generous funding from various bodies. To say that such funding made this thesis possible is an entirely accurate statement, it was definitely a necessary condition. Thus, I would like to thank the ESRC Oxford-Radcliffe body as well as the Department of Economics at Oxford.

It does truly take a village — and I have been very lucky to draw on the insights of many other academics included among these are: Abi, Barbara, Ian, Hamish, Howard, Banu, Alison, Sarah, Paola, Dennis, Romu, Cheryl, Simon, and Marco. I would particularly like to thank Johannes Abeler for first showing how to do research, and Douglas Gollin for his sage advice, and being proof that kindness and academic success can indeed go hand in hand. Outside of Oxford, my academic interest in economics owes itself to Mrs Auld at Hardenhuish Comp. and was continued by Prof. Schafgans, and Dr. Shapiro at the LSE, helped along by classmates such as Lukas Bolte and Charlotte Woodacre.

One of the greatest blessings of my P.h.D. has been those that I have completed it with. A truly special and incredibly supportive cohort — together we overcame a P.h.D. and a pandemic! Sam, Shihang, Binta, Guilio, Sangha, Verena, Lisa, Nick, and Vatsal — not only have you made the P.h.D. experience a joy, but you also greatly improved this thesis through many discussions, SLSS, and JSS seminars.

To my parents, thank you for, well, everything — there is no one to which I (or this thesis) owe more. Thank you for your love, and for valuing what I did/do, whatever that may be. To my grandparents, thank you for your continued support — now, finally, I've got a job! To grandad, Brian, you told me to continue in my studies, that I wouldn't regret it, I did, and I didn't, thank you. To KASP, Marty, and the many other friends and family who have supported me throughout, thank you.

And, finally, above all, I would like to thank Hannah. Without you, there is no P.h.D., nor indeed any point in one. Thank you for being you, and for bringing me continuous joy,

purpose, and happiness.

Contents

Declaration of Authorship	iii
Abstract	v
Acknowledgements	vii
Introductory Remarks	xxi
1 Spatial Inequality of Opportunity in West Africa	1
1.1 Introduction	2
1.2 Setting and data sources	5
1.2.1 Why focus on Benin, Cameroon and Mali	5
1.2.2 Data description	5
1.2.3 Education as a measure of opportunity across space in Benin, Cameroon, and Mali	7
1.3 How unequally distributed is opportunity across space in Benin, Cameroon, and Mali?	10
1.3.1 Estimating the overall convergence rate	12
1.3.2 Validity of the movers design	15
1.4 Estimating locality-year specific place effects	16
1.5 Characteristics of place effects	19
1.6 Heterogeneity by gender and the impact of place on gender inequality of op- portunity	22
1.7 Conclusion	26
Appendices	27
1.A Additional results	27
1.B Validating Causal Place effects	29
1.B.1 Is there evidence of varying selection effects on observables	30
1.B.2 Focusing on within household variation	31

1.B.3	Time varying selection effects — increasingly likely to be exogenous moves	32
1.B.4	Over-identification outcome-based test	34
1.B.5	Locality-year specific identifying assumption	36
1.C	Signal to noise ratio of estimated place effects	36
1.D	Robustness	37
1.D.1	Linear model	37
1.D.2	Robustness of the LPM	38
1.D.3	Considering different measures of location quality	39
1.E	Relationship between observed primary completion rates and estimated causal place effects	40
1.F	Locality-year specific identifying assumption	41
1.G	Standard errors	42
1.H	Comparability of movers and stayers	43
1.I	Descriptive statistics	45
1.I.1	Temporal variation in primary completion	45
1.I.2	Sub-national migration patterns	47
1.J	Maps	49
1.K	Heterogeneity by country analysis	52
1.L	Investigation of inequality of opportunity in Cameroon in 2005	53
2	Moving Opportunity. Local Connectivity and Spatial Inequality.	57
2.1	Introduction	58
2.2	A general spatial theory linking changes in connectivity to local educational opportunity	65
2.2.1	Developing the sufficient statistic result	65
2.2.2	The sufficient statistic result	70
2.3	Taking the sufficient statistic result to the data	72
2.3.1	The setting, data, and descriptive evidence: Benin, Cameroon, and Mali since 1970	73
	The changing geography of connectivity: Digitizing historical Michelin road maps	74
	Descriptive evidence on the relationship between connectivity and local educational opportunity	79

2.3.2	Overcoming the endogeneity of road placement	83
	Sorting across locations	85
2.3.3	Results from estimating the sufficient statistic relationship	86
	Non-random exposure to plausibly exogenous variation in road building	88
	Top-coding of primary completion rates and non-linearities in market	
	access and causal place effect changes	89
	Inference	90
	Interpreting the sufficient statistic results	91
2.4	Counterfactual analysis	91
2.4.1	Finding parameter estimates	93
2.4.2	The impact of road building since 1970 on the distribution of opportunity and spatial inequality of opportunity	94
2.4.3	The impact of future road investments on spatial inequality of opportunity	99
	Efficiency-equity trade off	107
2.5	Conclusion	108
Appendices		111
2.A	Estimating local educational opportunity [Heath Milsom, 2021]	111
2.B	Empirics appendix	113
2.B.1	Descriptive statistics	113
	Spatial variation in primary completion	113
	Secondary source showing variation in road building	115
2.B.2	Maps of the spatial distribution in market access terms	116
2.B.3	Spatial inequality in local returns to education and opportunity	116
2.B.4	Correlating wage proxies with available wage data	119
2.B.5	Not-on-least-cost path identification strategy in figures	120
2.B.6	Local clientelism and public good provision	122
2.B.7	Top-coding and non-linearities	124
2.B.8	Endogenous network formation	126
2.B.9	Koranic schools and Medersas	127
2.B.10	Inference	128
2.B.11	Mean shifts in the spatial distribution of opportunity	131
2.B.12	Decomposing effects	132

2.B.13	Predicting the impact of connecting any two locations	134
2.B.14	Additional results and robustness	138
2.B.15	Replicating results from Faber [2014]	146
2.C	Data construction	150
2.C.1	Market Access data construction	150
	Calculating expected travel times	150
	Calculating incomes Y_{it}^s	151
	Estimating bilateral transportation and migration costs	153
2.C.2	Digitizing Maps	154
2.D	Theory appendix	157
2.D.1	Solving the spatial model	157
2.D.2	Model extensions	159
	Endogenous education supply	159
	Generalizing consumption patterns	162
	Explicit agglomeration forces	165
	Explicit endogenous amenities	166
	Land in consumption (fixed supply)	166
	Land in production (fixed supply)	167
	Land in consumption and production (endogenous supply)	167
	Intermediate goods	168
	Non-homotheticities and structural transformation	168
	Generalizing the education choice problem	170
2.D.3	Quantitative Spatial Economics model example	171
2.D.4	Future road upgrading counterfactuals details	173
2.D.5	Road-locality level analysis	173
2.D.6	No-roads counterfactual results with alternative parameters	176
	Using parameter values taken from the literature	176
	Alternative values of β	177
3	The Butcher, the Brewer, or the Baker: The Role of Occupations in Explaining Wage Inequality	179
3.1	Introduction	180
3.2	Econometric model	183
3.2.1	Limited mobility bias	185

3.2.2	Endogeneity and specification tests	186
3.3	Data and empirical setting	187
3.4	Log wage decomposition	189
3.4.1	Overall wage variance decomposition	189
3.4.2	Between and within-occupation log-wage decompositions	192
3.4.3	Evidence on varying wage premia across occupations within firms	194
3.5	Differences in the log-wage decomposition across occupations	199
3.6	Understanding spatial wage variation through worker-job decompositions	202
3.7	Conclusion	207
Appendices		209
3..1	Details on data cleaning	209
3..2	Testing the identifying assumption	210
3..3	Is the log-additive functional form a good specification for wages?	212
3..4	Review of limited mobility bias	213
3..5	Analysing the result of misspecifying the model with worker-firm FE when the true DGP is with worker-job FE	214
3..6	Comparing log-wage decomposition results against other results in the literature	216
3..7	Direct evidence that occupational changes are relatively unlikely	217
3..8	Additional results on the relation between job FE as wage premia re- lated to rent-sharing	218
3..9	Appendix figures and tables relating to wage inequality across geog- raphy	225
3..10	Is the relation between size of travel to work area and variance due to occupational composition?	227
Closing Remarks		i
Bibliography		iv

List of Figures

I	Education is associated with positive later life outcomes	9
II	Locality level observed primary completion rates	11
III	Plotting $\hat{\beta}$ from estimating equation 1.2	14
IV	Distribution of recovered causal place effects	18
V	Correlates with observed primary completion rates and causal place effects . .	21
VI	Correlates with $\mu_{lt}^m - \mu_{lt}^f$	25
1.A.1	Correlates with observed primary completion rates and causal place effects: Multivariate regression	27
1.B.1	Children with more educated mothers don't move to better areas earlier . . .	31
1.B.2	Considering increasingly likely to be exogenous moves	34
1.D.1	Causal place effects in other variables	40
1.E.1	Relationship between observed primary completion rates and estimated rela- tive place effects	41
1.F.1	Placebo vs actual causal place effects	42
1.G.1	Permutation inference	43
1.I.1	Change in the primary education completion distribution	46
1.I.2	Proportion of the population who have completed primary school by age . . .	47
1.I.3	The proportion of migration to urban areas	48
1.J.1	Mapping estimated place effects	49
1.I.4	Migration chord diagrams	50
1.J.2	Proportion of the population following Islam	51
1.J.3	Female primary completion rates as a fraction of male completion rates	51
1.K.1	Correlates with overall and causal place effects by country	52
1.K.2	Correlates with $\mu_{lt}^m - \mu_{lt}^f$ by country	53
2.3.1	Digitizing Michelin road maps — Cameroon 2019	75
2.3.2	Distribution of expected pairwise travel times between localities	78
2.3.3	Correlation between connectivity and local returns to education	81

2.3.4	Correlation between local returns to education and opportunity	82
2.3.5	Correlation between connectivity and local opportunity	83
2.4.1	The distributional impact of roads built since 1970 on the causal effect of place on primary education completion	96
2.4.2	Relationship between the total relative effect of roads and 1970 expected travel time	97
2.4.3	Comparing the effect of counterfactual random networks to the realized network	99
2.4.4	Counterfactual impact on inequality of opportunity of upgrading each road segment	102
2.4.5	Comparing counterfactual effects in the actual and re-scaled network for a random road upgrade	106
2.4.6	Equity-Efficiency trade off	108
2.B.1	Locality level primary completion rates	114
2.B.2	Variation in paved roads	115
2.B.3	Market Access — Benin 2012	116
2.B.4	Spatial variation in the local benefits of education over birth location	118
2.B.5	Estimates of local educational opportunity from Heath Milsom [2021]	119
2.B.6	Not-on-least-cost-path identification strategy in figures	122
2.B.7	Changes in the distribution of primary completion rates	125
2.B.8	Proportion of children enrolled in Koranic schools or Medersas	128
2.B.9	Randomization inference on OLS results	131
2.B.10	Aggregate distributional shifts due to road building since 1970	132
2.B.11	Distribution of effects	133
2.B.12	Results from estimating the sufficient statistic relationship	138
2.B.13	Correlation between connectivity and local returns to education: pagerank centrality	142
2.B.14	Correlation between connectivity and local opportunity: pagerank centrality .	143
2.B.15	Distribution of effects	144
2.B.16	Efficiency-equity trade off by road type	146
2.B.17	“Experiments” identifying variation similar to that used in Faber [2014]	147
2.C.1	Original maps	155
2.C.2	Digitized road networks from historical Michelin maps	156
2.D.1	Changes in the distribution of primary completion rates	164
2.D.2	Defining road types	173

2.D.3	Compare the distribution effects of road building since 1970 under the baseline and alternative parameterization	176
2.D.4	Comparing the distributional effects of road building since 1970 under various values of β	178
3.4.1	Distribution of the variance of job fixed effects across occupations within firms	195
3.4.2	Contribution of worker-job match effects by firm and occupation mean pay quantiles	197
3.4.3	Plotting estimated match effects	198
3.5.1	Within occupation mean vs SD of wages	200
3.5.2	Variance of worker FE, job FE and correlation between the two components .	201
3.6.1	Within TTWA mean of worker FE and job FE (2011-19)	203
3.6.2	Within TTWA variance components vs local LM size	204
3.6.3	Decomposing the within TTWA job variance component into within and between occupations	205
3.6.4	Decomposing the within TTWA wage variance sorting component into within and between occupations	206
3..1	1 to 4 (and 4 to 1) job switches	211
3..2	Average log wage changes of moves between firm clusters	212
3..3	Mean of residuals by worker FE and job FE quantiles (25 groups each)	213
3..4	Distribution of the variance of mean JFE within occupations over individuals .	218
3..5	Relation between job fixed effect and firm revenue per worker, by occupation group (2011-19)	223
3..6	Relation between job fixed effect and firm revenue per worker, by skill group (2011-19)	225
3..7	Within TTWA log population vs moments of log pay (2011-19)	225
3..8	Within TTWA log population vs variance of job FE between and within occupation	226
3..9	Within TTWA log population vs covariance components decomposed into between and within occupation parts	226

List of Tables

I	Quantifying inequality of opportunity	19
1.A.1	Impact of the proportion following Islam on observed primary completion rates and causal place effects (CPE).	28
1.A.2	Impact of the proportion following Islam on observed gender differences in primary completion rates and causal place effects (CPE).	29
1.B.1	Parametric estimates of place effects including HH fixed effects	32
1.B.2	Outcome based placebo test	35
1.D.1	Parametric estimates of place effects	38
1.D.2	Robustness of the main results to the linear probability assumption	39
1.L.1	Causal effect of place on primary and secondary completion in Cameroon in 2005	54
1.L.2	Estimating sample primary education variation	55
2.3.1	Results from estimating the sufficient statistic relationship	87
2.4.1	Impact of different types of road on spatial inequality of opportunity	104
2.B.1	Controlling for ethnicity by leader	124
2.B.2	Relation between base level and changes (combined regressions)	125
2.B.3	Relation between base level and changes (individual regressions)	126
2.B.4	Including lagged market access terms	127
2.B.5	Inference on OLS results	130
2.B.6	Decomposing the impact of changes to the transport network since 1970 on spatial inequality of opportunity	134
2.B.7	Predicting the Impact of Roads on Local Opportunity	137
2.B.8	Results combining instruments and implementing the Borusyak and Hull [2020] correction.	139
2.B.9	Impact of network characteristics average impact on inequality	140
2.B.10	Impact of different types of road on spatial inequality of opportunity in the re-scaled networks	141

2.B.11	Sufficient statistic result by country	142
2.B.12	Robustness of the main sufficient statistic result	143
2.B.13	Alternative outcome variables	145
2.B.14	Replicating the results from Faber [2014]	149
2.C.1	Asset demand equations using DHS data	152
2.C.2	Gravity regressions to recover thetas	154
2.D.1	Endogenising Schooling Supply	162
2.D.2	Relationship with remoteness, comparing baseline and generalising consumption extension results	165
2.D.3	The impact of locality-road level characteristics on μ_{ir}	175
2.D.4	Relationship between the total relative effect of roads built since 1970 and locality remoteness in 1970. Comparing the baseline and alternative parameterization	177
3.3.1	Summary statistics	189
3.4.1	Comparison of different specifications (2011-2019)	190
3.4.2	Variance decomposition from jobs model over time	193
3..1	Largest connected set of common specifications (2011-19)	210
3..2	Comparison of basic AKM and clustered correlated fixed effect decomposition results with select estimates from previous studies.	217
3..3	Estimates of β in equation 3.15 (2011-19)	220
3..4	Coefficients on log revenue per worker interacted with 1 digit occupation group	222
3..5	Coefficients on log revenue per worker interacted with skill groups	224
3..6	Coefficients on log population from regression of occ-ttwa moments	228

Introductory Remarks

Inequality of opportunity is undesirable, although ubiquitous. This statement is the binding motivation behind the three chapters of this thesis. In the first chapter, I measure the extent to which such inequality of opportunity exists over space in three developing countries: Benin, Cameroon, and Mali. In the second chapter, I ask what causes some locations to offer greater opportunities than others, and specifically how the common policy of road building influences this spatial distribution of opportunity. The third chapter then explores the role of occupations in explaining wage inequality between individuals focusing on the UK, furthering our understanding of the drivers of inequalities in the labor market.

In this thesis, I focus on two types of inequality, opportunity over space and wage inequality. Although both are in my estimation of crucial importance, it is worth noting that many other important dimensions of inequality also exist — and to these, I cannot speak. Inequality of opportunity over space is perhaps an unfamiliar concept and one that deserves a short explanation. The key question that captures this concept is: To what degree does where you happen to be born and grow up affect your life outcomes? In this thesis, I narrow this definition further by focusing on lesser-studied within-country inequalities and proxying for “life outcomes” by education completion.

The first chapter confirms that opportunity is unevenly distributed across space in Benin, Cameroon, and Mali. I find that moving to a one standard deviation better location at birth increases a child’s probability of completing primary school by 15 percentage points. I additionally find that place matters more for girls than boys and is correlated with local characteristics such as urbanization, access to electricity, and employment in agriculture. Unlike other local characteristics, I find that the predominant local religion is only weakly correlated with average causal place effects. However, this zero-average effect hides considerable cross-gender differences as the predominant local religion is correlated with the gender gap in the effect place has on educational attainment. One output of this work is a locality-by-year data set of the causal effect growing up in a given location has on a child’s probability of completing primary school. I use this data in my second chapter.

The second chapter goes further and asks how governments can influence this uncovered spatial distribution of local opportunity. In particular, in this chapter, I consider how changes in connectivity over space influence inequality of opportunity using data on local opportunity from the first chapter. Spatial inequality within countries is a pervasive and persistent phenomenon. How does the connectivity of place determine underlying spatial inequality of opportunity? To answer this question, I derive a sufficient statistic result linking local opportunity to market access terms, developing a framework consistent with a broad class of spatial general equilibrium models. I empirically validate this result using a novel not-on-least-cost-path identification strategy and data from historical road maps in Benin, Cameroon, and Mali covering 1970 to 2020, that I digitize. Using these estimates to parameterize a structural case of the spatial model, I show that road building alters the spatial distribution of opportunity. By considering each possible road upgrade I show that although some roads decrease the standard deviation of opportunity by more than 2%, others increase inequality by a similar amount. By fixing the spatial distribution of opportunity I show that to achieve similar reductions in inequality by moving people, a prohibitively large number would need to migrate from low- to high-opportunity areas — between 13% and 44% of the low-opportunity areas' population. Policymakers also face an equity-efficiency trade-off: 4 out of the top 10 aggregate opportunity-increasing roads also *increase* spatial inequality of opportunity.

Whereas the first two chapters of this thesis are concerned with understanding the causes and consequences of inequality of opportunity over space, in the third chapter I move to discuss a different dimension of inequality. In the third chapter, I focus instead on labor market inequality, and how we can better understand its drivers. This is a crucial question both in helping our theoretical understanding of how labor markets operate, but also in asking what policies might be effective at tackling such large inequalities.

The third chapter aims to better understand wage inequality. As motivation, we note the following: Google does not pay its canteen chefs the same premium as it does its software engineers; there is heterogeneity in pay premia within a firm. To study the implications of this observation for wage inequality, we estimate a two-way, worker-job, fixed effect model that builds upon the canonical model by [Abowd et al. \[1999\]](#) by allowing for within-firm heterogeneity. Estimating our model on UK administrative data, we find that even if workers, and firms, were identical, 25% of current log-wage variance would remain, due solely to heterogeneity between occupations. We further document that worker heterogeneity accounts for a relatively small proportion of log-wage variance, and that between-firm pay

heterogeneity is more important in higher-wage occupations and larger labor markets.

Chapter 1

Spatial Inequality of Opportunity in West Africa

Abstract

In this paper, I study spatial inequality of opportunity measured in terms of the causal effect place has on primary completion, in Benin, Cameroon, and Mali. Using a movers design I find that on average moving to a one standard deviation better location at birth increases a child's probability of completing primary school by 15 percentage points. Estimating locality-year and gender-locality-year place effects I show that place matters more for girls than boys and place effects are correlated with local characteristics such as urbanization, access to electricity, and employment in agriculture. Religion is weakly correlated with average causal place effects but strongly correlated with the gender gap in the effect place has on educational attainment.

1.1 Introduction

Since 1970 in Benin, Cameroon, and Mali, GDP per capita has doubled in real terms, child mortality has halved and the human development index has also doubled.¹ However, the rising tide has not lifted all boats. Inequality within countries, and across space is pervasive. More worryingly, this spatial inequality might indicate differing opportunities accruing to individuals growing up in different locations, rather than the result of sorting across space. Inequality of opportunity determined by where a child grows up is intrinsically undesirable, could interact with gender inequality, and may have negative consequences on aggregate development.

In this paper, I quantify the magnitude of spatial inequality of opportunity, analyze the correlates of local opportunity, and study its consequences on gender inequality in Benin, Cameroon, and Mali. To do this I focus on one important dimension of inequality of opportunity across space — inequality in the causal effect growing up in a location has on a child’s probability of completing primary school. I focus on primary education as it is the most salient margin of education in this setting (a third of people in my sample complete primary school whereas only 7% go on to complete secondary school) and is correlated with greater later life outcomes broadly defined such as not working in agriculture, better housing quality, and higher wages. Inequality in causal place effects measured in this manner captures the notion of inequality of opportunity as simply by growing up in one location rather than another a child is more or less likely to complete primary education.

I use a movers design to estimate the causal effect of spending an additional year of childhood on primary school completion, in every locality²-year cell, both overall and by gender. The movers design focuses on a sub-sample consisting of those who move across localities, and leverages variation in the exact age of a child when a family moves to identify place effects.³ This approach builds on the place effects literature which maps the causal effect of growing up in different areas on child outcomes. In this, I follow seminal work by

¹Statistics based on author’s calculations using data from the World Development Indicators, and the World Bank, [Feenstra et al. \[2015\]](#) and [Prados de la Escosura \[2021\]](#), between 1970 and 2015.

²In this paper, I refer to second administrative divisions (Communes in Benin, Departments in Cameroon, and Circles in Mali) as localities. These have a median population of 267,000 and can be thought of as similar in scale to commuter zones in the US.

³Intuitively, this strategy considers children born in the same period t and in the same location o , who move to the same destination location d , but do so at different ages. I then compare the outcomes of children who spend a larger proportion of their childhood in d vis-à-vis o to infer the causal effect of d relative to o on primary school completion under the assumption that other determinants of children’s outcomes are unrelated to the age at which they move. I then make all such possible comparisons across localities to uncover the (relative) causal effect of spending an additional year of childhood in each location on the probability of completing primary education.

Chetty et al. [2016], Chetty and Hendren [2018a,b] and Chyn [2018] in the US context, as well as work by Damm and Dustmann [2014] in Denmark, Deutscher [2020] in Australia, Laliberté [2021] in Montreal, and Alesina et al. [2021] and van Maarseveen [2021] in Sub-Saharan Africa.

On average I find that children converge to the outcome of permanent residents in a given location at a rate of 2.7% per year of childhood spent in the location.⁴ The validity of the movers design rests on the key identifying assumption that selection effects do not vary over the age of a child when they move.⁵ I validate this assumption in five ways. First, I show that there is no evidence of varying selection effects on observables. Second, I demonstrate that results are robust to including household fixed effects, that is leveraging cross-sibling age variation. Third, I consider increasingly likely to be exogenous move events and show that results remain stable. Fourth, I perform an over-identification test using variation across gender in locality quality. Finally, I employ a placebo exercise and show that estimated place effects for those who move *after* they've completed primary education are not correlated with those found in my main specification.

Estimating locality-year effects, I find that opportunity is indeed unequally distributed across space. Moving to a location with one standard deviation greater opportunity at birth increases the probability of a child completing primary school by 15 percentage points. I consider heterogeneity by country and time period finding high levels of inequality in all countries in all years. Characterizing areas of high opportunity, I show that urbanization, electricity provision, sanitation, and housing quality are correlated with higher observed primary completion rates and opportunity. Conversely, employment in agriculture is negatively associated with both observed primary completion rates and the causal effect of place. However, in all cases, less than half of the correlation between characteristics and observed completion rates operates through the causal effects channel.

A key characteristic, often associated with education completion, is the prevailing religion in a locality [Alesina et al., 2020, Panin et al., 2021]. I find that, although the proportion of individuals in a locality indicating Islam as their religion is on average strongly negatively correlated with observed primary completion rates, religion impacts the causal effect of place to a much lesser degree. Indeed, religion stands out among the characteristics considered as one where a strong observational correlation doesn't translate into a significant

⁴This is consistent with results from Chetty and Hendren [2018a] in the US and Alesina et al. [2021] in Sub-Saharan Africa, both of which look at intergenerational mobility. Chetty and Hendren [2018a] find a convergence rate of about 4% and Alesina et al. [2021] find a rate between 1.5% and 3%.

⁵This assumption allows households with better unobservables to move to better areas, but is violated if they systematically do so earlier or later depending on their child's age.

impact on the causal effect place has on primary education completion.

Place could also play a role in mitigating or exacerbating prevailing gender inequalities.⁶ To study the relationship between spatial and gender inequality of opportunity, I estimate separate overall convergence rates for boys and girls, as well as gender-specific locality-year causal place effects. I find that girls converge to the outcomes of permanent (female) residents at a rate over twice that of boys (converging to the outcomes of permanent male residents). Even once differences in permanent residence variability in primary completion rates have been taken into account, this still implies that place matters more for female than male outcomes. By correlating place characteristics with the difference between relative male and female causal place effects, I test whether place characteristics are relatively better for male than female outcomes and therefore contribute to, or detract from, gender inequality of opportunity. Urbanization and electricity availability increase gender inequality of opportunity, but employment in agriculture reduces it. I also find that the average null correlation of local religion on causal place effects hides considerable gender heterogeneity.

This paper is most closely related to [Chetty and Hendren \[2018a\]](#) and [Chetty and Hendren \[2018b\]](#) who study the impact of place on income mobility in the US and estimate locality-specific effects, and [Alesina et al. \[2021\]](#) who study observational education mobility in Africa and estimate average place effects. I consider a different outcome, setting, and time frame from [Chetty and Hendren \[2018a,b\]](#) and go further than [Alesina et al. \[2021\]](#) by estimating place-specific effects over time, considering heterogeneity across countries, and studying the ramifications of inequality of opportunity on gender inequality. My main contributions relative to these papers are fourfold. First, I quantify the time-varying spatial inequality of opportunity in Benin, Cameroon, and Mali, and find that opportunity is unequally distributed in all countries and time periods. Second, I characterize place effects, describing what areas of high opportunity look like, and show that religion does not impact causal place effects on average. Third, I study the confluence of spatial and gender inequality and find that place has a potentially significant role to play in reducing gender inequality, but that characteristics associated with higher opportunity may also be associated with greater gender inequality. Fourth, I find that although religion is weakly correlated with average causal place effects, it is strongly correlated with the gender gap in the effect place has on educational attainment.

⁶Although overall primary completion rates have increased substantially over the sample period, gender gaps have remained persistently high: In Benin (1992-2013) increasing from 17 to 20 percentage points, in Cameroon (1976-2005) decreasing from 21 to 12 percentage points, and in Mali (1998-2009) increasing from 6 to 8 percentage points.

This paper proceeds as follows: Section 1.2 describes the data and empirical setting; section 1.3 estimates overall causal place effects; section 1.4 estimates locality-year specific place effects; section 1.5 characterizes place effects and considers the role of religion; section 1.6 studies heterogeneity by gender and place variables associated with greater (or lower) gender inequality. Finally, section 1.7 concludes.

1.2 Setting and data sources

1.2.1 Why focus on Benin, Cameroon and Mali

Benin, Cameroon, and Mali, whilst representing important countries to study in and of themselves, also provide a particularly interesting setting to examine how the location of upbringing affects primary education completion and thus later life outcomes for three main reasons. First, Benin, Cameroon, and Mali are developing countries aiming to grow rapidly in the near future, and ensuring equitable growth is key to individual well-being, as well as potentially aggregate success. In addition although exhibiting some similarities Benin, Cameroon, and Mali also have many differences enabling this paper to speak to issues of external validity and cross-country heterogeneity. Second, in these countries, primary completion is a key measure of the opportunities available to individuals broadly defined as discussed in more detail in subsection 1.2.3. Third: data requirements. Benin, Cameroon, and Mali are the only countries in Sub-Saharan Africa that meet the high data requirements of this paper which are explained in more detail in subsection 1.2.2.

1.2.2 Data description

This paper uses data on 8 million observations from 1976 to 2013, across 164 localities from every available census from Benin, Cameroon, and Mali.⁷ To estimate causal place effects using a movers design it's necessary to observe where an individual is when the census was taken, their previous location, their birth location, and how long they have resided in their current location.⁸ In addition to migration information, I require data on primary

⁷Data is accessed from [IPUMS \[2020\]](#) and consists of 10% samples, with thanks to the National Institute of Statistics and Economic Analysis in Benin, the Central Bureau of Census and Population Studies in Cameroon and the National Directorate of Statistics and Informatics in Mali, who provided the underlying data.

⁸In the censuses described, I can geo-locate individuals at the second administrative unit level. These localities have a median population of 267,000 overall samples. Benin's Communes have a median population of 103,000 with an inter-quartile range of 71,000 to 173,000. Mali's circles have a median population of 308,000 with an inter-quartile range of 197,000 to 520,000. Cameroon's departments have a median population of 456,000 with an inter-quartile range of 225,000 to 907,000. A broadly comparable geographical unit in the US would be commuter zones. To uncover causal place effects previous research [[Chetty and Hendren, 2018a](#), [Laliberté, 2021](#), [Deutscher, 2020](#)] has mainly relied on administrative data that is not typically publicly available. This data is

completion,⁹ age, and gender, at a fine and consistent geographic level over multiple cross-sectional waves.

Only three countries in Sub-Saharan Africa — Benin, Cameroon, and Mali fit the stringent data requirements. In Benin, I use censuses from 1992, 2002, and 2013, encompassing over 2.2 million individual-level observations in 77 consistent localities (Communes). In Cameroon, I use censuses from 1976, 1987, and 2005 covering over 3.4 million individual-level observations in 39 consistent localities (Departments). Finally, in Mali, I use censuses from 1998, and 2009 giving over 2.4 million individual-level observations in 48 consistent localities (Circles). In total, this gives me a sample of over 8 million individual-level observations across 164 localities and 444 locality-year cells. A benefit of using this data is that it is freely and publicly available from IPUMS International.

Using available variables on the current location, previous location, and birth location of individuals, I can define three groups: (i) Permanent residents whose birth location equals their current location. This group covers 81% of 14 to 18-year-olds in my sample. (ii) One-time movers whose birth location does not equal their current location but is their previous location. This group covers 13% of 14 to 18-year-olds in my sample. (iii) Multi-time movers whose birth location doesn't equal their current location or their previous location. This group covers 6% of 14 to 18-year-olds in my sample. Thus, for all but 6% of the sample, I can construct each 14 to 18-year-old individual's full location history, including the age and year of moving.

The data described above comes from the same source as that used in [Alesina et al. \[2021\]](#). However, whereas [Alesina et al. \[2021\]](#) consider 27 counties in Africa for their main analysis, and 16 countries when they analyze place effects, I focus only on three: Benin, Cameroon, and Mali. This is because I have stricter criteria for inclusion in my sample. I only consider censuses that include information on previous location as well as birth and current location allowing me to trace out entire migration histories of one-time movers. To the best of my understanding [Alesina et al. \[2021\]](#) discern age-moved using only birth and current location coupled with information regarding time spent at the current location. Thus [Alesina et al. \[2021\]](#) measure previous location with error bundling in those who indeed have moved from their birth location with those who may have moved from some intermediate location (32% of all movers in my sample). As [Alesina et al. \[2021\]](#) show, moves are on average

necessary for these studies to observe full migration histories and to match child and parent outcomes over long time frames. Such rich data is not needed to estimate place effects, in the censuses I use, it's possible to discern migration histories from cross-sectional evidence.

⁹In Benin, Cameroon, and Mali, education is compulsory for the first six years between the ages of 6 and 11/12 which covers primary school.

to better locations thus this has the potential to introduce non-classical measurement error into movers' change in location quality. In addition, I only include countries with multiple cross-sections to enable an analysis of place effects over time.

The Online Appendix provides some descriptive statistics from the census data. I show that the distribution over localities of primary completion rates in Benin, Cameroon, and Mali has shifted over time. There are considerable gains in Cameroon and Benin although intra-national variation remains large, while in Mali the gains have been smaller. Additionally, although most students complete primary education by age 12/13, (in all three countries primary school *should* be completed by this age) a significant minority complete it only at 14 or 15. In the online appendix, I also analyze migration patterns and find that migration in Benin, Cameroon, and Mali is not easily characterizable by to-capital migration, nor by rural-to-urban migration.

1.2.3 Education as a measure of opportunity across space in Benin, Cameroon, and Mali

Education is one of the most important dimensions of development and for the majority of individuals in Benin, Cameroon, and Mali, primary education is the most salient margin of variation. In my sample, among those 19 or older 34% have completed primary education and only 7% secondary education. Schooling has also been found to have high Mincerian returns, [Psacharopoulos and Patrinos \[2018\]](#) estimate the return to an additional year of education to be 10.5% in Sub-Saharan Africa — the highest of any region in the world.¹⁰

Rather than relying on estimates of returns to education from other studies which consider different samples, I can indirectly estimate the returns to education in my data. Although I don't observe wages or wealth, I do observe detailed information about housing quality. In particular, I observe the material walls, floors, and roofs are made from, the level of sanitation, the availability of clean running water, and whether or not the household has access to electricity. From this, I create a composite housing quality index as the standardized first principle component over these variables. Returns to education in each locality are measured as the average housing quality those with primary education have above those without, thus controlling for overall location-level differences in building materials and general housing quality. Figure [1a](#) displays the distribution of local returns to education by country over localities. In each country returns to primary education are on

¹⁰Similarly [Barouni and Broecke \[2014\]](#), and [Schultz \[2004\]](#) find evidence of high returns to education in Sub-Saharan Africa.

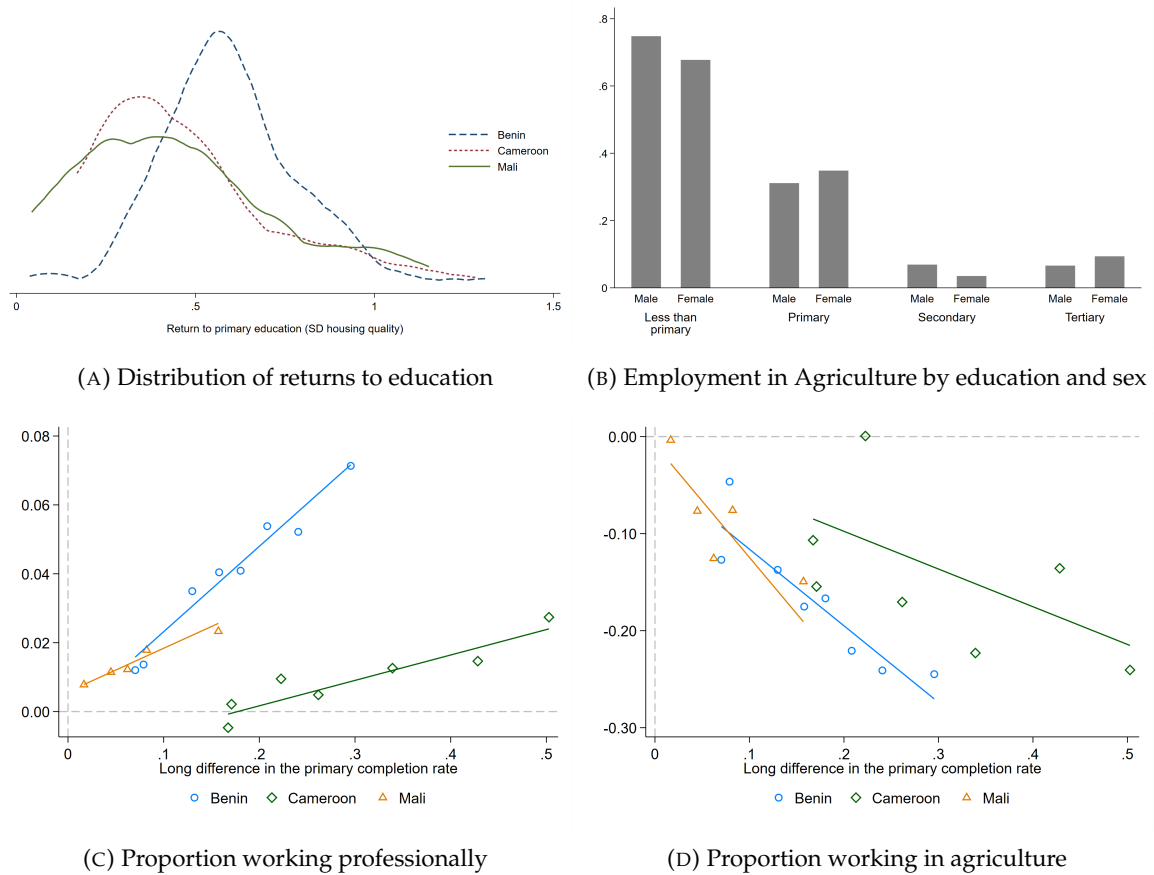
average roughly half a standard deviation of housing quality. This high mean hides significant variation across localities within a country. In all three countries, there are some locations with almost no returns to primary education while in others they are as large as a full standard deviation.

Primary completion is also highly correlated with employment prospects. Figures [1c](#) and [1d](#) show the long differenced relationship between locality-level average primary completion and the proportion working in professional services or agriculture respectively. All slopes of the lines of best fit are significant at the 5% level. Both figures tell the same, albeit correlational, story that areas in which primary completion rates most increased saw the greatest transition away from agriculture and towards employment in the professional sector.

At the individual level, there is also evidence linking education and occupation. Figure [1b](#) shows that the proportion of individuals working in agriculture decreases substantially with education completion. Primary education is a key route out of agricultural work (much of which is in subsistence agriculture¹¹). However, it shouldn't be taken as a panacea. Many individuals with primary education are still employed in agriculture, and of course the relationships documented here are purely correlational in nature.

¹¹Measures of the proportion of individuals employed in subsistence agriculture are by their very nature difficult to ascertain. [Medina et al. \[2017\]](#) estimates the size of the informal economy (part of which is subsistence agriculture) to be 20-40% of GDP in Cameroon and greater than 40% in Mali and Benin, however, a much larger proportion of the labor force will be employed in the informal economy due to its relatively lower productivity.

FIGURE I Education is associated with positive later life outcomes



Note: Panel (a) shows the distribution of returns to primary education across localities for Benin, Cameroon, and Mali. Returns to primary education are measured as the average housing quality of those with primary education minus the average housing quality of those without primary education. Housing quality is measured as the standardized first principle component over observables describing housing quality including floor, wall, and roof material, sanitation provision, access to running water, and access to electricity. Panel (b) shows the proportion employed in agriculture by education and sex. In both figures, the sample is restricted to include only those of core working age, 25 to 55. Panels (c) and (d) display binscatter plots with corresponding linear regression lines of the association between the locality level change in primary completion on the x-axis vs the change in the proportion employed in a professional capacity in the left-hand panel and the change in the proportion employed in agriculture in the right-hand panel. Each line represents a different regression, the slope of each is significant at the 5% level with robust standard errors. The long difference for Benin is from 1979-2013, for Cameroon from 1976-2005, and for Mali from 1987-2012.

In sum, the evidence presented in this section suggests that completing primary education is correlated with greater opportunity defined more broadly. Therefore, in this paper, unequal access to primary educational opportunities across space will be taken as indicative of unequal access across space to opportunities more generally defined. Of course, there will be other dimensions of opportunity that are not captured by primary education completion and notions of inequality of opportunity that go beyond that which is across space. In this

paper, I will focus on inequality of opportunity defined across space and in terms of primary education completion as this is a particularly salient margin of inequality of opportunity in this setting.

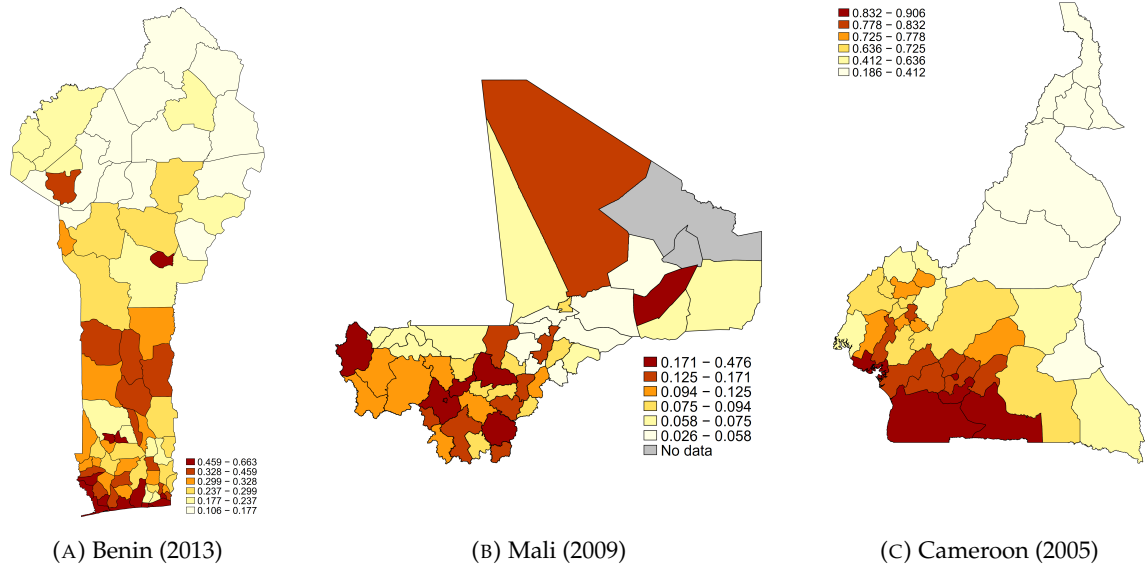
Primary school completion is also one of the most commonly available, and comparable variables across countries and decades. Due to its universality and ease of recognition, it is less likely to suffer from substantial measurement error. Primary school completion also has the advantage of being a historically determined variable that is observable in a cross-section. Most individuals have completed primary schooling by age 12 to 14, but it continues to be observed no matter what age the individual is at the time data is collected. This allows me to consider the impact of place *after* schooling has been determined constituting a natural placebo test, place cannot causally impact primary school completion after the event.

1.3 How unequally distributed is opportunity across space in Benin, Cameroon, and Mali?

Figure II shows that the observed distribution of primary education completion over space is highly unequal in Benin, Cameroon, and Mali. However, it's unclear from the raw variation to what extent these discrepancies reflect differing characteristics of individuals living in different locations or the underlying causal impact of place on primary completion rates. The distinction has normative as well as policy implications: inequality in causal effects is inequality of opportunity — simply by growing up in one location rather than another children are more or less likely to complete primary education.

1.3. How unequally distributed is opportunity across space in Benin, Cameroon, and Mali?

FIGURE II Locality level observed primary completion rates



Note: This figure shows the spatial distribution of observed primary education completion rates for those above the age of 12 in Benin, Cameroon, and Mali. Each figure has its own scale and corresponding legend where darker orange/red indicates higher completion rates. The data for Benin comes from the 2013 census, for Mali from the 2009 census, and for Cameroon from the 2005 census.

To set ideas consider a simple place effects model.

$$y_i = \sum_{a=1}^{13} \mu_{c(i,a)} + \theta_i \quad (1.1)$$

Where y_i is a dummy variable taking the value one if an individual i has completed primary education, and $c(i, a)$ denotes the location c individual i resided in at age a . I split the determinants of y into two parts. First, θ_i represents the individual-specific effects such as raw ability which are orthogonal to location effects. Second, $\sum_{a=1}^{13} \mu_{c(i,a)}$ captures the sum of all place effects individual i has been exposed to during childhood, where place effects can be thought of as location-specific fixed effects μ_c . Intuitively, μ_c captures any locality-specific characteristic which may impact y , for example: weather, religio-cultural factors, or the prevailing local returns to education.

This formalization introduces two assumptions. First, I assume that place impacts each individual i in the same manner. I relax this when considering differing effects by gender, but in general, do not probe this assumption. Instead, results reflect the average over potentially heterogeneous effects. Second, I assume additive separability of effects in what is effectively a linear probability model. In the results presented below, I retain this assumption but replicate qualitatively and quantitatively similar results using probit, logit, or

pseudo-Poisson approaches in the online appendix.

In the absence of random variation assigning children to locations, to estimate causal place effects I use a movers design building on work by Chetty and Hendren [2018a], Aaronson [1998], Weinhardt [2014], Finkelstein et al. [2016], Molitor [2018] and, Bronnenberg et al. [2012] among others. This approach uses observational variation in the outcomes of children that move across localities at different ages. If children who move to better areas¹² earlier have a higher probability of completing primary school, I conclude that the quality of a location matters. However, individuals who move earlier to better places could also have better unobserved characteristics. To disentangle causal place effects and selection I make, and interrogate, one key assumption: that selection effects do not vary with the age of the child at move.

1.3.1 Estimating the overall convergence rate

Formalizing the above identification strategy within the place effects framework, I estimate models of the form given in equation 1.2.

$$y_i = \rho_{os} + \alpha_m + \beta_m \cdot \Delta \bar{y}_{ods}^p + \varepsilon_i \quad (1.2)$$

Where i denotes an individual, o their origin (birth) location, d their destination (current) location, s their cohort and m the age at which they move. y_i is an indicator variable taking the value one if primary education has been completed. ρ_{os} is a set of origin by cohort fixed effects. α_m is a set of age-at-move fixed effects which flexibly absorb movement costs. $\Delta \bar{y}_{ods}^p = \bar{y}_{ds}^p - \bar{y}_{os}^p$ is the difference in primary completion rates between the destination and origin locations' permanent residents who are in the same cohort as i and ε_i is an idiosyncratic error term.¹³ This equation is estimated on a sample of one-time movers between the ages of 14 and 18. The decision to use a sample of 14 to 18-year-olds is the result of a balancing act, on one hand, I want to be able to match as many children as possible to their parents and can only do so if they cohabit, thus a younger sample is better. On the other hand, I want to capture children after their primary completion status has been determined,

¹²In general *better* can be defined on any dimension of interest, but in this setting, a *better* area is one with a higher primary completion rate among permanent residents, that is those who don't move from their birth location. I consider other definitions of location quality in the online appendix and find similar results.

¹³In the online appendix, I show that the distribution of the difference in location quality among one time movers, that is the distribution of $\Delta \bar{y}_{ods}^p$, is symmetric around a mean close to zero. Implying that it is not the case that all, or even a large majority, of moves, are to better locations. This gives some indicative evidence against large selection effects. I also show that male and female children are similarly likely to move to better or worse locations.

so later is better. β_m are the coefficients of interest and can be interpreted as the percentage point impact of moving at age m to a location with a 1 percentage point higher primary completion rate among permanent residents.

We can decompose β_m into the causal effect of place b_m and the selection effect of place δ_m as follows: $\beta_m = b_m + \delta_m$. b_m is the object of interest but cannot, in general, be separately identified from δ_m . To identify b_m , I make the additional assumption that selection effects do not vary over the age at move. This is a potentially strong assumption. In sub-section 1.3.2 I relax it, and test its validity, finding considerable evidence in support and thus, for the main analysis, take it as given. Under the identifying assumption, I can estimate the causal effect of spending an additional year of one's childhood at age m in a one percentage point better location as $\beta_{m+1} - \beta_m = b_{m+1} - b_m$. As primary schooling *should*¹⁴ be completed by age 12, it must be that $b_m = 0$ for $m > 12$ and so in this range, β_m estimates the selection effect, $\beta_m = \delta$ for all m greater than 12.

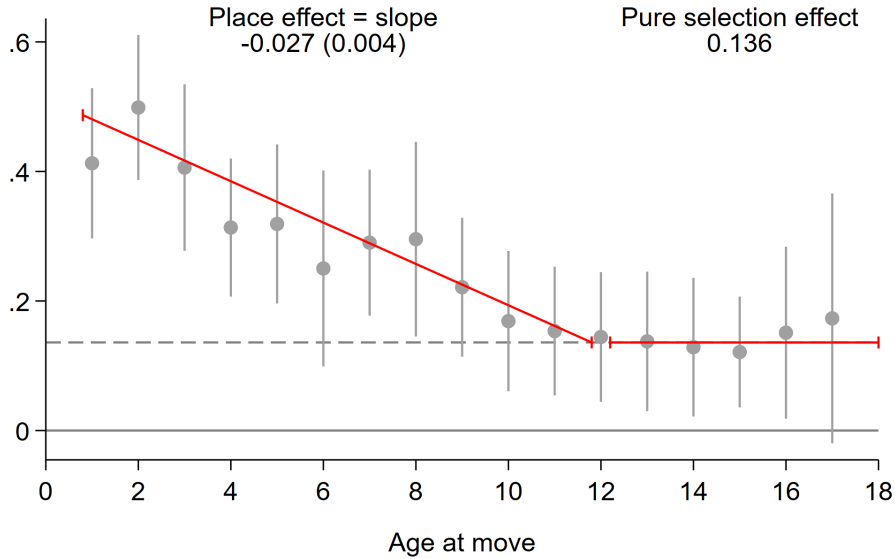
Figure III shows the results from estimating equation 1.2 where $\hat{\beta}_m$ is plotted against age at move m in grey with corresponding 95% confidence intervals. Standard errors are two-way clustered at the birth locality and current locality level.¹⁵ After age 12, $\hat{\beta}_m$ is constant, providing initial evidence in favor of the identifying assumption, as this shows selection effects are constant for all $m > 12$ when primary completion is already mainly determined. Between ages 1 and 12, I find a positive and increasing (with earlier ages) effect of place which is well approximated by a linear trend (as illustrated by the red overlay). This implies that where children grow up in Benin, Cameroon, and Mali, *does* exert a causal impact on their probability of completing primary education. Additionally, the linearity of the effect implies that there are no tipping points — or particularly important years [Cunha and Heckman, 2008]. Each additional year spent in a better location exerts roughly the same positive impact as any other before age 13.

¹⁴In each of the countries studied and over the whole sample period, primary school is expected to end at age 12. The online appendix documents that although many individuals complete primary school after age 12 by age 14 most who will complete have done so.

¹⁵In the online appendix I discuss inference in this setting and provide permutation-based inference results which affirm the conclusions reached here.

FIGURE III Plotting $\hat{\beta}$ from estimating equation 1.2

Effect of moving at year m to a 1pp. better location on the probability of completing primary education



Note: This figure shows the estimated β_m coefficients from the regression: $y_i = \rho_{os} + \alpha_m + \beta_m \cdot \Delta \bar{y}_{ods}^p + \varepsilon_i$ estimated on a sample of one time movers between the ages of 14 and 18 for each available census in Benin, Cameroon, and Mali. Corresponding 95% confidence intervals are indicated by gray bars where standard errors have been clustered at the birth locality and current locality level. Superimposed on this in red are two lines of best fit with a break at age 12. The horizontal gray dashed line indicates the post 12 mean of $\hat{\beta}_m$.

Due to the linearity apparent in causal place effects, I consider the pre-13 slope as the single key parameter. This slope, of -0.027 (0.04), is interpreted as the rate of convergence. That is, for each additional year spent in a new location, how *fast* do movers converge to permanent residents in that location? As a robustness test, in the online appendix, I estimate a parametric analog to equation 1.2 allowing a different slope and intercept for children moving between the ages of 1 and 12 and those moving after age 12. Over various fixed effects specifications, this recovers a very similar convergence rate and a precisely estimated zero slope for post-12 moves.¹⁶

Figure III is analogous to figure 4 in Chetty and Hendren [2018a] and figure 11 in Alesina et al. [2021]. It recovers remarkably similar results to these studies which consider different settings, the US in Chetty and Hendren [2018a], or different outcome variables in a larger sample in Alesina et al. [2021]. Chetty and Hendren [2018a] find a convergence rate of about

¹⁶Note that this estimate may be a lower bound for the convergence rate to the extent to which the “ladder” mechanism espoused by Card et al. [2022] is at play here. That is the extent to which those moving across locations are likely to move to/from sub-locations that are closer in outcomes to those of the larger geography.

4% and Alesina et al. [2021] find a rate between 1.5% and 3%. Relative to these studies, the main contribution of this paper lies in estimating and analyzing locality-year specific place effects which I turn to in section 1.4. Analyzing locality-year specific place effects allows me to quantify and characterize the changing spatial distribution of opportunity and study the ramifications of unequal opportunity across space on gender equality.

1.3.2 Validity of the movers design

I perform five validation exercises of the identifying assumption that selection effects do not vary with the age of a child when the family decides to move. They are described in detail in appendix 1.B and summarised here.

First, there is no evidence of varying selection effects on mothers' observed primary education. Mothers with primary education are more likely to move to better areas, but they do not systematically do this to a greater or lesser extent with respect to the age of their child at move. Second, results are robust to including household fixed effects and therefore only considering cross-sibling age variation. By including household fixed effects, I shut down any time-invariant dimensions of possible selection effects, for example, socioeconomic bracket, the child's revealed pre-school potential academic ability, or parents' attitudes towards education.

Third, I restrict the sample to only include increasingly likely to be exogenous move events and show that results remain stable. To do this, I define statistically increasingly unusually large out-move events from localities and restrict the sample to include only increasingly unlikely events — capturing an increasing proportion of exogenous moves. I find that the estimate of the causal effect of place is stable as I continue to restrict the sample, providing evidence for the identifying assumption.

Fourth, I perform an over-identification test using variation across gender in locality quality. This exercise utilizes differences in predictions from the place effects vs selection effects explanations. The place effects model postulates that individuals should converge to outcomes of permanent residents in their gender-specific sub-group in the destination location. Whereas, to the extent that households cannot observe gender-specific outcomes, selection effects should only operate at the aggregate level. By including own-gender-specific location quality (average primary completion rate of permanent residents with the same reported gender as individual i) and *other* gender quality in one regression, I can explicitly test the predictions of the two possible scenarios.¹⁷ If place effects are the driving force behind

¹⁷In the censuses I use only two genders are recorded, male or female.

the observed effects, then individuals will only react to the outcomes of children of their own gender. Whereas, if selection effects were predominant, individuals will react equally to the outcomes of children of both sexes. When performing this test, the coefficient on own gender remains positive and significant while that on other gender attenuates to zero. Giving further evidence in support of the place effects explanation.

In sum, any omitted variable bias or selection effects would have to operate within families proportionally to exposure times, persist in the presence of moves induces by displacement shocks, and precisely replicate permanent residents' outcomes by gender in proportion to exposure times in addition to not being observed on mothers education. I find it unlikely that any qualitatively or quantitatively important threats to identification operate in this manner.

In the online appendix, I also discuss the generalisability of my results which are based on a sample of movers to the whole population. The assumption that place effects are the same for movers and stayers can be decomposed into two components. First, that movers and stayers do not systematically differ in a manner that interacts with place effects [Lagakos, 2020, Hamory et al., 2021]. Second, movers and stayers could be similar but due to the nature of having moved to a new location, movers may interact differently with their surroundings. In the online appendix, I describe and provide some evidence against both of these dimensions by considering increasingly likely to be exogenous moves and replacing origin fixed effects with destination fixed effects.

1.4 Estimating locality-year specific place effects

To estimate locality-year specific effects, I generalize the specification presented in equation 1.2 by estimating equation 1.3 following Chetty and Hendren [2018b]:

$$y_i = \alpha_{odt} + \mu_{lt} \cdot e_{il} + \varepsilon_i \quad (1.3)$$

Where $l \in \mathcal{L}$ is a location, $o \in \mathcal{L}$ is the birth location of individual i , $d \in \mathcal{L}$ is i 's destination location and t is the sample year in which i is observed.¹⁸ α_{odt} are origin by destination by sample year fixed effects and ε_i is an idiosyncratic error term. Finally, e_{il} is a variable equal to the years of childhood (1 to 13) i spends in location l and is given by equation 1.4.¹⁹ Where

¹⁸ t is defined as the year the census is taken, however, it can be thought of as capturing place effects across the childhoods of the 14 to 18-year-olds in the sample at time t .

¹⁹This formulation assumes linearity of place effects, an assumption that is made based on the evidence provided in section 1.3.1.

m_i denotes the age at move.

$$e_{il} = \begin{cases} 13 - m_i & \text{if } l = d(i) \\ m_i & \text{if } l = o(i) \\ 0 & \text{otherwise} \end{cases} \quad (1.4)$$

I estimate equation 1.3 separately for each country on a sample of (o, d, t) cells with more than 20 observations.²⁰ The causal place effects, $\{\mu_{lt}\}_{l \in \mathcal{L}, t \in \mathcal{T}}$, can only be recovered in a relative sense as I can only compare across localities within a given census. Intuitively this approach considers all individuals who move from a given origin o to a given destination d in a given census period t and compares the outcomes of those who move earlier relative to later. If it is the case that those who moved from o to d earlier have better outcomes (are more likely to complete primary education) I conclude that $\mu_{dt} > \mu_{ot}$. I then combine information from all such comparisons to estimate each locality's causal place effect. By including origin-destination-year fixed effects, α_{odt} , I only use variation in the age of a child at move rather than comparing the outcomes of families that move to/ from different areas. Normalising, μ_{lt} can be interpreted as the causal effect of spending an additional year of childhood in location l in a given year t , on primary completion rates relative to the average location.

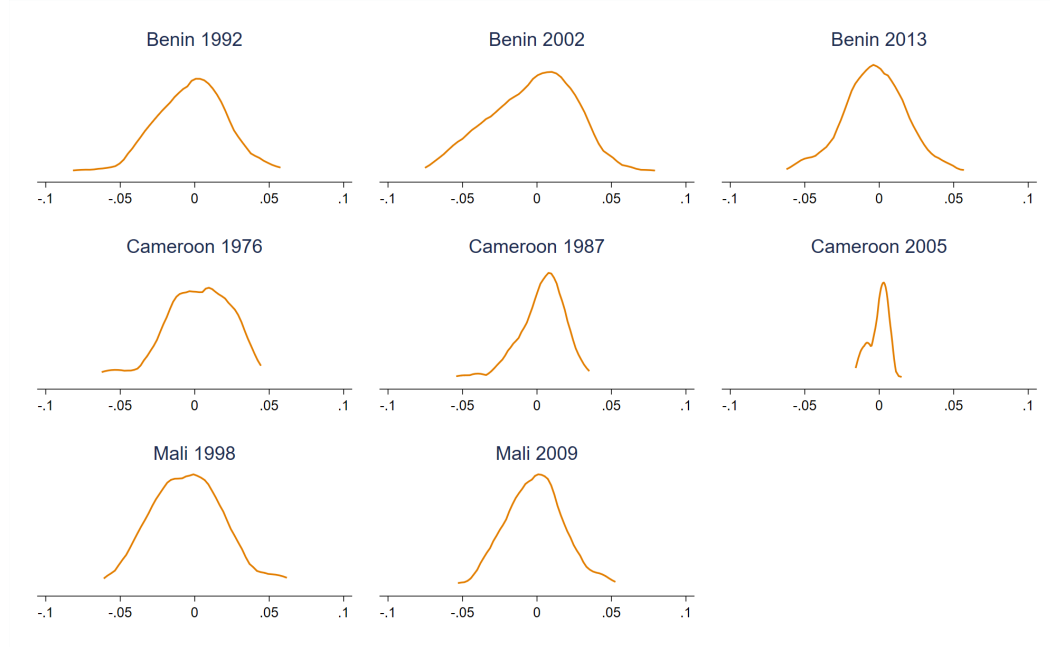
Figure IV shows the recovered causal place effects in each country-year sample. It shows considerable variation across space within each country. Across all censuses, I find that the average signal standard deviation²¹ of opportunity across localities is $\sigma_\mu = 1.2$ percentage points per year. This means that on average moving to a location with one standard deviation higher opportunity at birth causes a 15.0 percentage point (1.2pp. a year times 13 years of childhood) increase in the probability a child completes primary education. Figure IV shows significant variation in all country-year cells with the exception of Cameroon in 2005 which has a considerably tighter distribution. This is because, in contrast to the other countries and other years, spatial variation in observed primary completion rates in Cameroon in 2005 is relatively small due to a rightward shift in the distribution (and obvious truncation at 1). This does not necessarily mean that opportunity is equally distributed in Cameroon in

²⁰20 observations in the 10% samples implies an expected 200 movers. Restricting the estimating sample to include only cells with more movers would decrease measurement error but increase potential selection. Results are not sensitive to varying the cut-off.

²¹From the set of estimated place effects I can back out an estimate of $\sigma_{\hat{\mu}}$ but in actuality, we're interested in σ_μ , the signal standard deviation. As place effects are measured with error I can write $\hat{\mu}_{lt} = \mu_{lt} + \eta_{lt}$, and then find the signal standard deviation as follows: $\sigma_\mu^2 = \sigma_{\hat{\mu}}^2 - \sigma_\eta^2$. The standard deviation of the error, σ_η is estimated as the average standard error of $\hat{\mu}_{lt}$ denoted as s_{lt} . Then $\sigma_\eta = \sqrt{\mathbb{E}[s_{lt}^2]}$, and so an estimate of σ_μ can be recovered. Details of this procedure are presented in the appendix section 1.C.

2005, in the online appendix I show that considerable variation remains in the causal effect of place on secondary school completion in Cameroon in 2005.

FIGURE IV Distribution of recovered causal place effects



Note: This figure shows the distribution of recovered causal place effects in each census. Causal place effects are estimated from equation 1.3 and plotted relative to a 0 baseline. The x-axis is measured in yearly percentage-point effects, that is the impact of moving to a given location a year earlier on a child's probability of completing primary school.

Table I reports the total and signal variance of causal place effects in each country-year cell. I do not report estimates for Cameroon in 2005 for the reasons described above. Table I confirms what figure IV shows graphically, that the considerable average inequality of opportunity uncovered is not restricted to a specific country or specific time period. That being said, there are considerable differences in the signal variance of place effects across censuses. Table I also reports the within-country cross-census correlation in estimated causal place effects. Although the average within-country cross-time correlation is positive, this is not the case for each consecutive period. In Benin locations with high place effects in 1992 on average had higher place effects in 2002 but locations with high place effects in 2002 had on average lower place effects in 2013. In Cameroon, the correlation between locations' place effects in 1976 and 1986 is high, whereas in Mali the relationship between 1998 and 2009 remains positive but considerably weaker. The online appendix provides maps of the resulting estimated spatial distribution of opportunity.

TABLE I Quantifying inequality of opportunity

	Benin			Cameroon		Mali	
	1992	2002	2013	1976	1986	1998	2009
Total Variance	4.80	7.01	3.76	3.94	2.67	5.25	2.51
Signal Variance	0.72	4.64	1.75	2.04	1.47	2.74	1.21
Correlation with previous census		0.29	-0.23		0.62		0.09

Note: This table reports statistics from causal place effects recovered from equation 1.3 across censuses, that is country-year pairs. The first row reports the total variance measured per year of exposure. The second row reports the signal variance measured per year of exposure. The third row reports the within-country cross-year correlation between estimated effects.

1.5 Characteristics of place effects

Previous work has considered various correlates of observed primary completion (or educational inter-generational mobility) rates in Sub-Saharan Africa (e.g. [Majgaard and Mingat \[2012\]](#), [Alesina et al. \[2020\]](#), [Shafiq et al. \[2019\]](#), [Shifa and Leibbrandt \[2022\]](#)). These correlations could reflect sorting or they may be predictive of a characteristics' correlation with the causal effect a place has on individuals. In this paper, I can ask whether such known relationships indicate that an area with a given characteristic is correlated with higher opportunity or simply attracts people more likely to complete primary education anyway. To do this, I use the estimated place effects, $\hat{\mu}_{lt}$, as the dependent variable in regressions with locality-level covariates on the right-hand side. Specifically, I decompose the observed location-year primary completion rate, $\bar{y}_{lt}^p$, into an estimated causal effect of place, $\hat{\mu}_{lt}$, and a selection effect, $\bar{\theta}_{lt}$.

$$\bar{y}_{lt}^p = \alpha_{census} + 13 \cdot \hat{\mu}_{lt} + \bar{\theta}_{lt} + v_{lt} \quad (1.5)$$

Where α_{census} are census fixed effects, and v_{lt} is an idiosyncratic error term. Consider some explanatory variables X_{lt} then suppose the following relationships hold where β_μ and β_θ

are the objects of interest and $\varepsilon_{lt}^1, \varepsilon_{lt}^2, \varepsilon_{lt}^3$ are idiosyncratic error terms.

$$\begin{aligned} 13 \cdot \hat{\mu}_{lt} &= \beta_{\mu} \cdot X_{lt} + \varepsilon_{lt}^1 \\ \bar{\theta}_{lt} &= \beta_{\theta} \cdot X_{lt} + \varepsilon_{lt}^2 \\ v_{lt} &= \varepsilon_{lt}^3 \end{aligned}$$

As $\bar{\theta}_{lt}$ is not observed we cannot directly estimate β_{θ} . Instead, I combine the above terms with equation 1.5 and estimate β from equation 1.6.

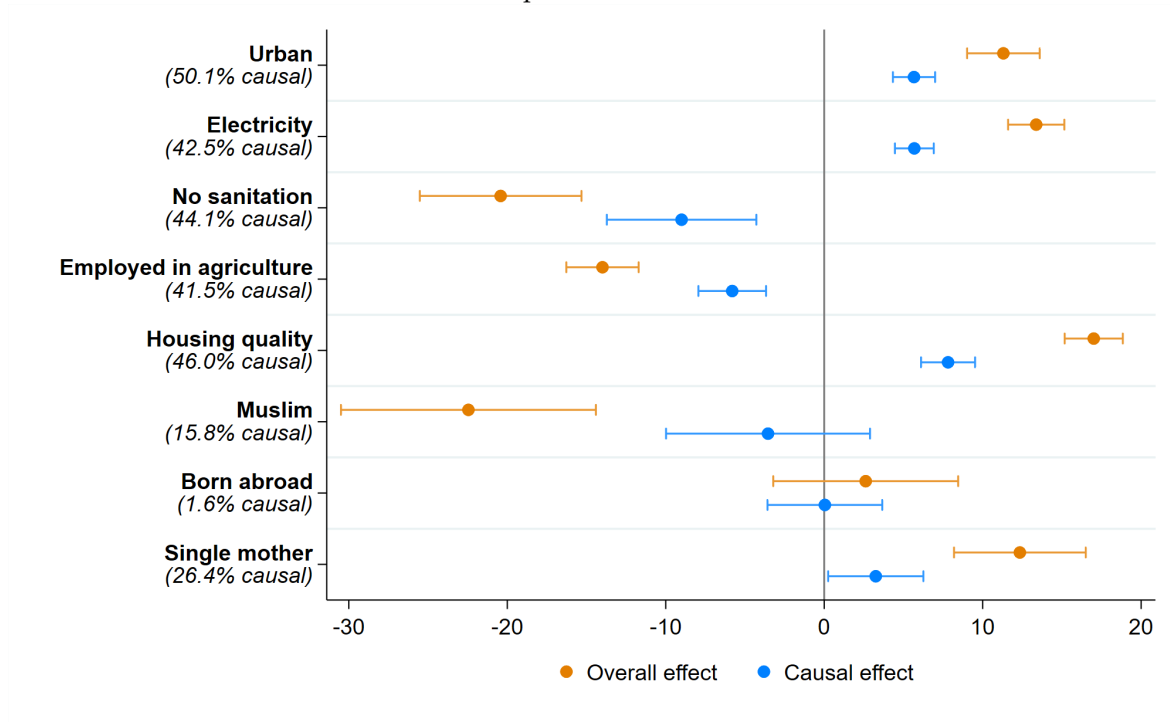
$$\bar{y}_{lt}^p = \alpha_{census} + \underbrace{(\beta_{\mu} + \beta_{\theta})}_{\beta} \cdot X_{lt} + \underbrace{\varepsilon_{lt}^1 + \varepsilon_{lt}^2 + \varepsilon_{lt}^3}_{\varepsilon_{lt}} \quad (1.6)$$

From this, I can then back out β_{θ} . Note that β_{θ} or β_{μ} can only be interpreted as capturing correlations. That is, a significant positive estimated $\hat{\beta}_{\mu}$ implies only that a characteristic of a location is correlated with higher local opportunity, not that it causes greater opportunity in a location. When estimating β_{μ} I use an outcome variable (the estimated causal place effects $\hat{\mu}_{lt}$) which is observed with error. However, classical measurement error in the left hand side variable will not impact estimates of β_{μ} and thus is not a concern.

Figure V plots the impact of standardized locality characteristics on observed primary completion rates ($\hat{\beta}$ reported in orange) and causal place effects ($\hat{\beta}_{\mu}$ reported in blue). Greater urbanization, greater access to electricity, better housing quality, and perhaps surprisingly a greater proportion of single-mother households, are all correlated with higher observed primary completion rates, and causal place effects. Similarly, a lack of sanitation, a higher proportion employed in agriculture, and a higher proportion of Muslims, are all negatively correlated with both observed primary completion rates and the causal effect of place. The proportion of individuals born abroad has little impact. In blue, causal place effects share the same associations, but to a lesser degree. In figure V percentages below characteristic labels indicate the proportion of the overall variation that operates through the causal effect channel. For example, a one standard deviation higher urbanization rate is correlated with an 11.3 percentage point higher observed primary completion rate and a 5.7 percentage point higher causal effect of place on primary completion. The online appendix reports results from a multi-variate version of these regressions in figure 1.A.1 and due to the demanding specification and considerable cross-characteristic correlated, few relationships remain statistically significant.

For most characteristics, at least half of the correlation between them and primary completion rates can therefore be attributed to selection or the differing characteristics of individuals across space. This implies that such intuitively known correlations like “education rates are higher in urban areas” are picking up significant effects that operate through the selection channel rather than purely correlations with the underlying distribution of local opportunity.

FIGURE V Correlates with observed primary completion rates and causal place effects



Note: This figure shows the results from running regressions of observed permanent resident primary completion rates and estimated causal place effects against place characteristics including census fixed effects. Each dot (with corresponding confidence intervals) represents a coefficient from an individual regression with the explanatory variable indicated on the y-axis. In orange, the dependent variable is observed permanent residents' primary completion, and in blue it is estimated causal place effects multiplied by thirteen. In italics, beneath characteristic labels, the proportion of the total correlation that operates through the causal place effects channel is noted. Standard errors are clustered at the locality level. Regressions are weighted by locality population.

I can also consider heterogeneity in characteristics across the three countries I study. In the online appendix, I present a figure analogous to figure V allowing each coefficient to vary by country. Although the broad qualitative conclusions from the pooled analysis hold, important differences arise. I find that, whereas the causal impact is very close to the overall impact for all variables in Mali, in Benin, and Cameroon it the observed correlations operate significantly less through the place effects channel.

Of the place-level characteristics considered in figure V, the proportion of Muslims in a location has the strongest relationship with observed primary completion rates, but one of the weakest with causal place effects. In Benin, Cameroon, and Mali, as well as many other countries across the world, religion, and education intersect [Auriol et al., 2020, Panin et al., 2021, Stoop et al., 2019]. The reasons behind known large differences in primary completion rates between those following different religions are often attributed to identity and legitimacy Manglos-Weber [2017], culture and persistence from colonization Platas [2018], the role of religious schools Tsimpo and Wodon [2014], Boyle [2014], Izama [2014], or religious violence Bertoni et al. [2019]. Armed with locality-specific causal place effects in figure V I go further than previous work and find that the spatial relationship between observed local primary schooling outcomes and the predominant local religion operates mainly through the selection effects channel. Table 1.A.1 in the appendix shows that the relationship between religion and causal place effects is robust, and considers heterogeneity by country.²² In Benin, the association is much weaker and mainly driven by place effects. On the other hand, in Cameroon, the association is stronger and it appears that place effects have a small, but not insignificant, role to play. Finally, in Mali, no statistically significant relationship can be discerned due to a lack of religious variation.

In sum, these results show that the previously observed relationship between development/ opportunity and religion is more nuanced than broad cross-country averages allow. Although the predominant religion in a locality does impact the opportunities children growing up experience, observational correlations across space will significantly overestimate the role of place, and averages across countries will obscure significant cross-country heterogeneity.

1.6 Heterogeneity by gender and the impact of place on gender inequality of opportunity

In my sample, 39.82% of males complete primary school but only 25.68% of females do. Whereas there have been aggregate increases in education completion, the gender gap has remained stubbornly high, in Benin (1992-2013) increasing from 17 to 20 percentage points, in Cameroon (1976-2005) decreasing from 21 to 12 percentage points and in Mali (1998-2009)

²²Information on religion is only recorded in 2005 in Cameroon and 2009 in Mali and all years in Benin, but I will focus on 2013. Therefore, for this analysis, I restrict my sample to include only the most recent census waves for which I have data. In Benin and Cameroon 25.8% and 13.9% follow Islam respectively and 46.8% and 77.7% are Christians. In Mali, however, there is much less variation with 94.9% Muslim.

increasing from 6 to 8 percentage points. There is a large and well-developed literature studying the importance of female education in Sub-Saharan Africa, and the reasons behind this large gender education gap. Among the most common reasons cited are cultural norms, early marriage, the disproportionate burden of household work, teacher attitudes, teenage pregnancy, and religion [Ombati and Ombati, 2012, Baten et al., 2020, Tuwor and Sossou, 2008, Anderson, 2005, Eloundou-Enyegue, 2004, Johannes and Noula, 2011, Aikman and Unterhalter, 2005, Cooray and Potrafke, 2011]. However, little attention has been paid to the potential role of place.

In this section, I ask: does the locality in which a child grows up play as large a role in determining male as female outcomes? In addition, I consider which characteristics of locations are associated with relatively greater opportunity for females vis-à-vis males, and therefore contribute to narrowing the gender gap in primary education completion. To answer these questions I replicate the analysis in section 1.3 estimating overall and locality-specific causal place effects for each gender separately.

I find that females are more responsive to locality than males. Females converge to the outcome of female permanent residence at the rate of 3.39% a year, which is 45% faster than males who converge to the outcome of male permanent residence at the rate of 1.52% a year — a highly statistically significant difference. As the standard deviation of the change in location quality of movers is 21% larger for females than males, this implies that most of the difference in convergence rates observed is due to actual differences rather than just increased variability of female outcomes. The question of why females respond more to place effects is left for future research, but one suggestive correlation is that areas with higher primary completion rates also have lower fertility rates. Women in locations with 1SD higher observed primary completion rates have 16% fewer children than those in average primary completion rate locations.

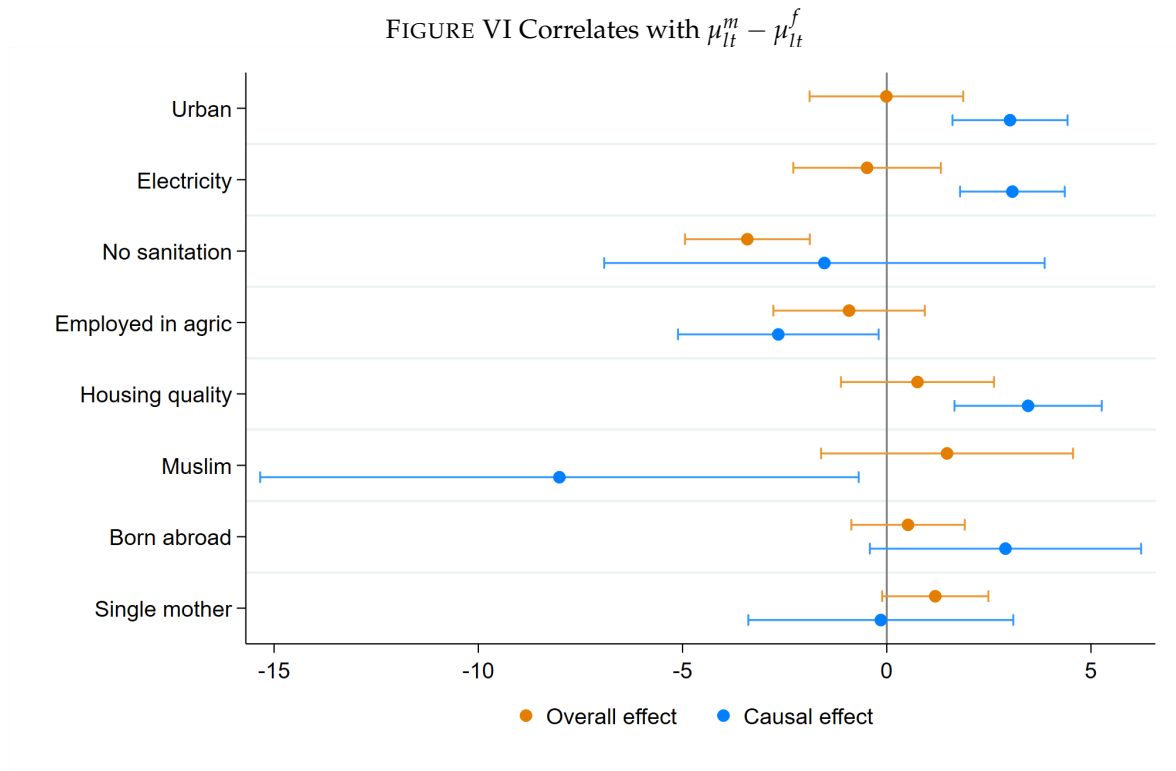
The finding that place matters more for female than male outcomes, opens up the possibility that place may have a role to play in determining the gender primary completion gap. To study this I replicate the analysis in subsection 1.4 on each gender-specific sub-sample. This procedure recovers two place-effect estimates for each locality-year cell, one for males and one for females. Each estimate is relative to the mean gender-specific causal place effects within that census. That is, a high female-specific causal place effect estimate in Benin in 2013 means that growing up in this location increases the probability that females complete primary school relative to females growing up in other localities in Benin in 2013. This does not allow a direct comparison between the level of female and male place effects as

gender-specific means are not identified, however, I can consider whether characteristics of locations affect relative gender effects differently.

First, I consider whether places that are relatively good for males are also relatively good for females by estimating the following relationship: $\mu_{lt}^f = \beta \cdot \mu_{lt}^m + \gamma_{census} + \varepsilon_{lt}$, where γ_{census} are census fixed effects. Which recovers $\hat{\beta} = 0.36$ (0.06). This significant and positive coefficient indicates that places that have relatively more opportunity for males also hold relatively more opportunity for females. However, the coefficient is significantly less than one, so although there is a strong correlation, it's far from a perfect relationship, leaving scope for locality characteristics to be differently correlated with gender-specific causal place effects.

In figure VI, I consider whether the locality characteristics explored in section 1.5 differently impact male and female relative place effects. I run regressions of the following form: $13 \cdot (\hat{\mu}_{lt}^m - \hat{\mu}_{lt}^f) = \beta \cdot X_{lt} + \gamma_{census} + \varepsilon_{lt}$. A β coefficient greater than 0 indicates that a given characteristic is more strongly correlated with a high male relative to female causal place effect. For example, figure VI shows that the coefficient on urban is around two, and statistically different from zero (at the 95% significance level). This means that a locality with one standard deviation higher urbanization rate is associated with opportunity for males 2 percentage points higher than that for females. Therefore, locations with higher urbanization rates are correlated with greater within-location gender inequality of opportunity. Figure VI shows a similar story for electricity availability and housing quality, but the opposite story for agricultural employment, and the proportion of the population who are Muslim. That is, if a larger proportion of the population is employed in agriculture, this is associated with better relative causal place effects for females vis-à-vis males and thus indicates that agricultural employment is associated with lower gender differences in opportunity across space.

Variables most commonly associated with development such as the urbanization rate, electricity provision, and housing quality, are found to be associated with worse gender equality of opportunity. This cautions against a uni-dimensional view of development and suggests that care needs to be taken for policy aimed at increasing development not also increasing gender inequality. This result is reminiscent of the U-shaped female labor supply development curve [Goldin, 1994, Ngai et al., 2022] whereby as a country moves away from agriculture female labor supply first decreases before increasing with the rise in services.



Note: This figure shows the results from running regressions of the difference between relative male and female causal place effects against various place-level characteristics. Regressions are weighted by population, include country-year fixed effects, and standard errors are clustered at the locality level.

Characteristics may also have differing impacts on local gender inequality of opportunity across each of the countries studied. In the online appendix, I present an analogous figure to figure VI allowing each coefficient to vary by country. These results show broad similarities across countries but also highlight some areas of difference. Urbanization and electricity provision have a similar effect, as to increase gender inequality, in Benin, Cameroon, and Mali. Sanitation appears to have little impact in all three countries. Agricultural employment and housing quality display some differences. For both, Cameroon displays a weaker relationship than either Benin or Mali — however, power concerns preclude any strong conclusions from being drawn.

Finally, in section 1.5 I found that although the predominant local religion is highly predictive of observed primary completion rates, it has little correlation with average causal place effects. Figure VI shows that this average null effect hides considerable gender heterogeneity. Perhaps surprisingly, the negative coefficient suggests that a greater Muslim population is relatively better for female vis-à-vis male opportunity in a region. In table 1.A.2 in the appendix I consider robustness of this result and heterogeneity by country. I find that this effect is driven by Benin, with Cameroon showing little impact and the estimate for Mali

being imprecise no doubt due to the lack of religious variation in Mali.

In this section, I have shown that place matters more for female than male outcomes and that although place effects are correlated across genders, some locality characteristics are associated with exacerbating or reducing the gender gap in opportunity. Spatial inequality has a role to play in reducing gender inequality, however as variables commonly associated with development are also associated with increasing gender inequality, future research is needed to identify how place-based interventions may best be harnessed by policymakers to reduce gender imbalances whilst promoting development.

1.7 Conclusion

In this paper, I have quantified and explored the impact of spatial inequality of opportunity relating to primary school completion in Benin, Cameroon, and Mali using a movers design. I find that moving to a better location at birth has a significant impact on a child's likelihood of completing primary school. Specifically, moving to a one standard deviation better location at birth increases the probability of completing primary school by 15 percentage points. Gender is an essential factor in this analysis, with place having a more significant effect on girls than boys. Furthermore, local characteristics such as urbanization, access to electricity, and employment in agriculture are correlated with place effects.

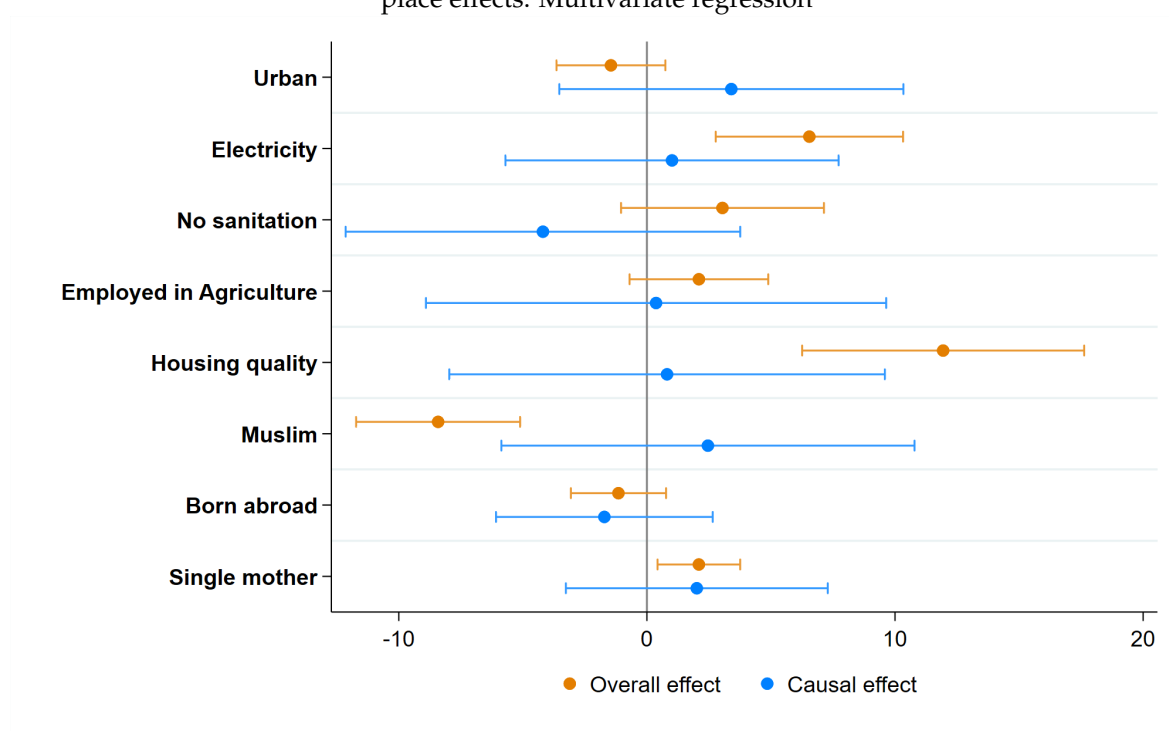
Religion also plays a role in the educational outcomes of children in these countries. Although it is weakly correlated with average causal place effects, it is strongly correlated with the gender gap in the effect of place on educational attainment. These findings highlight the importance of considering both gender and local characteristics in addressing spatial inequality of opportunity and improving access to education. This work illustrates the need for, and potential scope of, place-based policies in reducing inequality of opportunity. In future work, it will be crucial to understand the underlying causes of this variability, and arm policymakers with the tools to mediate spatial inequality of opportunity most effectively.

Appendix

1.A Additional results

In this appendix, I present additional results referenced in the main text. Figure 1.A.1 shows the results from running a multivariate regression of place characteristics on either observed primary completion rates or estimated causal place effects. Tables 1.A.1 and 1.A.2 show robustness to the documented correlation between religion and place effects or the difference between gender place effects respectively.

FIGURE 1.A.1 Correlates with observed primary completion rates and causal place effects: Multivariate regression



Note: This figure shows the results from running regressions of observed permanent resident primary completion rates and estimated causal place effects against place characteristics including census fixed effects. Each dot (with corresponding confidence intervals) represents a coefficient from a single multivariate regression of all characteristics. In orange, the dependent variable is observed permanent residents' primary completion, and in blue it is estimated causal place effects multiplied by thirteen. Regressions are weighted by locality population.

TABLE 1.A.1 Impact of the proportion following Islam on observed primary completion rates and causal place effects (CPE).

	(1)	(2)	(3)	(4)	(5)	(6)
	Primary comp.	CPE	Primary comp.	CPE	Primary comp.	CPE
Prop Muslim	-0.519*** (0.0906)	-0.123 (0.0772)	-0.562*** (0.0924)	-0.207*** (0.0658)		
Prop Muslim x Benin					-0.257*** (0.0664)	-0.252* (0.134)
Prop Muslim x Cameroon					-0.795*** (0.172)	-0.172*** (0.0499)
Drop Mali			X	X	X	X
R^2	0.641	0.0356	0.514	0.126	0.584	0.129
N	146	146	102	102	102	102

Note: This table shows the results from regressing either observed primary completion rates or estimated causal place effects against the proportion who indicate that they are Muslim in a locality. Regressions include country fixed effects, are weighted by locality population, and standard errors are robust. In Columns one, three, and five the dependent variable is the observed primary completion rate, in columns two, four, and six it is the estimated causal place effect. In columns, three to six Mali has been dropped from the sample. Due to data constraints I only use the 2005 census in Cameroon, the 2009 census in Mali, and the 2013 census in Benin when estimating these regressions.

TABLE 1.A.2 Impact of the proportion following Islam on observed gender differences in primary completion rates and causal place effects (CPE).

	(1)	(2)	(3)	(4)	(5)	(6)
	Male - Female CPE	diff_prim	Male - Female CPE	diff_prim	Male - Female CPE	diff_prim
Prop Muslim	-0.192** (0.0888)	0.0354 (0.0374)	-0.185** (0.0899)	0.0344 (0.0383)		
Benin × Prop Muslim					-0.372* (0.191)	-0.165*** (0.0223)
Cameroon × Prop Muslim					-0.0517 (0.0578)	0.176*** (0.0563)
Mali × Prop Muslim					-0.495 (0.475)	0.0785 (0.0662)
Drop Mali			X	X		
R ²	0.0315	0.270	0.0444	0.212	0.0504	0.506
N	137	137	96	96	137	137

Note: This table shows the results from regressing either the gender difference in primary completion rates or estimated causal place effects against the proportion who indicate that they are Muslim in a locality. Regressions include country fixed effects, are weighted by locality population, and standard errors are robust. In Columns one, three, and five the dependent variable is the observed gender difference in primary completion rates, in columns two, four, and six it is the estimated gender difference in causal place effects. In columns, three and four Mali has been dropped from the sample. Due to data constraints I only use the 2005 census in Cameroon, the 2009 census in Mali, and the 2013 census in Benin when estimating these regressions.

1.B Validating Causal Place effects

The identifying assumption underpinning the above analysis, that selection effects into place do not vary with the age of move, is potentially strong. For example, this prohibits those with better unobservables from systematically moving to better areas earlier. However, it doesn't preclude selection overall. I allow $\delta > 0$, that is for those with better characteristics to move to better places, just that this does not vary systematically with the child's age at move. This is also a common identifying assumption used in the literature [Chetty et al., 2016, Chetty and Hendren, 2018a,b, Chyn, 2018, Damm and Dustmann, 2014, Deutscher, 2020, Alesina et al., 2021, Laliberté, 2021].

I interrogate this assumption in four ways. Firstly, I consider whether it holds for characteristics of individuals we can observe. Secondly, I relax the assumption by including household fixed effects absorbing any time-invariant selection and only using cross-sibling

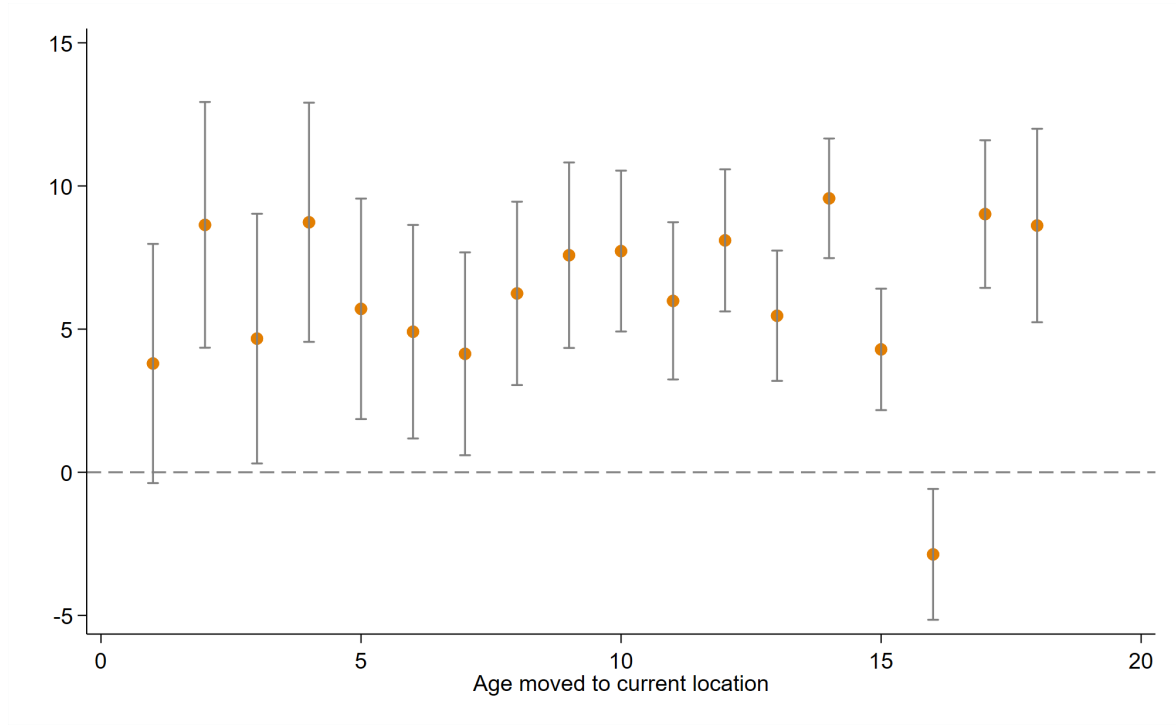
variation in the age at move. Thirdly, I consider a sample of increasingly likely to be exogenous moves by looking at increasingly unusual migration episodes. Lastly, I consider an over-identification outcome-based test where I show that individuals respond to their own-groups outcomes in line with the place effects, rather than selection effects, story. In all four approaches, I find evidence to support the identifying assumption. This means that any omitted variable bias or selection effects would have to operate within families proportionally to exposure times, persist in the presence of moves induces by displacement shocks, and precisely replicate permanent residents outcomes by gender in proportion to exposure times in addition to not being observed in observable dimensions such as mothers education. I find it unlikely that any qualitatively or quantitatively important threats to identification operate in this manner.

1.B.1 Is there evidence of varying selection effects on observables

For characteristics of individuals that are observed, I can directly consider the identifying assumption. Of course, not finding selection on some dimension doesn't preclude selection on another potentially unobserved dimension. Due to the cross-sectional nature of my data, I do not observe time-varying characteristics of individuals, however, as I can construct households I can consider selection on time-invariant characteristics of the household such as parents' education level. To investigate selection I first partition the distribution of observed differences in location quality due to moves, that is $\Delta \bar{y}_{ods}^p$ into below-average moves and above-average moves. I then compare the difference in the average quality of individuals between those who made *good* vs *bad* moves at each age at move. If I find that overall those who made better moves had better observables this is evidence of selection. If I additionally find evidence of systematic variation in this relationship by age at move - this would be evidence against the identifying assumption of constant selection effects.

Figure 1.B.1 shows the results from performing this kind of analysis where we consider the individual mothers' primary completion rate as our observable measure of individual quality. Consistent with figures in the main text I find evidence of significant positive selection effects overall. In addition, and also consistent with evidence from post-12 moves I find no slope in selection effects over child's age at move. That is, I do not find evidence that individuals with more educated mothers systematically move to better areas earlier or later. On this dimension at least, there is no evidence of selection effects varying by age at move.

FIGURE 1.B.1 Children with more educated mothers don't move to better areas earlier



Note: This figure plots the average difference in mover children mothers' primary completion rates between those who move to better vs worse locations by the age of the child when the move occurs. Vertical lines indicate 95% confidence intervals.

1.B.2 Focusing on within household variation

By including household fixed effects I can relax the assumption of constant selection by allowing selection effects to vary across, but not within, households and only consider variation in age at move between children within the same household. This specification estimated the following equation: $y_i = \rho_{os} + \alpha_m + \beta_m \cdot \Delta \bar{y}_{ods}^p + \psi_h + \varepsilon_i$. Where h denotes i 's household. This specification imposes considerable data requirements, firstly I can only consider households with multiple children between 14 and 18 at home²³ and secondly ψ_h is a high dimensional object²⁴ to be estimated. To reduce the data burden I estimate the linear parametric model, results reported in table 1.B.1.

²³This cuts the sample with estimating variation from 79,778 to 27,807.

²⁴In my main regressions there are 12,175 households.

TABLE 1.B.1 Parametric estimates of place effects including HH fixed effects

	Year born - birth location and age at move FE	Year born, year born by $\bar{y}_{os}$ and age move all by sample FE	Year born, year born by $\bar{y}_{os}$ and age move FE	Year born and age moved FE
$\hat{\beta}_1$	-0.0181** (0.00834)	-0.0180*** (0.00618)	-0.0201** (0.00785)	-0.0349*** (0.0114)
$\hat{\beta}_2$	0.00245 (0.0131)	-0.0141 (0.0145)	-0.0141 (0.0130)	0.00878 (0.0197)
R^2	0.779	0.745	0.741	0.727

Note: This table shows the robustness of the parametric convergence rate estimates over various fixed effects specifications where in all specifications I control for household fixed effects and therefore only use between sibling variation.

Although recovering a smaller coefficient than when estimating on the full sample without household fixed effects, the difference is not statistically significant and may be a function of attenuation bias from the demanding specification (incidental parameter problem). This points to the conclusion that across household selection is unlikely to be driving a meaningful portion of causal place effects I find. The results are also robust to alternative specifications as shown in table 1.B.1.

1.B.3 Time varying selection effects — increasingly likely to be exogenous moves

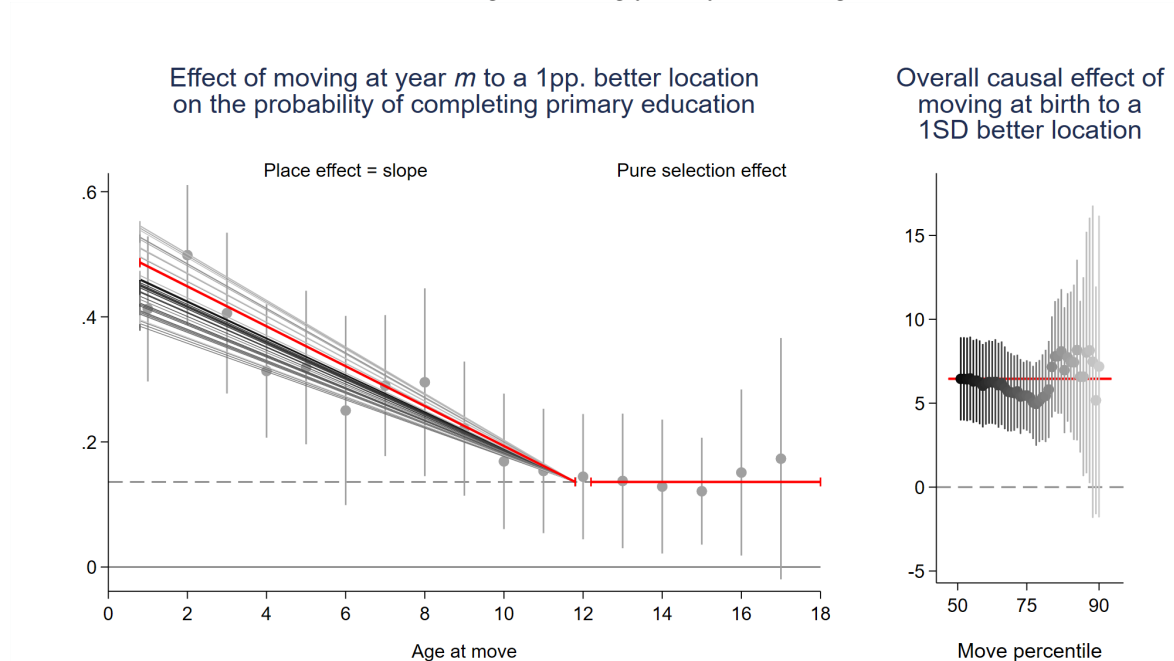
Time varying shocks could still invalidate my identifying assumption. A household could receive a windfall simultaneously both and move to a better location and increase investment in their child's education. To investigate whether this kind of force invalidates my identifying assumption I consider plausibly exogenous moves in two ways.

I use the data-driven strategy adopted by Chetty and Hendren [2018a,b] to identify location-year cells that exhibited likely to be exogenous movement events. The data-driven method of Chetty and Hendren [2018a,b] first constructs a *shock* variable k_{lt} for each location l and year t which is the size of that years out movers relative to that regions average number of out movers: $k_{lt} = \frac{Out_{lt}}{\overline{Out}_l}$. Where Out_{lt} is the number of out-movers from region l at time t and $\overline{Out}_l = \sum_t Out_{lt}$ is the locality average over time t . High values of k_{lt} represent

unusually large outflows in a given year from that given area. This method hypothesizes that higher values of k_{lt} represent unusual episodes and that these moves are more likely to be exogenous. Note that we don't require some cut off after which we determine that all such moves are undoubtedly exogenous. Instead, we only suppose that the proportion of moves due to exogenous reasons is increasing with k_{lt} . If place effect estimates remain stable as the sample is restricted to larger and larger values of k_{lt} , that is more and more likely to be exogenous moves, this is evidence in favor of the identifying assumption.

Figure 1.B.2 reports $\hat{\beta}_1^{k_r}$ found by estimating equation 4 in the main text on a sample of location years for which $k_{lt} > k_r$ where k_r is the r 'th percentile of k_{lt} . In both figures, I find little systematic evidence of our estimated exposure effects being driven by time-varying selection. That is, there is no significant evidence of the effect eroding as we restrict variation to moves more likely to be exogenous. In this graphical analysis I set selection effects equal to aid in the comparability of the slope, but when running the regressions allow this intercept to vary flexibly.

FIGURE 1.B.2 Considering increasingly likely to be exogenous moves



Note: This figure shows in the left hand side panel the estimated β_m coefficients from the regression: $y_i = \rho_{os} + \alpha_m + \beta_m \cdot \Delta \bar{y}_{ods}^p + \varepsilon_i$ estimated on a sample of one time movers between the ages of 14 and 18 in Benin, Cameroon and Mali. Additionally, corresponding 95% confidence intervals are indicated by gray bars where standard errors have been clustered at the birth locality and current locality level. Superimposed on this in red are two lines of best fit with a break at age 12. The horizontal gray dashed line indicated the post 12 mean of $\hat{\beta}_m$. Results from the linearised specification are also plotted in increasingly light gray where the sample is increasingly restricted to evermore unusually large movement events. In the right-hand side panel, the slope of the 1 to 12 line of best fit has been translated into the overall impact of moving at birth to a one standard deviation better location where over the whole sample a one standard deviation improvement in location quality is equal to a 20pp increase in primary completion rates. The overall effect is represented by the red vertical line and the corresponding points and standard errors in gray calculated over increasing likely to be exogenous move events are plotted for comparison.

1.B.4 Over-identification outcome-based test

Finally, I use heterogeneity in permanent residents' outcomes across reported gender to consider the validity of the identifying assumption. Intuitively in the causal place effect model movers' outcomes should converge to permanent residents' outcomes for their own subgroup in this case gender. On the other hand, a selection effects model would predict convergence to general location quality. The second statement requires either that households do not observe gender-specific location quality above overall location quality or that if they

do observe it they don't select on it. The former seems a reasonable informational assumption, especially when you consider the high correlation between gender-specific location qualities, and that I don't find that individuals move disproportionately to better areas as shown in the main text.

To formally test the hypothesis proposed above, I estimate the main parametric specification firstly including the own-gender measure of location quality — that is the primary completion rate of permanent residents who have the same gender as i , secondly with the other-gender measure of quality and finally with both. It's worth noting two things, first male and female primary completion rates across localities are highly correlative with correlation 0.82. Second, by reducing the number of permanent residents in each location we introduce potential measurement error in the main explanatory variable which could result in attenuation bias. Table 1.B.2 shows the results.

TABLE 1.B.2 Outcome based placebo test				
	(1)	(2)	(3)	(4)
Base	0.0209*** (0.0029)			
Same reported gender specific		0.0176*** (0.0025)		0.0123*** (0.0041)
Other reported gender specific			0.0164*** (0.0025)	0.00664 (0.0041)
N	65560	65560	65560	65560
R^2	0.358	0.356	0.358	0.358

Note: This table shows the results from the outcome-based over-identification test. Column one replicates the main analysis on the slightly smaller sample where gender information is available. Column two uses a gender-specific measure of location quality, that is the primary completion rate of permanent residents of the same gender born in the same year. Column three uses the opposite gender-specific measure of location quality. In column four both measures are included in one regression. Standard errors are clustered at the origin and destination locality level. Age at move and origin by census year fixed effects are included in all regressions.

Column one of table 1.B.2 re-estimates my main equation on the reduced sample for which gender is reported to provide a relevant comparison unit. Column two reports results using the permanent resident outcomes of those individuals who have the same gender as the moving child. Column three reports results using permanent residents outcomes for those who have the opposite gender to the moving child, it's unsurprising here that the

coefficient changes little from column one to column two given the high correlation between the two measures. Column four shows the results from including both measures in the same regression. Here, if both coefficients were similar and attenuated that would be evidence for the selection effects model. However, that's not what I find, instead, the same-gender coefficient remains positive and highly significant whereas the other-gender coefficient falls to a precisely estimated 0. Therefore, instead, I find evidence in favor of the causal place effects model and thus the main identifying assumption.

In sum, across all four approaches taken, I find consistent and strong evidence in favor of the key identifying assumption. Any omitted variable bias or selection effects would have to operate within families proportionally to exposure times, persist in the presence of moves induces by displacement shocks, and precisely replicate permanent residents outcomes by gender in proportion to exposure times, and not show up in observable such as mothers education. I find it unlikely that any qualitatively or quantitatively important threats to identification operate in this manner.

1.B.5 Locality-year specific identifying assumption

When estimating locality-year specific fixed effects I require a stronger identifying assumption: that selection effects are constant over age at move for each location l in each period t , rather than just on average. Formly I require that: $\mathbb{E}[e_{il} \cdot \varepsilon_i | \alpha_{odt}] = 0 \quad \forall l \in \mathbb{L}$.

The results supporting the validity of the on-average identifying assumption detailed above also provide evidence for this location-specific assumption. In addition, I perform a placebo test estimating equation 1.3 on a sample of those who move between 14 and 18, *after* primary education completion is determined. In this case, place effects μ_l should just be noise and unrelated to the actual estimated place effects. A scatter diagram in the online appendix summarizes the relationship between actual and placebo place effects, statistics are weighted by locality population. Actual and placebo estimates exhibit little relationship with a regression slope coefficient indistinguishable from zero (0.07 (se=0.05)) and a very small R-squared (0.01). In addition, less than five percent of the individual placebo coefficients are significant at the 5% level.

1.C Signal to noise ratio of estimated place effects

I can estimate the signal to noise ratio by decomposing estimated place effects into actual place effects and orthogonal measurement error, $\hat{\mu}_{lt} = \mu_{lt} + \eta_{lt}$, and estimating the true

variation, $\hat{\sigma}_{\mu_{it}}^2 = \sigma_{\mu_{it}}^2 - \sigma_{\eta_{it}}^2$. A natural estimator of $\sigma_{\eta_{it}}^2$ is given by the average squared standard errors s_{it}^2 recovered when estimating equation 1.3: $\sigma_{\eta_{it}}^2 = E[s_{it}^2]$. Using this decomposition I estimate the proportion of the raw variance which is due to signal not noise as $\frac{\hat{\sigma}_{\mu_{it}}^2}{\sigma_{\mu_{it}}^2} = \frac{\sigma_{\mu_{it}}^2 - \frac{1}{N} \sum s_{it}^2}{\sigma_{\mu_{it}}^2} \times 100 = 41.4$, that is 41.4% of the raw variance is due to signal not noise.²⁵ The signal standard deviation of place effects is estimated to be 15 indicating substantial spatial inequality of opportunity.

1.D Robustness

1.D.1 Linear model

Building on the linearity of place effects, I estimate a parametric version of my main equation allowing a different slope and intercept for children moving between the ages of 1 and 12 and those moving after age 12:

$$\begin{aligned}
 y_i = & \rho_{os} + \alpha_m \\
 & + (\alpha_1 + \beta_1 \times \text{age move}_i) \cdot \mathbb{1}[\text{age move}_i \in [1, 12]] \cdot \Delta \bar{y}_{ods}^p \\
 & + (\alpha_2 + \beta_2 \times \text{age move}_i) \cdot \mathbb{1}[\text{age move}_i \in [13, 18]] \cdot \Delta \bar{y}_{ods}^p + u_i \quad (1.7)
 \end{aligned}$$

Table 1.D.1 shows the results from estimating equation 1.7. The rate of convergence, β_1 , is as found in the non-parametric specification. Additionally, the post-12 slope is a precisely estimated 0. The first column of table 1.D.1 estimates the baseline specification, and columns 2, 3, and 4 estimate similar equations with varied fixed effects specifications. Column two replaces the cohort (year born) by birth location fixed effects with individual cohort fixed effects, allows cohort-specific slopes on birth location quality $\bar{y}_{os}^p$, and allows age at move fixed effects to vary by census. Column three removes the more flexible at age of move fixed effects of column 2. Finally, column four removes the year by $\bar{y}_{os}^p$ fixed effects and so no longer controls for origin location in any way. In all specifications, I find little change in results from the main specification given in column one.

²⁵Compared to 16% for the analogous commuter zone estimates in [Chetty and Hendren \[2018b\]](#).

TABLE 1.D.1 Parametric estimates of place effects

	Year born - birth location and age at move FE	Year born, year born by $\bar{y}_{os}$ and age move all by sample FE	Year born, year born by $\bar{y}_{os}$ and age move FE	Year born and age moved FE
$\hat{\beta}_1$	-0.0270*** (0.00530)	-0.0272*** (0.00500)	-0.0284*** (0.00554)	-0.0319*** (0.00729)
$\hat{\beta}_2$	0.00302 (0.00914)	0.00893 (0.00631)	0.00553 (0.0120)	-0.00742 (0.0234)
N	79778	79956	79970	79970
R^2	0.348	0.311	0.303	0.155

Note: This table shows the results from estimating equations of the form $y_i = FE + (\alpha_1 + \beta_1 \times \text{age move}_i) \cdot \mathbb{1}[\text{age move}_i \in [1, 12]] \cdot \Delta \bar{y}_{ods}^p + (\alpha_2 + \beta_2 \times \text{age move}_i) \cdot \mathbb{1}[\text{age move}_i \in [13, 18]] \cdot \Delta \bar{y}_{ods}^p + u_i$, where FE are the fixed effects specification which varies by column. In column one I include year born by birth location and age at move fixed effects (baseline specification); in column two I include year born, (slopes of) year born by origin location quality and age at move by sample fixed effects; in column three I include year born, (slopes of) year born by origin quality and age at move fixed effects; in column four I include year born and age at move fixed effects. Standard errors clustered at origin and destination.

1.D.2 Robustness of the LPM

In the main text I implicitly assume a linear probability model when performing the non-parametric or linearly parameterized regressions. In this section, I consider whether results substantively change if instead an alternative functional form is considered. In table 1.D.2 I replicate the results from table 1 in the main body in columns 1 and 2, where *few fixed effects* refers to year born, year born by locality quality, and age at move fixed effects. Columns 1 and 3 of table 1.D.2 use the full fixed effect specification of year born by birth location and age at move fixed effects. The other columns use the few FE specification to mitigate incidental parameter bias known to be pervasive in probit and logit settings. The coefficients all have the same sign and tell the same story of an exposure effect up to age 12 and no impact thereafter.

TABLE 1.D.2 Robustness of the main results to the linear probability assumption

	(1)	(2)	(3)	(4)	(5)	(6)
	LPM	LPM	PPML	PPML	Probit	Logit
1 to 12 slope	-0.0270*** (0.00530)	-0.0283*** (0.00545)	-0.0452*** (0.0113)	-0.0478*** (0.0121)	-0.0871*** (0.0113)	-0.145*** (0.0189)
13 to 18 slope	0.00302 (0.00914)	0.00469 (0.0119)	0.00210 (0.0185)	0.00951 (0.0248)	0.0220 (0.0246)	0.0295 (0.0418)
1 to 12 constant	0.486*** (0.0496)	0.514*** (0.0611)	0.856*** (0.122)	0.950*** (0.133)	1.557*** (0.178)	2.574*** (0.292)
13 to 18 constant	0.0932 (0.122)	0.0805 (0.166)	0.208 (0.238)	0.174 (0.346)	0.135 (0.367)	0.339 (0.624)
Year born by birth location and age moved FE	X		X			
Year born, year born by birth location quality and age moved FE		X		X	X	X
R ²	0.348	0.304				
N	79778	79778	78950	79776	79776	79776

Note: This table shows the results of re-estimating the linearly parametrized specification relaxing the linear probability assumption. Column one shows the baseline results using a linear probability model and the full fixed effects. Column two also uses a linear probability model with *easy* fixed effects. Columns three and four use pseudo poisson maximum likelihood estimation with the full and easy fixed effects respectively. Column five estimates a probit model with easy fixed effects and similarly column six estimates a logit model.

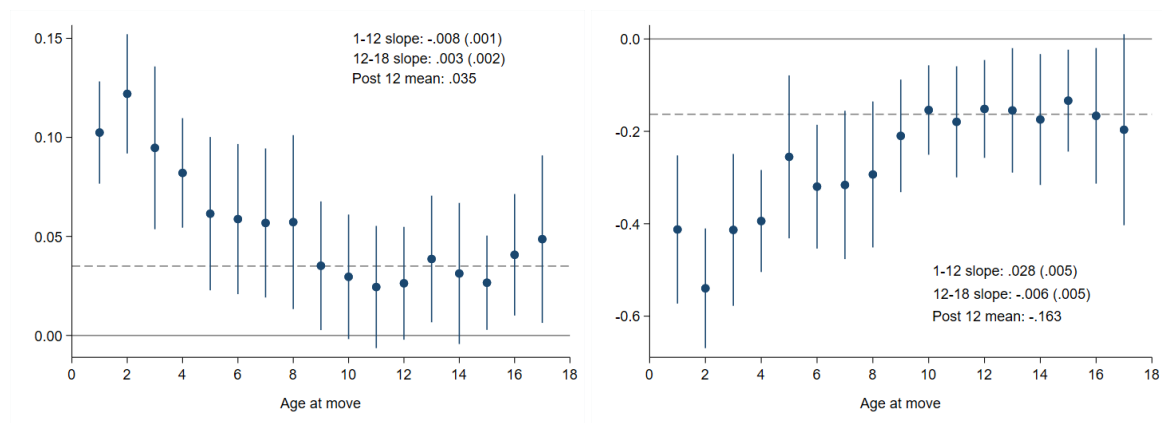
Of course, one cannot directly compare coefficients across each type of model. PPML approaches in columns (3) and (4) estimate a semi-elasticity and so can be converted into percentage point effects (as with the linear probability model) by multiplying by the mean of the left-hand side variable, which in the estimating sample is 0.64. This results in very similar estimates to those in columns (1) and (2). As the logistic and normal distributions are similarly shaped up to a scaling factor, it is known that logit coefficients will roughly be 1.82 times the probit coefficients - which is indeed what we find. Logit coefficients can in turn be interpreted as the percent impact on the odds ratio which is on average 1.78.

1.D.3 Considering different measures of location quality

In the above results, I've used primary completion as the outcome variable and average primary completion among permanent residents as the measure of location quality. It's also interesting to consider whether different measures of location quality generate similar results. In figures [1.D.1a](#) and [1.D.1b](#) I show analogous results to the main result presented in

figure III where instead of average primary completion among permanent residents I consider average housing quality and employment in agriculture among permanent residents respectively as my measure of location quality (the outcome variable remains primary completion as before).

FIGURE 1.D.1 Causal place effects in other variables
(A) Housing quality (B) Agricultural employment



Note: This figure replicates the main results in the paper using different measures of location quality. Instead of measuring location quality as the primary completion rate among permanent residents, in the left panel, I use housing quality, and in the right agricultural employment (both for permanent residents). The dependent variable remains primary completion in all cases.

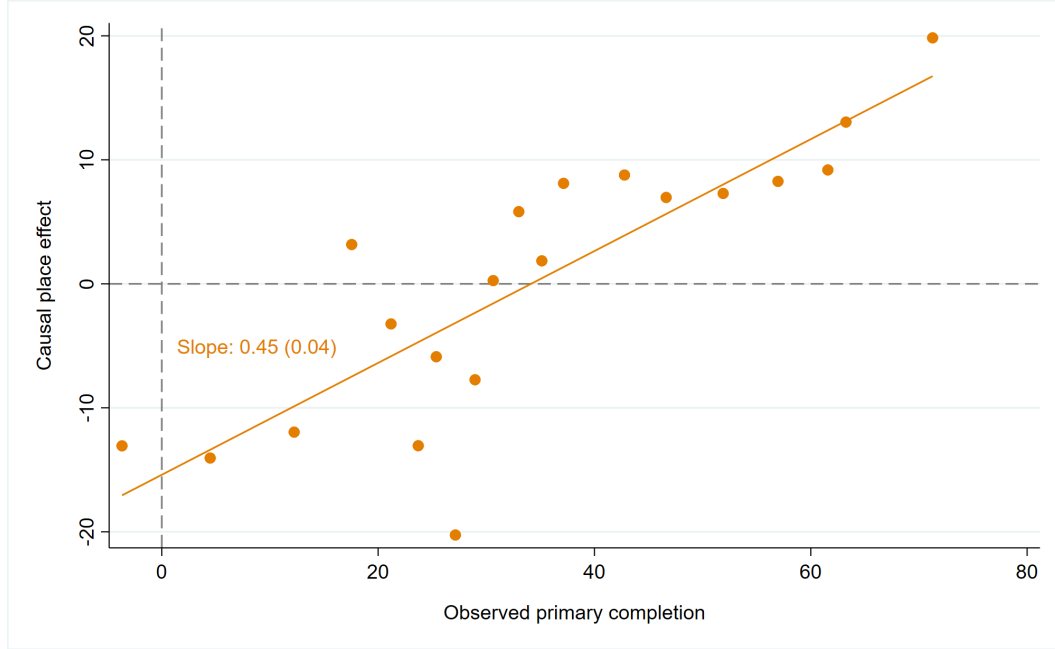
This analysis serves two purposes. First, it acts as a robustness/ sense check — indeed given that I find that location quality when measured in terms of primary completion matters it would be very surprising if place measured differently didn't matter. In this way, this exercise serves as a robustness check on my main results. It's also reassuring to find similar patterns, with convergence up until age 10 to 12 or so and a constant selection effect thereafter.

1.E Relationship between observed primary completion rates and estimated causal place effects

A simple sense-check on the estimated place effects is to study their relationship with observed primary completion rates. Figure 1.E.1 shows a (population-weighted) binscatter with the corresponding line of best fit both of which are residualized on census-level fixed

effects. There is a strong positive and tight relationship between observed primary completion rates and causal place effects with a slope coefficient of 0.45. Figure 1.J.1 maps the estimated place effects.

FIGURE 1.E.1 Relationship between observed primary completion rates and estimated relative place effects



Note: This figure shows the binscatter relationship between the observed primary completion rates and estimated causal place effects (which have been multiplied by 13 to put into *moved at birth* terms). Census fixed effects are included, as are population weights. Standard errors are clustered at the locality level.

1.F Locality-year specific identifying assumption

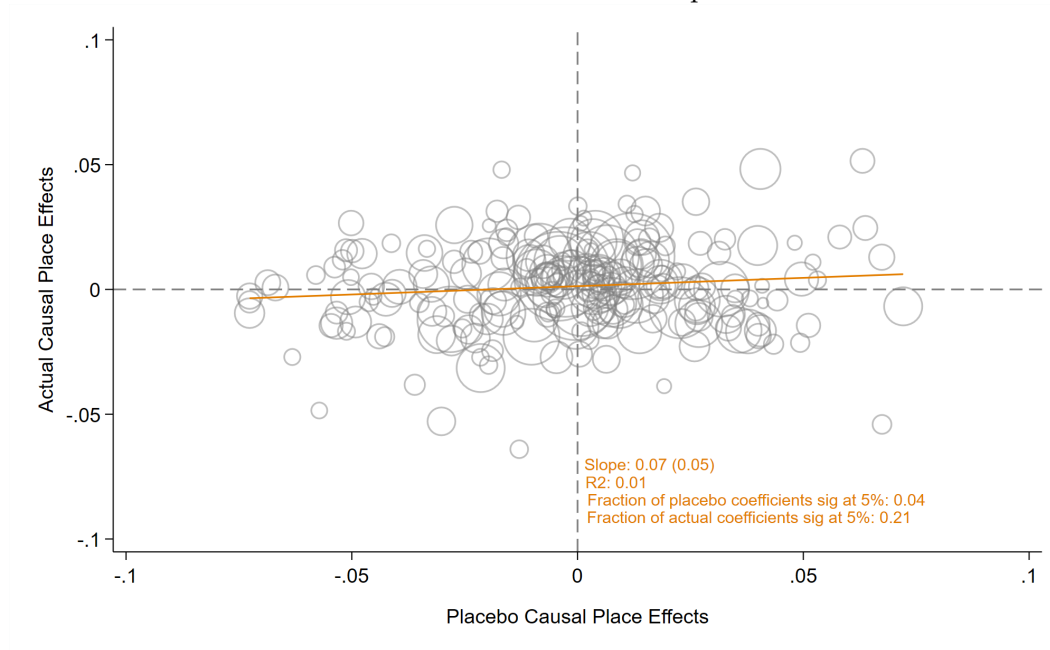
When estimating locality-year specific fixed effects I require a stronger identifying assumption: that selection effects are constant over age at move for each location l in each period t , rather than just on average.

Assumption 1. *Place specific identifying assumption:* $\mathbb{E}[e_{il} \cdot \varepsilon_i | \alpha_{odt}] = 0 \quad \forall l \in \mathbb{L}$.

The results supporting the validity of the on-average identifying assumption summarized in the main text also provide evidence for this location-specific assumption. In addition, I perform a placebo test estimating causal place effects on a sample of those who move between 14 and 18, *after* primary education completion is determined. In this case, place effects μ_l should just be noise and unrelated to the actual estimated place effects. Figure 1.F.1 summarizes the relationship between actual and placebo place effects, statistics are weighted by locality population. As is clear from this figure they exhibit little relationship

with a regression slope coefficient indistinguishable from zero and a very small R-squared. In addition, less than five percent of the individual placebo coefficients are significant at the 5% level.

FIGURE 1.F.1 Placebo vs actual causal place effects



Note: This figure shows the scatter plot relationship with a corresponding linear line of best fit between the actual and placebo-estimated place effects. Place effects in both instances have been residualised on year fixed effects. Circle size indicated population and the linear regression line has also been weighted by population.

1.G Standard errors

In all individual-level regressions, standard errors are clustered at the origin locality and destination locality level (as in [Chetty and Hendren \[2018a,b\]](#)) allowing correlation within both units. In this section, I additionally consider a simple permutation-based test of the main linearized regression given in equation 1.7. I run 1,000 simulations where I randomly re-sort journeys, defined as birth-destination location pairs, across movers and re-estimate equation 1.7 on each new data set. That is, I do not permute between movers and stayers nor across the year of move, but instead across journeys within movers. This allows me to retain any bias in the data that may be stemming from the non-permuted dimensions i.e. mover vs stayer and year of the move. The resulting slope coefficients from 1,000 replications are plotted in figure 1.G.1 and show clear evidence of significant effects.

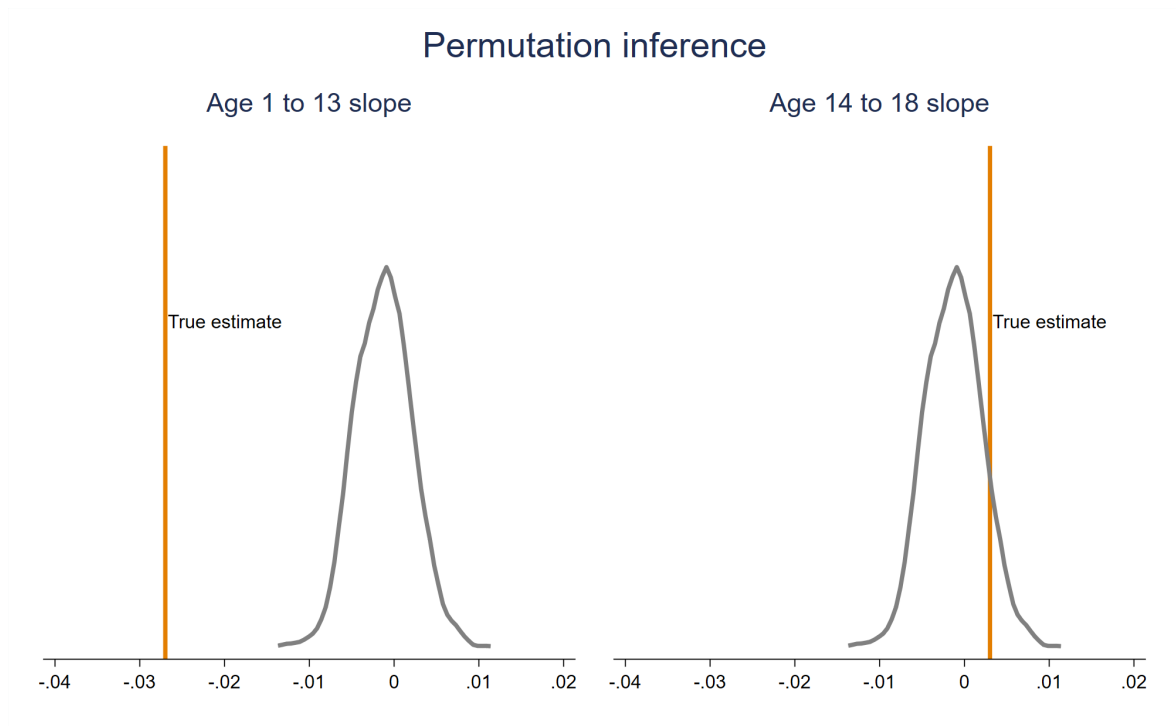


FIGURE 1.G.1 Permutation inference

Note: This figure shows the resulting distribution of estimated coefficients from a permutation test. I permute journeys (birth-destination pairs) across movers maintaining the observed year of move in 1000 replications estimating the linearised equation at each iteration.

1.H Comparability of movers and stayers

Those who move, and those who stay, are not the same. Those who migrate are often more wealthy (as migration is costly) and may self-select, especially on the urban-rural dimension [Lagakos, 2020, Hamory et al., 2021] to areas where they can be most productive. Throughout this analysis, I use variation in movers to estimate the causal effect of place on primary education completion. To generalize results to the whole population, I implicitly assume areas that exhibit high place effects for movers would also do so for stayers. This does not mean that I assume movers and stayers must look the same, but more specifically that the impact of where a child grows up on their probability of completing primary education must be similar. It is not necessary to make this assumption, as the results would remain of interest if they only pertained to movers who comprise close to 20% of 14 to 18 year-olds. However, in this section, I consider whether the results apply more broadly to both movers and stayers.

The assumption that place effects are the same for movers and stayers can be decomposed into two components. First, that movers and stayers do not systematically differ in

a manner that interacts with place effects. For example, it could be that movers are more entrepreneurial and therefore benefit relatively more from a high-quality location compared to stayers, and this is reflected in their propensity to educate their children. Secondly, it could be that movers and stayers are similar but due to the nature of having moved to a new location, movers interact differently with their surroundings. For example, a mover in a new location does not have connections or localized knowledge that a permanent resident may have. Therefore, although stayers and movers may be geographically co-located, they could interact with very different aspects of society.

To consider the second possible difference, I return to the main specification given in equation 1.2 and its linear parameterized version in equation 1.7. Due to the inclusion of origin by cohort fixed effects, this specification only uses variation in destination quality enjoyed after the move has occurred. Instead, I replace these fixed effects with destination by cohort fixed effects. This new specification uses variation in *origin* quality which movers enjoy *before* they moved. If the place effects I estimate when using variation in (to-be) movers' origin location quality are similar to those I obtain when using variation in their destination location quality, this is evidence that movers do not interact systematically differently with their new surroundings. Comparing convergence rates, I find that they are almost identical both statistically and economically. This implies that place exerts the same impact on those who are in their birth location and have yet to move as it does on those who have already moved and are residing in their destination location.

However, this result doesn't preclude the first concern, that movers and stayers are fundamentally different and that this interacts with place. To investigate this, I turn again to increasingly likely to be exogenous move events. In these cases, there is increasingly little selection into moving. Therefore, if I find similar causal place effects among the sample of more likely to be exogenous movers as I do in my overall sample, this will be evidence against the concern that movers are fundamentally different from stayers in a way that interacts with place effects. This is exactly what I find when performing this exercise in the online appendix.

I've highlighted two possible issues that can be ameliorated by considering increasingly likely to be exogenous moves: varying selection effects and the potential differences between movers and would-be stayers. It could be that these two effects are canceling each other out and that considering increasingly likely to be exogenous moves sheds no light on either issue. For example, suppose that selection effects are increasing in the child's age at move and that would be stayers (those who were it not for an exogenous move event,

would not have moved) experience much weaker place effects than movers. These two fundamental issues would be hidden by just looking at increasingly likely to be exogenous move events. However, it is unlikely that selection effects vary with the child's age at move in this manner. It is more intuitive to suppose that households with better unobservables move earlier to better places. For the observed result to hold it would have to be that stayers exhibit much *stronger* place effects than movers which seems unlikely, but if so implies that my analysis is a lower bound for place effects. Given the additional validation checks performed above, I find it unlikely that either force is at play here but cannot rule it out entirely.

1.I Descriptive statistics

1.I.1 Temporal variation in primary completion

Figure 1.I.1 shows the distribution of primary completion rates across localities for six countries in Sub-Saharan Africa, Benin Mali and Cameroon which are the main focus of this study as well as Tanzania, Zambia and Kenya for comparison²⁶. Evident in all countries is a considerable rightward shift in primary completion rates from the early to the late period, although these changes are not directly comparable as the size of the time difference varies between countries. What is also striking is the heterogeneity between Benin, Cameroon, and Mali highlighting the stark differences between these countries as well as variation that can be exploited when I consider heterogeneity. Cameroon has a far more diffused distribution as well as both relatively high means and a relatively large mean shift. Mali on the other hand is more bunched at the lower levels of completion and displays a more modest shift. Benin shows considerable gains seemingly across the whole distribution.

Although not crucial for my analysis, it's helpful for the interpretation of results that indeed most individuals who will at some point receive primary schooling do so by aged 14. This is evidence from figure 1.I.2 which shows the age fixed effects from a regression of primary completion against age and year fixed effects. This figure clearly shows, as anticipated, that most primary education is indeed completed by 14. However, it's worth noting that many children who haven't completed primary education by 12, which is the year officially primary schooling ends in the countries I study, go on to do so in the next few years.

²⁶This sample was chosen purely for practical data availability reasons, Tanzania Kenya and Zambia are the only additional countries in Sub Saharan Africa which fill the data requirements, although sadly they don't have sufficiently rich migration information to be part of my main sample.

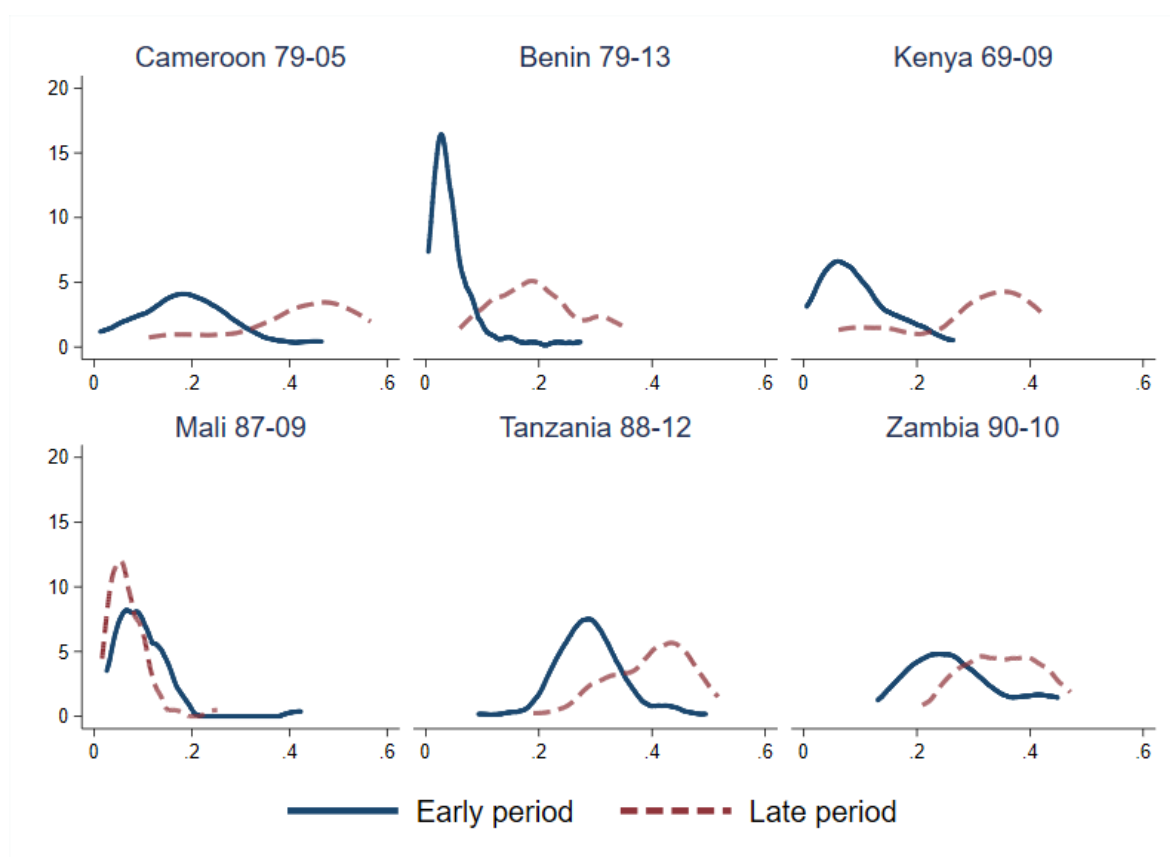


FIGURE 1.I.1 Change in the primary education completion distribution

Notes: This figure shows kernel density plots over locality level primary completion rates by country and year on a sample of over 12 year olds. In each case the solid blue line refers to the distribution in the earliest available census year in each country and the dashed red line the corresponding distribution in the most recently available census.

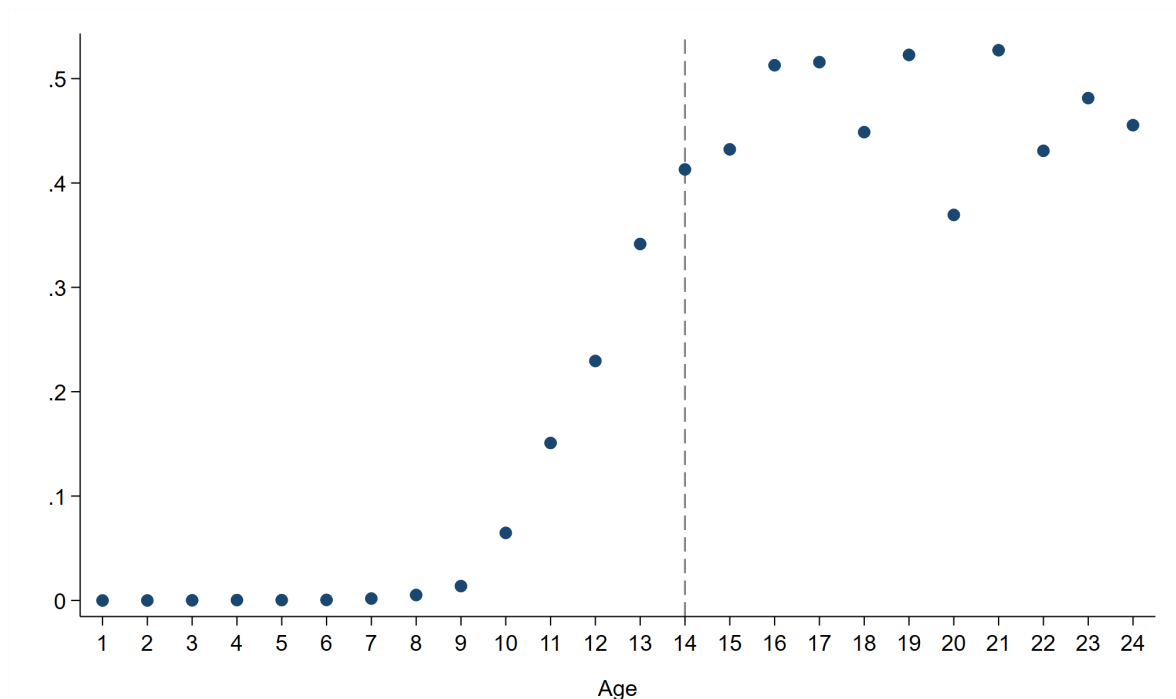


FIGURE 1.I.2 Proportion of the population who have completed primary school by age

Notes: This figure shows the proportion of the sample who report having completed primary education at the time of each census against their age. This figure uses data from the full sample Benin (1992, 2002, 2013) Cameroon (1976, 1987, 2005) and Mali (1998, 2009).

1.I.2 Sub-national migration patterns

The rich migration information collection in the censuses I use in this paper allows a detailed analysis of migration patterns at the sub-national level over a long time span. These patterns are important for my analysis because to estimate place effects I necessarily use variation in *movers* outcomes alone. If, for example, the majority of migration was towards the primate city I would have little external validity in considering counterfactual, or actual, changes in market access in the hinterland. Additionally, it would be challenging or impossible to estimate heterogeneity in effects across location characteristics. Figure 1.I.3 shows the proportion of migration directed towards localities with over 50% of residents self-reporting as living in an urban area. There are considerable differences across countries and time, allaying fears that we are only capturing one type of mover variation.

Additionally, figures 1.I.4a, 1.I.4b and 1.I.4c show flows in and out of each locality within each country over a 10 year period. Blue regions are classified as urban and red as rural. The size of the arc associated with each locality is proportional to the total movement in and out of the locality. One can see from these figures that sub-regional migration in Benin,

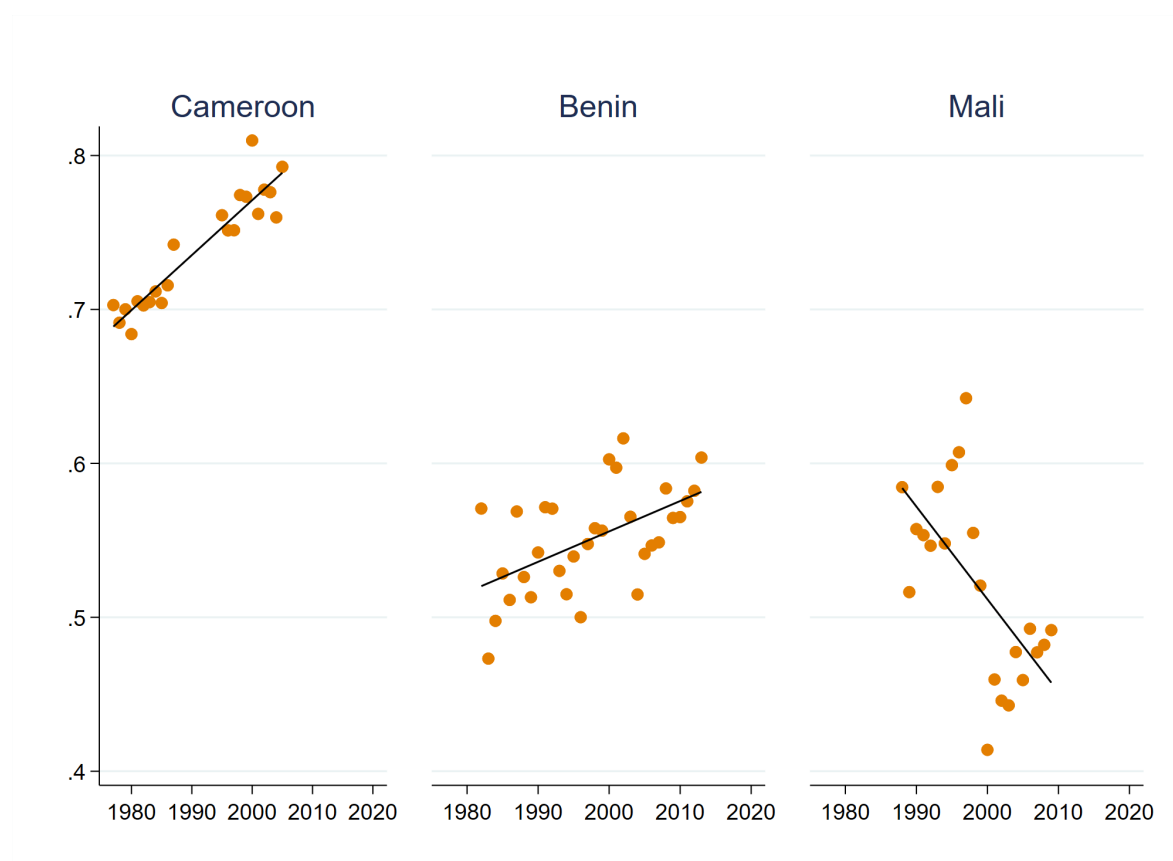


FIGURE 1.I.3 The proportion of migration to urban areas

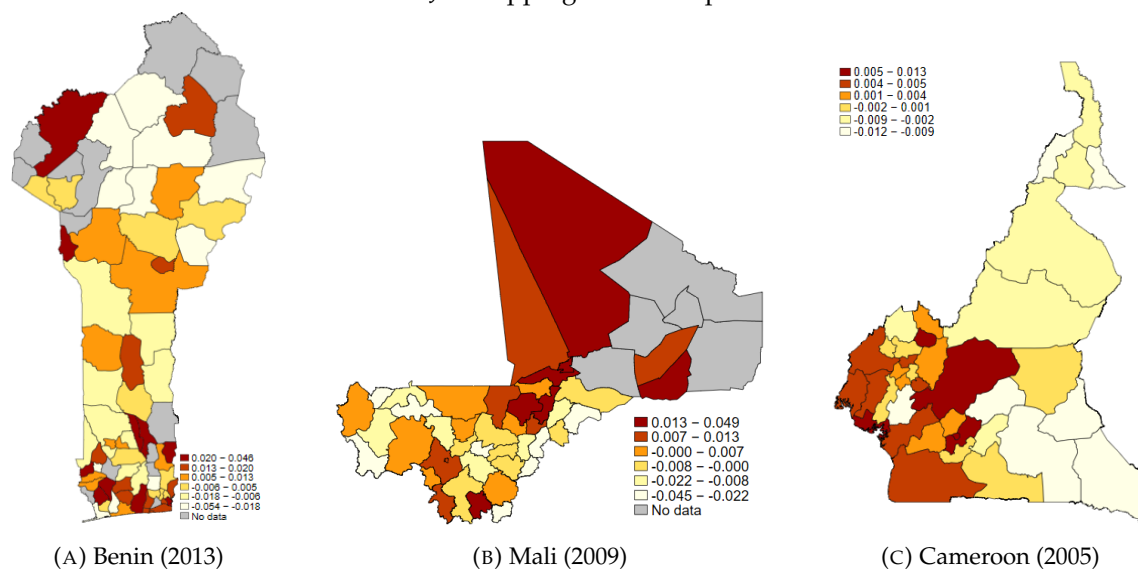
Notes: This figure shows the proportion of movers in a given year that move into an urban location where urban locations are defined as areas where over 50% of the residents report living in an urban locality. On each scatter plot a linear regression line has been added in red.

Cameroon, and Mali cannot be easily characterized. There is primate-to-hinterland migration, hinterland-to-primate migration, urban-to-rural migration, rural-to-urban migration, urban-to-urban migration, and rural-to-rural migration. The objective of these figures is not to trace out-migration patterns specifically but rather to stress the overall heterogeneity in migratory decisions.

In Benin, migration is dominated by the two main cities, Porto-Novo (the larger band) and Cotonou. These cities are geographically very close to each other and a clear pattern of moving away from the capital to the more dynamic *up and coming* Cotonou. We also see a more general exodus from the capital and influx into Cotonou from other localities within Benin. Mali is the most primate-heavy country of the three, the capital Bamako (largest arc) has a population of almost 2 million compared to the next biggest city Sikasso (second biggest arc) where only 250,000 or so people live. Figure 1.I.4b shows that unlike Benin the capital of Mali shows far greater inflow than outflow attracting Malians from across the country. Cameroon is a tale of two cities, Douala (top right) and Yaoundé (bottom left) both home to about 2.5 million people. There is far greater inflow than outflow to these two areas, but that's not to say that there isn't significant movement not involving these locations.

1.J Maps

FIGURE 1.J.1 Mapping estimated place effects



Note: These figures map the estimated yearly causal place effects. The left-hand side image shows Benin in 2013, the middle image Mali in 2009, and the right-hand side image Cameroon in 2005. Missing data is the result of dropping localities with too few movers to accurately measure place effects.

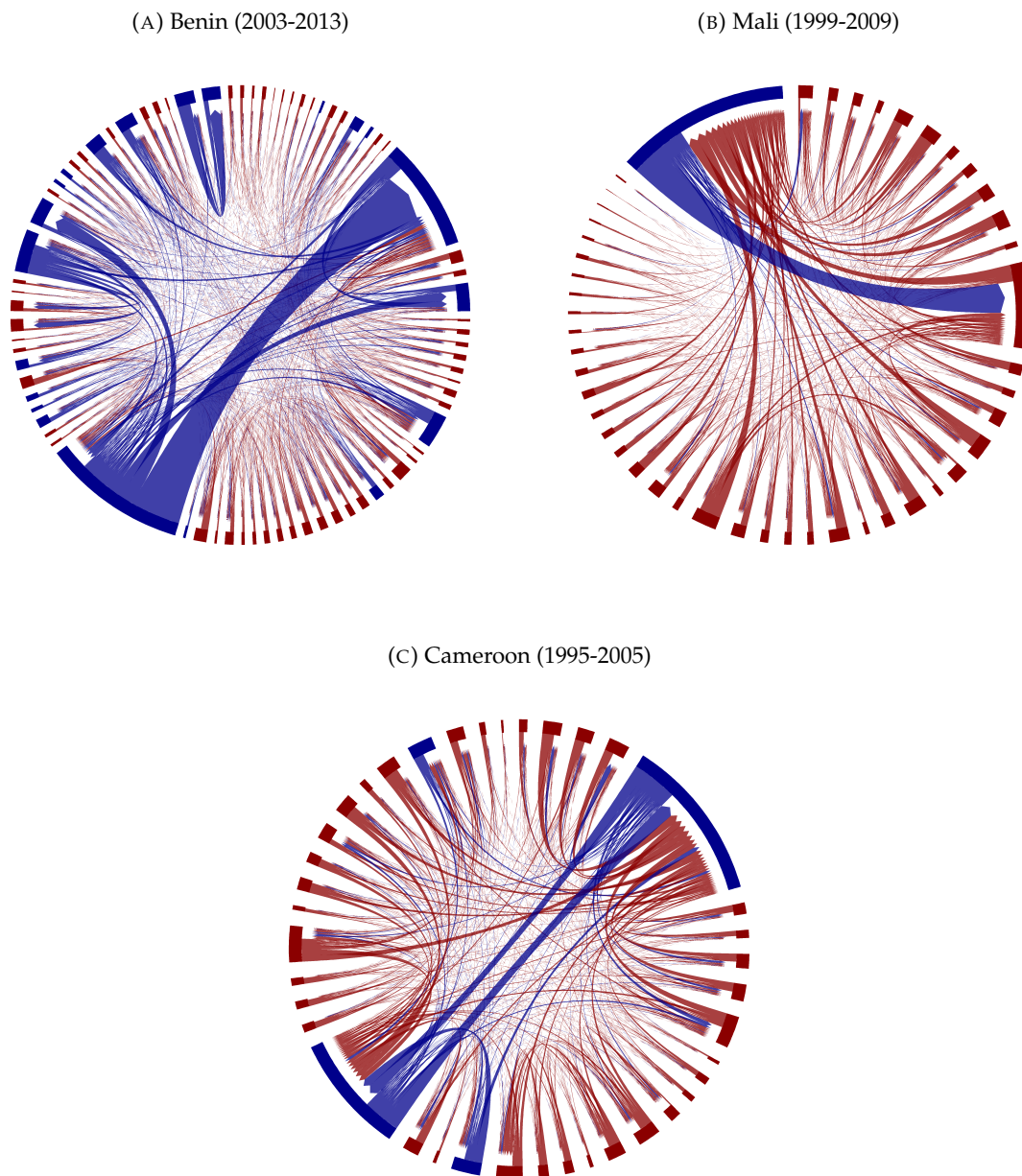
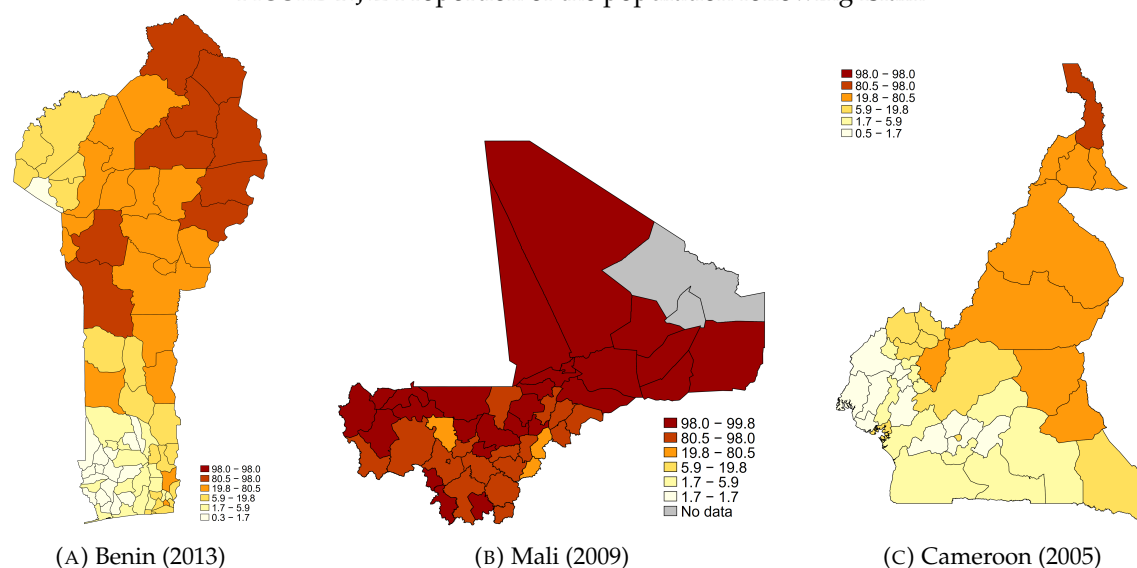


FIGURE 1.I.4 Migration chord diagrams

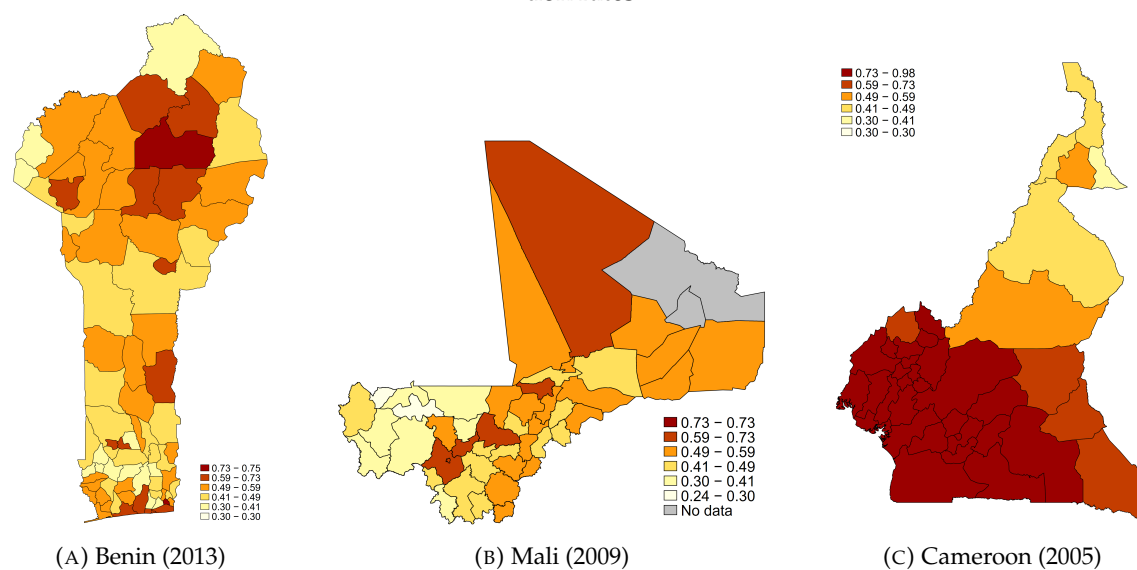
Notes: These figures are Chord diagrams showing population movements from and to each locality in my data within each country aggregating over the last observed 10 years. These figures will underestimate total migration as for individuals that move multiple times in the 10 year period only the most recent move is captured in my data. Each section of the circumference corresponds to a locality with size equal to the total in plus out migration observed. Arcs and flows colored in blue correspond to localities and movements out of said localities that are characterized as urban, that is greater than 50% of the individuals living in these localities report living in an urban area. Similarly red arcs and flows correspond to rural localities and movements.

FIGURE 1.J.2 Proportion of the population following Islam



Note: These figures map the observed proportion who report being Muslim in a locality. The left-hand side image shows Benin in 2013, the middle image Mali in 2009, and the right-hand side image Cameroon in 2005.

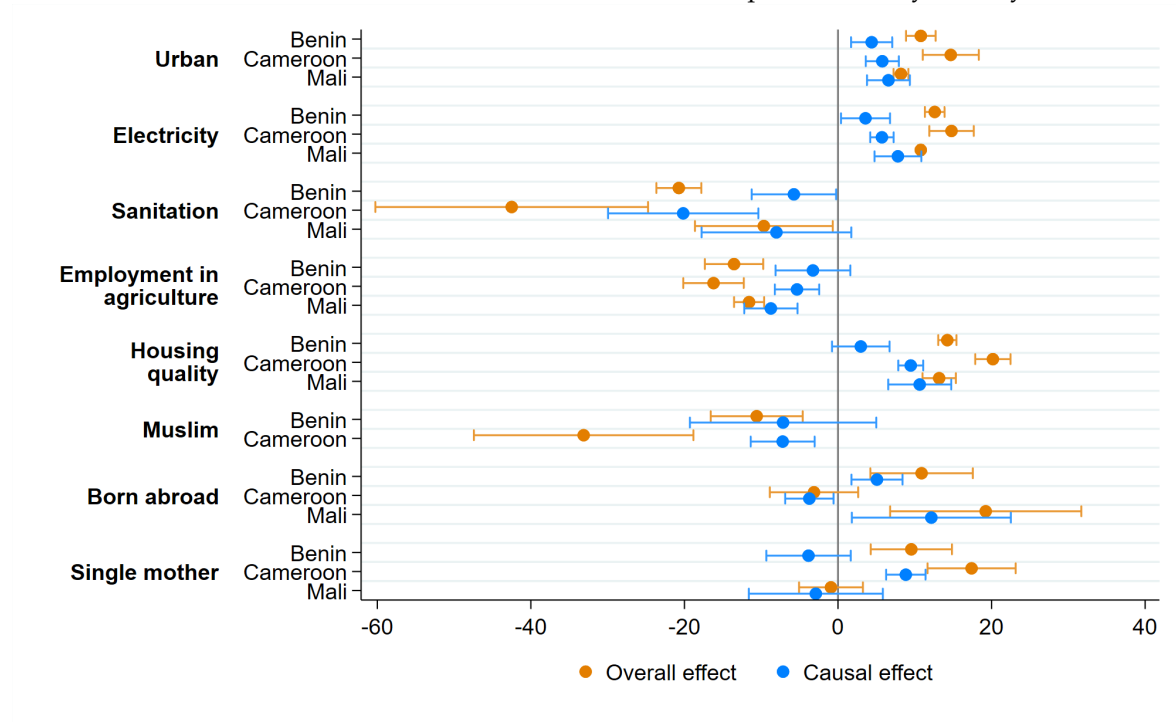
FIGURE 1.J.3 Female primary completion rates as a fraction of male completion rates



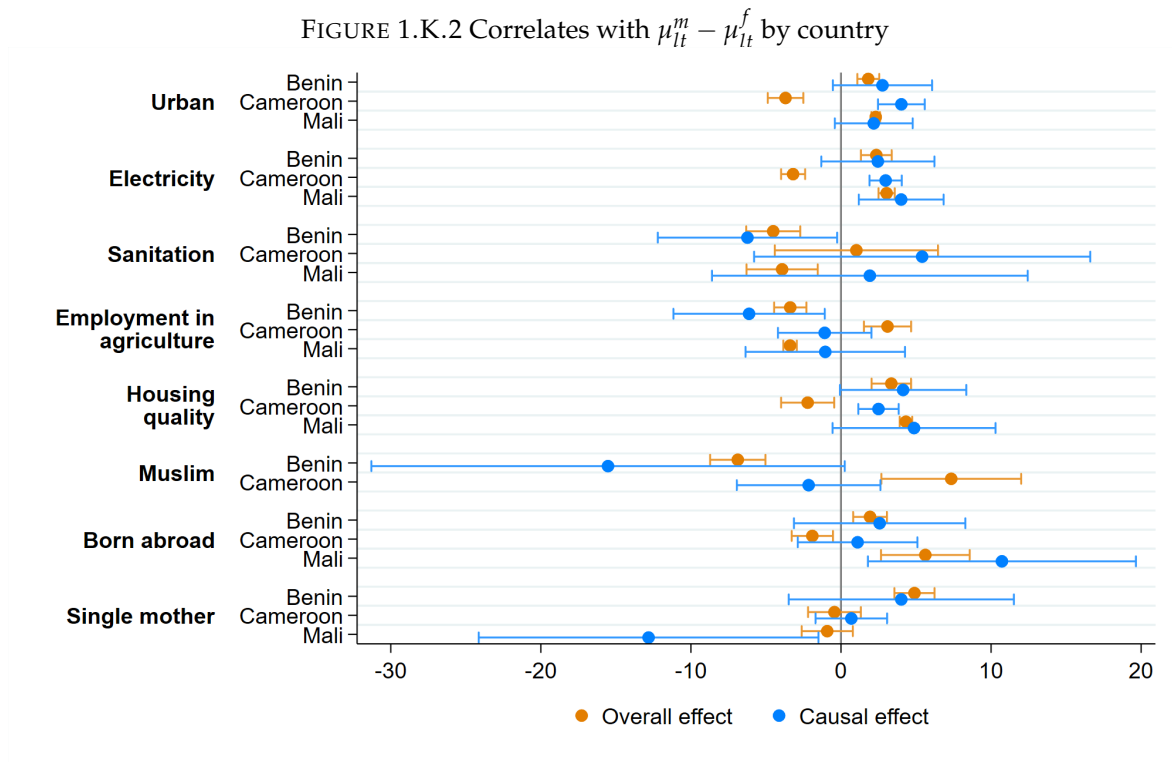
Note: These figures map the observed ratio of reported female to male primary completion in each locality. The left-hand side image shows Benin in 2013, the middle image Mali in 2009, and the right-hand side image Cameroon in 2005.

1.K Heterogeneity by country analysis

FIGURE 1.K.1 Correlates with overall and causal place effects by country



Note: This figure shows the results from running regressions of observed permanent resident primary completion rates and estimated causal place effects against place characteristics including census fixed effects. Each dot (with corresponding confidence intervals) represents a coefficient from an individual regression with the explanatory variable indicated on the y-axis. In orange, the left-hand side variable is observed permanent residents' primary completion, and in blue it is estimated causal place effects multiplied by 13. Standard errors are clustered at the locality level.



Note: This figure shows results from running regressions of the difference between male and female relative causal place effects on various place characteristics, allowing coefficients to vary by country. Each dot (with corresponding confidence intervals) represents a coefficient from an individual regression with the explanatory variable indicated on the y-axis. Census fixed effects are included and standard errors are clustered at the locality level.

1.L Investigation of inequality of opportunity in Cameroon in 2005

Figure 5 in the main text doesn't plot a variance for Cameroon in 2005. This is because there is no significant variation in the causal effect of place on primary completion rates in Cameroon in 2005. This result begs the question of whether Cameroon has suddenly become a much more (spatially) egalitarian society, or if something else is going on. Over time, in all three countries studied, primary completion rates are increasing quite rapidly, leading to less and less variation in primary completion. In Cameroon in 2005, among the individuals who provide identifying variation over 85% completed primary education (as compared to 34% in the entire sample). This means that there simply isn't sufficient variation in outcomes to estimate the variance of individual place effects with any precision. To investigate to what extent place does still matter in Cameroon in 2005, I perform a similar analysis to that in section 3 in the main text but on a sample of 18 to 22-year-olds and looking at secondary completion as the outcome of interest. Table 1.L.1 shows the results where I compare the overall causal effect of spending an additional year of childhood in Cameroon in 2005 on

primary and secondary school completion. The slope on primary completion, although significantly different from zero, is about half found using the full sample. I find that place also matters for secondary school completion. This indicates that the low variance of place effects for primary school does not suggest that in 2005 where you grow up doesn't matter in Cameroon, but that it matters less for primary school completion as the overall rate of completion approaches one.

TABLE 1.L.1 Causal effect of place on primary and secondary completion in Cameroon in 2005

	(1)	(2)
	Primary	Secondary
1 to 12 slope, primary	-0.0145** (0.00697)	
13 to 18 slope, primary	0.0134 (0.0100)	
1 to 12 slope, secondary		-0.0173** (0.00602)
13 to 18 slope, secondary		-0.00108 (0.0206)
Observations	26035	25888

Note: This table shows the results of estimating the linearly parameterized place effects model in Cameroon in 2005 where the outcome of interest is either primary school completion (sample of 14 to 18-year-olds) or secondary school completion (sample of 18 to 22-year-olds).

The above analysis also asks the bigger question: to what extent does the variation discussed in the main text capture true variation in opportunity across space and time, and to what extent does it merely reflect our ability to measure opportunity through variation in primary completion as observed rates increase? Table 1.L.2 shows summary statistics of the identifying variation across each census used where the sample has been restricted to that giving identifying variation. This table shows that the standard deviation of primary completion in the identifying sample is large and stable for all samples except Cameroon in 2005 — a lack of identifying variation may not be driving differences found in figure 5 in the main text. Table 1.L.2 also shows that primary completion is an important margin in all samples with Cameroon in 2005 being by far the highest, suggesting that it remains associated with opportunity in all other samples.

TABLE 1.L.2 Estimating sample primary education variation

Sample	Mean	Sd	N
	Primary completion	Primary completion	
Cameroon 1976	0.61	0.49	32,550
Cameroon 1987	0.72	0.45	61,270
Cameroon 2005	0.86	0.35	157,280
Benin 1992	0.35	0.48	23,950
Benin 2002	0.48	0.50	36,970
Benin 2013	0.72	0.45	41,790
Mali 1998	0.27	0.44	16,630
Mali 2009	0.42	0.49	47,450
Total	0.67	0.47	417,890

Note: This table shows summary statistics of the variation in primary school completion for each census. The sample has been restricted to one-time movers between the ages of 14 and 18 in each case.

Chapter 2

Moving Opportunity. Local Connectivity and Spatial Inequality.

Abstract

Spatial inequality within countries is a pervasive and persistent phenomenon. How does the connectivity of place determine underlying spatial inequality of opportunity? To answer this question, I derive a sufficient statistic result linking local opportunity to market access terms, developing a framework consistent with a broad class of spatial general equilibrium models. I empirically validate this result using a novel not-on-least-cost-path identification strategy and data from historical road maps in Benin, Cameroon, and Mali covering 1970 to 2020, that I digitize. Using these estimates to parameterize a structural case of the spatial model, I show that road building alters the spatial distribution of opportunity. By considering each possible road upgrade I show that although some roads decrease the standard deviation of opportunity by more than 2%, others increase inequality by a similar amount. By fixing the spatial distribution of opportunity I show that to achieve similar reductions in inequality by moving people, a prohibitively large number would need to migrate from low- to high-opportunity areas — between 13% and 44% of the low-opportunity areas' population. Policymakers also face an equity-efficiency trade-off: 4 out of the top 10 aggregate opportunity-increasing roads also *increase* spatial inequality of opportunity.

2.1 Introduction

Inequality of opportunity is undesirable, but across space it is ubiquitous. The literature has previously suggested the policy response of moving people to areas of higher opportunity [Chetty et al., 2016, Bryan et al., 2014], but this assumes that the underlying causal effect of place is immutable. In this paper, I relax this assumption and consider the scaleable policy alternative of moving opportunity to people. Such inequality across space is often rooted in locations being cut off from markets and the economic opportunities they foster [Allen et al., 2020c]. Thus, I study a large policy in a development setting, that affects a location's connectivity and position within the urban network: Road building.¹ Connecting any two locations increases the degree of local market integration and ease of migration, but could also divert trade and opportunity away from other areas. Additionally, as observed spatial differences in outcomes may reflect sorting rather than causal effects of place, it is unclear ex-ante whether and how such investments influence spatial inequality of opportunity.

This paper studies the effects of transportation infrastructure investment on spatial inequality of opportunity in Benin, Cameroon, and Mali. These three countries have substantially upgraded their road networks since 1970 and are likely to make significant additional investments, due to rapid urbanization and population growth [UN, 2018, Gwilliam, 2011, Foster and Briceño-Garmendia, 2010]. By studying three countries I am also able to identify the importance of network-level characteristics in determining the effect of any given road on the spatial inequality of opportunity — a crucial dimension when assessing the external validity of findings from any given setting. I use data on the causal effect growing up in a given location has on individuals from Heath Milsom [2021], and the changing connectivity of place from historical Michelin road maps that I digitize covering the period 1970-2020.

There are three main empirical challenges to credibly estimating the effects of road investments on spatial inequality of opportunity. First, building a road in any given location will have spillover and general equilibrium effects on other locations. Connecting an area may negatively affect other locations if trade is diverted away from them — or positively if trade is diverted towards a location they were previously well connected to. Additionally, migration patterns may shift, causing an excess or shortage of labor, affecting wages. Such changes to trade or migration patterns will alter prices across space — causing further adjustments. Second, roads may be built to galvanize flagging areas or service expanding ones, that is, they may be built for endogenous reasons. Third, the effect of any given road

¹In 2021 alone over 10% of the World Bank's lending was to finance transportation infrastructure investment.

will depend on the entire preexisting road network and distribution of economic activity. This means that network-level characteristics will be important determinants of effects — and a threat to external validity if one were to consider a single network.

To overcome the first challenge, of spillover and general equilibrium effects inherent in road building, I prove a sufficient statistic result, applicable to a large class of models with costly migration and trade and two sectors/types of workers: educated and non-educated. This provides a way of measuring the impact of roads on local opportunity allowing changes to the network to affect all other locations in a manner that is consistent with a broad class of plausible data-generating processes and micro-foundations. This result states that the effects of transportation infrastructure investments on educational opportunity are summarized by four measures of market access, two that capture demand for each sector's goods and two that capture the supply of each type of worker. The model builds upon [Allen, Arkolakis, and Li \[2020a\]](#) extending their framework by introducing costly migration, educational investments, and two sectors/types. Therefore, the model endogenizes the local incentives and costs of education, and thus local opportunities.

To estimate the sufficient statistic result, and counter the second challenge (endogeneity of road building), I develop a novel “not-on-least-cost-path” identification approach. For each focal location, I freeze the least cost path to all other locations and only use indirect variation in that location's market access that stems from how changes to the road network affect other locations' market access. Intuitively, if a planner builds roads to better connect two locations, I don't use this endogenous variation in connectivity and instead leverage the indirect effects through changes to other locations' market access. This procedure can be iterated to consider second-order indirect effects, or third-order, etc. — providing increasingly more restrictive instruments and enabling an implicit test that most endogenous variation has been removed, by checking if coefficients become stable. I take this approach as existing alternative identification strategies can not be employed in this setting. [Donaldson \[2018\]](#), [Faber \[2014\]](#), [Moneke \[2020\]](#), [Banerjee et al. \[2020\]](#) and others use placebo lines which are not available or consider incidentally connected areas which require a set of locations planned to be connected, which is also not available to me.²

²A placebo line strategy uses information of planned but not-built lines as a set of control connections under the identifying assumption that the locations they connect are otherwise similar to those actual roads connected ex-ante. Alternatively, the incidentally-connected approach leverages knowledge of the road-builders objective function. Usually, it credibly argues that planners aim to connect hubs and therefore locations that happen to be between such hubs become incidentally connected whereas relatively close alternative locations which do not lie on the route do not. Due to the heterogeneous treatment effects of roads stressed in this paper, these strategies will also likely suffer from bias due to essential heterogeneity and spillover effects (SUTVA violations). My strategy can also be thought of as improving upon the previously employed far-away variation strategy [[Jedwab and Storeygard, 2021](#), [Donaldson and Hornbeck, 2016](#)]. The intuition behind such strategies is that road building

Every road will affect every location differently, and this impact will depend on the entire preexisting road network and distribution of economic activity. This network dependence makes it difficult to quantify the impact of a specific road project on inequality of opportunity — the third main empirical challenge. To counter this challenge, I develop a structural spatial equilibrium model taken from the class of models that are consistent with the sufficient statistic result. By specifying a structural model, I can quantify the counterfactual effects of building/ upgrading any possible road, and understand how network-level characteristics impact the effects. I solve the model in changes using exact hat algebra [Dekle et al., 2008] and estimate parameters using the sufficient statistic coefficients, which can now be interpreted as exact bundles of structural parameters.

To take the sufficient statistic relationship and not-on-least-cost-path identification strategy to the data, I use variation in roads since 1970 from historical Michelin road maps that I digitize. These maps are available roughly every decade and consistently distinguish between three road types: dirt tracks, partially improved roads, and paved roads — providing both intensive and extensive margin variation in road building. To capture variation in spatial opportunity rather than the characteristics of individuals over space, I use a measurable dimension of place effects: local educational opportunity. Following Heath Milsom [2021] I define local opportunity as the causal effect of an individual growing up in a location³ on their probability of completing primary education.⁴ Although there are many dimensions of local opportunity, primary schooling is particularly salient⁵ and correlated with other indicators of later life success such as earnings, housing quality, and not working in agriculture. Many factors could influence these causal effects of place, but improvements to the road network will impact local market integration and the ease of migrating, altering the returns to, and costs of, education.

I find that changes in the connectivity of place in Benin, Cameroon, and Mali altered the spatial distribution of opportunity. Expansion in access to demand for goods produced

that occurs far from a given location is less likely to be completed for reasons endogenous to said location. This suffers from two drawbacks, first, it's unclear how far away is sufficiently far away, and second long connections may still be built for endogenous reasons. My approach counters both of these potential drawbacks.

³Regions or localities in this paper refer to second administrative divisions: Communes in Benin, Departments in Cameroon, and Circles in Mali. These have a median population of 267,000 and can be thought of as similar in scale to commuter zones in the US.

⁴This follows a burgeoning literature shows that the locality you grow up in shapes later life outcomes see for example Chetty and Hendren [2018b], Laliberté [2021], Chyn [2018], Deutscher [2020], Alesina, Hohmann, Michalopoulos, and Papaioannou [2021], Heath Milsom [2021]. See appendix 2.A for details regarding the estimation strategy used in Heath Milsom [2021].

⁵Primary-school completion is the most important, though not the only, dimension of educational success in this context: 30 percent of adults 25 to 55 have completed primary school in my sample, while just 7 percent have completed further schooling.

by educated workers increases opportunity by increasing the returns to education. Conversely, increases in access to the supply of educated workers decreases educated wages and the returns to education and so opportunity. This result is robust to controlling for clientelism, potential non-linearities, the presence of Koranic schools, endogenous schooling supply/quality, and expected changes in market access [Borusyak et al., 2018]. These results validate the theoretical approach and show a strong link between changes in connectivity and the spatial distribution opportunity, whilst remaining robust to a broad range of modeling approaches. However, to go further and answer policy-relevant questions, I turn to counterfactual estimation using the structural spatial general equilibrium model.

First, I consider the counterfactual impact of road building since 1970 on spatial inequality of opportunity in each of Benin, Cameroon, and Mali. Comparing the most and least affected locations within each country, I find a 7.5 p.p. gap in the impact of road building since 1970 on the causal effect of growing up in a given location on primary school completion. A third of the within-country variation in effects is explained by larger gains being associated with greater remoteness in 1970. By studying three countries, I am also able to investigate network-level heterogeneity. While road building since 1970 had little impact on the inequality of opportunity in Benin, it significantly increased inequality in Cameroon and decreased inequality in Mali. There are two possible reasons for cross-country differences. First, they could be explained by varying constraints faced by policymakers in 1970 such as geography, or the pre-existing road network. Second, conditional on such initial conditions they could be due to differences in policymakers' road-placement decisions. To understand the role played by factors outside policymakers' control, I solve the model for 250 randomly generated road networks of the same overall length in each country. The resulting impact on inequality of opportunity from the random networks is tightly clustered around the impacts realized by the actual networks. This finding indicates that varying constraints faced by policymakers played a considerable role in explaining the observed differences in effects.⁶

A first-order policy question is how future changes in connectivity may affect spatial inequality of local educational opportunity. I consider each of the 570 possible individual road upgrades as separate counterfactual exercises. For each road, I calculate the resulting change to the spatial distribution of opportunity and inequality of opportunity due to

⁶An alternative explanation is that policymakers targeted an objective that was only weakly correlated with reducing inequality or had poor knowledge of the mapping between road placement and impacts on inequality. These possibilities cannot be ruled out. However, to the extent to which they are common across policymakers do not diminish the importance of initial conditions in determining outcomes.

upgrading⁷ it whilst keeping the rest of the network constant at 2019 levels. I find significant heterogeneity both across roads within a given network and across networks within a given road (that is through counterfactually changing the network, keeping the road fixed). Some roads decrease the standard deviation of opportunity by 2%, whereas others increase it by 2%. Roads that connect two periphery areas on average increase inequality of opportunity by more relative to those that connect periphery areas to a main city. The intuition is that periphery locations have more to gain from the main city by being better connected to it than vice-versa.⁸ By counterfactually changing the distance between locations within a given network, I can study cross-network within-road heterogeneity, and find that the effect of any given road is greater in networks where localities are closer together. Intuitively, by increasing the distance between locations, it is harder for other areas not directly connected to make use of any given improvement in connectivity — attenuating the impact of given road on opportunity. The uncovered importance of network-level characteristics highlights potential pitfalls in extrapolating the effects of road building from one setting to another.

Finally, I use the 570 estimated road-level results and consider the possible equity-efficiency trade-off of building roads. I find that roads that increase aggregate opportunity the most do not necessarily decrease inequality of opportunity the most. On average only 6 of the top 10 aggregate opportunity-increasing roads also decrease inequality of opportunity. Policymakers are likely to give weight to inequality of opportunity due to normative concerns around fairness. Indeed individuals have recently been shown to value equity across space in and of itself [Gaubert et al., 2021]. My findings underpin that if policymakers consider equity of opportunity to be an important objective, this will change where roads should be built in the future.

This paper focuses on moving opportunity to people in order to decrease spatial inequality of opportunity and improve outcomes. However, previously the literature has focused on the other side of the coin — moving people to areas of high opportunity (see Bryan et al. [2014] and Chetty et al. [2016] for two prominent examples). To benchmark my results against the alternative policy of moving people to areas of high opportunity, I perform a back-of-the-envelope calculation of how many people would have to be moved to achieve

⁷In my data I can distinguish between road types: dirt track, partially improved road, and paved road. I upgrade each road to a “highway” with an average travel speed of 80km/h, considerably faster than the fastest previous category, paved roads which have an average speed of 60km/h.

⁸I show formally and empirically that the first-order effects of connecting any two locations can be summarized by their initial trade and migration flows. If a given location i exports more goods produced by educated than those produced by non-educated workers to a given location j , and/or if a larger proportion of i 's in-migration from j comes from non-educated than educated workers, then connecting i and j is likely to increase opportunity in i .

reductions in inequality similar to those found by road upgrading projects. In a policy counterfactual analogous to the Moving To Opportunity experiment [Chetty et al., 2016], I move people from the lowest to the highest opportunity areas. Fixing the spatial distribution of opportunity, I find that to reduce the population-weighted standard deviation of opportunity by 1%, a reduction achieved by many roads, between 71,000 and 466,000 individuals would have to move. Under any reasonable assumptions, the associated movement costs would be far in excess of the cost of upgrading a road, which [Buys et al., 2006] estimates to be roughly 12.8m USD for a 100km stretch.

This paper contributes to the literature in three main ways. First, it contributes to the literature studying spatial inequality and place effects. Previously the literature has focused on exploiting the possibilities represented by spatial variation within country borders by moving people to areas of opportunity (see Bryan et al. [2014] and Chetty et al. [2016] for example). In this paper, I instead consider how policymakers could move opportunity to people via road building. I add to the place effects literature [Chetty and Hendren, 2018a,b, Deutscher, 2020, Laliberté, 2021, van Maarseveen, 2021, Alesina et al., 2021, Rojas Ampuero, 2022], which previously has focused on estimating the causal effect of place, by going a step further and asking how policy can alter the distribution of causal place effects. A recent strand of this literature has considered place effects within a general equilibrium setting [Chyn and Daruich, 2022, Eckert et al., 2021]. I build on this literature by: considering changes in connectivity a low and middle income country setting, analysing cross-country heterogeneity, and developing a framework that allows greater location-heterogeneity (relative to Chyn and Daruich [2022]) and takes a market access approach with costly trade as well as migration.

Second, I contribute to the literature on quantitative spatial economic modeling, by deriving a novel sufficient statistic relationship from a broad class of models exhibiting: education demand and supply, multiple sectors, and costly trade and migration. This builds on recent theoretical work on quantitative spatial general equilibrium models and the generality of the gravity-based approach [Redding and Rossi-Hansberg, 2017, Allen et al., 2020b, Donaldson and Hornbeck, 2016, Donaldson, 2018, Allen and Arkolakis, 2022]. Empirically I consider the effect of roads following (among others) Kebede [2020], Sotelo [2020], Adamopoulos [2019], Castaing Gachassin [2013], and Morten and Oliveira [2021]. A smaller literature considers the interaction between observed educational attainment and trade [Fujimoto et al., 2019, Khanna, 2022, Hsiao, 2022]. Edmonds et al. [2010] studies the impact of the Indian tariff reform of the 1990s and find that the most impacted areas saw the smallest

increases in schooling. [Atkin \[2016\]](#) looks at the impact of growth in export manufacturing in Mexico and similarly finds that more affected areas saw greater declines in schooling. Most related to this work, [Adukia et al. \[2020\]](#) and [Asher and Novosad \[2020\]](#), consider the impact of connecting villages in India to the main road network on educational and economic outcomes. They find evidence of higher attainment in connected villages with enrollment increasing by more in locations where the returns to education are the highest. I build on this work by considering the impact of more large-scale inter-city road-building in a different empirical setting and explicitly considering the impacts on the inequality of opportunity rather than observed primary completion and allowing for crucial spatial general equilibrium effects.

Lastly, I contribute to the literature on identifying the impact of changes in connectivity by developing a novel identification strategy using an iterative not-on-least-cost-path approach. An established literature looks at the causal impacts of colonial railways in Sub-Saharan Africa such as [Jedwab and Moradi \[2016\]](#) and [Jedwab, Kerby, and Moradi \[2017\]](#). This has more recently been supplemented by work looking at roads, ([Moneke \[2020\]](#), [Jedwab and Storeygard \[2021\]](#), [Banerjee, Duflo, and Qian \[2020\]](#), [Faber \[2014\]](#)), and bridges ([Brooks and Donovan \[2020\]](#), [Zant \[2022\]](#)). I contribute to this literature by expanding on the existing *far-away* variation strategy and developing an alternative to the straight-line instrument, or incidental-middle approach [Redding and Turner \[2015\]](#), [Michaels \[2008\]](#) — one which can be applied in all settings that result in a market access relationship.⁹

This paper proceeds as follows. Section 2.2 overcomes the first main empirical challenge of inherent spillover and general equilibrium effects in road building by turning to theory, but remaining as general as possible, and developing the sufficient statistic result. Section 2.3 then describes my setting and data before developing a novel identification strategy to

⁹I also speak to the related literature that has considered the more reduced-form impacts of transport infrastructure on local outcomes. The conclusion from this literature is somewhat ambiguous. [Faber \[2014\]](#) finds that incidentally connected periphery areas in China have lower GDP relative to similar not-connected areas. On the other hand, [Baum-Snow et al. \[2017\]](#) and [Banerjee et al. \[2020\]](#) find positive impacts of being connected in a similar setting, and [Jedwab and Storeygard \[2021\]](#) similarly in SSA. [Baum-Snow et al. \[2020\]](#) attempts to reconcile this, again considering roads in China, by showing that impacts depend on the ex-ante urban hierarchy: Core cities benefit at the expense of periphery ones. This paper can be seen as taking the implicit heterogeneity seriously by allowing the impact of every road on every location to differ, and for this to depend on the entire pre-existing road network and distribution of economic activity. This highlights the importance of not generalizing from one road to another within a country, but also crucially also not pre-supposing impacts found in one network will port to another setting. As well as allowing for such heterogeneity, I also build on this literature by considering a very different setting, focusing on the causal effect of place (which may not be related to observed outcomes such as GDP due to the prevalence of sorting), studying aggregate impacts using a spatial model, and stressing the impact of roads on country-wide outcomes such as inequality of opportunity over space. In appendix 2.B.15, I additionally show that by focusing on variation of the form considered by [Faber \[2014\]](#), I can replicate his main results in my setting.

overcome the second main challenge — that of the endogeneity of road placement, and estimating the sufficient statistic result. Finally, to understand the importance of network-level characteristics (the third main challenge), and answer questions of policy relevance, section 2.4 estimates a structural spatial economics model and performs counterfactual analysis. Section 2.5 then concludes.

2.2 A general spatial theory linking changes in connectivity to local educational opportunity

In this section, I overcome the first empirical challenge, that of spillover and general equilibrium effects, by turning to theory but remaining as general as possible. The theory describes how to measure the impact of road building on each location, accounting for spillovers and general spatial equilibrium effects in a manner that is consistent with a broad class of spatial models. Connecting a given location may negatively affect others if trade is diverted away from them — or positively if trade is diverted towards a location they were previously well connected to. Additionally, migration patterns may shift, causing a glut or drought of labor in a given location and affecting wage rates. Such changes to trade or migration patterns will then feed into varying prices across space — which may themselves cause further adjustments.

I account for these forces within a general framework without having to tie my hands to a specific micro-foundation or given set of modeling features. In this manner, I overcome the challenge represented by spillover and general equilibrium effects by putting minimal structure on the data.

2.2.1 Developing the sufficient statistic result

In this section, I develop a sufficient statistic result in the sense described by Donaldson [2022]. This result states that, within a broad class of data generating processes, market access terms capture all of the effects of changes in connectivity on the spatial distribution of local opportunity. The framework I develop nests various micro-foundations, including: perfect competition Armington-based differentiated products (e.g. Allen and Arkolakis [2014]); Eaton-Kortum models of economic geography with economies of scale in production (e.g. Bartelme [2015]); Melitz-type frameworks with heterogeneous firms (e.g. Di Giovanni and Levchenko [2013]); monopolistic competition and economies of scale with differentiated products (e.g. Krugman [1980]); approaches based on a multi-region Helpman

framework (e.g. Redding and Sturm [2008]); recent quantitative spatial economics models¹⁰ (e.g. Santamaria [2020], Tsivanidis [2019]). The model represents the most parsimonious set-up that captures key features such as multiple sectors, costly movement of goods and people, and education choice. In addition, the sufficient statistic result is robust to various extensions such as including land in production, land in consumption, endogenising land/housing provision, including explicit agglomeration forces or endogenous amenities, generalizing preferences, generalising the factor content of production, including intermediate goods, and allowing other factors to influence education choice. Appendix 2.D formally shows that each of these extensions results in the same sufficient statistic.

This theory builds on work that notes the general nature of gravity spatial models (Allen, Arkolakis, and Takahashi [2020b], Allen, Arkolakis, and Li [2020a], Allen and Donaldson [2020], Bartelme [2015], Yotov, Piermartini, Monteiro, and Larch [2016]), by including costly movement over space [Morten and Oliveira, 2021], two sectors, and education completion.

Consider an economy populated by households (which consist of one child and one adult) $\omega \in \mathcal{I}$ who reside in discrete locations $i \in \mathcal{L}$ over periods $t \in \mathcal{T}$. For now, I present the simplest version of the model with one sector, no education, and myopic agents. Each region produces a representative good (which will be some bundle of underlying products). Denote by $Q_i \geq 0$ the output produced by region i and similarly $p_i \geq 0$ as the factory gate price of said output. Then $p_i Q_i = Y_i$ is the income in i . Now consider another region j , denote by $Q_{ij} \geq 0$ the quantity of i 's representative good that is consumed in j , and similarly $p_{ij} \geq 0$ is the price of i 's representative good in j . Then $X_{ij} = p_{ij} Q_{ij}$ is the value of trade flows from i to j . Denote the total value of imports i.e. the expenditure in i by $E_i = \sum_j X_{ji}$. Finally denote the overall price level in a locality as P_i .

Assume iceberg trade costs $p_{ij} = \tau_{ij} p_i$ for some $\tau_{ij} \geq 1$, and that aggregate demand takes a constant elasticity of substitution form: $E_i = (\sum_{j \in \mathcal{L}} p_{ij}^{-\phi})^{-1/\phi}$. These together with Shephard's lemma give the aggregate demand equation for i 's goods in j of the familiar gravity form.

$$X_{ij} = \frac{(\tau_{ij} p_i)^{-\phi}}{\sum_k p_{kj}^{-\phi}} E_j \quad (2.1)$$

Define market access as the inverse of the familiar price index $MA_i = P_i^{-\phi} = \sum_{j \in \mathcal{L}} p_{ij}^{-\phi}$ then we can write $X_{ij} = (\tau_{ij} p_i)^{-\phi} MA_j^{-1} E_j$. Now strengthen the assumption of iceberg trade

¹⁰In appendix section 2.D.3 I show that a quantitative spatial economics framework following Morten and Oliveira [2021], Tsivanidis [2019], Ahlfeldt et al. [2015], among others, can be considered a special case of the more general set up described here.

costs to that of symmetric iceberg trade costs¹¹, that is $\tau_{ij} = \tau_{ji}$. As [Allen, Arkolakis, and Takahashi \[2020b\]](#) shows symmetry implies that $p_i^{-\phi} = MA_i^{-1}E_i$. Assume that labor is the only factor of production then $p_i Q_i = Y_i = w_i L_i$ where w_i is the local wage rate and L_i is the local employed population. Goods market clearing implies $w_i L_i = Y_i = E_i$, and thus we have that $w_i = p_i^{-\phi} MA_i L_i^{-1}$.

Now moving onto labor markets, first assume they clear: that is, the total work force equals the sum of those who move in (including those who stay) $L_i = \sum_j M_{ij}$ where M_{ij} denotes the quantity of movers from j to i . Define π_{ij} as the proportion of individuals in j who move to i then $M_{ij} = \pi_{ij} L_j$. Assume that this proportion takes a gravity form, analogous to that of trade¹² $\pi_{ij} = \frac{u_{ij}^\lambda}{\sum_{k \in \mathcal{L}} u_{kj}^\lambda}$ where u_{ij} is the utility derived from moving from j to i . Suppose that migration costs also take a symmetric iceberg form $u_{ij} = \frac{1}{\kappa_{ij}} u_i$. Iceberg migration costs κ_{ij} capture both the pecuniary cost of moving, but also the cost of rebuilding social capital or of moving large cultural or social distances (which are correlated with physical distance) [[Glaeser et al., 2002](#), [Falck et al., 2012](#), [Bailey et al., 2018](#)] which are commonly estimated to, in sum, exceed the equivalent of annual income [[Kořar et al., 2021](#)]. Together this implies that movement across space, obeys the following equation (in aggregate):

$$M_{ij} = \frac{\kappa_{ij}^{-\lambda} u_i^\lambda}{\sum_k \kappa_{kj}^{-\lambda} u_k^\lambda} L_j \quad (2.2)$$

Using this I can show that $L_i = \sum_j \pi_{ij} L_j = \sum_j \kappa_{ij}^{-\lambda} u_i^\lambda L_j EMA_i^{-1}$ where employer market access (EMA) is defined as $EMA_i = \sum_j \kappa_{ij}^{-\lambda} WMA_j^{-1} L_j$ where $WMA_i = \sum_j \kappa_{ij}^{-\lambda} u_j^\lambda$, is worker market access. Employer market access is increasing in the number of potential workers a firm could draw on, and worker market access is increasing in the number of potential jobs a worker could work at. Thus we have $L_i = u_i^\lambda EMA_i$. Substitute this back into WMA_i to find that $WMA_i = \sum_j \kappa_{ij}^{-\lambda} L_j EMA_j^{-1}$. The only solution to this system of equations under symmetric migration costs is $EMA_i = \rho_L WMA_i = LMA_i = \sum_j \kappa_{ij}^{-\lambda} L_j LMA_j^{-1}$ for some constant ρ_L [[Donaldson and Hornbeck, 2016](#)]. Labor market access is higher in a location i if i is well connected (low κ_{ij}) to many locations which have large labor markets (large L_j) and few alternative sources of employment (low LMA_j). Therefore we can write $L_i = u_i^\lambda LMA_i$. This equation is very intuitive: the total population of a location i is larger if the location can

¹¹In reality this doesn't matter. We can instead suppose quasi-symmetric trade costs where $\tau_{ij} = \tau_i^A \tilde{\tau}_{ij} \tau_j^B$ where only $\tilde{\tau}_{ij}$ is symmetric, and everything follows through with slightly more algebra. Alternatively we can *not* assume symmetric costs at all and we find ourselves with inward and outward market access terms as sufficient statistics. In practice these would be so highly correlated as to make little empirical difference.

¹²This can be micro-founded from example by assuming type two extreme value preferences over locations as is common.

draw on/ attract a larger labor market (LMA_i is higher) or is a particularly good place to live (u_i is high).

By an analogous argument on the trade side we can write $MA_i = \sum_j \tau_{ij}^{-\phi} Y_j MA_j^{-1}$, the goods market access in a location i is higher if i is close to (has low τ_{ij}) many locations which have large markets (large Y_j) but aren't themselves well connected to alternative markets (low MA_j). Combining the goods (labor demand) and migration (labor supply) sides we have wages increase in goods market access, but decrease in labor market access:

$$w_i = \frac{p_i^{-\phi} MA_i}{u_i^\lambda LMA_i} \quad (2.3)$$

To proceed, we have to take some stance on the relationship between u_i , p_i and w_i . The following two assumptions are robust to the class of models discussed above¹³: $u_i = A_i \left(\frac{w_i}{p_i}\right)^a$ and $w_i = B_i p_i^b$. Where A_i, B_i are some exogenous shifters and a, b are exogenous constants. Effectively this gives three equations in three unknowns w, u, p and so can solve for w to find:

$$w_i = \Omega_i \cdot MA_i^{\frac{1-\frac{a\lambda}{\phi}}{x}} \cdot LMA_i^{-\frac{1}{x}} \quad (2.4)$$

where $\Omega_i = A_i^{-\lambda/x} B_i^{\phi/bx}$ is a collection of exogenous terms and $x = 1 + a\lambda + \phi/b$. Equation 2.4 recovers the basic sufficient statistic relationship between wages in a location and good/ labor market access terms.

To make this basic framework more amenable to my analysis I introduce multiple sectors/ types and education. Denote sector and type by $s \in \{E, N\}$ where sectors differ in all exogenous components and constants described above, E indicates educated individuals (those who have completed primary school), N indicates not educated individuals. The s sector produces s -type goods and only uses s -type workers in production¹⁴. Each assumption described above is now assumed to hold at the sector level, it's well known that gravity models are separable in sectors which allows us to write sector-specific market access terms as follows¹⁵.

$$MA_i^s = \sum_j (\tau_{ij})^{-\phi^s} \frac{Y_j^s}{MA_j^s} \quad LMA_i^s = \sum_j (\kappa_{ij}^s)^{\lambda^s} \frac{L_j^s}{LMA_j^s} \quad (2.5)$$

I model education as part of a locations amenity value, decompose $A_i = \bar{A}_i E_i$ where $\bar{A}_i$ is

¹³Note that I can define u_i and w_i to be any multiplicative function of endogenous or exogenous model variables and the sufficient statistic result which follows will hold, see appendix 2.D for details.

¹⁴The sufficient statistic result is robust to this rigid production set up, shown formally in appendix 2.D.

¹⁵The implied assumption on consumption patterns is generalized in the appendix section 2.D, and results are found to not be qualitatively different

an exogenous amenity shifter orthogonal to education considerations. E_i captures the utility value of sending your child to complete primary school. Following [Edmonds et al. \[2010\]](#), [Adukia et al. \[2020\]](#) and others I suppose that the value of education is increasing in the returns to education and write $E_i = r_i^\beta = (w_i^E/w_i^N)^\beta$. I can also allow the value of education to be decreasing in the cost of education, the opportunity cost of education, decreasing in how congested schools are, or increasing in parental income — these extensions will not affect the sufficient statistic result and so for simplicity, here I focus on returns to education. The amenity value of education determines with some idiosyncratic primary completion shock ε_i , whether an individual completes primary school in a location or not.

$$\mathbb{P}[\text{complete primary school}_i] = \mu_i = \beta_1 \cdot \ln(r_i) + \varepsilon_i \quad (2.6)$$

This is a simple and general approach to modeling primary completion which can be micro-founded in a number of ways, and relates the model object, $\mathbb{P}[\text{complete primary school}_i]$ to it's empirical observed counterpart, μ_i , local educational opportunity.

In the above exposition, I have assumed that the supply of education is perfectly elastic and that the cost is constant across localities and time. Of course, a feature of the increasing primary completion rate over the study period has been a large expansion in the number of schools. This will decrease the cost associated with traveling to school which is one of the largest impediments to school attendance in Sub-Saharan Africa [[DeStefano et al., 2007](#), [Evans and Mendez Acosta, 2021](#)]. Therefore, one may worry that I have omitted an important dimension of variation both over time and space. Here I postulate a simple way of endogenizing education supply which does not impact the sufficient statistic result. In appendix section [2.D.2](#) I also develop a more detailed model of education supply, introducing a central planner, following [Khanna \[2022\]](#) and show that the data suggests this is also captured by the sufficient statistic result.

Perhaps the simplest way of allowing the cost of education to vary endogenously is to suppose that it is inversely related to the distance one has to travel to school (on average) in a locality: $c_{it} = cd_{it}^\chi$, where d_{it} is the average distance children have to travel to school in locality i and $c > 0$ and $\chi > 0$ are constants. Then suppose that this distance is decreasing in local population as more populous locations have a higher density of schools¹⁶ such that, $d_{it} = dL_{it}^\gamma$ for some constants $d > 0, \gamma < 0$. Together this gives, $c_{it} = \tilde{c}L_{it}^{\tilde{\chi}}$ where $\tilde{c} = cd^\chi$

¹⁶Perhaps it's more intuitive to consider this in terms of population density, but as locality size is time invariant, and the sufficient statistic relationship is defined in logs, the result will be isomorphic once we allow locality level fixed effects.

and $\tilde{\chi} = \chi\gamma$. As this set-up endogenizes c_{it} as a log-linear function of constants and pre-existing endogenous variables, the sufficient statistic result remains unchanged, although the interpretation of coefficients will vary.

Intensive margin variation in schooling quality, at least partly due to teacher quality and availability, has also played a significant role in recent decades. This can be allowed to vary in response to road building within my framework by noting that increased E-type labor market access will influence the wages of graduates, and thus also those of teachers. This dimension is already captured directly within the sufficient statistic result. It may be, however, that schooling quality reacts to road-building through other channels. In an attempt to allow for this, in the appendix I include school quality explicitly in the sufficient statistic regression as an additional variable, finding that it does not significantly impact opportunity, nor does its inclusion change the coefficients on market access terms. Measuring school quality at the second administrative division level, over time, is challenging in any setting, in this analysis I use a proxy which is the proportion of those who complete primary school but are not literate. This is far from an ideal proxy, as there is relatively little variation (standard deviation over all countries and years is 3pp), but the results give some evidence to suggest that changes to roads are not impacting opportunity through channels not already captured by the sufficient statistic result. I leave a more detailed analysis of the role changes in education quality play for future work.

2.2.2 The sufficient statistic result

The exposition above supplies us with a general theory to consider how changes in connectivity affect endogenous variables including local opportunity. This set up allows us to side step the common empirical issues outlined at the beginning of this section, such as spillover effects, essential heterogeneity and general equilibrium considerations, whilst remaining agnostic with regards to the underlying data generating process. In appendix 2.D.1 I show how the equations above can be combined to arrive at a log linear relationship given in equation 2.7 which relates local educational opportunity to the four market access terms, and can be directly taken to the data.

$$\mu_i = \gamma_1 \cdot \ln(MA_i^E) + \gamma_2 \cdot \ln(MA_i^N) + \gamma_3 \cdot \ln(LMA_i^E) + \gamma_4 \cdot \ln(LMA_i^N) + v_i \quad (2.7)$$

As noted above this result is robust to a broad set of modeling assumptions and micro-foundations and this is formally shown in appendix 2.D.2. It is robust in the sense that the

form of the sufficient statistic result will remain unchanged, local opportunity remains a function of the four market access terms only.

Depending on the exact specification used the interpretation of the coefficients will vary. When taking equation 2.7 to the data, I don't need to take an exact stance on the interpretation of the coefficients, beyond that which significance indicates changes in the transport network do alter the spatial distribution of opportunity, which is true of all interpretations. However, when I turn to using this framework to consider counterfactual road networks I will have to take a more precise stance, and will use the parsimonious model as described in this section.

In appendix 2.D.1 I formally derive the coefficients in terms of structural parameters following the parsimonious model set out above. The interpretations are very intuitive. For example, E -type market access affects μ through five main channels. First, it directly increases the demand for E -type goods and therefore increases w_t^E and so the returns to education. Second, it directly decreases the price level causing greater E -type migration and therefore putting downward pressure on E -type prices. The first impact will be stronger if trade is more elastic than the effective migration elasticity, $\phi_E > (1 - \beta)\lambda_E$, when estimating the model I find the inequality is satisfied, thus in sum increases in E -type market access increase E -type wages. Third, as increases in local wages increase the cost of employing local labor the price of locally produced goods rises, this in turn decreases demand for local goods and therefore acts as a dampening affect on wages. Fourthly, and similarly, increases in local wages further induce migration of E -types, increasing the supply of local workers and acting again as a dampening affect on wages. Lastly, as discussed above changes in MA_t^E will impact local opportunity, however if this impact is positive this will cause greater migration to the area, reducing local wages and thus mitigating and positive affect. Additionally, although we haven't discussed this here, E -type market access will also influence N -type wages which may put downward pressure on local opportunity. The channel through the feedback of how changes in local opportunity impact migration and therefore local opportunity, is complex and depending on parameter values may act as to mitigate or multiple effects.

Understanding the structural relationships underlying the simple interpretation of the coefficients allows me to make strong empirical predictions on the signs of the estimated coefficients. Assuming $\phi_s > (1 - \beta)\lambda_s$, for $s = E, N$, the theory predicts that $\hat{\gamma}_1 > 0$, $\hat{\gamma}_2 < 0$, $\hat{\gamma}_3 < 0$, $\hat{\gamma}_4 > 0$. The coefficient on E -type market access is predicted to be positive ($\hat{\gamma}_1 > 0$) as higher E -type market access increases demand for E -type goods, this in turn

increases E -type wages (in real terms) which increases the returns to education and so the incentives to educate. Analogously and oppositely the theory predicts $\hat{\gamma}_2 < 0$. Turning to labor market access terms, the intuition is again analogous. An increase in E -type labor market access increases the supply of E -type labor, and so puts downward pressure on E -type wages decreasing the returns to education and so incentives to educate. Thus, the theory predicts $\hat{\gamma}_3 < 0$ and oppositely, $\hat{\gamma}_4 > 0$.

These predictions depend on the exact model specified, if for example income effects were included in the education choice problem, an increase in N -type wages due to an increase in N -type market access could have a positive effect on the incentives to educate. Therefore, by formally testing for these inequalities when taking the model to the data, I can implicitly test if the parsimonious set up described in this section can adequately explain behavior. I find that all sign predictions described in the paragraph above are indeed born out in the data.

One limitation of this approach is that the model described in this paper is static, households make one-off migration decisions. If households are forward looking they make take future expectations of migration (and potentially of their children's migration) into account when making migration decisions. Much like in [Morten and Oliveira \[2021\]](#) for models of this type the continuation value is captured as part of a localities amenity and therefore the above described estimation strategy and sufficient statistic result are robust to concerns of this nature. However, it does mean that I am forced to hold the continuation value fixed in counterfactuals, this removes a potential additional source of gains, especially when considering the inter-generational perspective.

2.3 Taking the sufficient statistic result to the data

In this section, I describe how I take the sufficient statistic result developed to the data. I focus on the setting of Benin, Cameroon, and Mali since 1970, and local educational opportunity. To estimate the sufficient statistic result I first require data on market access terms and opportunity over space and time. I use historical Michelin road maps I digitize over the period 1970-2020, and data on local educational opportunity due to [Heath Milsom \[2021\]](#). Additionally, although the sufficient statistic gives an equation to estimate, market access terms will inherit the endogeneity of road placement. This identification problem is the second main empirical challenge that must be overcome to make progress. I provide identification by developing a novel instrumental variables approach which utilizes the structure

of market access terms and not-on-least-cost-path variation.

2.3.1 The setting, data, and descriptive evidence: Benin, Cameroon, and Mali since 1970

Benin, Cameroon, and Mali together provide a near-perfect setting to study how changes in connectivity affect spatial inequality. Whilst also being a particularly pertinent one from a policy perspective. Each country displays considerable past variation in local connectivity due to road building, which I will leverage when estimating the effects of changes in connectivity on spatial inequality. Additionally, due to low preexisting levels of paved road coverage [Gwilliam, 2011, Foster and Briceño-Garmendia, 2010] and high anticipated urbanization and population growth [UN, 2018] considerable investment in road infrastructure is expected in the near future, making this a particularly direct and policy-relevant setting to be studying the impacts of road building.

I use the causal effect of growing up in a given location on the probability of completing primary school as my main measure of local opportunity for three main reasons. First, due to the large informal sector, it is unclear whether the more traditional measure of later life income, as used in other contexts, is appropriate here. Additionally, data on primary completion rates are available at a fine geographic level over time and are unlikely to suffer from significant measurement error. Second, primary completion rates are correlated with opportunity more broadly defined in later life. In this setting individuals who have completed primary school are less likely to work in agriculture, have better housing quality, and greater returns to education [Psacharopoulos and Patrinos, 2018]. Finally, primary schooling is the most salient margin of education, in my sample about a third of individuals have completed primary school, but only 7% have completed secondary school. Thus, although defensible as a relevant and general measure of local opportunity, primary education completion cannot speak to all potential dimensions and so opportunity in this paper should be taken to mean local *educational* opportunity. In appendix section 2.B.3 I show that there is considerable variation across space in the local returns to education and educational opportunity.

I study three countries as, on the one hand, data limitations prevent widening the net and including more countries from Sub-Saharan Africa. In order to study the impact of changing connectivity on local opportunity, I need data at a fine geographic level, over time, on local opportunity — which is only available in Benin, Cameroon, and Mali due to Heath Milsom

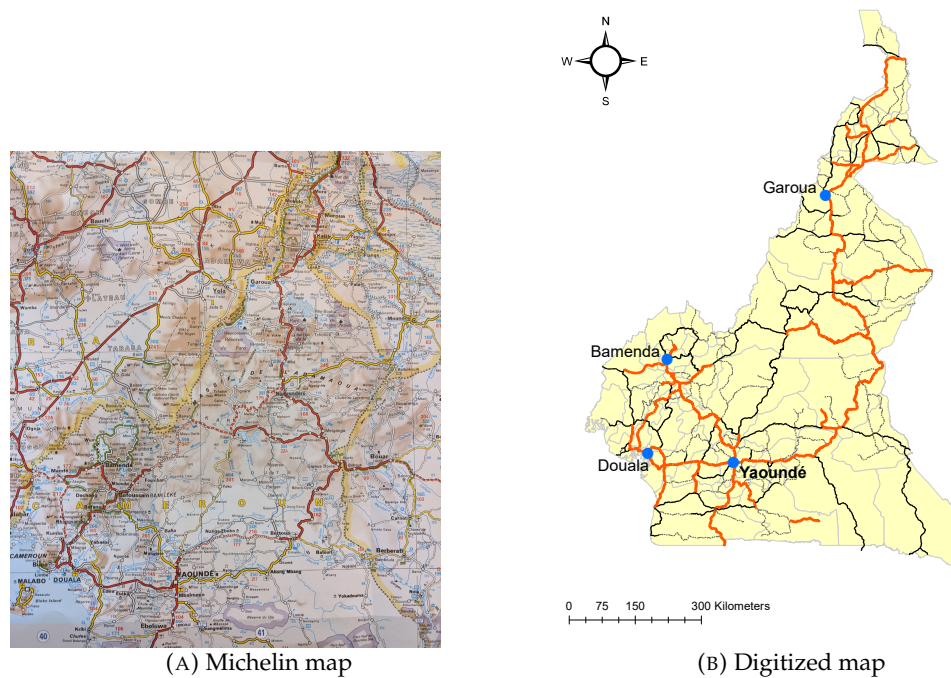
[2021]. On the other hand, I consider as many countries as possible, as it allows me to investigate the importance of network-level, that is country-level, characteristics in determining the effect of changes to connectivity over space. Such characteristics have been neglected previously in the literature but will turn out to play an important role — highlighting the danger of extrapolating results from one network (country) to another.

In this section, I describe the main data sources and institutional context, summarize key dimensions of variation, and present some descriptive correlational evidence. This evidence shows that places that became more connected as the result of road building saw larger increases in local returns to education which is then shown to positively correlate with greater increases in in local educational opportunity.

The changing geography of connectivity: Digitizing historical Michelin road maps

Road data comes from historical Michelin maps which I have digitized using GIS software from the following years: 2019, 2012, 2003, 1986, 1976, and 1969. In these maps, it's possible to consistently classify roads into highways, paved roads, improved roads (laterite or gravel), and dirt roads. This provides a full description of the (main inter-city) roads over time since 1969 in each country. The ability to distinguish road type is of particular importance as much of the variation in connectivity in later years comes from upgrading roads rather than building new ones. Figure 2.3.1 gives an example of this process. Panel 2.3.1a shows an image of the raw Michelin map of Cameroon in 2019 and panel 2.3.1b shows the digitized version.

FIGURE 2.3.1 Digitizing Michelin road maps — Cameroon 2019



Notes: This figure shows in panel 2.3.1a the original Michelin road map of Cameroon in 2019 and in panel 2.3.1b the digitized version. In panel 2.3.1b thick red lines are paved roads, dark black lines are improved roads and gray lines are dirt tracks. Note that in panel 2.3.1a the color of roads denotes their importance/ frequency of use for long-range trucking and does not necessarily reflect their size or quality which is denoted instead by the thickness of outlines.

Michelin maps are themselves constructed using four main sources (Jedwab and Storeygard [2021]): the previous Michelin map, government road maps, local information from Michelin tire stores across Africa, and finally direct correspondence from users. It is generally thought that this process leads to consistent and accurate road mapping over time, and certainly is the only known source providing such information. Indeed, the success of Michelin maps relied on them being a trustworthy source of information, and so Michelin had a vested interest in providing accurate maps. However, it's still very possible that not every change is noted, and even if a change is noted it could only be included with some lag. Additionally, although I can use variation in road upgrading, this can only be observed when a road changes categories. That is, road maintenance or changes that would not count as upgrades across categories (such as pothole filling) are not captured. This may be a particular issue in the more recent years as it is expected a greater proportion of road spending reflects this unobserved variation.

Constructing a network data set

I use the digitized road maps to construct a network data set for each country, where roads correspond to the arcs in the network. To complete the construction of a network data set, however, I need to know node locations and weights, that is the location and size of agglomerations. I obtain agglomeration populations and locations from Africapolis that combines various primary data sources to estimate the population of all agglomerations above 10,000 in 2015 on a decadal basis since 1960. To this dataset of known large agglomerations, I add a set of location centroids for each sub-national geography with weights equal to the estimated remaining urban population found using Census data for each country. Further details can be found in the appendix sub-section 2.C.1. With nodes and arcs in hand, I construct a network data set for each country in each year a map is available and run Dijkstra's algorithm to find the fastest path along the known network from each node to each other node. I follow [Jedwab and Storeygard \[2021\]](#) who also digitize Michelin maps in Africa assuming travel speeds of 60km/hour on paved roads, 40km/hour on improved roads, and 12km/h on dirt roads.

This paper uses census data (10% samples accessed from [IPUMS \[2020\]](#)) from every available census in Benin, Cameroon, and Mali covering 1976 to 2013¹⁷ with a total of 8 million observations across 164 localities and 444 locality-year cells. Individuals are geo-located at the second administrative unit level which has a median population of 267,000 across all samples¹⁸. This data gives me information on migration histories, local population, local employment compositions, housing quality, and education completion. In my data, an individual has completed primary education if they have completed the mandatory 6-year primary cycle¹⁹. This does not include Koranic schools but does include Medersas in Mali as they follow the national curriculum [[Boyle, 2014](#)], in appendix 2.B.9 I show that both of these non-traditional schools play a minor role.

Census data, although rich on many dimensions, and covering a broad geography/ time frame, does not include information on incomes. To recover income estimates at the locality-census year-education level, I use asset data from censuses coupled with available income data from Development Health Surveys (DHS) using an Engle curve approach [[Young,](#)

¹⁷In Benin censuses were conducted in 1992, 2002, and 2013 covering 77 geographies. In Cameroon, censuses were conducted in 1976, 1987, and 2005 covering 39 geographies. In Mali, censuses were conducted in 1998, and 2009 covering 48 geographies.

¹⁸Benin's Communes have a median population of 103,000 with an inter-quartile range of 71,000 to 173,000. Mali's circles have a median population of 308,000 with an inter-quartile range of 197,000 to 520,000. Cameroon's departments have a median population of 456,000 with an inter-quartile range of 225,000 to 907,000. A broadly comparable geographical unit in the US would be commuter zones.

¹⁹In each country over the entire sample period the primary school cycle is 6 years with the exception of the English speaking regions of Cameroon where the primary cycle is 7 years. To remain consistent in these localities children are marked as having completed primary school if they have completed 6 years.

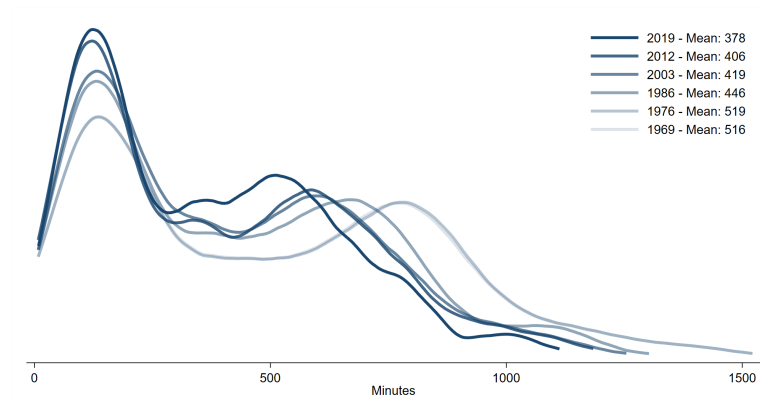
2012]. DHS surveys are available in Benin in 1995 and Mali in 1996. DHS surveys are relatively large-scale representative surveys that include information on earnings (not all DHS surveys include earnings information, but the two mentioned above do) as well as the same assets as one can observe in the survey data. Details of this procedure can be found in appendix 2.C.1.

Variation in connectivity over time and space

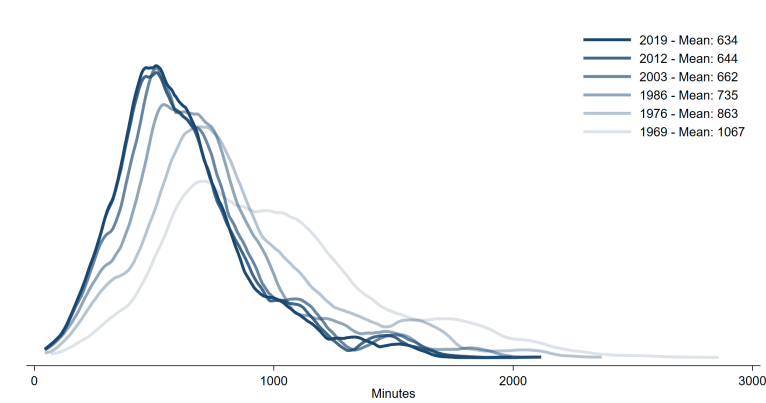
In figure 2.3.2, I plot the distribution of expected travel times for each location, in each year maps are available over the digitized road network²⁰. The expected travel time for a given location is defined as the time an individual should expect to travel for, if they were to pick a person at random to travel to from the rest of the country. To calculate this over time I fix the population distribution to 1970 levels and calculate each localities expected travel time using the road network in each year. All three figures show considerable leftward shifts in the distribution of travel times over the study period, with mean travel times decreasing by 27%, 41%, and 44% in Benin, Cameroon, and Mali respectively.

²⁰Over my study period, in Benin, Cameroon, and Mali, other forms of transport such as railways or waterways exhibited little variation and are not modeled.

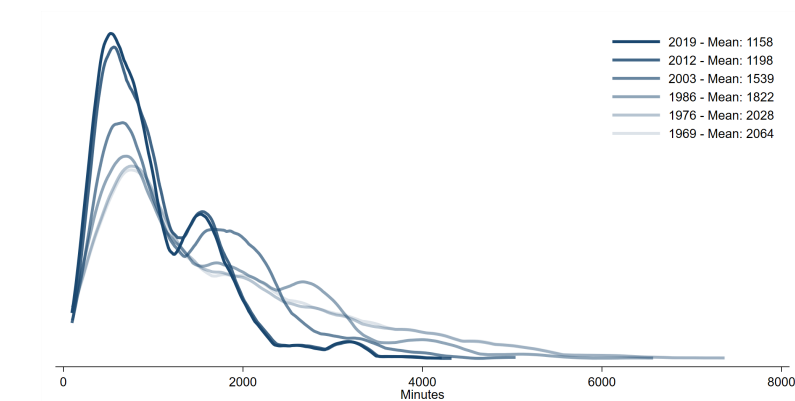
FIGURE 2.3.2 Distribution of expected pairwise travel times between localities
(A) Benin



(B) Cameroon



(C) Mali



Notes: These figures show the density of expected travel times from each locality in a given country-year. Expected travel time in a given location is defined as the time an individual in the location should be expected to be traveling if one chooses an individual at random in the same country to travel to. The population distribution is kept fixed at 1970 levels, but the road network is allowed to vary. More recent years are denoted in a darker shade of blue. Population-weighted means across localities for each year are given in the top right.

Measuring market access

The sufficient statistic results suggest that within a broad class of spatial economics models, all of the complex forward and backward network effects road building may have on causal place effects, are captured by the sufficient statistics of labor and trade market access for E and N types, given in equation 2.5. Each market access term is a series of non-linear simultaneous equations that can't be solved until the bilateral cost terms, $\{\tau_{ijt}^{-\phi_s}\}$, $\{\kappa_{ijt}^{-\lambda_s}\}$, have been recovered. In addition, I don't directly observe local output, $Y_{it}^s = w_{it}^s L_{it}^s$, as I don't observe wages at a fine geographic level over time in each country. In order to overcome these issues I make the following simplifications and assumptions. First, I model bilateral iceberg transport costs as depending on the logarithm of the calculated least cost path travel times estimated via Dijkstra's algorithm from the digitized Michelin road maps, t_{ijt} , as follows $\phi \ln(\tau_{ijt}) = \tilde{\phi} \ln(t_{ijt})$, $\lambda_s \ln(\kappa_{ijt}) = \lambda_s \ln(t_{ijt})$ similar to Allen and Donaldson [2020]. Second, as I don't observe wages in the census data I use an Engle-curve based imputation relying on asset data available in censuses, and wage and asset data from Demographic and Health Surveys (DHS), using a method similar to Young [2012]. Details on both procedures can be found in appendix 2.C.1.

As I have data on bilateral internal migration I can estimate $\tilde{\lambda}^s$ using the gravity relationships postulated in the theory. However, I don't have data on locality-level bilateral trade and so am forced to take a value of $\tilde{\phi}$ from the literature.

Once the above has been estimated, equation 2.5 is a series of simultaneous non-linear equations which have a unique positive solution [Donaldson, 2018] that can be found numerically. Calculating market access terms only requires uncovering the bilateral transportation cost terms $\{\tau_{ijt}^{-\phi_s}\}$, $\{\kappa_{ijt}^{-\lambda_s}\}$ not τ_{ijt} , κ_{ijt} nor ϕ_s , λ_s independently. This means that I don't have to heavily rely on the structure of the model or introduce further identification restrictions, as I will have to do when estimating the full model, in order to calculate market access terms.

Descriptive evidence on the relationship between connectivity and local educational opportunity

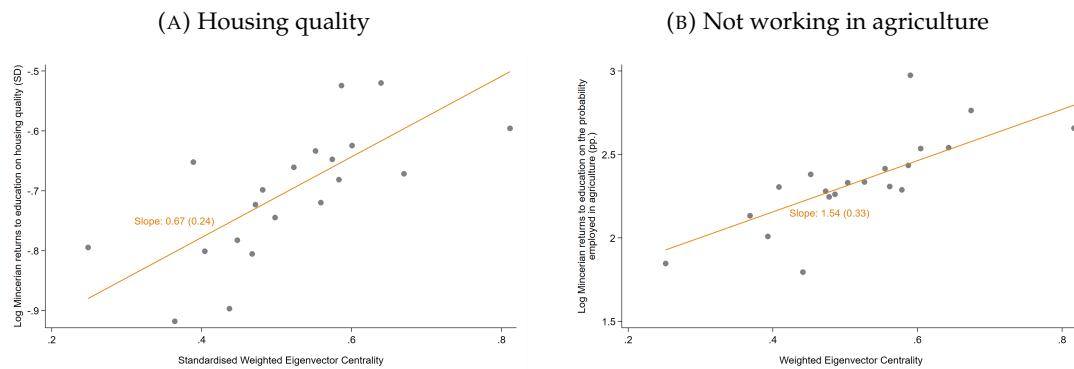
Having established that there is considerable variation in causal place effects, returns to education, and connectivity, I turn to providing some indicative correlational evidence on the key relationships between these sources of variation before adding structure to the data. To measure changes in connectivity I use a well-known centrality measure that most closely

resembles the theoretical concept of market access developed later in the paper. That is, I use Eigenvector centrality where the centrality assigned to each node is taken from the eigenvector associated with the largest eigenvalue of the adjacency matrix between nodes calculated as the least-cost path across the digitized road network. Formally, $\lambda e = Ae$ where e is a vector of eigenvector centralities (i.e. the eigenvector), λ is the largest eigenvalue and A is the adjacency matrix. Eigenvector centrality explicitly takes the centrality of places a locality is connected to into account when calculating the importance of a node, much like market access — however results are similar for other appropriate centrality measures.

In the binscatter plots and corresponding OLS regressions which follow I include locality and year fixed effects as well as weighting by locality population and clustering standard errors at the locality level. Because of this, coefficients can be interpreted as the association between changes in centrality (due to changes in the road network) and changes in local returns to education in figure 2.3.3 and changes in local returns to education and causal place effects in figure 2.3.4.

First, I show that there is a strong correlation between local connectivity and the local benefits of completing primary education. Figure 2.3.3 shows that locations which saw greater increases in centrality due to road building also saw larger increases in the benefits due to completing primary school measured in terms of housing quality (2.3.3a) or the probability of not working in agriculture (2.3.3b).

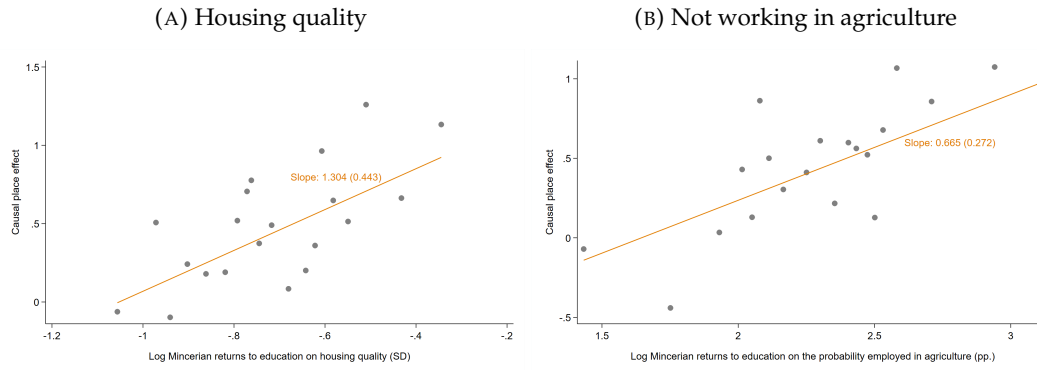
FIGURE 2.3.3 Correlation between connectivity and local returns to education



Notes: This figure shows in panel 2.3.3a the correlative relationship between the log Mincerian returns to education in terms of housing quality on the y -axis and eigenvector centrality on the x -axis. Panel 2.3.3b shows the correlative relationship between the log Mincerian returns to education in terms of the probability of not being employed in agriculture on the y -axis and eigenvector centrality on the x -axis. In each case Mincerian returns, β_l are calculated using the following regression $y_i = \beta_l^y \text{Primary}_i + \beta_{1l} \text{age}_i + \beta_{2l} \text{age}_i^2 + \varepsilon_i$ for each locality l separately and for y equal to housing quality or a dummy variation equaling one if not employed in agriculture. Housing quality is calculated as the first principle component in a PCA analysis of floor, wall, roof material, access to electricity, and sanitation. In a second stage, the above binscatter plots are constructed by comparing Cent_l with β_l^y controlling for locality and year fixed effects. Slope coefficients are indicated in orange on the figures and have been calculated from the analogous linear regression. Associated standard errors are given in parenthesis clustering at the locality level.

Figure 2.3.4 then shows that there is a strong correlation between local returns to education either measured in terms of housing quality, or propensity to not be working in agriculture, and local educational opportunity. A one percent of a standard deviation increase in housing quality is associated with a 1.3pp increase in the causal effect of place on primary completion, and similarly, a one percent increase in the probability of not being employed in agriculture is associated with a 0.67pp. increase in causal place effect. This gives some suggestive evidence both that local returns to education matter, and that they are correlated with causal place effects.

FIGURE 2.3.4 Correlation between local returns to education and opportunity



Notes: This figure shows in panel 2.3.4a the correlative relationship between the log Mincerian returns to education in terms of housing quality on the x-axis and causal place effects on the y-axis. Panel 2.3.4b shows the correlative relationship between the log Mincerian returns to education in terms of the probability of not being employed in agriculture and causal place effects on the y-axis. In each case Mincerian returns, β_l are calculated using the following regression $y_i = \beta_l^y Primary_i + \beta_{1l} age_i + \beta_{2l} age_i^2 + \varepsilon_i$ for each locality l separately and for y equal to housing quality or a dummy variation equaling one if not employed in agriculture. Housing quality is calculated as the first principle component in a PCA analysis of floor, wall, roof material, access to electricity, and sanitation. In a second stage, the above binscatter plots are constructed by comparing μ_l with β_l^y controlling for locality and year fixed effects. Slope coefficients are indicated in orange on the figures and have been calculated from the analogous linear regression. Associated standard errors are given in parenthesis clustering at the locality level.

Finally, figure 2.3.5 puts the previous two figures together and shows that there is also a direct, positive, correlational relationship between connectivity and causal place effects. This is the relationship that I will explore in more detail in the remainder of the paper, putting some more structure and theory on what is meant by *connectivity*, providing causal evidence, and finally building a structural model to quantify the counterfactual effects of past and future road building on spatial inequality of opportunity.

FIGURE 2.3.5 Correlation between connectivity and local opportunity



Notes: This figure shows the correlative relationship between locality eigenvector centrality on the x-axis and causal place effects on the y-axis. The binscatter plots are constructed by comparing μ_l with $cent_l$ controlling for locality and year fixed effects. The slope coefficient is indicated in orange on the figure and is calculated from the analogous linear regression. Associated standard errors are given in parenthesis clustering at the locality level.

2.3.2 Overcoming the endogeneity of road placement

A road may be built to galvanize a previously flagging area or serve a rapidly expanding one. Whereas turning to theory and taking a market access approach allows me to overcome the first core challenge of how to measure the complex effect changes in the road network may have on outcomes, it doesn't immediately give direction on how to estimate the causal impacts thereof as market access terms inherit the endogeneity of road placement. In this subsection I thus turn to overcoming the second core empirical challenge presented by my research question — the endogeneity of road placement.

Previous identification strategies designed to overcome the endogeneity of road placement include using placebo lines from planned but unbuilt routes Donaldson [2018], Okoye, Pongou, and Yokossi [2019], using straight line or least-cost path spanning tree instruments²¹ Moneke [2020], Michaels [2008], Ghani, Goswami, and Kerr [2016], Faber [2014], or leveraging *far-away* variation in road changes Donaldson and Hornbeck [2016], Jedwab and Storeygard [2021]. In my setting, it's difficult to see how the first two approaches can be implemented. First, I don't have data on unbuilt but planned placebo lines. Second, there is no clear set of locations that are being connected, and in addition, a significant proportion of

²¹Locations that just happen to lie between two cities that are being connected by a road may be plausibly described as exogenous. This is also known as the *inconsequential units* IV.

the variation in travel times comes from road upgrading rather than the building of entirely new roads, in this setting it's unclear how localities are "incidentally" connected.

The third strategy, leveraging far away variation in roads, is appealing but suffers from a number of known drawbacks. First, it's unclear how far "far-away" should be; and although researchers can present many distances, it is ultimately an ad-hoc choice. Second, and more fundamentally, variation due to large projects which may be far away but for endogenous reasons, or relatively far away connections that are built to ease transport to, or encourage trade to a given location, remain threats to identification.

In this paper, I propose a novel identification strategy that builds upon the far-away variation approach, by considering not-on-least-cost-path variation. Not-on-least-cost-path variation only uses changes in a locality i 's market access that stems from indirect changes to all other locations' market access freezing the least cost path from i to all other locations²². This approach has two intuitive explanations. First, one can consider it as the same as using far-away variation, but whereas far-away variation defines distance over Euclidean space, not-on-least-cost-path defines distance over network space, where a "distance" of one refers to one-degree removed indirect variation. Under this interpretation, we take the network structure seriously and resolve the ad-hoc nature of what "far-away" may mean by appealing to the theory. Secondly, one can think of not-on-least-cost-path variation as approximating the decision-making process of the policymaker building roads, and using the residual variation. If a central planner builds roads to/ from i in order to directly improve its connectivity, we don't use that variation and instead consider the residual variation in market access.

As is evident from the intuitive explanations of not-on-least-cost-path variation, the process can be iterated and one can consider n -th order indirect variation. Intuitively this is just considering n -th order removed indirect variation or extending the allowed complexity of the policymakers' decision-making process. For example, 2nd order not-on-least-cost-path variation allows a policymaker to build roads to improve the direct connection from i to any other location, as well as, indirect connections from j to k that may be built to ease congestion, or improve trade, to i . The iterative nature of the resulting instruments presents an opportunity as we consider increasingly likely to be exogenous variation one can test

²²This approach is not to be confused with one variation of the incidental connection approach which uses an estimated least cost route over an estimated cost surface to instrument for actual connections. I don't take this approach in this paper for the same reasons that incidental connection approaches, in general, are unsuitable, and in addition, because a significant proportion of the variation comes from within road quality and it is unclear how to leverage this intensive margin dimension using this approach.

whether at some point coefficients stabilize giving evidence to suggest that sources of endogeneity have been removed²³.

To formalize the above discussion, consider a generic market access variable $MA_{it} = \sum_{j \in \mathcal{L}} \tau_{ijt}^\phi Y_{jt} (MA_{jt})^{-1}$. Following Donaldson [2018] I first remove own-location market access (therefore sum over other locations $j \in \mathcal{L}/i$) and freeze the market size variable at the initial level (L_{j0}) as these objects are co-determined with the outcome variable. Then I can use not-on-least-cost-path variation (freeze the least cost path to the initial value τ_{ij0}) of degree n by constructing the following instrument.

$$MA_{it}^{IV(n)} = \sum_{j \in \mathcal{L}/i} \tau_{ij0} Y_{j0} \left(MA_{jt}^{IV(n-1)} \right)^{-1}$$

Where $MA_{it}^{IV(0)} = \sum_{j \in \mathcal{L}/i} \tau_{ijt} Y_{j0} \left(MA_{jt}^{IV(0)} \right)^{-1}$ gives actual market access terms. See appendix section 2.B.5 for a graphical explanation of the far-away-variation approach.

This approach does, however, have some limitations. Instruments will fairly quickly become weak as one iterates, in my application, this happens after the four-order iteration. In addition, this strategy will not be able to overcome endogeneity that occurs at the level of large geographies, for example, a program to build more roads in the south of the country to stimulate growth there. However I show in appendix 2.B that clientelism is not of first order concern in this setting. In addition, this approach, much like any that relies on market-access type measures, will suffer from the Borusyak and Hull [2020] critique of endogenous exposure to exogenous shocks. However, this is relatively easily overcome by permuting over possible roads a procedure which is described in more detail below.

Sorting across locations

A remaining threat to identification is that changes in market access might induce selection into migration and migrating location and therefore sorting. Perhaps it is the case that higher market access areas simply induce those who are more likely to complete primary school anyway due to family characteristics, to disproportionately locate there. Such a story would preclude any claims regarding the impact of road-building on the underlying causal effect of place on primary completion rates. To counter such a possibility I use data directly on the causal effect of place from Heath Milsom [2021] which captures only the part of the

²³As with any analysis of this type, if one allows policy makers to be infinitely sophisticated and consider the entire general equilibrium impacts of their actions it will be very difficult to identify any exogenous sources of variation. However, due to the iterative nature of this approach one can consider stability of estimates as indicating that, as least using this procedure, we've reached the limits of policy makers sophistication.

local variation in observed primary completion rates which is due to causal place effects rather than differing characteristics of individuals over space. Note that local demographics will still influence place effects through affecting local labor markets and so wages. By considering causal place effects I isolate this channel separately from sorting.

2.3.3 Results from estimating the sufficient statistic relationship

With the identification strategy in hand, I can estimate the theory-informed sufficient statistic result derived above. I estimate specifications of the form given in equation 2.8 where recall μ_{it} is the locality level measure of the causal effect growing up in a location has on the probability of completing primary school, τ_i denotes locality fixed effects, α_t time fixed effects and v_{it} an idiosyncratic error.

$$\mu_{it} = \gamma_1 \cdot \ln(MA_{it}^E) + \gamma_2 \cdot \ln(MA_{it}^N) + \gamma_3 \cdot \ln(LMA_{it}^E) + \gamma_4 \cdot \ln(LMA_{it}^N) + \tau_i + \alpha_t + v_{it} \quad (2.8)$$

I estimate equation 2.4 on a sample of 334 locality-year pairs for which I have data on local opportunity. Market access terms are calculated based on the road map year closest to 14 years before local opportunity estimates are calculated. In this manner I allow the road network to influence the entire (pre-primary completion) childhood of individuals. Table 2.3.1 and figure 2.B.12 display the main results from this section. In each column of table 2.3.1 I build up my identification strategy including instruments that consider more and more plausibly exogenous variation in road building/ upgrading. Column (1) displays the OLS results. Column (2) instruments each market access and labor market access term by its counterpart removing a locations own market. Column (3) additionally keeps all markets at a constant size in 1970. In subsequent columns I take column three as the baseline and individually add restrictions to the instrument used. Column (4) removes variation in the least cost paths for each locality. Column (5) removes least-cost path variation and only considers far-away variation. Column (6) removes first and second-order least-cost path variation and far-away variation. Column (7) removes up to third-order least-cost path variation and far-away variation. Column (8) removes up to fourth-order least-cost path variation and far-away variation²⁴. Removing higher than 4th order least-cost path variation results in weak instruments. For each regression (column) the Kleibergen-Paap under-identification

²⁴In this specification far-away is defined as 20km although results are not sensitive to the distance used. In general, a smaller distance than normal is appropriate to my setting when coupling far-away variation with not-on-least-cost-path variation.

rank Lagrange multiplier statistic and associated p -value are reported as well as the [Sanderson and Windmeijer \[2016\]](#) first-stage under and weak identification statistics for each individual regressor are reported. Under/ weak identification tests appear satisfactory in all specifications except column (5). Standard errors are clustered at the locality level. For a full discussion of inference in this setting see sub-section 2.3.3 and appendix section 2.B.10.

TABLE 2.3.1 Results from estimating the sufficient statistic relationship

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	OLS	No including own market	Constant market size	Not including least cost path variation	1st order least cost + only far away variation	2nd order least cost + only far away variation	3rd order least cost + only far away variation	4th order least cost + only far away variation
Log(LMA Educ)	-0.0598 (0.0509)	-0.0569 (0.0564)	-0.170 (0.107)	-0.0312 (0.125)	-0.264 (0.356)	-0.192* (0.108)	-0.179* (0.106)	-0.175 (0.110)
Log(MA Educ)	0.0434** (0.0201)	0.0961*** (0.0271)	0.137*** (0.0290)	0.0545 (0.0598)	0.0904 (0.237)	0.0795** (0.0396)	0.0699 (0.0425)	0.0696 (0.0453)
Log(LMA No Educ)	0.151*** (0.0249)	0.0767 (0.0850)	0.265** (0.124)	0.156* (0.0879)	0.619 (0.475)	0.252* (0.131)	0.215** (0.0997)	0.218** (0.106)
Log(MA No Educ)	-0.0814*** (0.0212)	-0.132** (0.0529)	-0.185*** (0.0403)	-0.136** (0.0563)	-0.111 (0.293)	-0.151** (0.0682)	-0.123** (0.0555)	-0.118* (0.0601)
Locality by year FE	X	X	X	X	X	X	X	X
Kleibergen-Paap stat		7.39	8.01	5.65	0.52	2.93	4.23	4.74
Kleibergen-Paap p-value		0.007	0.005	0.017	0.470	0.087	0.040	0.029
SW under ID stat LMA(E)		108.53	34.45	7.84	2.61	15.97	26.00	23.65
SW under ID stat MA(E)		15.48	14.02	10.49	0.84	10.28	12.31	10.33
SW under ID stat LMA(NE)		13.40	13.21	12.59	1.16	4.57	9.81	11.49
SW under ID stat MA(NE)		12.93	27.01	23.74	0.69	5.00	12.45	11.25
SW weak ID stat LMA(E)		112.78	35.80	8.15	2.71	16.59	27.02	24.57
SW weak ID stat MA(E)		16.09	14.57	10.90	0.87	10.68	12.79	10.73
SW weak ID stat LMA(NE)		13.92	13.73	13.08	1.21	4.75	10.20	11.94
SW weak ID stat MA(NE)		13.43	28.07	24.67	0.71	5.19	12.94	11.69
# localities	127	127	127	127	127	127	127	127
N	334	334	334	334	334	334	334	334

Notes: This table shows the results from running regressions of the form given in equation 2.8. In column (1) OLS is employed and in subsequent columns (just identified) 2SLS methods are used with instruments as indicated by the column titles. Locality by year fixed effects are included in each specification. Standard errors are clustered at the locality level. The dependent variable is the causal effect of spending an additional year of childhood in a given location on the probability of completing primary education as estimated in [Heath Milsom \[2021\]](#). Kleibergen-Paap and Sanderson-Windmeijer weak-/ under- identification tests are reported below for each regression.

Individual coefficients can be interpreted as percent to percentage point change, i.e. a 1% increase in market access causes a β percentage point change in the causal effect of spending an additional year of childhood in a given location on the probability of completing primary education. Table 2.3.1 shows results in line with the predictions from the theory: E-type labor market access has a negative coefficient, E-type market access has a positive coefficient, N-type labor market access has a positive coefficient and N-type market access has a negative coefficient.

Figure 2.B.12 in the appendix displays the same results in a graphical format omitting

the weak IV case in column (5). It's clear from this figure that results remain stable across specifications. Given the cumulative nature of the IVs employed this suggests that most substantive sources of endogenous variation which can be addressed using this iterative procedure, have been removed. In figure 2.B.12 coefficients significant at the 5% level have been colored in, and those not remain transparent, however, we can improve efficiency by combining instruments and so do in table 2.B.8 in the appendix.

To provide as precise results as possible I combine instruments using 2nd to 4th order least-cost-path-variation and report my main sufficient statistic coefficient estimates in column (1) of table 2.B.8 in the appendix. In column (2) of table 2.B.8 I show results controlling for expected market access (in all four variables) calculated as described in subsection 2.3.3. The coefficient estimates are similar across the two columns and so I retain estimates from column (1) as my main results when considering quantification and counterfactuals.

In appendix table 2.B.11 I show results estimating the sufficient statistic for each country individually. Testing for the equality of coefficients between the country-specific regressions and the pooled regression — I find no statistically significant difference (although power concerns prevent strong conclusions). Differences in coefficients across countries can only reflect varying structural parameters for each setting. Therefore similarity of coefficients suggests that using the same set of parameters for each country is appropriate when considering counterfactual analysis.

In addition table 2.B.13 in the appendix shows results using returns to education or the proportion employed in agriculture as alternative outcome variables. I find significant effects of market access terms on these variables, giving further evidence on some of the main channels posited.

Non-random exposure to plausibly exogenous variation in road building

In recent work, [Borusyak and Hull \[2020\]](#) illustrate that even in the case of random shock assignment spatial regressions such as those considered here may be biased. This is due to the network theoretic nature of the variable market access, some regions, even under random road placement, are more likely to receive higher increases in market access than others. For example, locations near a country's border will mechanically see smaller increases in market access, and thus if there is any correlation between centrality and the variable of interest, estimates will be biased. Similarly, geography matters: areas that are particularly inhospitable, or close to inhospitable regions, are mechanically less likely to see large increases in market access. Even more concerning the initial urban structure of a country matters in

a similar fashion. In order to overcome this [Borusyak and Hull \[2020\]](#) suggest constructing a *expected instrument* and demeaning market access from this in regressions. The expected instrument is exactly the (trade or labor) market access a region would expect to receive on average over many possible realisations of the road-building data generating process.

In order to construct the expected instrument, one must first specify a data generating process for the as-good-as-random shocks to the transport network. [Borusyak and Hull \[2020\]](#) suggest a number of ways to do this such as permuting over proposed but not built roads or using local policy discontinuities. I take a similar approach, permuting over all possible individual road upgrades. That is, in any given year, I iterate through all roads and upgrade them by one category (or build yet-to-be-built roads). I then take the new, hypothetical, road network and re-calculate market access values. Averaging market access calculated in this fashion over every road (I do not upgrade paved roads as faster roads are not observed in my data set), in a given country-year I calculate expected market access terms.

Top-coding of primary completion rates and non-linearities in market access and causal place effect changes

Two potential concerns arise from the above specification. The first is that in reality, primary completion rates cannot move above 100%, which represents a potential top-coding issue. However, as figure [2.B.7](#) in the appendix section [2.B.7](#) shows, in my sample period no locations reach this upper bound, therefore top-coding is unlikely to be affecting results.

The issue of top-coding however, is not necessarily so clear-cut. It's likely that areas with pre-existing low primary completion rates saw larger increases over time as such changes were more possible or easier. As causal place effects are defined relative to the country average, the same issue is apparent. Areas far below a country average, could mechanically see greater increases. In and of itself, this doesn't necessarily represent a problem for identification, but it is in addition possible that the same phenomenon affects changes in market access terms. Areas with pre-existing low market access could more easily see large increases (in log terms) than areas with pre-existing high market access. Combined, these could create a spurious correlation between causal place effects and market access in the above regression which includes locality and time fixed effects.

To consider whether this is a serious concern I regress base levels on long differences for each market access term, primary completion rates, and causal place effects. Tables [2.B.2](#) and [2.B.3](#) in the appendix section [2.B.7](#) show the results. In table [2.B.2](#) I regress changes

for each variable on levels of all variables and in table 2.B.3 I regression changes for each variable on the level of that variable only. The results show that although changes in market access terms are negatively correlated with observed primary completion rates, the same is not true for causal place effects, which have no relation. In addition, for the majority of market access variables when both individually included or combined, there is no relationship between changes and levels, and if a relationship does exist it is more often positive than negative. The tables also show a *positive* relationship between pre-existing primary completion rates and changes in primary completion rates, giving further evidence to suggest that in the sample I consider, the potential concavity of the ease of increasing primary completion rates is not biting. Additionally, these tables replicate the results in [Heath Milsom \[2021\]](#) by showing strong persistence in causal place effects. In sum, these results suggest that the concerns raised are not significant in my sample and setting.

One may also be concerned if market access variables depend on past market access terms, especially if such auto-correlation is also found in market access instruments. This could be the case if the network endogenously reacts to previous market access changes, or there are long-run projects (or connected projects) taking multiple decades to complete. I test for this possibility in appendix section 2.B.8 by including lagged market access terms in the main sufficient statistic regression — and find no evidence that this is a relevant channel for concern in this setting.

Inference

Inference surrounding the coefficients recovered from estimating equation 2.8 is complicated by four main factors. First, serial correlation within geographical units over time is undoubtedly pervasive. Second, spatial correlation. Third, the dependent variable has itself been estimated. Fourth, due to the nature of market access terms, they likely encode complex dependencies over space [[Borusyak and Hull, 2020](#)]. To overcome these potential problems I respectively: cluster standard errors at the locality level, use Conley standard errors, note that classical measurement error on the left hand side works to attenuate coefficients, and use the randomization inference procedure suggested by [Borusyak and Hull \[2020\]](#). Details of each approach and results are given in appendix section 2.B.10. I find that standard errors remain similar to those reported in this section, never significantly inflating.

Interpreting the sufficient statistic results

Certainly, the results from this section show that changes in connectivity, through market access terms, have causally altered the spatial distribution of opportunity in Benin, Cameroon, and Mali. Market access terms are a sufficient statistic in the sense that they capture all of the impact of roads on local opportunity, but there may be other factors which also affect the causal effect of place. I find that in an R^2 sense changes to market access terms explain 63% of the residual signal variation in μ_i conditioning on locality and time fixed effects.²⁵ Therefore, although they explain a sizable proportion of the explainable variation, there remains significant scope for other factors to play an important role.

Going further, and asking what the effect of a specific road was, or discussing aggregate effects, is not possible without leveraging the full structure of the model. This is because the results from estimating the sufficient statistic relationship encode significant heterogeneity. They allow every road, to affect every location differently, and for this to depend on the entire pre-existing road network and distribution of economic activity. That is — network-level characteristics matter. To overcome this third and final challenge I use the estimated coefficients from the sufficient statistic result to back out structural parameters, and then leverage the full structure of the model to answer policy-relevant questions.

2.4 Counterfactual analysis

The sufficient statistic result and not-on-least-cost-path identification strategy allow me to overcome the first two challenges presented by my research question — spillovers and general equilibrium effects as well as the endogeneity of road placement. However, as highlighted by the results from estimating the sufficient statistic equation, one challenge remains: network-level characteristics matter. To know the impact of building any given road, or set of roads, you have to take into account the entire preexisting network and distribution of economic activity. To allow me to do this, and study how network-level characteristics matter for the impact of any given road, I set up the parsimonious model described above for counterfactual analysis.

Relative to the sufficient statistic result, I must now take a stance on the exact ingredients of the model but can remain agnostic as to the micro-foundations. I use the framework described in the main text, with costly migration and trade, two sectors/ types, education

²⁵This estimate includes any changes in market access terms over this period, not just those due to road building.

choice, and exogenous education costs. I use this parsimonious specification because it's the simplest which still captures the main objects of interest, and can be directly estimated from the identified sufficient statistic coefficients, closely linking the empirical and structural aspects of the paper.

The model is as set out in section 2.2 and can be summarized by the seven equations given below for each locality i , period t and sector/type s . Combining this structure with exogenous variables $\{B_{it}^E, B_{it}^N, A_{it}^E, A_{it}^N\}$ (location characteristics) and $\{\tau_{ijt}, \kappa_{ijt}\}$ (transportation costs), and parameters $\{\phi_E, \phi_N, \lambda_E, \lambda_N, \beta\}$, I can solve for the endogenous quantities $\{u_{it}^s, E_{it}, w_{it}^s, Y_{it}^s, L_{it}^s, MA_{it}^s, LMA_{it}^s\}$.

1. $u_{it}^s = A_{it}^s \left(\left(\frac{w_{it}^s}{P_{it}^s} \right)^{1-\beta} E_{it}^\beta \right)^{\lambda_s}$
2. $E_{it}^\beta = \left(\frac{w_{it}^E}{w_{it}^N} \right)^\beta$
3. $w_{it}^s = \frac{Y_{it}^s}{L_{it}^s}$
4. $Y_{it}^s = B_{it}^s (w_{it}^s)^{-\phi_k} MA_{it}^s$
5. $L_{it}^s = u_{it}^s LMA_{it}^s$
6. $MA_{it}^s = (P_{it}^s)^{-\phi_s} = \sum_j \tau_{ijt}^{-\phi_s} \frac{Y_{jt}^s}{MA_{jt}^s}$
7. $LMA_{it}^s = \sum_j \kappa_{ijt}^{-\lambda_s} \frac{L_{jt}^s}{LMA_{jt}^s}$

Appendix section 2.D.1 shows that these equations can be simplified into a series of simultaneous non-linear equations given by equation 2.9. Note that this again recovers the sufficient statistic result, all endogenous quantities can be written in terms of market access variables which themselves are recursively defined depending on some kernel.

$$MA_{it}^r = \sum_j K_{ijt}^r \prod_{h=1}^4 \left(MA_{jt}^h \right)^{b_{rh}} \quad (2.9)$$

Where MA_{it}^r denotes market access of type $r \in \{1, 2, 3, 4\}$ such that $MA_{it}^1 = MA_{it}^E$, $MA_{it}^2 = MA_{it}^N$, $MA_{it}^3 = LMA_{it}^E$, $MA_{it}^4 = LMA_{it}^N$. K_{ijt}^r denotes the kernel associated with market access of type r and is a bundle of exogenous shifters and iceberg costs, for example $K_{ijt}^1 = \tau_{ijt}^{-\phi_E} (B_{jt}^E)^{a_{11}} (B_{jt}^N)^{a_{12}} (A_{jt}^E)^{a_{13}} (A_{jt}^N)^{a_{14}}$. The scalars $\{a_{rh}\}$ and $\{b_{rh}\}$ are known functions of structural parameters.

Equation 2.9 is of the exact form studied by Allen, Arkolakis, and Li [2020a]. Thus, following Allen, Arkolakis, and Li [2020a] existence of equilibrium is guaranteed, but uniqueness depends on the spectral radius of the matrix $B = (b_{rh})$. As stated in Allen, Arkolakis,

and Li [2020a] theorem 1 if the spectral radius of B corresponding to the system in 2.9 is less than 1, uniqueness is guaranteed. For the parameters I estimate I indeed find a spectral radius less than one implying a unique solution.

In order to use equation 2.9 to study counterfactual road networks, I need to find values of $\{b_{rh}\}_{r,h=1,2,3,4}$ as well as overcome the problem that exogenous location specific shifters A_{it}, B_{it} , are not observed. Turning first to the latter issue, I solve the model in changes using exact-hat algebra (Dekle, Eaton, and Kortum [2008]), here the exogenous shifters drop out and therefore do not need to be estimated. Denote by a hat variables written in changes e.g. $\hat{x} = x'/x$, where x' indicates the counterfactual and x the observed values. In this paper, I'm interested in counterfactuals in terms of alternative road networks, thus denote by $\hat{\rho}_{ijt}^r$ the change in iceberg trade costs for market access of type r associated with some counterfactual road network. This change is known given the parameterization discussed and estimated in section 2.2. Then we can write the system given in equation 2.9 (see appendix section 2.D.1 for a formal derivation) in changes where the only unknown objects are the parameters $\{b_{rh}\}_{r,h=1,2,3,4}$.

$$\widehat{MA}_{it}^r = \sum_j \hat{\rho}_{ijt}^r \lambda_{ijt}^r \prod_{h=1}^4 \left(\widehat{MA}_{jt}^h \right)^{b_{rh}} \quad (2.10)$$

λ_{ijt}^r is the proportion of i 's market access (of type r) in t , which is due to location j , and is a known quantity which due to the assume gravity relationships can be directly recovered from observed migration and trade flows. Given a counterfactual road network and therefore $\hat{\tau}_{ijt}$ and $\hat{\kappa}_{ijt}$, and parameter estimates, I can then solve this non-linear system of equations to find $\widehat{MA}_{it}$ (vector of market access terms) and then recover the change in opportunity in each location due to the counterfactual network, using the estimated coefficients (stack to form the vector $\hat{\gamma}$) from the sufficient statistic relationship $\Delta\mu_l = \hat{\gamma} \ln(\widehat{MA}_l)$.

2.4.1 Finding parameter estimates

I first set β (the utility value of education relative to consumption) to equal to 0.2²⁶, although quantitative results presented in section 2.4.2 are not sensitive to other reasonable values of β as shown in appendix section 2.D.6. There are four remaining parameters, $\phi_E, \phi_N, \lambda_E, \lambda_N$. In appendix 2.D.1 I show that each coefficient from the sufficient statistic result can be written as a function of these parameters $\gamma_i = f_i(\phi_E, \phi_N, \lambda_E, \lambda_N)$ for $i = 1, 2, 3, 4$. Thus given information on $\hat{\gamma}_i$ and knowledge of f_i I can numerically solve for the structural parameters.

²⁶Eckert et al. [2021] estimate a similar parameter and find $\beta = 0.39$ using this value doesn't change my results.

Finally, I also show in appendix 2.D.1 that $b_{rh} = g_{rh}(\phi_E, \phi_N, \lambda_E, \lambda_N)$ and so can use estimates of the structural parameters to back out the b_{rh} terms in equation 2.10. This process is attractive, as it doesn't require any further identification strategies beyond those already employed, and provides a strong link between theory and the empirical results.

This procedure gives $\phi_E = 2.86$, $\phi_N = 1.79$, $\lambda_E = 1.44$, $\lambda_N = 0.89$. ϕ_s can be interpreted as the negative of the elasticity of wages to factory gate prices operating through trade sensitivity, and λ_s as the negative of the elasticity of wages to locality utility, operating through migration sensitivity. Therefore, these results suggest that wages are more sensitive to price changes for *E*-type goods and that migration is more sensitive to local utility for *E*-type workers.

For robustness, I also consider a second set of parameter estimates taken as from the existing literature. These parameters have not been estimated in Benin, Cameroon, or Mali before, or anywhere in Sub-Saharan Africa to the authors' best knowledge, so the applicability of such parameters should be taken with a pinch of salt. To estimate movement costs I turn to Tsivanidis [2019] who estimates migration-iceberg cost elasticities in the setting of commuting within Bogota and distinguishes between high and low-educated workers. To make this more conformative with my cross-city setting I scale estimates such that the average across types is equal to that found by Morten and Oliveira [2021] who consider cross-city migration in Brazil. This approach leaves me with the following estimates: $\lambda_E = 1.74$, $\lambda_N = 2.11$. It's worth stressing that Bogota and Brazil are far from Benin, Cameroon, and Mali, and there is no particular reason to think that these estimates will pass over unadjusted. Turning to trade costs estimates of the elasticity of within-country trade with respect to iceberg trade costs in a developing country context, by worker type, are even rarer. I take the estimates from Zárate [2020] who studies this elasticity within Mexico city distinguishing between formal and informal workers, and once again rescale by Morten and Oliveira [2021]. Using this approach I find $\phi_E = 3.47$, $\phi_N = 4.52$. Neither of these settings are particularly close to Benin, Cameroon, or Mali, but nevertheless it is reassuring that when I use this alternative set of parameter values to calculate counterfactuals I find similar results.

2.4.2 The impact of road building since 1970 on the distribution of opportunity and spatial inequality of opportunity

I use the full model described in section 2.4 to estimate the counterfactual vector of market access terms in 2020 in the absence of any changes to the road network since 1970. Figure 2.4.1 shows the distribution over localities in each country of the total relative effect of roads

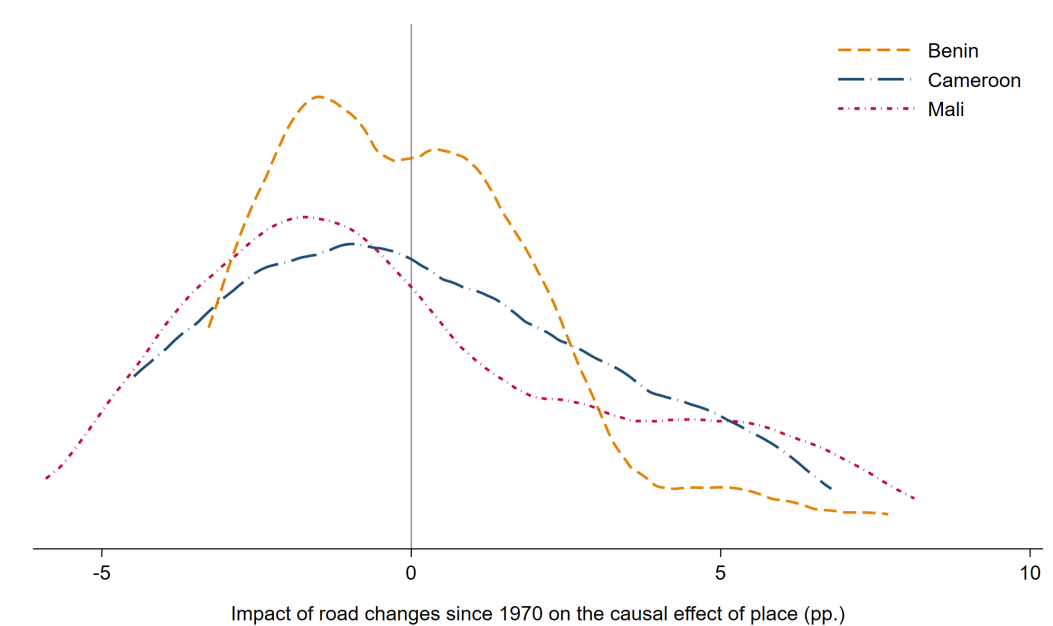
built since 1970 on the causal effect of place on the probability of completing primary school, given by $\Delta\mu_l = \hat{\gamma}(\ln(MA_{lt}) - \ln(MA_{lt}^{NoRoads}))$. Where MA stacks the market access terms, $\hat{\gamma}$ is a vector of estimated coefficients from section 2.2, and the log-operator is taken element-wise.²⁷

Figure 2.4.1 shows that changes to the road network since 1970 have increased the causal effect on primary education completion of growing up in the most affected areas (90th percentile) by 7.54pp. more than the least affected (10th percentile). Although the magnitude of relative changes is large, it is not unreasonable given the large decreases in travel times displayed in figure 2.3.2 which shows that on average travel times fell by over 37% during this time period. This average however, hides considerable heterogeneity across countries: the corresponding 90th-10th percentage difference in Benin is only 4.72pp., whereas in Cameroon and Mali it is 8.93pp., and 8.67pp. respectively.

Using these estimates I can study how changes in connectivity since 1970 directly impacted inequality of opportunity, measured as the change in the variance of local opportunity across space. Here, I also find significant differences across countries. In Benin the variance of opportunity over space was largely unaffected by road building since 1970 decreasing only by 0.04%, that is in the absence of any changes to the transport network since 1970 inequality of opportunity would have been 0.04% higher in Benin. However in Cameroon road building since 1970 *increased* the variance of opportunity over space by 5.81% and in Mali decreased the variance of opportunity over space by 1.44%.

²⁷In this paper I focus on the effects of roads on the distribution of opportunity and inequality of opportunity, rather than on levels shifts. For this reason I don't discuss mean shifts in opportunity in the main body of the paper, and relegate such analysis to appendix section 2.B.11.

FIGURE 2.4.1 The distributional impact of roads built since 1970 on the causal effect of place on primary education completion

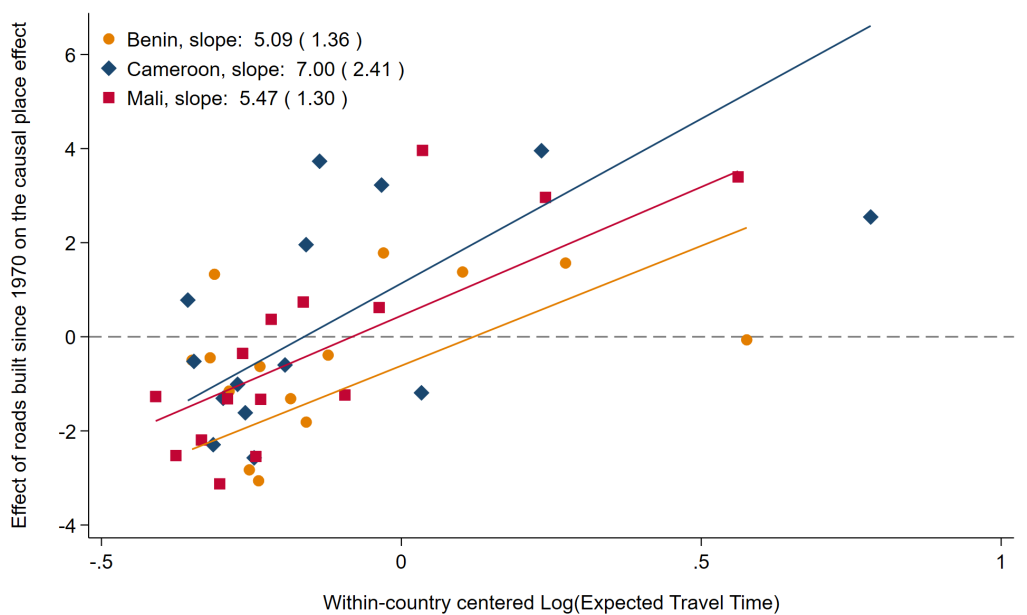


Notes: This figure shows the distribution of the total relative effects of road building from 1970 to 2020, on the causal effect of place on completing primary education in each location in each of Benin, Cameroon, and Mali.

Figure 2.4.1 shows considerable heterogeneity both across roads within countries, but also across countries that is across road networks. Turning first to understanding within-country cross-road heterogeneity, figure 2.4.2 shows the relationship between the total estimated effect and the 1970 expected travel time of each location (normalized by overall country size) where expected travel time is defined as on average how long one should expect to be on the road to travel to another random person within the same country. In all three countries we see a significant and positive relationship²⁸ overall 1970 remoteness explains 33% of the within-variation in the change in opportunity due to road building since 1970. Within-country, those locations that were initially more remote, saw larger increases in local educational opportunity as compared to less remote locations. In Benin for example a one percent increase in 1970 remoteness is associated with on average a 5% high place effect for roads built since 1970. Actual road building since 1970 has disproportionately benefited more remote locations in each country.

²⁸In figure 2.4.2 I omit departments in the Extrême-Nord province of Cameroon as they are significant outliers. With this departments included the Cameroonian slope remains positive but is diminished and no longer statistically significantly different from 0.

FIGURE 2.4.2 Relationship between the total relative effect of roads and 1970 expected travel time



Notes: This figure shows the correlation between the overall relative effect of roads on the causal effect of place on a localities remoteness in 1970. Remoteness is measured as expected travel time relative to the country average and the effect of roads built since 1970 is estimated using the structural model. The relationship is allowed to vary by country, results are weighted by 1970 locality population. Corresponding linear regression coefficients and robust standard errors are reported for each country in the top left of the figure. Departments in the Extrême-Nord province of Cameroon have been omitted as they are significant outliers.

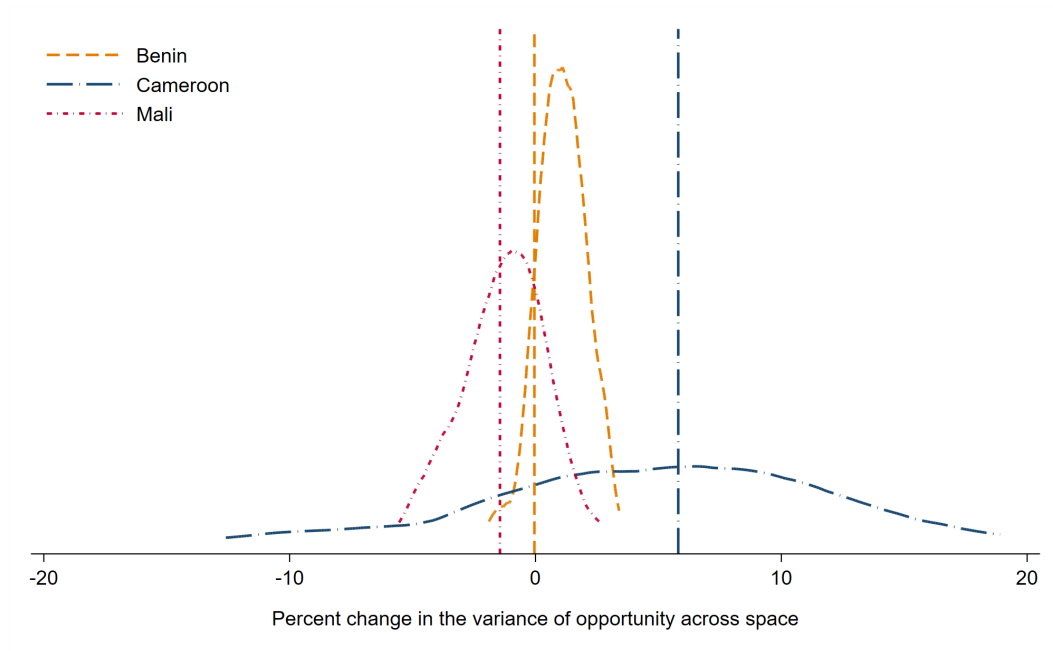
Second, I turn to investigating cross-country heterogeneity. But studying three countries I am uniquely able to investigate the extent to which network-level characteristics are important, a key ingredient in consider the external validity of any single-country study. In general, there is no particular reason to suppose that road-building in each of these countries since 1970 would have had similar impacts. Each country is unique in its geography, spatial distribution of economic activity, and existing 1970 network. In addition, each country built and upgraded its network in a different manner, potentially following different policy objectives. An advantage of my approach is that it allows effects to vary on such dimensions, and thus the study of what may be causing differences. One natural question to ask is to what extent the effect of road building since 1970 on inequality is the result of prudent policy, or reflect differences in immutable characteristics facing policymakers in each country.

To understand the role country (or network) specific initial conditions may be playing, I remove the effect of policy-makers decisions by asking what the impact on inequality of opportunity would have been had the road network (of the same size) instead been built

randomly in each country. The intuition is that by considering random networks, I remove the impact that policymakers themselves had, and am left with that due to constraints. Figure 2.4.3 shows the distribution of the change in inequality of opportunity (measured in terms of % change in variance) across localities from 250 such random counterfactual networks in each Benin, Cameroon, and Mali. Superimposed on this graph are horizontal lines showing the observed actual effect of roads built in each country on inequality of opportunity. Figure 2.4.3 shows that policy makers in each country faced environments that differed considerably in dimensions that mattered for determining the impact of road building on inequality. In Benin, effects are tightly clustered around a slight increase in inequality whereas in Mali effects are clustered around a slight decrease. Cameroon on the other hand has a dispersed distribution with some random networks decreasing inequality of opportunity by 10% whereas others increasing inequality of opportunity by 10%. Policy makers achieved outcomes close to the center of each distribution of random effects — implying that constraints facing policy makers are an important determinant of outcomes, particularly in Benin and Mali.

The finding that the preexisting road network and distribution of economic activity can limit the scope for discretion by policy makers when looking to upgrade the network speaks to the literature on persistence and path dependence in urban structure. Various authors have shown evidence in a developed country setting that path dependence plays a large role: Davis and Weinstein [2002, 2008], Bleakley and Lin [2012], Hornbeck and Keniston [2017], Allen and Donaldson [2020]. This finding contributes to the smaller literature showing similar evidence in a developing country setting [Miguel and Roland, 2011, Jedwab and Moradi, 2016, Bertazzini, 2022], by giving further evidence to suggest that path dependence plays an important role in determining the spatial distribution of economic activity in Benin, Cameroon, and Mali.

FIGURE 2.4.3 Comparing the effect of counterfactual random networks to the realized network



Notes: This figure compares, for each of Benin, Cameroon, and Mali, the impact on inequality of opportunity measured as the percent change in variance across space of roads built since 1970 on the causal effect of place on the probability of completing primary school (vertical lines) to the distribution of average effects over 250 random simulated networks. Random simulated networks have the same overall decrease in travel time as the observed change in networks from 1970 to 2020, but road upgrades have been randomly decided.

Finally, the structural model encodes three main features — I allow road building to decrease trade costs, decrease movement costs, and for location-specific utility, and so migration decisions be influenced by the returns to education. It's natural to consider which channel drives the distributional results. In sub-section 2.B.12 in the appendix, I shut down each feature in turn and re-run the counterfactual analysis. The main conclusion is that each channel plays an important role in determining the effects of road changes on spatial inequality of opportunity. Ignoring any one of them materially changes conclusions. The stresses the importance of considering both direct and general equilibrium effects of road building, highlighting the necessity of a framework which allows for such interactions.

2.4.3 The impact of future road investments on spatial inequality of opportunity

Section 2.2 showed that road building alters the spatial distribution of opportunity and section 2.4.2 showed that since 1970 changes to the transport network have had large, but heterogeneous, effects on the spatial distribution of local educational opportunity. Within this context, it's natural to ask the policy relevant question: what is the impact of future

road investments on spatial inequality of opportunity? The approach taken in this paper is uniquely suited to answer this question as I can allow every road to effect every locality differently, and for this impact to depend on the entire pre-existing road network. Additionally, as I study three countries, I can consider the importance of network-level characteristics in determining the effect of any given road — an important dimension when considering the applicability of country-specific results to an alternative setting.

To investigate the impact of future road investments I simulate the impact on the spatial distribution of local educational opportunity of upgrading each existing segment of the road network²⁹ to a *highway* (speed of 80km/h relative to the existing fastest roads which have travel speeds of 60km/h). Unlike some recently studied place-based policies such as place-based taxation [Gaubert et al. \[2021\]](#), or the large literature on opportunity-zones (see for example [Freedman et al. \[2021\]](#)), road building represents a public good provision problem with no place-blind alternative.

By considering the set of potential road upgrades I can build a road-locality level data set calculating the impact on each localities relative place effect of each road upgrade. In this manner I can study the aggregate impact on inequality of opportunity for each road upgrade, the characteristics of roads which lead to larger decreases in inequality, the characteristics of places which are related to greater increases in opportunity for a given road, and how the interaction between road and locality level characteristics shapes how the spatial distribution of opportunity changes with changes in connectivity over space. Uniquely, as I study multiple countries, I can additionally consider the importance of network-level characteristics. By looking at the set of all counterfactuals I don't have to rely on estimating the impact of changes in connectivity due to the selected-sample of actually built roads which suffers from endogeneity concerns, that the all-road sample sidesteps.

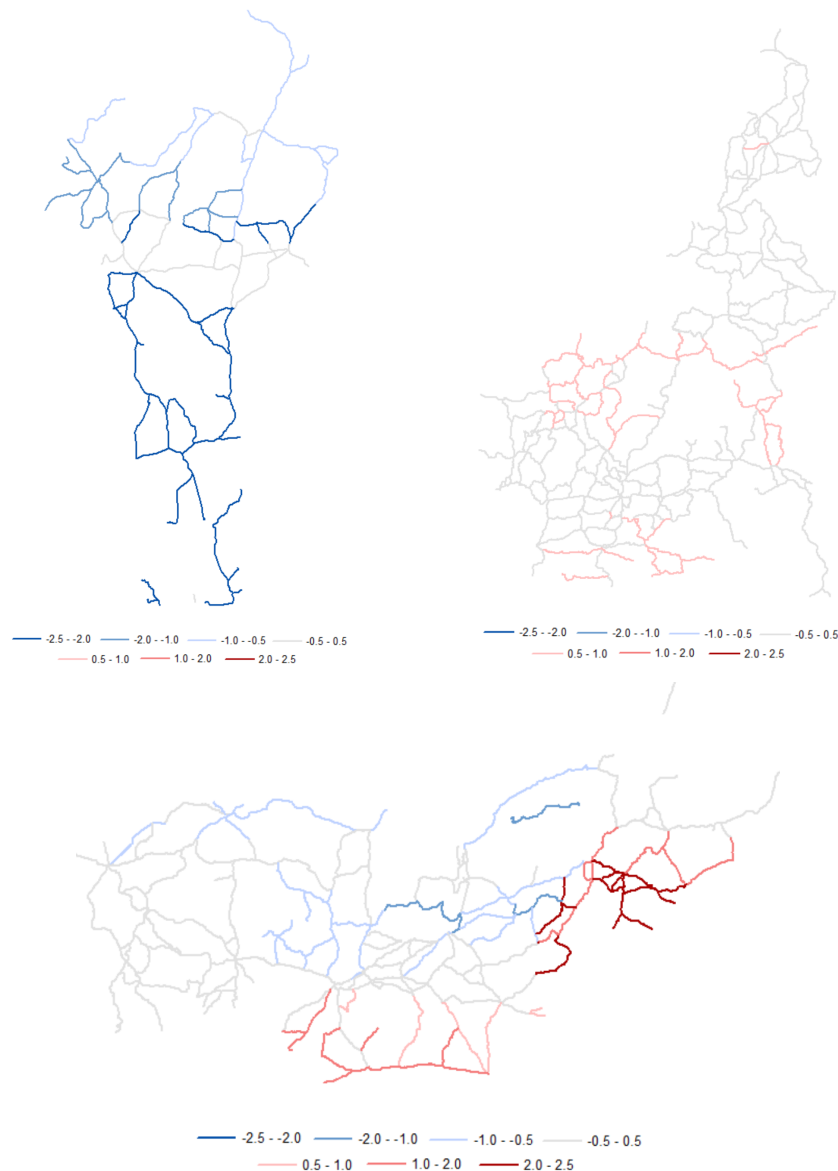
This approach can also be used to approximate in a computationally feasible manner the optimal network problem as discussed in [Fajgelbaum and Schaal \[2020\]](#), and is similar to the procedure taken in [Balboni et al. \[2022\]](#). That is, although I don't solve over the entire space of potential new road locations, I do consider the finite set of road segment upgrades — which given the majority of variation in the past 20 years has been in upgrading rather than building — is a particularly policy-relevant margin, whilst remaining computationally

²⁹A road segment is defined as part of a road that starts and ends at a settlement or cross-roads. Although road segments will have varying lengths and initial conditions, and therefore associated upgrade-costs, I do not take this into account when considering welfare. In this manner, and due to the substitutability of roads, this exercise should be considered as indicative of the area in which upgrading has a given effect, rather than pertaining to which specific road segment is most effective.

feasible. In this manner, instead of solving for the optimal, inequality-minimizing improvement, I can upgrade each road-segment in turn, and perform the counterfactual analysis. That is, if the road with the largest gains is built it will not necessarily be the case that the road with the reported second largest gains should be built next as one would need to re-run the analysis starting from the new current network. Indeed, many roads are likely to be substitutes, leading to an anticipated geographic clustering of *good* roads which will change once one has been built.

Figure 2.4.4 shows how the spatial inequality of opportunity changes in each counterfactual scenario. Where a counterfactual is upgrading each of the 554 road segments in turn. Bluer lines in these figures correspond to larger decreases in inequality of opportunity and redder lines to larger increases. Changes are measured in terms of standard deviations. Figure 2.4.4 shows considerable cross-country differences. In Benin, almost every road decreases inequality of opportunity, whereas in Cameroon almost every road (mildly) increases inequality and in Mali the picture is more nuanced with roads to the south-east causing large increases in inequality but central roads decreasing inequality. Country level differences indicate that it's not just road-type, or locality characteristics that matter for what impact changes in connectivity will be on inequality of opportunity — but rather that properties of the network as a whole play a key role.

FIGURE 2.4.4 Counterfactual impact on inequality of opportunity of upgrading each road segment



Notes: This figure shows the counterfactual impact of upgrading each road segment, keeping the remainder of the network fixed at 2019 levels, to have travel speed of 80km/h on the spatial inequality of opportunity measured as the change in the standard deviation of opportunity over space. Bluer lines represent larger decreases in inequality whereas redder lines represent larger increases in inequality.

This paper presents moving opportunity to people through road building as an alternative to the previously studied policy of moving people to opportunity. In order to benchmark my results, I perform a back-of-the-envelope calculation to estimate how many people would have to be moved to achieve similar reductions in inequality of opportunity to those found in figure 2.4.4. To do this I reshuffle the population of each country, moving people uniformly from the lowest 10% opportunity locations and reassigning them evenly to the

highest 10% of locations. This is a policy similar in targeting to the Moving To Opportunity experiment [Chetty et al., 2016]. I assume that population movements don't change the underlying affect of locations, and calculate the change in population-weighted inequality over space due to moving the population in this way. Many roads in figure 2.4.4 decrease the standard deviation of opportunity over space by 1%, so I take this as the target change for such a reshuffling of the population to achieve. In order to reduce overall inequality of opportunity following the program set out above one would need to move 71,000, 219,000, and 466,000 people from the lowest opportunity areas to the highest in Benin, Cameroon, and Mali respectively. This is 13%, 12%, and 44% of the population from the lowest 10% of opportunity areas. These numbers are very large, and under any reasonable estimate of movement costs are considerably in excess of the cost of building a road which Buys et al. [2006] estimate using data from the World Bank to be roughly 12.8m USD for a 100km road.

Although figure 2.4.4 highlights cross-country differences, it also displays considerable heterogeneity across roads within-country. To investigate this further I categorize roads into three types: those connecting periphery areas to each other (periphery), those connecting periphery cities to the main city (main) and those doing neither (other). Figure 2.D.2 in appendix 2.D.4 shows how each road is categorized. A road is categorized as connecting periphery cities if it doesn't enter any locality surrounding the main city. A road is categorized as connecting main to periphery cities if it enters a locality close to the main city. By grouping roads together in this fashion I can shed light on a common policy assertion that to bring opportunity to flagging areas one should better connect it to existing vibrant locations, such as capital cities vs the alternative that one should encourage areas to flourish independently of the capital.

Table 2.4.1 shows the results from estimating equations at the road r level of the following form: $\Delta SD(\mu_l)_r = \beta_t \text{RoadType}_t + \varepsilon_r$, where $\Delta SD(\mu_l)_r$ is the change in the standard deviation of opportunity over space due to upgrading road r and is my measure of spatial inequality of opportunity, and t denotes road type. The results in table 2.4.1 are given relative to the *other* road category. Therefore β_t can be interpreted as the overall effect of building a road of a given type relative to the baseline category on spatial inequality of opportunity. In column (1) I pool results across countries and include country fixed effects. Whereas building a road to the main city doesn't increase inequality relative to building a non-categorized road, building a road that connects two periphery cities increases the standard deviation of spatial inequality of opportunity by 0.3 on average. In columns (2), (3),

and (4) I restrict the sample to each country individually. In Cameroon, I find no differential impact by road type. In Benin and Mali, on the other hand I find that building roads connecting periphery cities relatively increases the standard deviation of opportunity over space by 0.95 and 0.42 respectively.

TABLE 2.4.1 Impact of different types of road on spatial inequality of opportunity

	(1)	(2)	(3)	(4)
	Overall	Benin	Cameroon	Mali
Main	0.145*	-0.0194	-0.0492	0.247
	(0.0831)	(0.0816)	(0.0598)	(0.166)
Periphery	0.310***	0.954***	0.00485	0.421***
	(0.0643)	(0.101)	(0.0435)	(0.126)
Observations	534	94	260	180
R^2	0.500	0.147	0.003	0.041

Notes: This table estimates the road-level impact of future road upgrades on inequality of opportunity measured as the standard deviation of local educational opportunity over space. Coefficients are from estimating the following equation: $\Delta SD(\mu_l)_r = \beta_l \text{RoadType}_r + \varepsilon_r$ and are relative to the left-out category *other*. A positive coefficient means that relative to the left out category upgrading roads of that type increased inequality of opportunity over space. Column one pools across countries and includes country fixed effects whereas columns (2), (3), and (4) restrict the sample to Benin, Cameroon, and Mali respectively. Standard errors are robust and reported in parenthesis below point estimates.

The intuition behind the result presented in table 2.4.1 is simple. A periphery location stands to gain more by trading with a relatively richer main location than vice-versa. The first order effects of connecting any two locations is to double down on their current trading and migration patterns — as periphery locations gain more from core locations than vice-versa, this affect acts as to improve their outcomes by more. This channel is shown theoretically and confirmed empirically in appendix section 2.B.13.³⁰

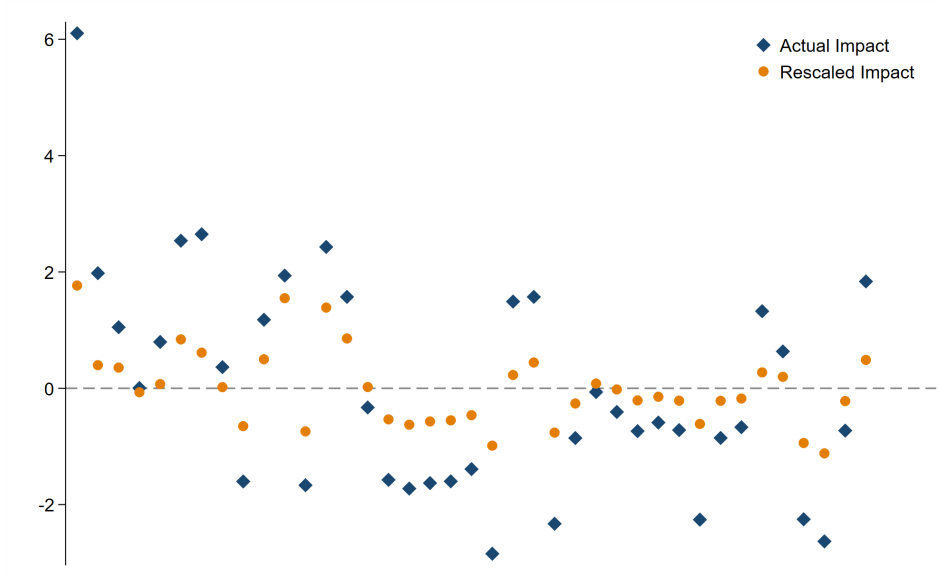
³⁰This result does not contradict the findings of Faber [2014]. In this paper, I compare all possible periphery-core connections to periphery-periphery ones focusing on the aggregate country-wide impact. On the other hand, Faber [2014] compares outcomes in incidentally connected periphery locations relative to non-connected ones in a significantly more industrialized setting, focusing on roads that by definition also connect two main locations. A key additional difference may be that in China educated workers are employed in manufacturing with increasing returns to scale, whereas due to the lack of structural transformation in many African locations this is less likely to be the case in my setting [Jedwab, 2013, Gollin et al., 2016, Jedwab et al., 2022]. However, in

Within-country cross-road-type heterogeneity is important, but figure 2.4.4 also highlights cross-country, within-road, differences. To help understand what might be driving these differences, I can alter network-level characteristics and re-run counterfactuals. One of the first-order differences between Benin, Cameroon, and Mali is their scale. In 2019 in Benin the average travel time from any given two locations is 369 minuets whereas in Cameroon it is 698 minuets and in Mali 1176 minuets. To understand what impact such stark differences in network scale might have I re-run each counterfactual in Benin and Cameroon altering their networks such that expected travel time is equalized to that in Mali. That is, I add a quantity S_c to each i, j connection such that $\mathbb{E}_c[t_{ij} + S_c] = \mathbb{E}_{\text{Mali}}[t_{ij}]$ where $\mathbb{E}_c(t_{ij})$ denotes an expectation over i and j within country c of the average pairwise travel time t_{ij} . Intuitively one expects that by expanding the network in this manner the impact of any given road will be muted, this is because it will be harder for locations further away to utilize this upgrade.

Figure 2.4.5 shows the locality-level effects of a randomly chosen individual road upgrade in Benin. In blue I show impacts on the actual network and in orange impacts on the re-scaled network. In all cases the re-scaled impacts are a muted version of the actual impacts attenuating effects towards 0. Intuitively as localities are effectively further away from eachother upgrading any given road is less effective as the rest of the network remains prohibitively costly to trade with or migration to/from.

appendix 2.B.15 I show that by focusing on the same type of variation that Faber [2014] uses I can replicate his main results in my setting.

FIGURE 2.4.5 Comparing counterfactual effects in the actual and re-scaled network for a random road upgrade



Notes: This figure compares the counterfactual relative effect of upgrading a random road in Benin or Cameroon on local educational opportunity in the actual network and in the re-scaled network. Each dot represents a locality, blue dots correspond to the effect in the actual network and orange dots the effect in the re-scaled network. The road network is re-scaled to have the same average expected travel time between any two locations as in Mali.

The impact of changing the scale of the network is perhaps most clearly seen in table 2.B.9 in appendix 2.B.14 which shows the average impact on spatial inequality of opportunity measured as the variance of opportunity over localities over all road upgrades in each country. In column one I give the average based on the actual network, and in column two the same average on the re-scaled network for Benin and Cameroon. As anticipated, rescaling the network in such a manner significantly attenuates the impact of roads.

Finally, table 2.B.10 in appendix 2.B.14 shows the impact of re-scaling on the road-type level effects reported in table 2.4.1. Overall the impact of different types of road upgrade on spatial inequality of opportunity relative to the *other* category is muted with the re-scaled networks, as expected. In sum, these results show that as well as road-level characteristics being of importance, network-level characteristics matter and are an important consideration when considering the external validity of results from a specific setting.

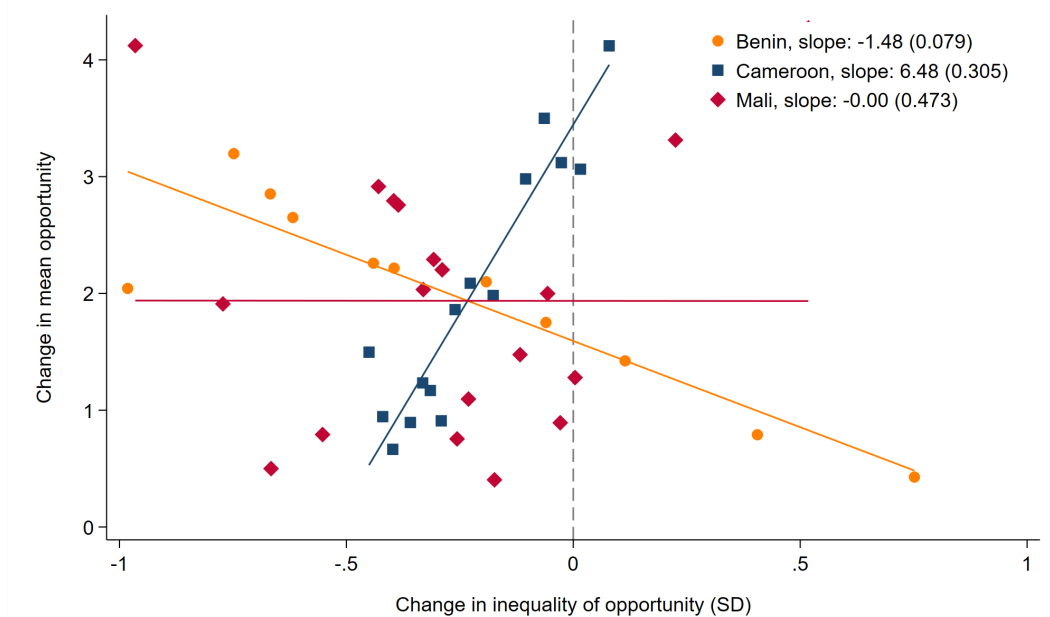
These results highlight that country (network) as well as road-type level heterogeneity matters. These exercises however, don't shed light on how different locations are differentially effected by different roads in appendix 2.D.4 I consider how effects vary by locality and road characteristics.

Efficiency-equity trade off

In this paper I focus on inequality of opportunity. However, when considering where to build or upgrade roads, policy makers are also likely to give weight to the mean effects of any such project. In this section I consider the relationship between the impact of any given road upgrade on the mean and variance of local educational opportunity across space.

Figure 2.4.6 shows the road-level relationship between the impact of upgrading a given road on inequality of opportunity (x-axis) and mean shifts in opportunity (y-axis). Results are separated by country. For Benin results are in orange with circle markers, for Cameroon in blue with square markers, and for Mali in red with diamond markers. There are considerable differences across countries. In Benin there is a significant negative relationship — areas where roads increase overall opportunity by the most see the largest decreases in inequality of opportunity. Thus, in Benin, both margins move in the same direction (assuming policy makers prefer reductions in inequality of opportunity). However, conditioning on a given mean shift, there is still considerable variation in affects on inequality which may influence road-project decisions. In Cameroon however, the story is reversed. Roads which cause the largest increases in mean opportunity also cause the largest increases in inequality of opportunity. Thus, in Cameroon policy makers face a real trade off and the relative weights given to mean shifts vs inequality reductions will have a large impact on which roads are deemed to have the greatest overall benefit. Finally, in Mali, there is almost no correlation between mean shifts and changes in inequality.

FIGURE 2.4.6 Equity-Efficiency trade off



Notes: This figure shows the binscatter relationship between the impact upgrading roads has on across country mean-shifts in opportunity and changes in inequality of opportunity, measured in standard deviations. In orange with circle markers the relationship for Benin is plotted. In blue with square markers the relationship for Cameroon is plotted, and in Red with diamond markets the relationship for Mali is plotted. Associated slope coefficient and (in brackets) robust standard errors are given in the legend.

The bottom line of this analysis is that by considering the impact on inequality of opportunity policy makers are likely to come to different decisions when evaluating which projects to embark upon. Policy makers may give weight to equality of opportunity concerns for normative reasons citing equity and fairness concerns, or to placate their citizens who may identify with such motives as recently found in the US [Gaubert et al. \[2021\]](#). Indeed, figure 2.B.16 in the appendix shows that as policy makers put more weight on equity relative to efficiency they on average value periphery-periphery connections more relative to core-periphery connections.

2.5 Conclusion

We know from previous work that large observed within-country spatial inequality betrays inequalities of opportunity. This paper studies how connectivity of space shapes this geographical distribution of opportunity, and therefore how policymakers can affect spatial inequality of opportunity through road building, in the setting of Benin, Cameroon, and Mali.

To study the impact of road building I develop an approach to measure the effect of any given road on locations across the entire network. This is challenging because not only does it matter what a given road connects, but roads will impact outcomes in all locations across the network. To overcome these challenges, but remain as general as possible, I turn to theory and develop a sufficient statistic approach that is consistent with a broad class of data-generating processes. This result endogenizes skill premia across localities in a many-location setting with costly movement of goods and individuals over space, two sectors/types, and education choice. It states that labor and goods market access terms capture all the potentially complex effects of roads on local opportunity.

The sufficient statistic result suggests an expression that can be directly taken to the data but requires an identification strategy to overcome the endogeneity of road placement which market access terms inherit. Existing strategies cannot be used in my setting, instead, I develop a novel instrumental variables approach, building on the “far away variation” strategy, using not-on-least-cost-path variation. Using this approach I take the sufficient statistic result to the data and find that changes in the road network do indeed influence the spatial distribution of opportunity. To go further and answer counterfactual questions, I write down a structural spatial general equilibrium model from the class consistent with the sufficient statistic result. Solving the model in changes and parameterizing it using the estimated coefficients from the sufficient statistic result I then use it to answer policy-relevant questions.

First, I find that road building in Benin, Cameroon, and Mali since 1970 significantly affected the spatial distribution of opportunity. Road investments raised the probability of primary school completion by 7.5p.p. more in the most-affected locations relative to the least-affected ones. I find that locations that were more remote in 1970 saw larger gains. Cross-country differences are also stark and using randomly generated networks I show that they can partly be explained by the initial constraints facing policymakers in each country. Turning to consider the impact of possible future roads, I calculate the resulting change in inequality of opportunity due to upgrading each of 570 roads. There is considerable heterogeneity across-roads within-country — roads that connect two periphery locations are more likely to increase inequality than those that connect a main and a periphery location. Intuitively this is due to the first-order effect of connecting any two locations being to strengthen their existing ties, and the demand for *E*-type goods predominantly comes from main, not periphery, locations. By considering the mean shifts in opportunity as well as changes in inequality of opportunity due to upgrading each road segment I highlight an equity-efficiency

trade-off — the bottom line for policy is that if a social planner values equity this will change where roads should be built in the future.

As I study three countries, I can consider cross-network variation in counterfactuals, and find that network-level variation plays an important role. By counterfactually changing the size of the network in Benin and Cameroon, to equal that of Mali, I show that the magnitude of effects can partly be explained by network-level density. This opens the door to a potentially exciting area of future research considering how local and global network-level characteristics determine the effect of any given change in connectivity over space — or indeed the impact of any place-based policy.

Appendix

2.A Estimating local educational opportunity [Heath Milsom, 2021]

Heath Milsom [2021] estimates the causal effect of place on primary education completing following the methodology of Chetty and Hendren [2018b], and various other authors. Employing a movers design Heath Milsom [2021] estimates the following specification.

$$y_i = \alpha_{odt} + \mu_{lt} \cdot e_{ilt} + \varepsilon_i \quad (2.11)$$

Where $l \in \mathcal{L}$ is a location, $o(i) \in \mathcal{L}$ is the birth location of individual i and $d(i) \in \mathcal{L}$ is i 's destination location, $t(i)$ is the period of i 's childhood. α_{odt} are origin by destination by period fixed effects and ε_i is an idiosyncratic error term. Finally, e_{ilt} is a variable equal to the years of childhood (1 to 13) i spends in location l in period t and is given by equation 1.4.

$$e_{ilt} = \begin{cases} 13 - m_i & \text{if } l = d(i), t = t(i) \\ m_i & \text{if } l = o(i), t = t(i) \\ 0 & \text{otherwise} \end{cases} \quad (2.12)$$

Where, y_i is an indicator variable taking the value one if individual i has completed primary school, and m_i is the year i moved. This equation is estimated on a sample of 14-18 year old one-time movers³¹ from Benin, Cameroon, and Mali using the census data described in the main text. This approach uncovers the causal effect of spending an additional year of childhood in a given location on the probability of completing primary education, μ_{lt} , relative to each country-year average as the equation is in practice estimated separately for each country-year census sample. These estimates can then be used to decompose observed primary completion rates, denoted by $\bar{y}_{lt}$, in each locality-year into the variation due to causal place effects, and that due to different characteristics of individuals across space denoted by $\bar{\theta}_{lt}$, where $\bar{y}_{lt} = \gamma_{ct} + 13\mu_{lt} + \bar{\theta}_{lt}$. Where γ_{ct} are country-year fixed effects.

³¹13% of 14-18 year olds in my sample can be classified as one-time movers.

Intuitively this approach considers all individuals who move from a given origin o to a given destination d and compares the outcomes of those who move earlier relative to later. If it's the case that those who moved from o to d earlier have better outcomes (more likely to complete primary education) I conclude that $\mu_d > \mu_o$. I then combine information from all such comparisons to estimate each place's causal effect. By including origin-destination-time fixed effects, α_{odt} , I only use variation in the timing of moves rather than comparing the outcomes of families that move to/ from different areas.

To interpret μ_{lt} as the causal effect of spending an additional year of childhood in a given location on the probability of completing primary school in that location, I make three assumptions. First, the above estimating equation implicitly assumes that place effects are linear, i.e. that spending an additional year of ones childhood in a given location has the same effect on outcomes irrespective of which specific year. Evidence in favor of this linearity assumption is presented in [Heath Milsom \[2021\]](#). Second, the above design estimates μ_{lt} using movers only. The interpretation above supposes that this can be extended to all children growing up in a location, including those who don't move. This assumption is not strictly necessary, one could rephrase and discuss the effects on movers only, who constitute a non-negligible 19% of the total population³². However, [Heath Milsom \[2021\]](#) provides evidence to suggest that μ_{lt} is also informative of stayers place effects, first by using only pre-move variation in locality quality and secondly by considering increasingly likely to be exogenous movers. Lastly, the above makes a causal claim, which requires the formal identifying assumption that $\mathbb{E}[e_{il} \cdot \varepsilon_i | \alpha_{odt}] = 0 \quad \forall l \in \mathbb{L}$. Intuitively, this is satisfied if selection effects do not systematically vary with the age at move in each location-pair-period cell. This is the classic movers design assumption used in [Chetty and Hendren \[2018b\]](#) as well as many other authors. [Heath Milsom \[2021\]](#) goes to some length to show that in this setting there is also evidence to suggest this assumption holds, by for example only considering within-household variation, or using a placebo test on 14-18 year old movers (who have mainly completed primary school). For details of all specification tests and a detailed investigation of assumptions see [Heath Milsom \[2021\]](#).

³²In addition to 13% of 14-18 year olds having moved once, 6% have moved multiple times.

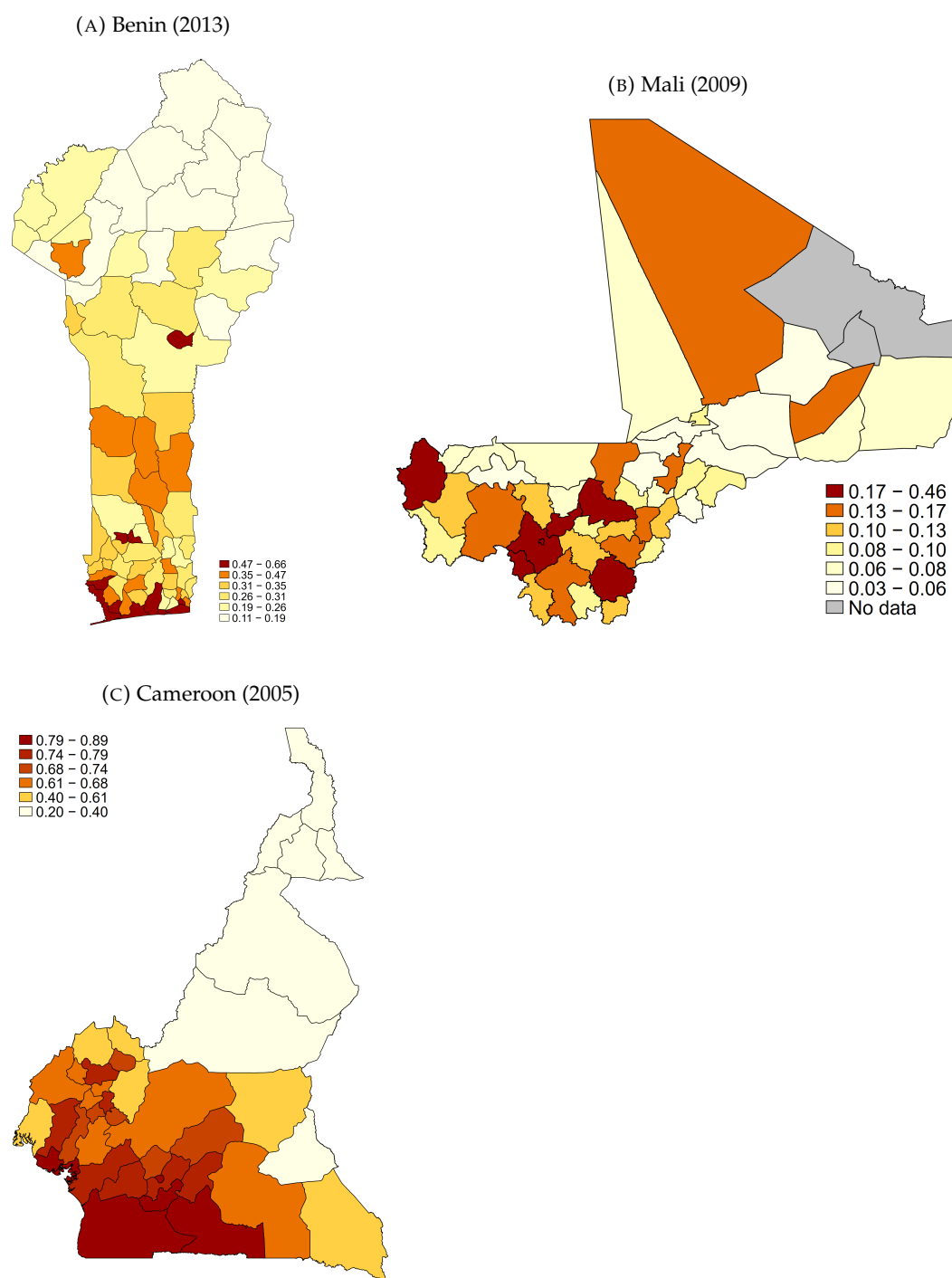
2.B Empirics appendix

2.B.1 Descriptive statistics

Spatial variation in primary completion

There is substantial variation in primary education completion within country across localities as can be seen by figures [2.B.1a](#), [2.B.1b](#) and [2.B.1c](#). In Benin in 2013 the proportion of individuals who had completed primary education in an area varied from 11% in the north to as high as 66% on the coast close to the capital. In addition, Parakou, a large city in the centre of the country had high completion rates. Mali, has consistently lower primary completion rates as compared to Benin or Cameroon. Figure [2.B.1b](#) also displays significant cross-locality variation with some areas completion rates as low as 3% and some, especially round the capital, closer to 50%. Cameroon shows a similar pattern, the most educated areas are around the capital, or close to the large coastal city of Douala. Cameroon also has the highest completion rates over all varying from 20% to almost 90%.

FIGURE 2.B.1 Locality level primary completion rates

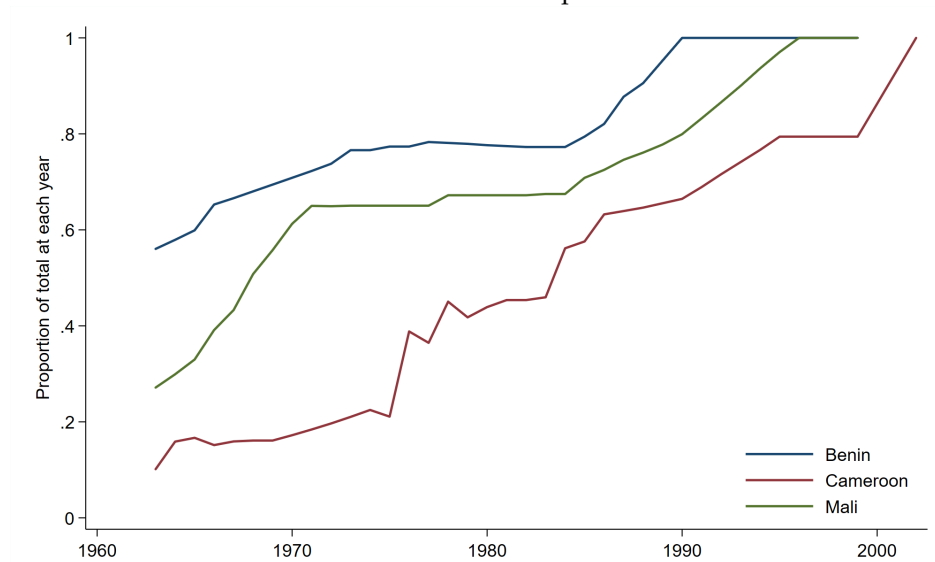


Notes: this figure shows the spatial distribution of primary education completion rates for all those above the age of 12 in each of Benin, Cameroon and Mali. Each figure has its own scale and corresponding legend where darker orange/red indicates higher completion rates. The data for Benin comes from the 2013 census, for Mali the 2009 census and for Cameroon the 2005 census.

Secondary source showing variation in road building

Figure 2.B.2 uses data from [Canning and Pedroni \[2008\]](#) to calculate the proportion of the completed paved road network existing in a given year over my study period. From these figures it's clear that, unlike railways in Benin, Cameroon, and Mali, roads were mainly a post-colonial technology displaying significant variation even in the recent past. In figure 2.B.2 it's clear that Cameroon has seen the most intensive increase in road stock since 1960 when it had less than 20% of the length it does today. Mali and Benin, however, are not too far behind with less than 30% and less than 60% respectively of their modern road stock in place by 1960.

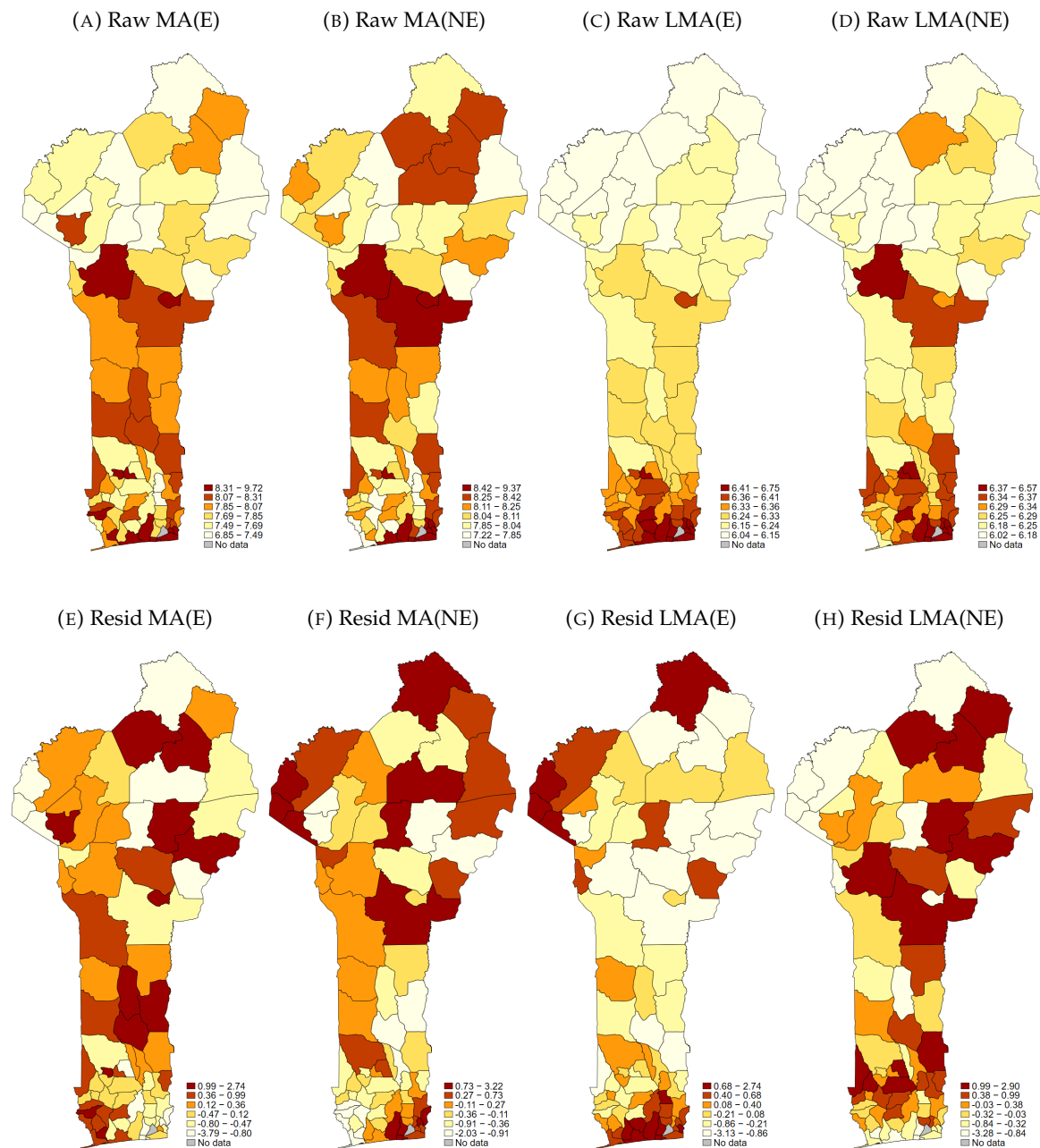
FIGURE 2.B.2 Variation in paved roads



Notes: This figure shows the proportion of the 2000 total paved road stock in place in each given year for Benin, Cameroon, and Mali. It uses data from [Canning and Pedroni \[2008\]](#).

2.B.2 Maps of the spatial distribution in market access terms

FIGURE 2.B.3 Market Access — Benin 2012



Notes: This figure shows spatial variation in the calculated market access terms in Benin in 2013. In the top row I map raw values. In the bottom row I residualise each market access variable on the other three variables and map the standardised residuals. The first column shows E-type goods market access. The second column shows NE-type goods market access. The third E-type labor market access and finally the fourth NE-type labor market access. Maps showing the corresponding figures for Cameroon or Mali, or in different years, are available upon request.

2.B.3 Spatial inequality in local returns to education and opportunity

Skill premia vary locally if relative skill demand is heterogeneous across space and within-country migration is costly. In appendix section 2.C.1 I show that within-country migration

costs in this setting are indeed high mirroring results found in the literature and emphasizing how migration costs, in this low formal barrier setting, reflect more than just the pecuniary costs of moving [Gollin et al., 2021, Bailey et al., 2018]. Figure 2.B.4 then shows considerable spatial variation in returns to education over birth locations. This is also unsurprising given the extensive literature which has documented and leveraged similar differences over space [Eckert et al., 2021, Chetty and Hendren, 2018b, Adukia et al., 2020, Atkin, 2016, Hsiao, 2022]. Figure 2.B.4 doesn't show returns to education in terms of income as I do not observe wages directly for each census year in each locality and so instead study returns to education using two proxy variables which capture returns to education in terms of housing quality and returns to education in terms of the probability of not being employed in agriculture. The housing quality variable is the standardized first principle component of a PCA analysis over the floor, wall, and roofing material, access to electricity, and sanitation. In this setting, with high informality rates and subsistence agriculture, defining returns to education in terms of wages makes less sense, and indeed these two proxies may better capture the notion of returns to education.

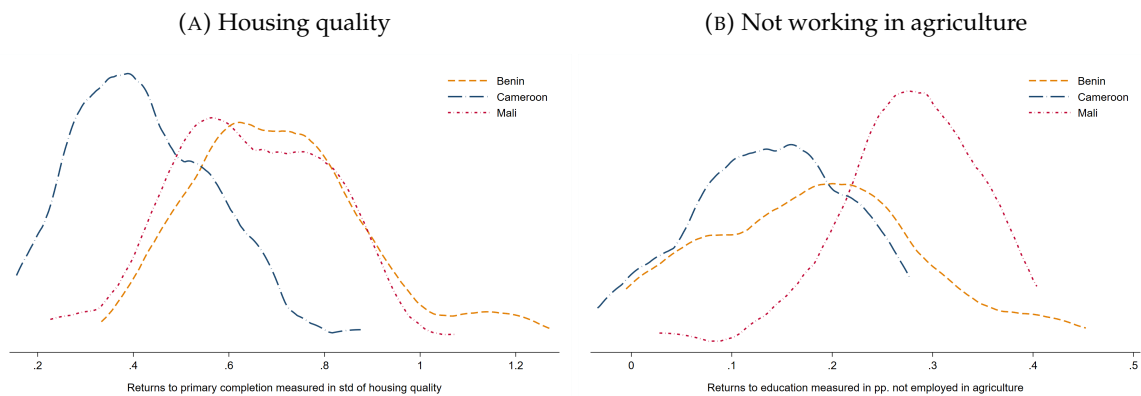
Using either proxy I run Mincerian-esque regressions for each locality³³ of the form given in equation 2.13 where y_ω denotes either housing quality or whether individual ω works in agriculture. Equation 2.13 is estimated on a sample of those between 25 and 55 separately for each location i , in the most recent year data is available. This results in a set of estimates $\{\hat{\beta}_i^y\}$ for each proxy y and each location i , which are then plotted in figure 2.B.4.

$$y_\omega = \beta_i^y \cdot \text{Primary}_\omega + \beta_{1i} \cdot \text{age}_\omega + \beta_{2i} \cdot \text{age}_\omega^2 + \varepsilon_\omega \quad (2.13)$$

Figure 2.B.4 shows considerable variation in returns: in some birth locations those who complete primary education have over 1SD better housing quality, whereas in others quality of housing is very similar to those who don't complete primary school. In addition in appendix 2.C.1 I show that moving across space is costly, lending credence to the non-equalization of returns.

³³Although wages may be imperfect in their ability to capture returns to education in this setting, it's natural to consider what correlation my proxies do have to what we can observe on wages. Census data doesn't have information on incomes, but for a similar time period in Benin and Mali I do have limited information on wages from the demographic and health surveys. In appendix 2.B.4 I show that in this limited sample observed wages are strongly correlated with my proxies at both the individual and regional level.

FIGURE 2.B.4 Spatial variation in the local benefits of education over birth location



Notes: This figure shows the distribution over birth locations of returns to primary education in each country. Returns are measured in the most recent period available, 2013 in Benin, 2005 in Cameroon, and 2009 in Mali. Panel 2.B.4a measures returns to education as the Mincerian return to education in terms of housing quality. Panel 2.B.4b measures returns to education as the Mincerian return to education in terms of the probability of not being employed in agriculture. In each case Mincerian returns, β_l^y are calculated using the following regression $y_i = \beta_l^y \text{Primary}_i + \beta_{1l} \text{age}_i + \beta_{2l} \text{age}_i^2 + \epsilon_i$ for each locality l separately and for y equal to housing quality or a dummy variable equaling one if not employed in agriculture. Housing quality is calculated as the first principal component in a PCA analysis of floor, wall, roof material, access to electricity, and sanitation. The distribution of the recovered β_l^y are then plotted.

The spatial distribution of local educational opportunity

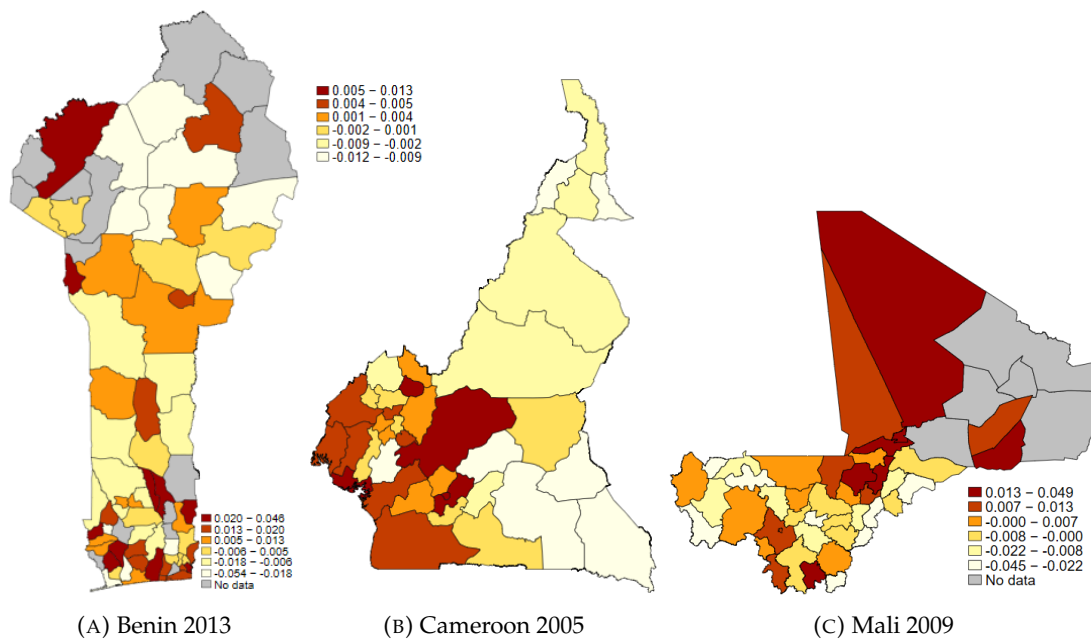
The causal effect of place on primary education completion is distinct from locality level information on observed primary completion rates as the latter conflates two forces, causal place effects and the differing characteristics of individuals over space. Heath Milsom [2021] disentangles these two forces and estimates causal place effects using a movers design following Chetty and Hendren [2018a] and others. Intuitively, this compares children who move from, and to, the same locations but at different ages and uses variation in exposure to either location to back out the effect of place. If the child who moved earlier has better outcomes this is taken as evidence to suggest that the location moved to exerts a higher causal effect than that moved from. Combining all such comparisons I can estimate the causal effect of growing up in each location on the probability of completing primary education in Benin, Cameroon, and Mali in each census year. This method relies on the identifying assumption that selection effects do not vary with the age at move. Evidence for this is provided in Heath Milsom [2021] by considering cross-sibling effects, increasingly likely to be exogenous move events, and placebo tests.

Using causal place effects as the outcome variable changes the interpretation of regressions: a statistically significant positive effect is interpreted as increasing the causal effect

growing up in a location has on primary completion. If instead my outcome variable was observed primary completion rates, a statistically significant positive effect would rather be said to increase observed completion rates in a location, which could conflate changes in the causal effect of place with changes in demographics or the selection of households into said location. Details of the estimation procedure, and the assumptions under which it is valid can be found in appendix 2.A.

Figure 2.B.5 shows the spatial variation in local educational opportunity as recovered in Heath Milsom [2021]. These figures show considerable variation: moving to a one standard deviation better location at birth increases a child's probability of completing primary school by 15 percentage points.

FIGURE 2.B.5 Estimates of local educational opportunity from Heath Milsom [2021]



Notes: These maps show estimates from Heath Milsom [2021] of the spatial distribution of local educational opportunity. Darker colors indicate areas of higher opportunity. Numerical values can be interpreted as the causal effect (relative to the country's mean) of spending an additional year of childhood in a given location on the probability of completing primary school.

2.B.4 Correlating wage proxies with available wage data

In the main text I use two variables to proxy for skill premia as wages are neither available in my data, nor is it clear that in this setting with significant informality and subsistence agriculture, that they are appropriate. Nevertheless one would expect these variables to bare some relation to wages when I can observe the returns to labor. Using data from all

available Demographic and Health Surveys in Benin, Cameroon, and Mali (Benin 1996, and Mali 1995) I recover yearly income as well as my measures of housing quality and whether an individual is working in agriculture or not.

First, I residualize these variables on age, age squared, sex, and country. I find that the correlation at the individual level between the residualized (log) yearly income and the residualized housing quality variable is 0.37, and between the residualized (log) yearly income and the residualized not working in agriculture variable is 0.43. Moving to the regional level, I find that the analogous correlations are 0.85, and 0.48. Additionally, these variables are strongly correlated with the DHS provided household wealth index. This is particularly true of housing quality, although this correlation is partly mechanical as the wealth index variables uses as part of it's inputs the variables used to construct the housing quality variable

2.B.5 Not-on-least-cost path identification strategy in figures

Figure 2.B.6 explains the novel not-on-least-cost-path identification strategy developed and used in the main text. Sub-figure 2.B.6a shows a stylized network with three locations represented by hexagons and the least-cost-path from each denoted by solid black lines. Through this exposition we shall consider sources of variation from the view point of location one which is colored orange. Sub-figure 2.B.6a shows the entire network without any restrictions on the source of variation used to instrument for location one's market access. Sub-figure 2.B.6b restricts variation to that far-away from location one, depicted by a red circle around location one. Parts of the least-cost-path which are no longer sources of variation for location one are colored red. Thus intuitively one can see that sub-figure 2.B.6b restricts variation to that occurring in sections of the road network far from location one. This sub-figure also betrays one of the main weaknesses with using the far-away variation strategy in isolation, that policy makers may wish to improve connections between location one and other locations which could occur outside of the red circle.

Sub-figure 2.B.6c additionally restricts the variation used to instrument location ones' market access to that which does not lie on the least cost path from location one to any other location. Again, this is depicted by coloring in red sections of the road which are "frozen", that is which we don't use variation in. Note that even in the case represented by figure 2.B.6c, only considering far-away variation introduces restrictions above those introduced by only considering not-on-least-cost-path variation as roads near location one which may not lie on its least cost path to any other location are additionally "frozen". The remaining

variation used to estimate locations one's market access comes from how changes in the road network impact the market access of other locations and therefore indirectly cause location one's market access to vary. Sub-figure 2.B.6d goes one step further by depicting second-order not on least cost path variation, again from location one's perspective. Here, I additionally freeze changes in the transport network when calculating every other locations market access which then in turn feeds into calculating location one's market access. In sub-figure 2.B.6d this second-degree remoteness is depicted from the perspective of location four only by coloring in blue sections of the least-cost-path network which are frozen from location four's perspective. The instrument for location one's market access represented in 2.B.6d therefore uses indirect variation in the indirect variation due to how changes in the road network effect market access terms which indirectly effects location four's market access (and indeed every location other than location one) which indirectly effects location one's market access.

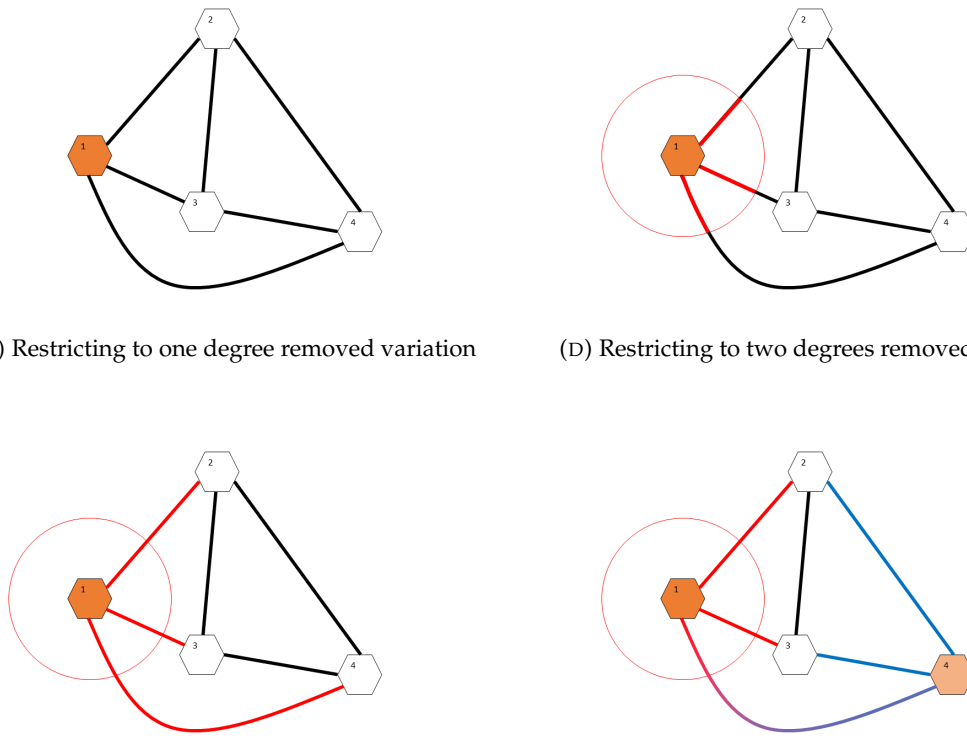
FIGURE 2.B.6 Not-on-least-cost-path identification strategy in figures

(A) Using all variation

(B) Restricting to far-away variation

(C) Restricting to one degree removed variation

(D) Restricting to two degrees removed variation



Notes: This figure graphically explains in a stylized manner the not-on-least-cost-path identification strategy used in the main text. Panel 2.B.6a depicts all variation being used. Panel 2.B.6b visually restricts variation in a circle around the focal location, indicating the use of only *far-away* variation. Panel 2.B.6c additionally restricts variation on the least cost path from the focal location to all other locations. Finally panel 2.B.6d illustrates restricting second order variation from the perspective of location 4.

2.B.6 Local clientelism and public good provision

One potential threat to the identification strategy used in this paper is that the provision of government services and public goods may vary over time and space. That is, if a government comes into power and builds roads and schools so as to benefit a given location in a potentially complex way that is not nullified by the not-on-least-cost-path identification strategy, this could bias the estimated coefficients. In my setting this concern is most manifest when considering the interaction between local ethnic groups and that of the current political leader as discussed in the Kenyan context by Burgess et al. [2015]. However, the situation in Benin, Cameroon, and Mali is very different to that in Kenya. In Cameroon Paul Biya Beti has been in power since 1982, and thus in Cameroon there has been no temporal variation over my study period. In Mali, although there has been considerable variation

in presidents since the 80's ethnic favoritism or clientelism has been found to play only a minor, or perhaps even non-existent role [Dunning and Harrison, 2010, Basedau et al., 2011, Basedau and Stroh, 2012, Franck and Rainer, 2012].

In Benin, however, there is some evidence of politics having an ethnic component and clientelism [Battle and Seely, 2010, Fujiwara and Wantchekon, 2013, Wantchekon, 2003] and some correlational evidence that this may lead to less road building in political marginalised locations Blimpo et al. [2013]. To investigate whether these forces are driving my estimated effects I construct a dummy variable equal to 1 if the ethnic majority in a location is equal to that of the leader of the time in Benin. Using the Geo-referencing of ethnic groups (GREG Weidmann et al. [2010]) database I assign each locality in Benin to one of the four major Beninese ethnic groups³⁴. Over my sample period Benin has had three political leaders in 1992 Nicéphore Soglo (Fon/ Ewe) was in power, in 2002 Mathiew Kérékou (Somba) was in power, and in 2013 Thomas Boni Yayi (Yoruba) was in power. In table 2.B.1 I show the results from my main sufficient statistic analysis in the first column and in the second column replicate this result additionally controlling for the variable described above. The coefficients on the market access terms are stable across columns and the coefficient on the same ethnicity variable is a precisely estimated 0. I take this as evidence to suggest that threats to identification of this nature are minimal.

³⁴Broadly defined as: Ewe, Yoruba, Somba or Barba.

TABLE 2.B.1 Controlling for ethnicity
by leader

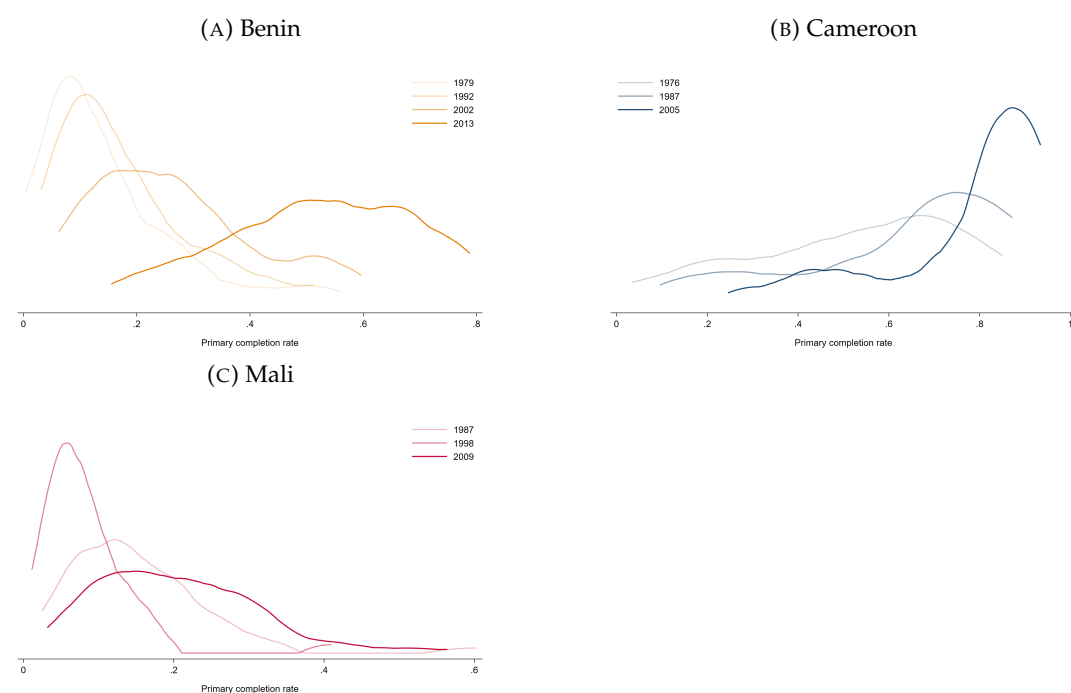
	(1)	(2)
	Baseline	Including ethnicity by leader
Log(LMA Educ)	-0.198** (0.0767)	-0.208*** (0.0780)
Log(MA Educ)	0.118*** (0.0281)	0.114*** (0.0268)
Log(LMA No Educ)	0.284*** (0.0581)	0.277*** (0.0549)
Log(MA No Educ)	-0.171*** (0.0329)	-0.166*** (0.0324)
Same ethnicity as leader		0.00137 (0.00584)
Locality by year FE	X	X
# localities	127	127
N	334	334

Notes: This table compares the baseline results to those including controls for clientelism. Column one replicates the main results from table 2.B.8 column one. Column two includes a dummy variable equal to one if the majority ethnicity of the locality is equal to that of the leader in Benin and zero otherwise.

2.B.7 Top-coding and non-linearities

In this section I provide figures and tables referred to in sub-section 2.3.3 of the main text. Figure 2.B.7 shows how the distribution over localities of primary completion rates for those aged between 15 and 20 has changed in each country over the sample period. These figures show considerable rightward shifts in the distribution, but don't display bunching around 100% — that is they show evidence that top-coding at the upper limit of 100% primary completion is not present. Tables 2.B.2 and 2.B.3 show evidence that changes are not particularly related to levels — exploring the possibility that although top-coding is not an issue, non-linearities may cause similar problems. That is, locations with low primary completion levels in 1970 may also have low market access, and potentially both of these variables are “easier” to increase when low. However, the tables show little evidence that this is a significant issue in this setting.

FIGURE 2.B.7 Changes in the distribution of primary completion rates



Notes: This figure shows the distribution of primary completion rates in each locality in each country in each census year of those between the ages of 15 and 20. Although primary schooling officially ends at 12 in each country over the time period I study, many children only complete in the years following, and thus I take 15 to be when most who will complete, have done so. I cap at 20 in an attempt to capture more recent dynamics, and to remove mechanical correlation across censuses by re-sampling the same individuals.

TABLE 2.B.2 Relation between base level and changes (combined regressions)

	(1)	(2)	(3)	(4)	(5)	(6)
	Long Diff $\text{Log}(LMA_E)$	Long Diff $\text{Log}(LMA_N)$	Long Diff $\text{Log}(MA_E)$	Long Diff $\text{Log}(MA_N)$	Long Diff Prim Educ	Log Diff μ
$\text{Log}(LMA_E)$	0.0546 (0.0935)	0.0895 (0.117)	0.611** (0.291)	0.652*** (0.231)	-0.186* (0.102)	-0.0141 (0.0253)
$\text{Log}(LMA_N)$	0.139 (0.141)	0.0235 (0.173)	0.149 (0.402)	-0.322 (0.312)	0.570*** (0.143)	0.0502 (0.0342)
$\text{Log}(MA_E)$	0.224** (0.0873)	0.230*** (0.0806)	0.306 (0.228)	0.373** (0.180)	-0.258*** (0.0819)	-0.00717 (0.0165)
$\text{Log}(MA_N)$	-0.315* (0.165)	-0.198* (0.116)	-0.503 (0.434)	-0.318 (0.315)	0.0152 (0.115)	-0.0113 (0.0237)
Prim Educ	-0.585*** (0.197)	-0.778*** (0.162)	-1.664*** (0.567)	-1.845*** (0.444)	1.027*** (0.184)	-0.0108 (0.0262)
μ	0.0667 (0.522)	0.109 (0.488)	0.00185 (1.661)	0.145 (1.260)	1.207** (0.588)	1.158*** (0.167)
N	136	136	136	136	136	114

Notes: This table shows the results from running regressions of the long difference of each variable on the initial period level of all other variables. These regressions are weighted by 1970 locality population and include country fixed effects. Standard errors are robust.

TABLE 2.B.3 Relation between base level and changes (individual regressions)

	(1)	(2)	(3)	(4)	(5)	(6)
	Long Diff $\text{Log}(LMA_E)$	Long Diff $\text{Log}(LMA_N)$	Long Diff $\text{Log}(MA_E)$	Long Diff $\text{Log}(MA_N)$	Long Diff Prim Educ	Long Diff μ
$\text{Log}(LMA_E)$	-0.0581 (0.0471)					
$\text{Log}(LMA_N)$		0.199*** (0.0728)				
$\text{Log}(MA_E)$			-0.0741 (0.0510)			
$\text{Log}(MA_N)$				0.0823 (0.0726)		
Prim Educ					0.124* (0.0735)	
μ						0.372* (0.190)
N	162	162	162	162	162	114

Notes: This table shows the results from running regressions of the long difference of each variable on the initial period level of that variable only. These regressions are weighted by 1970 locality population and include country fixed effects. Standard errors are robust.

2.B.8 Endogenous network formation

In this subsection I investigate the possibility that future road building responds endogenously to previous road building — a potential threat to identification. The main empirical specification resulting from the sufficient statistic result can be summarized as $\mu_{it} = \beta MA_{it} + v_i + u_t + \varepsilon_{it}$. Exogeneity of market access term implies that $\text{Cov}(MA_{it}, \varepsilon_{it}) = 0$. This will not be the case if lagged market access enters the equation i.e. $\varepsilon_{it} = MA_{it-1} + \zeta_{it}$, and indeed will remain an issue under my identification strategy if market access instruments display a similar dependence. This is potentially the case if areas that previously had high market access endogenously respond to this my increasing market access further, as maybe the case under endogenous network formation — or if market access terms adjust slowly in response to a long-run project or series of connected projects.

I can test directly for this temporal dependence by including lagged market access terms directly in my main sufficient statistic regression and considering results to the baseline either under OLS or using the efficient 2SLS instrument. Table 2.B.4 shows the results. In columns (1) and (3) I replicate column (1) in table 2.3.1 in the main text and column (1) of table 2.B.8 in the appendix. In column (2) I add lagged market access terms and estimate via OLS and in column (4) I add lagged market access terms and estimate via 2SLS instrumenting lagged terms by the counterpart lagged instruments. In all regressions I include locality and year fixed effects, weight by 1970 population and cluster standard errors at the locality

level. I find that lagged terms are never statistically significant at any reasonable level and that their inclusion doesn't change the main estimated coefficient. From this I conclude that such potential threats to identification are empirically unfounded in this setting.

TABLE 2.B.4 Including lagged market access terms

	(1)	(2)	(3)	(4)
	Baseline	Baseline	2SLS	2SLS
Log(LMA Educ)	-0.0598 (0.0509)	-0.0516 (0.0500)	-0.198** (0.0767)	-0.236* (0.131)
Log(MA Educ)	0.0434** (0.0201)	0.0454** (0.0209)	0.118*** (0.0281)	0.106 (0.0690)
Log(LMA No Educ)	0.151*** (0.0249)	0.263*** (0.0627)	0.284*** (0.0581)	0.324** (0.155)
Log(MA No Educ)	-0.0814*** (0.0212)	-0.110*** (0.0317)	-0.171*** (0.0329)	-0.126 (0.118)
Lag Log(LMA Educ)		0.0349 (0.0790)		0.0409 (0.240)
Lag Log(MA Educ)		0.0122 (0.0173)		0.0671 (0.0509)
Lag Log(LMA No Educ)		-0.0512 (0.0686)		-0.175 (0.168)
Lag Log(MA No Educ)		-0.00490 (0.0229)		-0.0277 (0.0406)
Observations	334	160	334	160

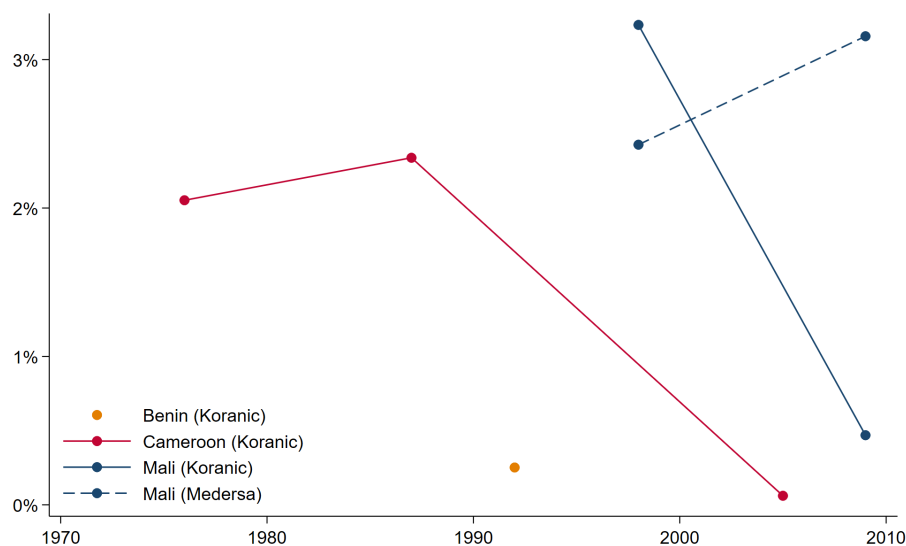
Notes: This table shows the results from estimating the sufficient statistic result described in the main text including lagged values of market access terms. In columns (1) and (3) I replicate column (1) in table 2.3.1 in the main text and column (1) of table 2.B.8 in the appendix. In column (2) I add lagged market access terms and estimate via OLS and in column (4) I add lagged market access terms and estimate via 2SLS instrumenting lagged terms by the counterpart lagged instruments. In all regressions I include locality and year fixed effects, weight by 1970 population and cluster standard errors at the locality level.

2.B.9 Koranic schools and Medersas

Koranic schools are a traditional method of educating which involves memorizing and reciting the Koran. They remain popular in many Muslim counties, and often offer a cheaper or more local method of schooling. In this paper, as well as in [Heath Milsom \[2021\]](#), I don't

count those who have solely had a Koranic education as having completed primary school, inline with the classification used by IPUMSi. Although these schools primarily concern themselves with memorizing and reciting the Koran, it maybe that they provide some opportunities to those who complete a course at them, and therefore this may be an important dimension I am missing from this analysis — or in the case where students switch from Koranic to state-sponsored schooling, I may be overstating the impact. Fortunately, in some of the censuses used I can distinguish between those at a Koranic school, from those at a secular school. Figure 2.B.8 plots the proportion of 6 to 14 year old's in each census where data is available, who are at a Koranic school. It's clear from figure 2.B.8 that Koranic education is in the vast minority (maximum 3% of children) and appears to be declining further. It's likely that many more students attend Koranic schools in the evening or on weekends in addition to attending state school — but this dimension is not covered in the data and is less consequential. Figure 2.B.8 also shows the proportion of students in Mali at a Medersas [Boyle, 2014] which is a religious school in Mali that follows the national curriculum, toeing the line between Koranic schools and state schools. These schools are on the rise, but still constitute a very small proportion of the overall education.

FIGURE 2.B.8 Proportion of children enrolled in Koranic schools or Medersas



Notes: This figure shows the proportion of primary school aged children (6 to 14) who report attending a Koranic school or a Medersa in the Census.

2.B.10 Inference

As discussed in the main text performing inference with regressions of the form given in equation 2.8 is complicated by four factors. First, as I follow locations over time, we expect

considerable auto-correlation within-location. To counter this one can cluster at the locality level, which is the approach taken in the regressions in the main text. However, as well as correlation across time within-location, it is possible that correlation across space within a given time period exists. To investigate the extent to which such correlation may be artificially attenuating standard errors I perform Conley inference [Conley, 1999] using various distance bands and lag lengths³⁵. Table 2.B.5 shows the results from each procedure on the OLS standard errors. In column one I report the coefficients from estimating equation 2.8 by OLS. In column two I report the unadjusted standard errors. In column three standard errors clustered at the locality level are displayed. In column four Conley standard errors are reported using a distance band of 100km and not taking serial correlation into account. In column five I additionally allow for serial correlation up to 30 years. Finally, in column six I increase the distance band to 1000km, effectively allowing high degrees of spatial correlation, and additionally allow unrestricted serial correlation which is equivalent to clustering at the locality level.

The main take away from table 2.B.5 is that even allowing very flexibly for spatial and serial correlation standard errors increase only modestly. Due to limitations of the statistical software I am not able to replicate these results for the combined instrument case, although similar conclusions are found using individual instruments. In general, it seems unlikely given this evidence that inferential conclusions will be overturned.

³⁵Implementation of Conley standard errors is performed using the Stata program *reg2hdspatial* Fetzner [2020], Hsiang [2010].

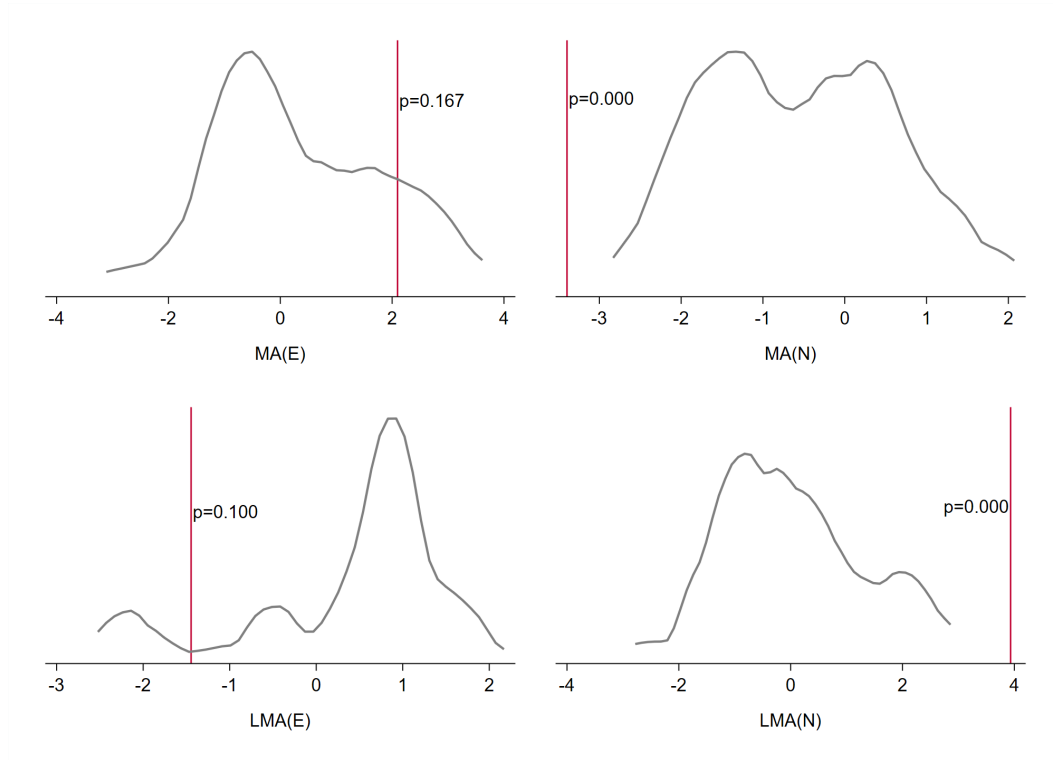
TABLE 2.B.5 Inference on OLS results

	Coefficient	Unadjusted	Clustered	Conley d=100	Conley d=100,l=30	Conley d=1000,l=infinity
Log(LMA Educ)	-0.0819	0.0442	0.0566	0.0513	0.0584	0.0588
Log(MA Educ)	0.0458	0.0163	0.0218	0.0186	0.0205	0.0254
Log(LMA No Educ)	0.1662	0.0450	0.0423	0.0331	0.0383	0.0561
Log(MA No Educ)	-0.0886	0.0214	0.0261	0.0183	0.0217	0.0280

Notes: This table shows the results from performing various inference procedures. In column one I report the coefficients from estimating equation 2.8 by OLS. Column two reports the unadjusted standard errors associated with these coefficients. Column three adjusts standard errors by clustering at the locality level. Column four reports Conley standard errors with a distance cut off of 100km. Column five reports Conley standard errors with a distance cut off of 100km and allowing auto-correlation up to 30 years. Finally, column six reports Conley standard errors with a distance band of 1000km and allowing general autocorrelation.

The third complication facing inference in this setting is due to the nature of the market access variables. As highlighted by [Borusyak and Hull \[2020\]](#) market access terms potentially encode complex dependencies across observations. This is because one shock, for example a road being built, will affect many locations — although these maybe spatially disparate. A natural remedy for inference in this setting is to apply a randomization inference approach, a solution suggested by [Borusyak and Hull \[2020\]](#). To implement this, I use the same data generating process as developed to re-center instruments and purge coefficients of bias resulting from endogenous exposure to plausibly exogenous shocks. At each simulation I estimate the test statistic (on the null of 0) for each market access term coefficient. Figure 2.B.9 plots the resulting distribution of test statistics and with a red vertical line the actual test statistic. This figure shows considerable deviations from normality, suggesting that indeed standard asymptotic inference suffers from complex inter-dependencies. However, the main take away is that again the inferential conclusions given in the main text hold. Due to computational constraints I am only able to show permutation inference results on the OLS coefficients, but this gives evidence to suggest that large deviations are unlikely in the 2SLS results either.

FIGURE 2.B.9 Randomization inference on OLS results



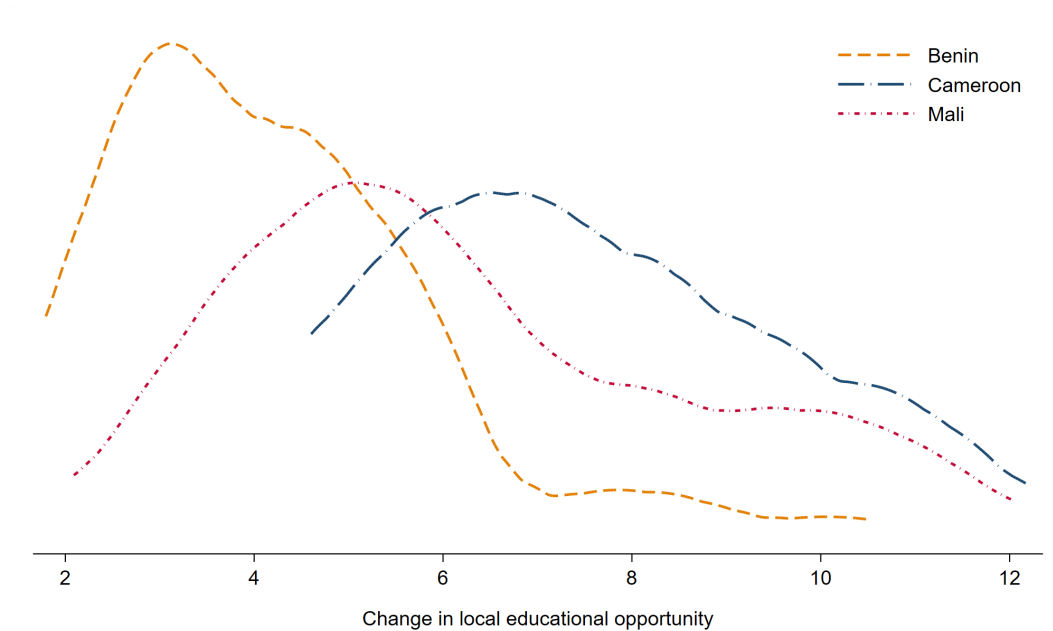
Notes: This figure shows the result from performing randomization inference on the coefficients in equation 2.4 (OLS results). In each simulation I randomly build a set of roads onto the existing road stock in each year, recalculate market access terms, and estimate the corresponding regression with the random MA terms. Reported is the distribution of t-test statistics over each simulation in gray, and the actual estimated t-test statistic with a red vertical line. Corresponding implied p-values are also reported.

The final complication is that the main outcome variable of interest μ_{it} , is itself an estimated quantity from an earlier paper [Heath Milsom \[2021\]](#). As a result of this, it will exhibit measurement error. However, there is little reason to suppose that this measurement error is not classical in nature, and therefore will only attenuate coefficients, thus imposing stricter requirements on rejecting the null of 0 effect.

2.B.11 Mean shifts in the spatial distribution of opportunity

Figure 2.B.10 shows the aggregate counterfactual shifts in the spatial distribution of opportunity due to road building since 1970 in each of Benin, Cameroon, and Mali. The figure shows the distribution of effects over locations in each country. Positive numbers reflect a positive impact of road building, and can be interpreted as the amount by which a location would have *lower* opportunity in the absence of changes to the road network since 1970.

FIGURE 2.B.10 Aggregate distributional shifts due to road building since 1970

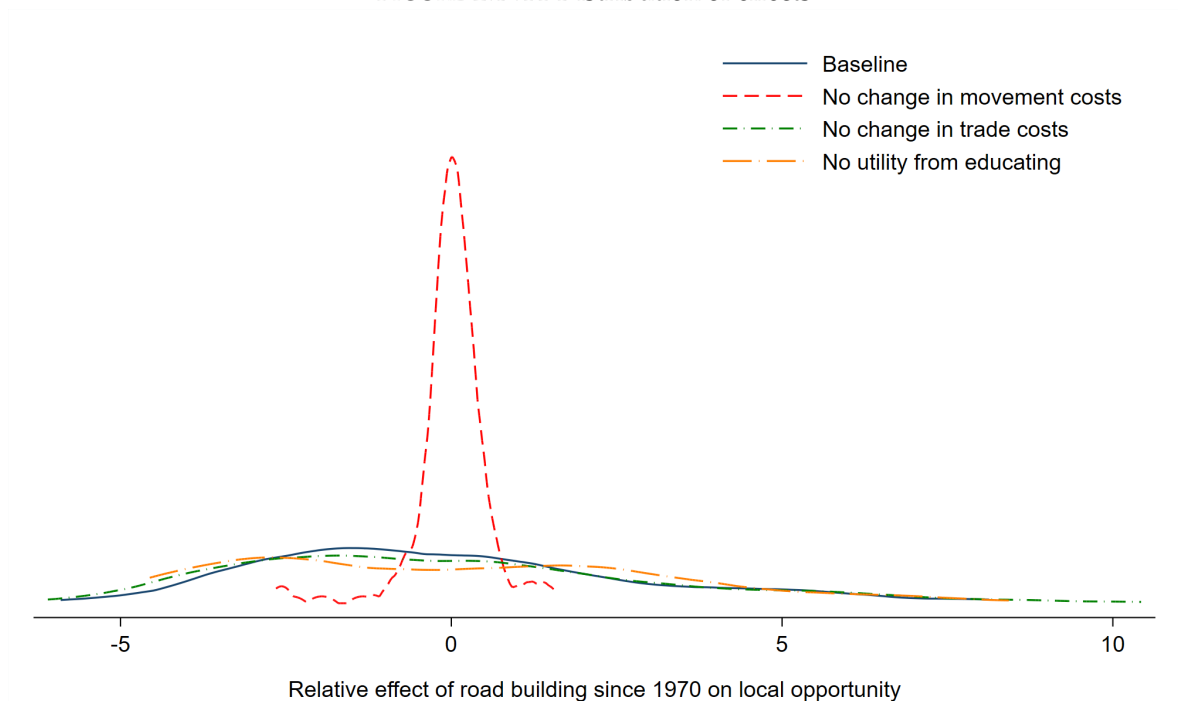


Notes: This figure shows the aggregate effects of changes to the road network since 1970 on spatial inequality of opportunity in each of Benin (yellow dashed), Cameroon (blue dash-dot), and Mali (red small dash-dot), including mean shifts.

2.B.12 Decomposing effects

I consider the same counterfactual exercise “what if no roads had been built since 1970” and in figure 2.B.11 show the distributional effect from shutting down each of the main channels. In blue I replicate the results from figure 2.4.1 averaging over all three countries. In red I don’t allow changes in the road network since 1970 to decrease movement costs. This prevents any direct effects through lower transport costs encouraging migration, but also diminishes the general equilibrium effects by which changes in trade patterns and so relative wages and skill premia induce migration. In green I don’t allow changes in the road network to decrease trade costs. Finally, in orange I set $\beta = 0$, meaning that individuals no longer put any weight in the utility function on their children completing primary school. The stark result from figure 2.B.11 is that impacts are much more concentrated when migration costs are held constant — implying that changes in migration costs are an important driver and amplifier of effects. Figure 2.B.15 in the appendix shows results without the “no change in movement costs” line to highlight remaining differences. In this figure the polarizing affect of shutting down utility from educating is clearer.

FIGURE 2.B.11 Distribution of effects



Notes: This figure shows the (centered) distribution over locations of effects of road building since 1970 on local education opportunity, pooled over Benin, Cameroon, and Mali. In blue (solid line) baseline results are plotted. In red (dashed line) movement costs are no longer allowed to adjust in response to road building since 1970. In green (dash-dot line) changes in goods trade costs are shut down. In orange (long dash dot) the utility value of education is set to 0.

Table 2.B.6 shows the country-specific effect of road building since 1970 on the variance of opportunity over space in each of the versions of the model set out above. Results are reported as percentage deviations from baseline 2020 inequality. In column one I report the baseline for the full model, replicating earlier results. In column two I shut down changes in migration costs. In this version of the model road building since 1970 decreased inequality of opportunity by more modestly more in Benin and significantly more in Cameroon, in Mali however the previous negative impact has been reverse. In column two I fix trade costs and find, relative to the baseline, modest increase in inequality in Benin and Cameroon, but decreases in Mali. Finally, in column four I set $\beta = 0$ and find significantly larger decreases in inequality in Benin and Cameroon, but not in Mali. In sum, these results highlight the importance of including all three channels, as well not extrapolating results from one country to another.

TABLE 2.B.6 Decomposing the impact of changes to the transport network since 1970 on spatial inequality of opportunity

	Baseline	Fixing Migration Costs	Fixing Trade Costs	Setting $\beta = 0$
Benin	-0.04	-0.78	0.29	-2.59
Cameroon	5.81	-2.55	6.37	-15.38
Mali	-1.44	0.13	-1.67	-0.18

Notes: This table shows the impact of road building since 1970 on spatial inequality of opportunity, measured as percentage deviation from the baseline 2020 variance. That is -0.04 means that in the absence of road building since 1970 the variance of opportunity over space would have been 0.04% higher in Benin using the full model. In column two I don't allow movement costs to adjust in response to changes in the road network since 1970. In column three I instead show down changes in trade costs. Finally, in column four I set $\beta = 0$ implying that households no longer derive utility from education choices.

2.B.13 Predicting the impact of connecting any two locations

The main analysis considers how infrastructure projects affect overall measures of spatial inequality of opportunity — although such considerations are important for policy, it is natural to also be concerned with how specific roads affect specific locations. Indeed, one motivation for this project is to study whether the commonly cited policy recommendation of connecting lagging to prosperous locations actually does move opportunity to low-opportunity areas. The approach used in this paper allows every road to affect every location differently, and for these effects to vary with exact structure of the network. This allows me to capture complex effects and provide a rich discussion of how upgrading a specific road influences inequality of opportunity. But, this flexibility comes at the price of complexity — it's difficult to tell without running the full counterfactual what the locality-level impacts of a given infrastructure investment will be. With the aim of providing some intuitive *rules of thumb* I simplify the focus to roads which connect two given locations i and k .

We know from the theory that the locality-specific proportionate change in opportunity due to a given road (r) upgrade is given by $\hat{\mu}_{ir} = \sum_{h=1}^4 \gamma_h \ln \left(\widehat{MA}_{ir}^h \right)$. If we make the first-order approximation, that decreasing travel times between i and k does not alter travel times between any other locations we find that the proportionate change in market access terms

simplifies to $\widehat{MA}_{ir}^h = \hat{\rho}_{ikr}^h \lambda_{ik}^h \left(\prod_{q=1}^4 \left(\widehat{MA}_{kr}^q \right)^{b_{qh}} \right)$ and therefore we can write the proportionate change in local educational opportunity as the following.

$$\hat{\mu}_{ir} = \sum_{h=1}^4 \gamma_h \left(\ln \left(\hat{\rho}_{ikr}^h \right) + \ln \left(\lambda_{ik}^h \right) + \ln \left(\prod_{q=1}^4 \left(\widehat{MA}_{kr}^q \right)^{b_{qh}} \right) \right) \quad (2.14)$$

The first term in this equation translates the change in travel time between i and k into iceberg travel costs for each r and is mechanical. The third term varies at the connecting locality k level, and for each r captures a notion of the change in the size of the available market one is connecting to indeed as shown in appendix 2.D.1 we have that $\prod_{h=1}^4 \left(\widehat{MA}_k^h \right)^{b_{rh}} = \hat{f}_k^r / \widehat{MA}_k^r$ where $J_k^1 = Y_k^1$, $J_k^2 = Y_k^2$, $J_k^3 = L_k^3$, and $J_k^4 = L_k^4$. To calculate this term we have to use the entire structure of the model and run the appropriate counterfactual. Finally, the second term, λ_{ik}^h captures the proportion of i 's market access of type h that is due to locality k , before any new road is built. Due to the underlying gravity framework lambda terms give current trade and migration flows. That is, λ_{ik}^h is observable without the need to run any counterfactual analysis, and is therefore a good candidate for the type of variable that might be of practical use.

To investigate whether $\ln(\lambda_{ik}^h)$ are useful variables for predicting the impact in i on local opportunity of improving the connection from i to k , I consider the following regression.

$$\hat{\mu}_{ir} = \beta_1 \cdot \ln \left(\lambda_{i,k(i,r)}^1 \right) + \beta_2 \cdot \ln \left(\lambda_{i,k(i,r)}^2 \right) + \beta_3 \cdot \ln \left(\lambda_{i,k(i,r)}^3 \right) + \beta_4 \cdot \ln \left(\lambda_{i,k(i,r)}^4 \right) + \varepsilon_{ir} \quad (2.15)$$

Where $k(i, r)$ is the locality connected to i via road r . This regression is run on a sub-sample of road upgrades which have different starting and ending localities. We expect the coefficients β_h to have the same sign as the sufficient statistic coefficients γ_h . Intuitively we expect that better connecting i and k is good for i if i gets more “goods” than “bads” from k . That is, if a large proportion of demand for E -type goods in i comes from k and a large proportion of the supply of N type workers in i comes from k , then improving the connection between i and k is likely to be good for i . Similarly if a large proportion of demand for N -type goods in i comes from k and a large proportion of the supply of E -type workers in i comes from k , then improving the connection between i and k is likely to be bad for i .

Table 2.B.7 shows the results from estimating equation 2.15. Coefficient signs are as expected from the theory, and in general it seems that the impact on local opportunity of better connecting i and k is significantly correlated with $\ln(\lambda_{ik}^h)$. Column one of table 2.B.7 controls for country fixed effects only, column two weights by population, column three

includes regional fixed effects, column four weights by population and includes regional fixed effects and finally column five weights by population, includes regional fixed effects and controls for log expected travel time in 2019. Results are stable across specifications. Taking column five as the baseline, the results suggest that a one percent increase in the proportion of i 's E -type goods market access due to k is associated with a road connecting i and k increasing the causal effect of growing up in i on the probability of completing primary school by 2.7pp. In sum, if i and k become better connected then this is more likely to improve relative local educational opportunity in i if i gets more *good* stuff than *bad* stuff from k .

TABLE 2.B.7 Predicting the Impact of Roads on Local Opportunity

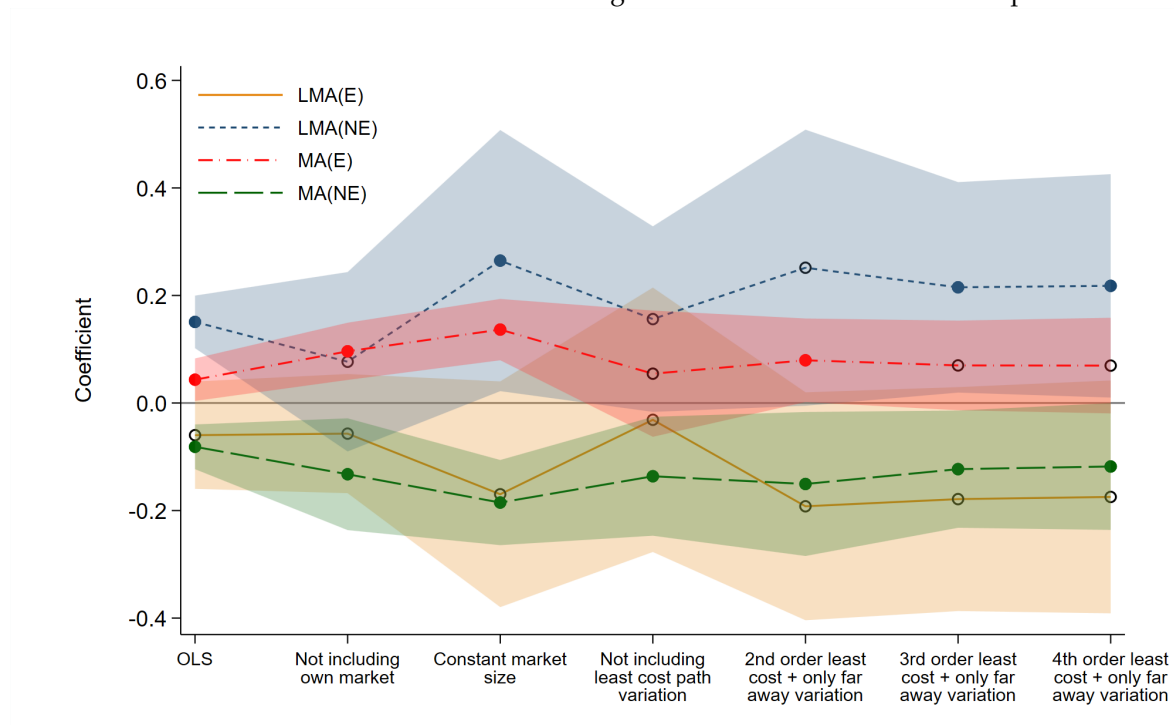
	(1)	(2)	(3)	(4)	(5)
$\ln(\lambda_{ik}^{MA(E)})$	3.175*** (0.914)	1.837* (1.070)	2.858*** (0.928)	2.475** (1.149)	2.771** (1.186)
$\ln(\lambda_{ik}^{MA(N)})$	-2.694** (1.298)	-0.858 (1.657)	-2.321 (1.432)	-2.718 (1.775)	-3.219* (1.811)
$\ln(\lambda_{ik}^{LMA(E)})$	-10.52*** (2.306)	-8.784*** (2.936)	-5.927*** (2.190)	-5.595** (2.662)	-4.507* (2.658)
$\ln(\lambda_{ik}^{LMA(N)})$	4.189* (2.403)	1.947 (2.865)	1.368 (2.959)	3.012 (2.979)	2.581 (2.777)
Log(Exp travel time)					-2.860** (1.311)
Country FE	X	X	X	X	X
Population weighted		X		X	X
Region FE			X	X	X
R^2	0.0881	0.117	0.248	0.232	0.247
N	396	368	368	368	368

Notes: This table shows the results from estimating equations of the form given in 2.15. The left-hand-side variable is the change in local educational opportunity in locality i due to upgrading road r . λ_{ik}^h is the proportion of i 's market access of type h due to locality k . Column one includes country fixed effects only. Column two includes country fixed effects and weights by locality population in 2019. Column three includes country fixed effects and region fixed effects. Column four includes country fixed effects, region fixed effects and weights by population. Column five includes country fixed effects, region fixed effects, weights by population and controls for log expected travel time (remoteness). The regressions are run on a sample of road upgrades which traverse two localities. Standard errors are robust and reported in parentheses.

These results shed some light on the previous result that road which connect main and periphery cities are more likely to decrease inequality than roads that connect two main cities. It likely that a periphery locality gets a higher proportion of its E -type goods trade from a main location, and vice-versa that a main location gets a higher proportion of its N -type goods trade from a given periphery location. Therefore, we would expect opportunity to increase by more in the periphery location relative to the main location — due to the first-order effect documented above that existing ties are strengthened.

2.B.14 Additional results and robustness

FIGURE 2.B.12 Results from estimating the sufficient statistic relationship



Notes: This figure graphically displays results from running regressions of the form as given in 2.8. Each tick on the x-axis corresponds to a different regression either using OLS or instrumenting as described in the tick labels. Dots refer to coefficients and 95% confidence intervals are given by the shaded area. Coefficients significant at the 5% level are colored whereas those that aren't are transparent (with a black border). Confidence intervals are constructed with standard errors clustered at the locality level. Coefficients referring to E-type labor market access are in orange with a solid line. Coefficients referring to NE-type labor market access are in blue with a short-dash line. Coefficients referring to E-type market access are in red with a dot-dash line. Coefficients referring to NE-type market access are in green with a long-dashed line.

TABLE 2.B.8 Results combining instruments
and implementing the [Borusyak and Hull](#)
[2020] correction.

	(1) Baseline	(2) Including average MA variables
Log(LMA Educ)	-0.198** (0.0767)	-0.228*** (0.0844)
Log(MA Educ)	0.118*** (0.0281)	0.136*** (0.0355)
Log(LMA No Educ)	0.284*** (0.0581)	0.334*** (0.0953)
Log(MA No Educ)	-0.171*** (0.0329)	-0.211*** (0.0536)
Locality by year FE	X	X
Kleibergen-Paap stat	13.97	18.56
Kleibergen-Paap p-value	0.235	0.070
SW under ID stat LMA(E)	35.81	26.97
SW under ID stat MA(E)	7.26	9.61
SW under ID stat LMA(NE)	4.39	8.71
SW under ID stat MA(NE)	15.96	9.58
SW weak ID stat LMA(E)	423.72	323.32
SW weak ID stat MA(E)	85.94	115.19
SW weak ID stat LMA(NE)	51.90	104.42
SW weak ID stat MA(NE)	188.91	114.79
# localities	127	127
N	334	334

Notes: This table shows the results from running regressions of the form given in equation 2.8. Column (1) represents my baseline results where the final three sets of instruments are all included (2nd, 3rd, and 4th order not-on-least-cost-path variation). The second column replicates these results but includes expected market access terms calculated from the average over random draws of possible network trajectories, using the method described in section 2.3.3.

TABLE 2.B.9 Impact of network characteristics average impact on inequality

	Actual	Same scale as Mali
Benin	-1.49	-0.38
Cameroon	0.18	0.00
Mali	0.26	

Notes: This table shows the average effect over all possible road upgrades, of up upgrading a road on inequality of opportunity measured as the variance of opportunity across space. Column one shows averages over the actual observed road network for each country, and column two shows averages over the re-scaled network where the network in Benin and Cameroon has been re-scaled such that the expected average time of traveling between any two locations is the same as that in Mali.

TABLE 2.B.10 Impact of different types of road on spatial inequality of opportunity in the re-scaled networks

	(1)	(2)	(3)	(4)
	Overall	Benin	Cameroon	Mali
Primate	0.0828 (0.0690)	0.00368 (0.0159)	-0.0768*** (0.0225)	0.247 (0.166)
Hinterland	0.243*** (0.0557)	0.234*** (0.0222)	0.0698*** (0.0214)	0.421*** (0.126)
Observations	534	94	260	180
R^2	0.160	0.175	0.115	0.041

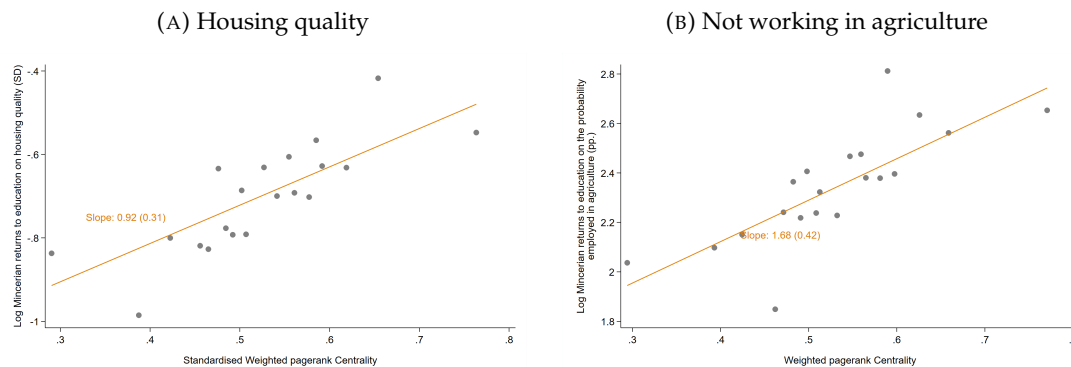
Notes: This table estimates the road-level impact of future road upgrades on inequality of opportunity measured as the standard deviation of local educational opportunity over space. It replicates the results of table 2.4.1 on the re-scaled networks. The networks of Benin and Cameroon have been re-scaled such that average expected travel time across each network is equal to that of Mali. Coefficients are from estimating the following equation: $\Delta SD(\mu_l)_r = \beta_r \text{RoadType}_r + \varepsilon_r$ and are relative to the left-out category *other*. A positive coefficient means that relative to the left out category upgrading roads of that type increased inequality of opportunity over space. Column one pools across countries and includes country fixed effects whereas columns (2), (3), and (4) restrict the sample to Benin, Cameroon, and Mali respectively. Standard errors are robust and reported in parenthesis below point estimates.

TABLE 2.B.11 Sufficient statistic result by country

	Overall	Benin		Cameroon		Mali	
	Coef (O)	Coef (B)	Pval (O-B)	Coef (C)	Pval (O-C)	Coef (M)	Pval (O-M)
LMA E	-0.198 (0.077)	-0.175 (0.153)	0.883	-0.298 (0.160)	0.535	-0.228 (0.117)	0.800
MA E	0.118 (0.028)	0.017 (0.082)	0.222	0.123 (0.023)	0.815	0.043 (0.050)	0.142
LMA N	0.284 (0.058)	0.527 (0.188)	0.203	0.325 (0.102)	0.688	0.388 (0.126)	0.413
MA N	-0.171 (0.033)	-0.069 (0.086)	0.238	-0.186 (0.040)	0.712	-0.213 (0.058)	0.479

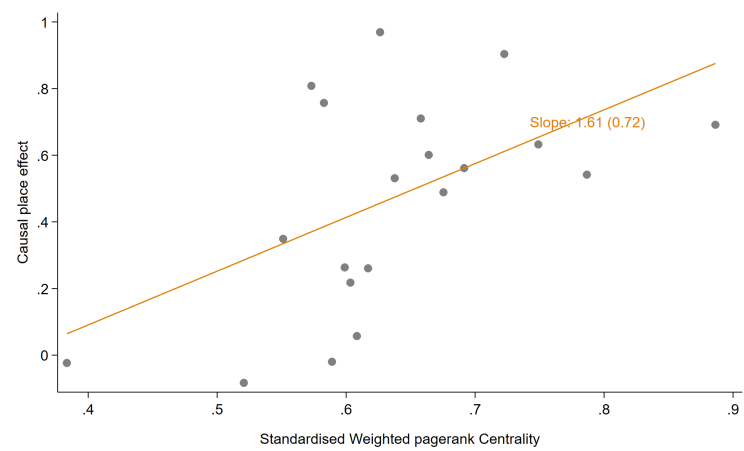
Notes: This table shows the results from re-estimating the sufficient statistic equation on samples from each country individually. The first column replicates the main (pooled) results for comparison. For each country I present a column of results for each market access term, and a column displaying the p-value on the test of equality between the country-specific and pooled coefficients.

FIGURE 2.B.13 Correlation between connectivity and local returns to education: pagerank centrality



Notes: This figure shows in panel 2.3.3a the correlative relationship between the log Mincerian returns to education in terms of housing quality on the y -axis and pagerank centrality on the x -axis. Panel 2.3.3b shows the correlative relationship between the log Mincerian returns to education in terms of the probability of not being employed in agriculture on the y -axis and pagerank centrality on the x -axis. In each case Mincerian returns, β_l are calculated using the following regression $y_i = \beta_1^y \text{Primary}_i + \beta_{1l} \text{age}_i + \beta_{2l} \text{age}_i^2 + \varepsilon_i$ for each locality l separately and for y equal to housing quality or a dummy variation equaling one if not employed in agriculture. Housing quality is calculated as the first principle component in a PCA analysis of floor, wall, roof material, access to electricity, and sanitation. In a second stage, the above binscatter plots are constructed by comparing Cent_l with β_l^y controlling for locality and year fixed effects. Slope coefficients are indicated in orange on the figures and have been calculated from the analogous linear regression. Associated standard errors are given in parenthesis clustering at the locality level.

FIGURE 2.B.14 Correlation between connectivity and local opportunity: pagerank centrality



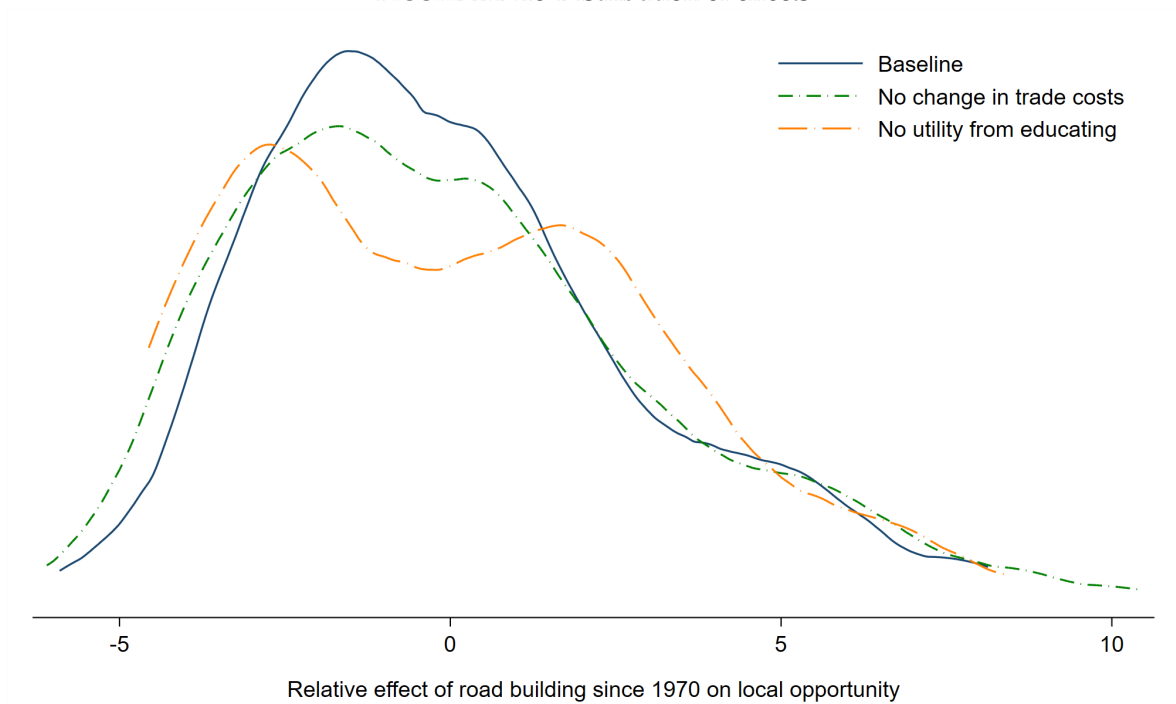
Notes: This figure shows the correlative relationship between locality pagerank centrality on the x-axis and causal place effects on the y-axis. The binscatter plots are constructed by comparing μ_l with $cent_l$ controlling for locality and year fixed effects. The slope coefficient is indicated in orange on the figure and is calculated from the analogous linear regression. Associated standard errors are given in parenthesis clustering at the locality level.

TABLE 2.B.12 Robustness of the main sufficient statistic result

	(1)	(2)	(3)	(4)	(5)
	Baseline	No weighting	Trimmed	Trimmed and no weighting	Weight by inverse SE
Log(LMA Educ)	-0.198*** (0.0656)	-0.424* (0.241)	-0.201*** (0.0540)	-0.333* (0.169)	-0.228 (0.166)
Log(MA Educ)	0.118*** (0.0287)	0.159** (0.0797)	0.115*** (0.0235)	0.110* (0.0570)	0.145** (0.0627)
Log(LMA No Educ)	0.284*** (0.0577)	0.575** (0.228)	0.239*** (0.0480)	0.342** (0.165)	0.444** (0.175)
Log(MA No Educ)	-0.171*** (0.0346)	-0.261*** (0.0962)	-0.150*** (0.0286)	-0.158** (0.0681)	-0.232*** (0.0775)
Observations	334	334	292	292	333

Notes: This table shows the robustness of my sufficient statistic estimation. Column one replicates my main results from table 2.B.8. Column two removes weighting by 1970 population. Column three trims by removing the 10% most extreme values. Column four trims and removes weighting. Column five weights by the inverse standard error in μ estimation.

FIGURE 2.B.15 Distribution of effects



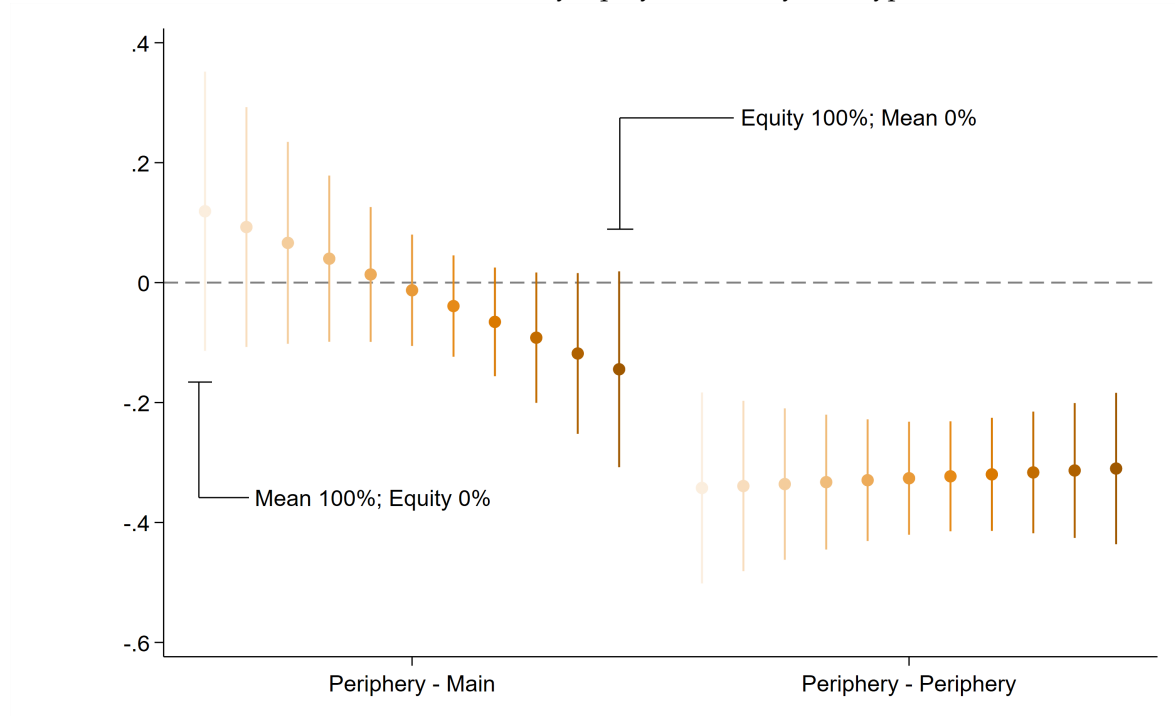
Notes: This figure shows the (centered) distribution over locations of effects of road building since 1970 on local education opportunity, pooled over Benin, Cameroon, and Mali. In blue (solid line) baseline results are plotted. In green (dash-dot line) changes in goods trade costs are shut down. In orange (long dash dot) the utility value of education is set to 0.

TABLE 2.B.13 Alternative outcome variables

	(1) Prop not employed in agriculture	(2) Return to education proxy: Housing Quality	(3) Return to education proxy: Not in agriculture
Log(LMA Educ)	-0.170 (0.295)	-1.373** (0.654)	0.124 (0.350)
Log(MA Educ)	0.392** (0.162)	0.699** (0.328)	0.172 (0.183)
Log(LMA No Educ)	-0.761*** (0.275)	2.042*** (0.675)	0.442 (0.312)
Log(MA No Educ)	-0.0385 (0.164)	-1.322*** (0.356)	-0.463** (0.187)
Observations	400	398	398

Notes: This table shows the results from estimating the main sufficient statistic equation on alternative outcome variables. Column one uses the local proportion not employed in agriculture (adults between 25 and 55), column two local returns to education proxied by housing quality, column (3) local returns to education proxied by not working in agriculture. In all regressions I include locality and time fixed effects, weight by 1970 population, and cluster at the locality level.

FIGURE 2.B.16 Efficiency-equity trade off by road type

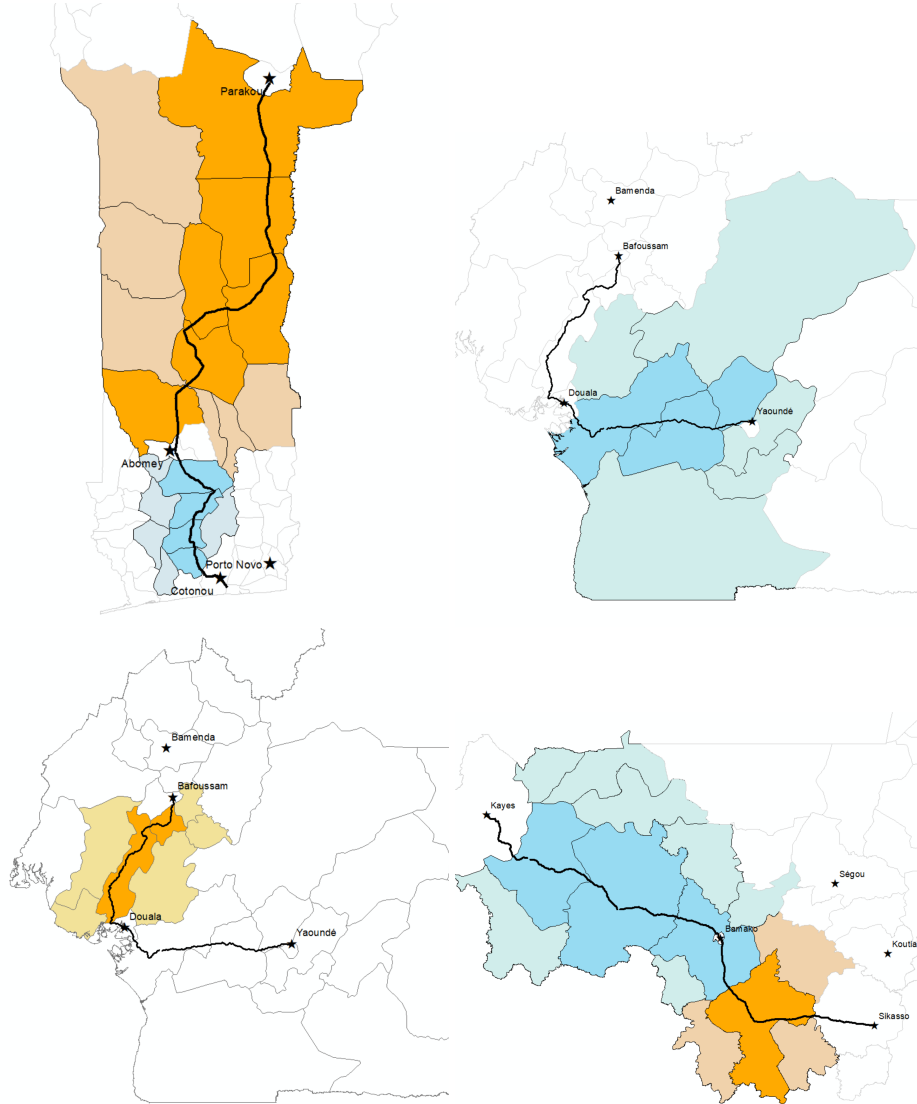


Notes: This figure shows the effects of building road of different types (periphery-main connections on the left-hand-side, periphery-periphery connections on the right-hand-side) on an outcome which is a convex combination of the road-upgrade induced mean opportunity shift and (the negative of the) road-upgrade induced change in standard deviation of inequality. Lighter orange colors weigh mean shifts more, darker orange colors weight equity more. Coefficients are relative to the omitted road category “other”.

2.B.15 Replicating results from Faber [2014]

In this section, I replicate the variation used in Faber [2014] in my data. I do this by focusing on counterfactual road-upgrade “experiments” where the road in question connects two main cities and compare the effect of such a road on locations on the route relative to those just off the route. As discussed in the main text there are many differences between my setting and that considered in Faber [2014], one of the results of this is that there are far fewer such “experiments” in Benin, Cameroon, and Mali, than Faber [2014] finds in China. Defining a main city as one with a population over 100,000 the experiments I focus on are depicted in 2.B.17 figure below.

FIGURE 2.B.17 “Experiments” identifying variation similar to that used in Faber [2014]



Notes: These figures depict the variation isolated to replicate the analysis in Faber [2014] in my setting. Black lines are the routes between two main cities, where main cities are defined as having a population over 100,000. Darker colors represent treated locations, ones which the roads pass through. Lighter colors represent control locations, ones adjacent to treated areas. Locations containing the main city are removed from the analysis. Roads between main cities where it wasn't possible to create reasonable treated and control groups have also been omitted.

Using only the variation depicted in figure 2.B.17 I estimate the empirical specification from Faber [2014], given in equation 2.16.

$$y_i = \beta \cdot \text{Connect}_i + \eta \cdot X_i + \varepsilon_i \quad (2.16)$$

The outcome, y_i is the change in local opportunity due to upgrading the road. Unlike Faber [2014] as this is a counterfactual exercise, I don't require an empirical identification strategy to estimate this. The treatment variable Connect_i indicates whether the road ran through i

or not. X_i denotes various controls. ε_i is an idiosyncratic error. Table 2.B.14 presents the results from estimating equation 2.16 on the sample depicted in figure 2.B.17. Column one includes no controls, column two controls for baseline values of the proportion employed in agriculture, the proportion living in an urban setting, and log population, column three additionally controls for the proportion who have access to electricity and the proportion who have completed primary school. Finally, column four weights the regression by pre-sample population. I include these controls in an attempt to more closely replicate Faber [2014] which includes controls for economic conditions. In addition they help make estimates more precise.

TABLE 2.B.14 Replicating the results from Faber [2014]

	(1)	(2)	(3)	(4)
Connect	-9.893 (12.54)	-15.66 (11.19)	-18.39** (5.728)	-12.70* (6.760)
Proportion employed in agriculture		2.036** (0.763)	3.130*** (0.824)	2.065** (0.707)
Proportion urban		1.556** (0.693)	-0.120 (0.252)	-0.200 (0.214)
Log(Population)		21.59 (14.52)	20.65* (10.34)	24.69** (10.14)
Proportion with access to electricity			1.523** (0.510)	1.273** (0.443)
Proportion with primary school			0.856 (0.659)	0.596 (0.602)
Observations	504	496	496	496
R^2	0.002	0.065	0.137	0.073

Notes: This table presents the results from estimating equation 2.16 on the sample depicted in figure 2.B.17. Column one includes no controls, column two controls for baseline values of the proportion employed in agriculture, the proportion living in an urban setting, and log population, column three additionally controls for the proportion who have access to electricity and the proportion who have completed primary school. Finally, column four weights the regression by pre-sample population.

The results are exactly inline with the main findings from Faber [2014]. Locations that happened to be on the road between two main cities had worse outcomes relative to those just off the road. This may be surprising as I consider a very different setting, and my modeling framework doesn't allow for economies of scale.

This exercise highlights that the type of variation used can lead one to come to very different conclusions as to the effect of changes in connectivity on local outcomes. The advantage of my framework is that I can capture all of the potentially complex effects of different changes in connectivity on outcomes.

2.C Data construction

2.C.1 Market Access data construction

Calculating expected travel times

A key object required to calculate market access terms are the iceberg style movement and trade costs $\kappa_{ijt}^{\phi_s}$, $\tau_{ijt}^{\lambda_s}$. Both of these are based on the fastest path from i to j in period t along the national transport network of the corresponding country, and so I turn first to estimating these travel times which I denote by t_{ijt} . However, my data is available at the locality level which means that t_{ijt} is an aggregate measure of travel times across regions. In order to fully utilize the available variation and keep as close to the actual road network as possible I don't just rely on centroid-to-centroid measures of distance across large localities. Instead I take the interpretation that transport costs from i to j are measured as the expected cost of a randomly chosen individual in i traveling to a randomly chosen individual in j . That is, consider individuals $p \in i$ and $q \in j$ and denote their travel time as d_{pqt} . Then I estimate, t_{pqt} as the following where $|i|$ and $|j|$ denote the population size of i and j respectively.

$$t_{ijt} = \frac{1}{|i|} \sum_{p \in i} \frac{1}{|j|} \sum_{q \in j} d_{pqt} \quad (2.17)$$

However, in order to estimate t_{ijt} in this manner I would need to observe the exact within locality distribution of the population. To focus on variation in road building rather than potentially endogenous changes in the population distribution, I estimate the within locality population distribution in the pre-sample year of 1970 and keep it fixed. To do this I introduce a new data source, Africapolis, which maps all agglomerations in Sub-Saharan Africa that will achieve a population of at least 10,000 in 2015 and backdates each agglomerations' population to 1970. I take all such available agglomerations, their exact coordinates and 1970 populations. To this I add the backdated approximate remaining population of each locality using census data and assign this to the locality centroid. This gives the most accurate possible within-locality and within-country population distribution in 1970 using available data.

Having completed the above steps I have a set of locations p within each locality i that is $p \in i$. For each location I associate a 1970 population $P_{p,1970}$ and time of travelling along the observed road network to each other point $q \in j$ for each j locations in the same country, d_{pqt} . Then the expected travel time of a randomly chosen household in i traveling to a randomly

chosen household in j in year t is given by the following.

$$t_{ijt} = \sum_{p \in i} \frac{P_{p,1970}}{P_{i,1970}} \sum_{q \in j} \frac{P_{q,1970}}{P_{j,1970}} d_{pqt} \quad (2.18)$$

This can be seen as a coarse discretisation of equation 2.17, the best that can be done with the data available.

Calculating incomes Y_{it}^s

Census data does not provide information about wages or the total income/ output of localities. However, it does provide some limited information on the assets households own such as flooring material, sanitation and electricity. I can use this information coupled with auxiliary regressions using income data from development health surveys (DHS) to impute approximate income at the locality-year-primary completion level. Intuitively this approach is similar to that of Young [2012] in that I use auxiliary Engle curve regressions to uncover parameters which are then used in a second stage with richer data to impute the outcome of interest at a broader and more granular geographic level. This approach requires some assumptions which are difficult to test, however given the paucity of data available on wages/ incomes at sufficient geographic and temporal desegregation I believe that this is approach the best that can be done. Additionally, due to high informality rates, it's unclear whether wages would be the most appropriate measure even if they were available.

Postulate that the (real) demand for an asset a by household h in locality i in year t is give by the following equation.

$$\ln(Q_{ahit}) = \alpha_a + \eta_a \ln(C_{hit}^N) + \xi_a \ln(P_{it}) + \beta X_{hit} + \varepsilon_{ahit} \quad (2.19)$$

Where α_a are product constants, η_a is the (quasi) income elasticity of demand, C_{hit}^N is nominal household consumption expenditure which is equal to household income in our setting, ξ_a is a vector of own and cross-price (quasi) elasticities of demand, $\ln(P_{it})$ is a vector of regional prices, X_{hit} and β are vectors of household characteristics and their coefficients. Finally ε_{ahit} is a white noise household-product preference shock. Elasticities are referred to as quasi above as for all assets considered I use an indicator variable rather than a logarithm. To estimate this equation I use data from the available DHS waves in Benin, Cameroon, and Mali, that report income. Sadly this is only two waves: Benin in 1995 and Mali in 1996. Additionally, these surveys don't include information on prices and so I estimate equation 2.19

using product-locality-year fixed effects (although year fixed effects are redundant given that localities are only observed once) which absorbs price variation. Results from running regressions are given in table 2.C.1.

TABLE 2.C.1 Asset demand equations using DHS data

	(1)	(2)	(3)	(4)
Concrete floor	0.00550*** (0.000914)	0.00557*** (0.000923)	0.00479*** (0.000820)	0.00496*** (0.000915)
Electricity	0.00212*** (0.000653)	0.00218*** (0.000668)	0.00205*** (0.000593)	0.00158** (0.000665)
Sanitation	0.00312*** (0.000890)	0.00319*** (0.000898)	0.00313*** (0.000814)	0.00258*** (0.000893)
Asset $\times$ Region FE	X	X	X	X
Age polynomial		X	X	X
Asset $\times$ Region $\times$ Urban FE			X	
HH members control				X
R^2	0.402	0.403	0.498	0.404
N	22586	22586	22586	22586

Notes: This table shows the results from running regressions of the form given in equation 2.19 using DHS data.

Focusing on column (4) these results suggest that a 1% increase in income is associated with a 0.5pp. increase in the probability of having concrete floor, a 0.16pp increase in the probability of having access to electricity and a 0.26pp. increase in having accessible sanitation.

In the second step, I use the inverted estimated coefficients from table 2.C.1 to approximate income differences by assets households own as indicated in census data. Imputed average income in a locality-year-education cell is then given by $\tilde{Y}_{it}^e = 1/N_{ite} \sum_{h \in \{i,t,e\}} \sum_a \frac{1}{\eta_a} Q_{hait} + base$ where *base* is the average income calculated from DHS data. Intuitively if we see households in an area with more assets than those in a different area we infer that those in the first area have more income with which to purchase such assets. The Engle curve auxiliary regressions in the first stage allow me to approximate how much more income owning an asset may signal, and thus translate regional and by-education differences into a money-metric form.

A simpler way to estimate incomes that uses more asset information from census data and no auxiliary data is to postulate a linear relationship between housing quality (measured as the first principle component over all available housing asset variables) and wages, $w_{it}^e = \beta HQ_{it}^e$. As my market access equations are invariant to constant multipliers such a

relationship implies that I can use $Y_{it}^e = HQ_{it}^e L_{it}^e$ as my measure of income in a locality-year-education cell. This formulation is simpler, but a-theoretic and relies on what appear to be stronger assumptions. However, results from using this specification to calculate incomes bring qualitatively similar conclusions to those using the Engle-curve, method. Therefore, in the main text I focus on using data from the Engle-curve approach.

Estimating bilateral transportation and migration costs

Migration costs

Recall that the theory delivered a gravity equation for the intra-national movement of households. I can estimate this equation using origin-year and destination-year fixed effects by psudeo-poisson-maximum-likelihood (PPML) as in [Yotov, Piermartini, Monteiro, and Larch \[2016\]](#) and others. The estimating equation is given in [2.20](#).

$$M_{ijt}^s = \alpha_{jt}^s \cdot \kappa_{ijt}^{-\lambda_s} \cdot \rho_{it}^s \cdot \varepsilon_{ijt}^s \quad (2.20)$$

i is the origin location, j is the destination location, and s denotes individual type. Where α_{jt}^s are destination-time fixed effects, ρ_{it}^s are origin-time fixed effects and ε_{ijt}^s is an idiosyncratic error term. I approximate bilateral migration costs by the following specifications: $\kappa_{ijt}^{-\lambda_s} = \ln(t_{ijt})^{\hat{\lambda}_s}$. As M_{ijt}^s and t_{ijt} are observed I can estimate this equation by PPML to recover $\hat{\lambda}_s$. Table [2.C.2](#) shows the results from estimating equations of the form given in equation [2.20](#). Columns (1) and (3) estimate [2.20](#) by OLS after first taking logs. Columns (2) and (4) estimate equation [2.20](#) by PPML. Column (4) represents my main specification and thus I use values $\hat{\lambda}_s = 1.48$ and $\hat{\lambda}_s = 0.985$ to derive labour market access terms.

TABLE 2.C.2 Gravity regressions to recover thetas

	No primary education		Primary education	
	(1)	(2)	(3)	(4)
	Log-linear	PPML	Log-Linear	PPML
Log(Travel time)	-1.270*** (0.0193)	-1.477*** (0.0356)	-0.933*** (0.0171)	-0.985*** (0.0375)
Destination-time FE	X	X	X	X
Origin-time FE	X	X	X	X
N	11005	26234	10941	26234

Notes: This table shows results from running gravity equations of the form given in 2.20 using Census data on internal migration across localities. Columns one and three use a log-linear specification whereas columns (2) and (4) use PPML.

Note that, those who haven't completed primary education, are significantly more sensitive than those who have to changes in transport costs. That is distance presents more of an impediment to travel, which isn't surprising. These tables also highlight the importance of correctly estimating gravity equations using PPML rather than OLS which gives significantly attenuated coefficients.

Bilateral trade costs

In the absence of intra-national, locality-level bilateral trade data I am unable to perform a similar calculation to estimate $\tilde{\phi}$. Instead I'm forced to turn to values commonly found in the literature and take as my baseline $\hat{\phi} = -3.8 \times 0.088$ from Donaldson and Hornbeck [2016], Donaldson [2018].

2.C.2 Digitizing Maps

Data on the changing connectivity of place comes from digitized historical Michelin road maps accessed from the Bodleian Library at the University of Oxford. In this section I detail the digitizing procedure taken. Figure 2.C.1 gives examples of the original maps used for Mali, reproduced with permission from Michelin, and figure 2.C.2 shows some examples of the finished digitized maps. Throughout I used the geographical mapping software ArcGIS. The procedure taken is detailed in the steps below.

1. Download the **Open Street Maps** shapefiles for Benin, Cameroon, and Mali, which representing the current road network in each country.
2. Remove minor roads, or other roads not represented on the most recent (2019) Michelin road maps.
3. Categorize all remaining roads as in the most recent Michelin road maps.
4. Add all settlements from Africapolis which include all agglomerations which have a population of at least 10,000 in 2015.
5. Using settlement level population estimates from Africapolis and back-dated locality level population estimates from Census data, calculate the remaining locality-level population not covered by the Africapolis settlements. Add this population to a location at the centre of each locality.
6. Make small adjustments to the 2019 road network so that roads hit the centroid of each settlement and form a connected network. To do this I used the topology tool in ArcGIS.
7. Iteratively delete or downgrade roads using maps increasingly in the past. In this way, for each year a map is available, create the complete road network.

FIGURE 2.C.1 Original maps

(A) Mali 1969



(B) Mali 2019



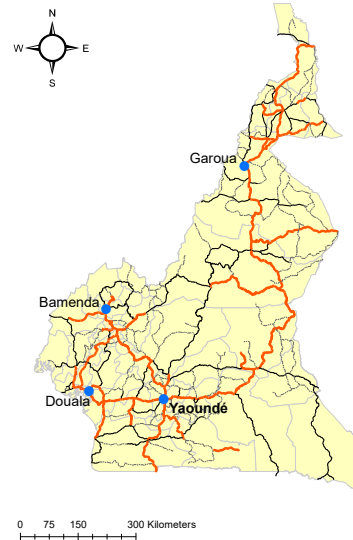
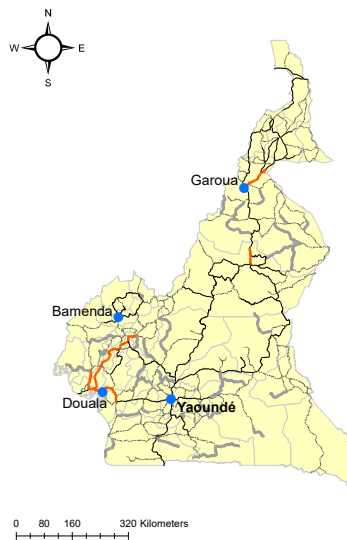
Notes: This figure shows pictures of the original Michelin road maps for Mali in 1969 on the left hand side and in 2019 on the right hand side.

FIGURE 2.C.2 Digitized road networks from historical Michelin maps
(A) Benin 1969 (B) Benin 2019



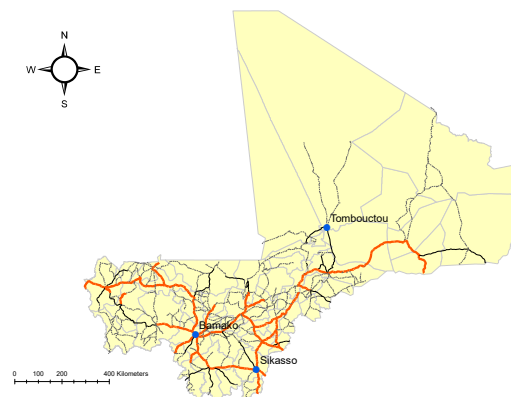
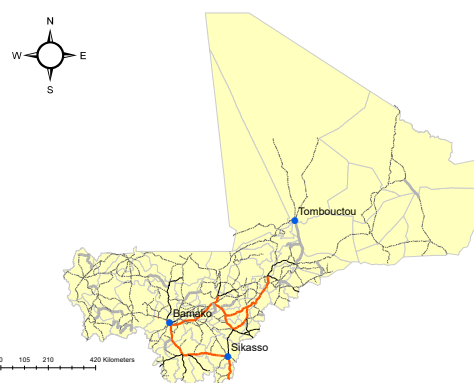
(C) Cameroon 1969

(D) Cameroon 2019



(E) Mali 1969

(F) Mali 2019



Notes: These maps show the digitized Michelin road maps for each country comparing the road network in 1969 to 2019. Roads are categorized into four categories. Red lines are paved roads, and the fastest. Black lines are partially improved (gravel) roads. Dotted lines denote tracks. Finally, thick gray lines indicate to-be-built roads.

2.D Theory appendix

2.D.1 Solving the spatial model

Theorem 1. *The equations given below:*

$$u_{lt}^s = T_{lt}^{B_s} \left(\left(\frac{w_{lt}^s}{P_{lt}^s} \right)^{1-\beta} E_{lt}^\beta \right)^{\lambda_s} \quad (2.21)$$

$$E_{lt} = \left(\frac{w_{lt}^E}{w_{lt}^N} \right) \quad (2.22)$$

$$w_{lt}^s = \frac{Y_{lt}^s}{L_{lt}^s} \quad (2.23)$$

$$Y_{lt}^s = T_{lt}^{Z_s} (w_{lt}^s)^{-\phi} MA_{lt}^s \quad (2.24)$$

$$L_{lt}^s = u_{lt}^s LMA_{lt}^s \quad (2.25)$$

$$(P_{lt}^s)^{-\phi} = MA_{lt}^s = \sum_k \tau_{lkt}^{-\phi} \frac{Y_{kt}^s}{MA_{kt}^s} \quad (2.26)$$

$$LMA_{lt}^s = \sum_k \kappa_{lkt}^{-\lambda_s} \frac{L_{kt}^s}{LMA_{kt}^s} \quad (2.27)$$

can be rewritten to be of the form:

$$MA_{it}^r = \sum_j K_{ijt}^r \prod_{h=1}^4 \left(MA_{jt}^h \right)^{b_{rh}} \quad (2.28)$$

where K_{ijt}^r is some (log-linear) bundles of fundamentals. Therefore, all endogenous variables can be written as a log-linear function of market access terms.

Proof. Subscripts will be suppressed unless strictly necessary to ease notation. Plug 2.21 into 2.25 to find $L_s = \left(A_s E^\beta \left(\frac{w_s}{P_s} \right)^{1-\beta} \right)^{\lambda_s} LMA_s$ then plugging 2.27 and 2.22 into this we find

$$L_s = \left(A_s \left(\frac{w_E}{w_N} \right)^\beta \left(\frac{w_s}{MA_s^{-1/\phi_s}} \right)^{1-\beta} \right)^{\lambda_s} LMA_s \quad (2.29)$$

. Then noting that by plugging 2.23 into 2.24 we have $w_s = B_s^{\frac{\phi_s}{1+\phi_s}} L_s^{-\frac{1}{1+\phi_s}} M A_s^{\frac{1}{1+\phi_s}}$. This can then be plugged into 2.29 to find (defining $x_s = 1 + \phi_s + (1 - \beta)\lambda_s$):

$$L_s = \left(A_s \left(\frac{w_1}{w_2} \right)^\beta \left(B_s^{\frac{\phi_s}{1+\phi_s}} L_s^{-\frac{1}{1+\phi_s}} M A_s^{\frac{1}{1+\phi_s}} \right)^{1-\beta} M A_s^{\frac{1-\beta}{\phi_s}} \right)^{\lambda_s} L M A_s$$

$$L_s = A_s^{\frac{\lambda_s(1+\phi_s)}{x_s}} \left(\frac{w_E}{w_N} \right)^{\frac{\beta\lambda_s(1+\phi_s)}{x_s}} B_s^{\frac{\phi_s\lambda_s(1-\beta)}{x_s}} M A_s^{\frac{(1-\beta)\lambda_s}{x_s} + \frac{(1+\phi_s)}{\phi_s} \frac{(1-\beta)\lambda_s}{x_s}} L M A_s^{\frac{1+\phi_s}{x_s}}$$

Using our expression for w_s we can then write (defining $x = x_E x_N - \beta(x_E \lambda_E - x_N \lambda_N)$).

$$\frac{w_E}{w_N} = B_E^{\frac{x_E x_N}{x}} \left(\frac{\phi_E}{1+\phi_E} - \frac{\phi_E}{1+\phi_E} \frac{\lambda_E(1-\beta)}{x_E} \right) B_N^{\frac{x_E x_N}{x}} \left(\frac{\phi_N}{1+\phi_N} \frac{\lambda_N(1-\beta)}{x_N} - \frac{\phi_N}{1+\phi_N} \right) A_E^{\frac{\lambda_E}{x_E} \frac{x_E x_N}{x}} A_N^{\frac{\lambda_N}{x_N} \frac{x_E x_N}{x}}$$

$$\times M A_E^{\frac{x_E x_N}{x} \left(\frac{1}{1+\phi_E} - \frac{(1-\beta)\lambda_E}{x_E(1+\phi_E)} - \frac{(1-\beta)\lambda_E}{\phi_E x_E} \right)} M A_N^{\frac{x_E x_N}{x} \left(-\frac{1}{1+\phi_N} + \frac{(1-\beta)\lambda_N}{x_N(1+\phi_N)} + \frac{(1-\beta)\lambda_N}{\phi_N x_N} \right)}$$

$$\times L M A_E^{-\frac{x_E x_N}{x} \frac{1}{x_E}} L M A_N^{\frac{x_E x_N}{x} \frac{1}{x_N}}$$

This equation gives the structural interpretation of the coefficients from estimating the reduced form equation in equation 2.7 in the main text. Plugging this into the above derived equation for L_s gives the following.

$$L_E = A_E^{\frac{\lambda_E(1+\phi_E)}{x_E} - \frac{\beta\lambda_E(1+\phi_E)}{x_E} \frac{\lambda_E}{x_E} \frac{x_E x_N}{x}} A_N^{\frac{\beta\lambda_E(1+\phi_E)}{x_E} \frac{\lambda_N}{x_N} \frac{x_E x_N}{x}}$$

$$\times B_E^{\frac{\phi_E\lambda_E(1-\beta)}{x_E} + \frac{\beta\lambda_E(1+\phi_E)}{x_E} \frac{x_E x_N}{x} \left(\frac{\phi_E}{1+\phi_E} - \frac{\phi_E}{1+\phi_E} \frac{\lambda_E(1-\beta)}{x_E} \right)}$$

$$\times B_N^{\frac{\beta\lambda_E(1+\phi_E)}{x_E} \frac{x_E x_N}{x} \left(\frac{\phi_N}{1+\phi_N} \frac{\lambda_N(1-\beta)}{x_N} - \frac{\phi_N}{1+\phi_N} \right)}$$

$$\times M A_E^{\frac{(1-\beta)\lambda_E}{x_E} + \frac{(1+\phi_N)}{\phi_N} \frac{(1-\beta)\lambda_E}{x_E} + \frac{\beta\lambda_E(1+\phi_E)}{x_E} \frac{x_E x_N}{x} \left(\frac{1}{1+\phi_E} - \frac{(1-\beta)\lambda_E}{x_E(1+\phi_E)} - \frac{(1-\beta)\lambda_E}{\phi_E x_E} \right)}$$

$$\times M A_N^{\frac{\beta\lambda_E(1+\phi_E)}{x_E} \frac{x_E x_N}{x} \left(-\frac{1}{1+\phi_N} + \frac{(1-\beta)\lambda_N}{x_N(1+\phi_N)} + \frac{(1-\beta)\lambda_N}{\phi_N x_N} \right)}$$

$$\times L M A_E^{\frac{1+\phi_E}{x_E} - \frac{\beta\lambda_E(1+\phi_E)}{x_E} \frac{x_E x_N}{x} \frac{1}{x_E}}$$

$$\times L M A_N^{\frac{\beta\lambda_E(1+\phi_E)}{x_E} \frac{x_E x_N}{x} \frac{1}{x_N}}$$

Therefore, using L_E as derived above I can write $L M A_{iE} = \sum_j \kappa_{ij}^{-\lambda_E} L_{iE} L M A_{iE}^{-1}$. Noting that 2.23 and 2.24 can also be combined to derive $Y_s = L_s^{\frac{\phi_s}{1+\phi_s}} B_s^{\frac{\phi_s}{1+\phi_s}} M A_s^{\frac{1}{1+\phi_s}}$ it's straight forward to see how analogous expressions for each market access term can be derived recovering the desired series of non-linear equations.

□

Theorem 2. The model in theorem 1 can be written in changes where $\hat{x} = x'/x$ in the following manner.

$$\widehat{MA}_{it}^r = \sum_j \hat{\rho}_{ijt}^r \lambda_{ijt}^r \prod_{h=1}^4 \left(\widehat{MA}_{jt}^h \right)^{b_{rh}}$$

Note the above system does not include fundamentals. $\lambda_{ijt}^{MA^E}$ denotes the fraction of i 's market access, due to j at time t .

Proof. Consider MA_i^E , where not-needed subscripts have been suppressed. From it's definition we know that $MA_i^E = \sum_j \tau_{ij}^{-\phi_E} \frac{Y_j^E}{MA_j^E}$, thus we can write:

$$\begin{aligned} \widehat{MA}_i^E &= \frac{(MA_i^E)'}{MA_i^E} = \frac{\sum_j (\tau_{ij}^c)^{-\phi_E} \frac{Y_j^{E,c}}{MA_j^{E,c}}}{\sum_k \tau_{ik}^{-\phi_E} \frac{Y_k}{MA_k^{E,c}}} \\ &= \sum_j \frac{(\tau_{ij}^c)^{-\phi_E} \frac{Y_j^{E,c}}{MA_j^{E,c}}}{\sum_k \tau_{ik}^{-\phi_E} \frac{Y_k}{MA_k^{E,c}}} \\ &= \sum_j \frac{\tau_{ij}^{-\phi_E} \frac{Y_j^E}{MA_j^E}}{\sum_k \tau_{ik}^{-\phi_E} \frac{Y_k}{MA_k^{E,c}}} \hat{\tau}_{ij}^{-\phi_E} \hat{Y}_j^E \left(\widehat{MA}_j^E \right)^{-1} \\ &= \sum_j \lambda_{ij}^{MA^E} \hat{\tau}_{ij}^{-\phi_E} \hat{Y}_j^E \left(\widehat{MA}_j^E \right)^{-1} \end{aligned}$$

Where $\lambda_{ij}^{MA^E} = (\tau_{ij}^{\phi_E} Y_j^E (MA_j^E)^{-1}) / MA_i^E$, is the proportion of i 's E-type market access due to j as is known. Noting that from the proof of theorem 1 we know that Y_j^k can be given as a log-linear function of market access terms and fundamentals and therefore $\hat{Y}_j^k$ is a log-linear function of market access variables defined in changes. Via an analogous expression for each of the other types of market access we arrive at the desired result, where the β_{ij} coefficients are the same as those given in theorem 1. $\square$

2.D.2 Model extensions

Endogenous education supply

Denote by A_i the number of students who have access to schooling in locality i , S_i as the total number of schools in a region, c as the constant cost of constructing a given school, and finally I is the total school construction budget. A social planner whos objective is to

maximise the number of people who can access a school solves the following problem.

$$\max_{S_i} \sum_i A_i \quad s.t. \quad c \sum_i S_i \leq I \quad (2.30)$$

The number of students with access to schooling in a region is assumed to be a function of three parameters. First, the number of schools in the region S_i , second local population density D_i , third how developed the local road network is R_i . The intuition is simple, suppose that for a given school, the number of people who have access to said school is given by everyone who can travel to it within some radius r minutes of travel time. Then this radius will effectively be bigger if the local road network is more developed, and capture more people if the population is denser. Model the production function for A_i in a Cobb-Douglas fashion.

$$A_i = \pi S_i^{\alpha_1} D_i^{\alpha_2} R_i^{\alpha_3} \quad (2.31)$$

We expect $0 < \alpha_1 < 1$ due to the potential of overlapping catchment areas and expect the other coefficients to be positive, α_2 is likely less than one due to congestion. We can sub this expression into the planners problem and find the first order conditions with respect to S_i of the corresponding Lagrangian: $S_i = \alpha_1 S_i^{\alpha_1-1} D_i^{\alpha_2} R_i^{\alpha_3} = c\lambda$ where λ is the Lagrange multiplier which can be found by subbing the first order conditions into the budget constraint to be $\lambda = \alpha_1 A/I$ where $A = \sum A_i$. Putting this together we have an equation determining the number of schools in a region.

$$S_i = \left(\frac{I}{cA} \right)^{\frac{1}{1-\alpha_1}} D_i^{\frac{\alpha_2}{1-\alpha_1}} R_i^{\frac{\alpha_3}{1-\alpha_1}} \quad (2.32)$$

Equation 2.32 shows that school supply will be higher in locations that are denser and or have better local transport networks.

Turning to the household decision to education their child or not, we now introduce a locality-varying cost of education that is increasing in the distanced required to travel to school. DeStefano et al. [2007] and Evans and Mendez Acosta [2021] find that distanced to primary school is a large impediment to education completion in Sub-Saharan Africa. A household will travel longer to their closest school if there are fewer schools in their area, or if their area has a less developed road network. Thus, I give education costs the following functional form: $c_i^E = \kappa S_i^{\gamma_1} R_i^{\gamma_2}$. We augment the probability a child completes primary school to be decreasing in the cost of education and increasing in the returns to education as

follows.

$$\mu_i = \beta_0 + \beta_1 \ln(r_i) + \beta_2 \ln(c_i^E) + \varepsilon_i \quad (2.33)$$

Subbing the above derivations in we come to the following expression.

$$\mu_i = \bar{\beta}_0 + \beta \ln(r_i) + \frac{\alpha_2 \gamma_1 \beta_2}{1 - \alpha_1} \ln(D_i) + \left(\beta_2 \gamma_1 \frac{\alpha_3}{1 - \alpha_1} + \gamma_2 \beta_2 \right) \ln(R_i) + \varepsilon_i \quad (2.34)$$

Where $\bar{\beta}_0$ captures all constants. As $D_i = L_i / \text{Area}_i$ and market access terms are a sufficient statistic for L_i , equation 2.34 will collapse to an equation similar to that which was found before, with the exception of the introduction of R_i and a change in coefficient interpretation.

I can take the additional predictions from this section to the data, using additional data on current school locations from the United Nations Office for the Coordination of Humanitarian Affairs. Table 2.D.1 shows the results from this. Column one uses the school data (which is only available in the most recent period) to estimate equation 2.32 and finds evidence that density increases school supply ($\alpha_2 > 0$) but that the local quality of the road network has no effect ($\alpha_3 = 0$). Column three replicates the main results from table 2.B.8 in the body of the paper. Column two runs the same specification but includes local road quality as suggested by equation 2.34. Again, I find a zero coefficient suggesting that $\gamma_2 = 0$ or $\beta_2 = 0$ as we have already established that $\alpha_3 = 0$. In the latter case we could conclude that endogenising education supply makes little substantive difference to results. The lack of difference in coefficients on the market access terms between columns two and three suggests that even the more complex model for education supply nests within the set of data generating processes captured for the sufficient statistic result. Due to this, and the zero coefficient on local road quality, when estimating counterfactuals, I take the view that endogenous education supply is of second order concern and to remain as parsimonious as possible it is omitted.

TABLE 2.D.1 Endogenising Schooling Supply

	(1)	(2)	(3)
	$\ln(\#Schools)$	μ	μ
Local Road Network	-0.0827 (0.141)	0.000528 (0.00620)	
Log Density	0.332*** (0.0839)		
$\ln(LMA - E)$		-0.211** (0.0768)	-0.198** (0.0767)
$\ln(MA - E)$		0.116*** (0.0280)	0.118*** (0.0281)
$\ln(LMA - N)$		0.291*** (0.0610)	0.284*** (0.0581)
$\ln(MA - N)$		-0.171*** (0.0333)	-0.171*** (0.0329)
N	126	334	334

Notes: This table shows the results from running regressions pertaining to the empirical implications from endogenizing education supply. In column one I show the result from running a regression of the form given in equation 2.32 in a log-linear form. In column two I show the results from estimating equation 2.34 and in column three I give the baseline sufficient statistic results for comparison.

Generalizing consumption patterns

In the main text I assume that types only consume their own goods. This is evidently an unrealistic assumption but is a simplification designed to emphasize the uncontroversial observation that educated workers spend a larger proportion of their consumption on goods produced by educated workers than non-educated workers do. The assumption is also made because it allows me to write the sufficient statistic result as an exact log-linear equation (although it will remain log-linear to a first order approximation) and appeal to known existence and uniqueness results when deriving equilibrium and performing counterfactual analysis. Allowing both types of workers to consume both types of goods breaks both of these convenient and parsimonious results. However, in this section I do perform this generalization and with the caveat that formal existence and uniqueness results have not been

provided, show that counterfactual results are qualitatively and quantitatively unchanged.

Denote by E_{is} expenditure in location i on sector s and S_{is} spending in i by group s . In the main text we have that $E_{is} = S_{is} = Y_{is}$ but now we allow for the possibility that $E_{is} \neq S_{is} = Y_{is}$. Suppose instead that each group k spends a constant fraction of their income on each sector s 's products and denote this by α_{ks} then we have:

$$E_{is} = \sum_k \alpha_{ks} Y_{ik} \quad S_{ik} = \sum_s \alpha_{ks} Y_{ik} = Y_{ik}$$

, and therefore $E_{is} = \sum_k \alpha_{ks} S_{ik}$. From this generalization we can see that the results in the main text are a special case where $\alpha_{ks} = 1$ if $k = s$ and 0 otherwise. The derivations change as we can no longer substitute income, Y_{is} , for expenditure, E_{is} , and must instead use the more complex expression for expenditure above. This breaks the log-linear nature of the model, and thus doesn't allow the same exact log-linear sufficient statistic result, although market access terms will remain sufficient statistics in a non-linear (or log-linear) fashion. The labor market side of the model remains unchanged. Working through the algebra, and using exact-hat techniques, we arrive at the following series of differenced equations which can then be used for counterfactual analysis.

$$\hat{L}_{is} = \left(\left(\frac{\hat{w}_{is}}{\hat{P}_{is}} \right)^{(1-\beta)} \left(\frac{\frac{\hat{w}_{iE}}{\hat{P}_{iE}}}{\frac{\hat{w}_{iN}}{\hat{P}_{iN}}} \right)^\beta \right)^{\lambda_s} \widehat{LMA}_{is} \quad (2.35)$$

$$\hat{E}_{is} = \hat{w}_{is}^{-\phi_s} \widehat{MA}_{is} \quad (2.36)$$

$$\hat{E}_{is} = \hat{w}_{iE} \hat{L}_{iE} \underbrace{\left(\frac{\alpha_{Es} Y_{iE}}{\alpha_{Es} Y_{iE} + \alpha_{Ns} Y_{iN}} \right)}_{\zeta_{iEs}^E} + \hat{w}_{iN} \hat{L}_{iN} \underbrace{\left(\frac{\alpha_{Ns} Y_{iN}}{\alpha_{Es} Y_{iE} + \alpha_{Ns} Y_{iN}} \right)}_{\zeta_{iNs}^E} \quad (2.37)$$

$$\widehat{MA}_{is} = \sum_j \lambda_{ij}^{MA_s} \hat{\tau}_{ij}^{-\phi_s} \hat{E}_{js} \widehat{MA}_{js}^{-1} \quad (2.38)$$

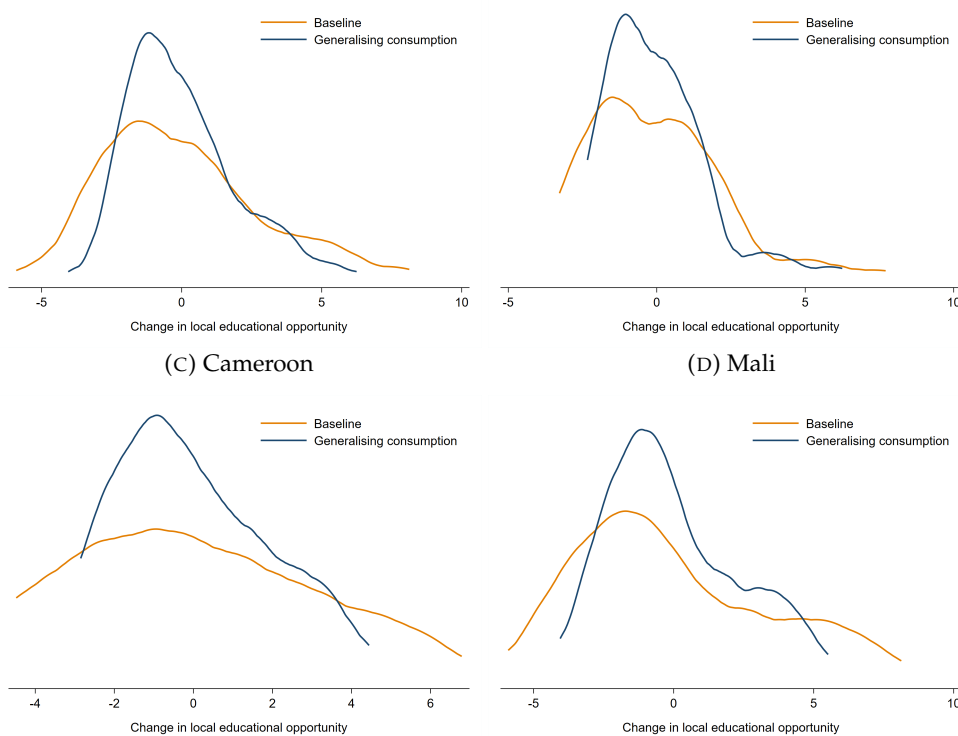
$$\widehat{LMA}_{is} = \sum_j \lambda_{ij}^{LMA_s} \hat{\kappa}_{ij}^{-\lambda_s} \hat{L}_{js} \widehat{LMA}_{js}^{-1} \quad (2.39)$$

The system of equations given in 2.35 together with a counterfactual change in the road network $(\hat{\tau}, \hat{\kappa})$, data on $\lambda_{ij}^{MA_s}, \lambda_{ij}^{LMA_s}, \zeta_{iEs}^E, \zeta_{iNs}^E$, and parameter estimates for $\phi_s, \lambda_s, \beta, \alpha_{Es}, \alpha_{Ns}$ allow me to solve for counterfactual changes in opportunity, as in the main text. $\lambda_{ij}^{MA_s}, \lambda_{ij}^{LMA_s}, \lambda_{iEs}^E, \lambda_{iNs}^E$ are all observed, so the only new obstacle lies in estimating α_{ks} . I assume that E -types spend 50% of their income on E -type goods, $\alpha_{EE} = 0.5$ and N -types spend 10% of their income on N -type goods $\alpha_{NE} = 0.1$. These parameter estimates seem reasonable, but given the lack of evidence with respect to spending patterns by education in Benin, Cameroon, or, Mali, can

only be taken as conjecture. Results are robust to other reasonable estimates.

I compare the results on the aggregate effect of road building since 1970 from the main text and from the above described *generalizing consumption* (gc) model extension. The baseline model generates the change in opportunity due to road building since 1970 for each locality i denoted as $\hat{\mu}_i$, and the extension derives the same objects denoted as $\hat{\mu}_i^{gc}$. First, I find that the two series are highly correlated, $Corr(\hat{\mu}_i^{gc}, \hat{\mu}_i) = 0.991$. Figure 2.D.1 then compares the distribution of (relative) effects in the baseline and in the generalizing consumption extension. Sub-figure 2.D.1a shows the overall distribution whereas sub-figures 2.D.1b, 2.D.1c, and 2.D.1d show country specific distributions for Benin, Cameroon, and Mali respectively. Although the two series are highly correlated it's clear that the results from the extension are less dispersed than those from the baseline with an overall variance some 25% smaller.

FIGURE 2.D.1 Changes in the distribution of primary completion rates
(A) Overall (B) Benin



Notes: These figures show the distributional results from estimating the model and running the no-roads counterfactual with generalized consumption preferences (in blue) compared to the baseline (in orange).

Turning to comparing the counterfactual estimates of how road building since 1970 affected the inequality of local educational opportunity across space, I find that in this extension the variance of opportunity across space due to road building increased by 0.24, 1.85, and -0.48 percent in each of Benin, Cameroon, and Mali respectively vs 0.04, 5.81, and -1.44,

in the baseline.

Finally, table 2.D.2 compares the correlation if the estimated effects from the baseline and extension to pre-period remoteness. I find unsurprisingly given the strong correlation but already noted lower dispersion, that the previously uncovered relationship remains, in all three countries more remote locations saw larger gains, but effects are somewhat attenuated.

TABLE 2.D.2 Relationship with remoteness, comparing baseline and generalising consumption extension results

	(1)	(2)
Benin ×	5.092***	3.478***
Log(Expected travel time)	(1.357)	(1.031)
Cameroon ×	6.995***	5.527***
Log(Expected travel time)	(2.409)	(1.591)
Mali ×	5.468***	3.555***
Log(Expected travel time)	(1.299)	(0.888)
Observations	156	156
R^2	0.334	0.343

Notes: This table shows the results from estimating the relationship between the counterfactual change in opportunity due to road building since 1970 and locality remoteness in 1970. Regressions include country fixed effects, and weighted by 1970 population, have robust standard errors, and drop departments in the Extreme-Nord province of Cameroon. Column one shows the baseline results and column two shows results from estimating the model with generalized consumption patterns.

Explicit agglomeration forces

There is considerable evidence for agglomeration economies [Puga, 2010] whereby firms located in denser areas are more productive. I extend the baseline model to allow for this in the standard way following the set up in Allen and Arkolakis [2014], by allowing productivity to be positively effected by population.

$$w_{it} = B_{it} L_{it}^{\alpha} p_{it}^b \quad (2.40)$$

Where $\alpha > 0$ determines the strength of the agglomeration economies. In equilibrium, all equations remain the same with the exception of the equation for output which becomes $Y_{it} = B_{it} L_{it}^\alpha w_{it}^{-\phi} M A_{it}$ which can be rearranged to $Y_{it} = B_{it}^{1/(1-\alpha)} w_{it}^{-\frac{\phi+\alpha}{1-\alpha}} M A_{it}^{1/(1-\alpha)}$. From this it is clear that the derived sufficient statistic result follows. When extending this set up to two sectors/ types one can allow sector specific agglomeration forces and for sector-specific agglomeration to depend on own-type population only or some combination of both types of population.

Explicit endogenous amenities

Following [Diamond \[2016\]](#) and various other authors I can also allow for amenities to be endogenously determined by local population, or as in [Diamond \[2016\]](#) the educated population ratio. I generalize the idiosyncratic component of utility to include a remaining idiosyncratic component and an endogenous component which depends on the educated population ratio: $A_{it}^s = \tilde{A}_{it}^s \left(\frac{L_{it}^E}{L_{it}^N} \right)^\gamma$. In equilibrium all equations remain as in the baseline model with the exception of utility which includes this new amenity term. γ determines the strength of endogenous amenity forces, with $\gamma = 0$ corresponding to the baseline case. It's clear from this set up that the sufficient statistic result will follow.

Land in consumption (fixed supply)

Following the set up in [Donaldson and Hornbeck \[2016\]](#) I allow workers to consume land (housing) H_{it} which is in fixed supply exogenously given in each location. Workers now consume goods and housing in fixed proportions in a Cobb-Douglas fashion and so indirect utility is given by $u_{it} = A_{it} (w_{it}/P_{it})^a (H_{it}/q_{it})^{(1-a)}$. Thus workers spend a constant fraction $1 - a$ of their income on housing and so we have $q_i H_i = (1 - a) Y_i$. The housing market clears so demand $H_i = \bar{H}_i$ supply, which is exogenously given. Therefore we can derive the price of housing in a given locality as $q_i = (1 - a) Y_i / \bar{H}_i$. Indirect utility in a location now depends on the price of housing.

$$u_{it} = A_{it} \left(\left(\frac{w_{it}}{P_{it}^a q_{it}^{1-a}} \right)^{1-\beta} E_{it}^\beta \right)^\lambda$$

$$u_{it} = A_{it} \left(\left(\left(\frac{w_{it}}{P_{it}} \right)^a \left(\frac{1-a}{\bar{H}_{it}} \right)^{-1} L_{it}^{a-1} \right)^{1-\beta} E_{it}^\beta \right)^\lambda$$

$(1 - a)/\bar{H}_{it}$ is exogenous and so loads onto the residual in the sufficient statistic result, and the additional L^{a-1} acts as a dispersion force. Intuitively, if more people live in a location with fixed land supply, the price of housing in this location will increase, raising the cost of living, and making the location less attractive. All other equations remain as in the baseline, thus it's clear from here that the sufficient statistic result will once again be recovered.

Land in production (fixed supply)

Following the set up in [Ahlfeldt et al. \[2015\]](#) I allow firms to use land in production. Thus I extend the firms production function to include land and labor in a Cobb-Douglas production function. Denote by q_i the cost of land in location i then the price of goods produced in i is now $p_i = \left(\frac{w_i}{B_i}\right)^b q_i^{1-b}$ where $1 - b$ is the land-intensity of production. Land is in fixed supply exogenously given and thus land market clearing implies that $(1 - \alpha)Y_i = q_i H_i$. All equations remain the same as in the baseline with the exception of output which now becomes $Y_i = \left((w_i/B_i)^b q_i^{1-b}\right)^{-\phi} M A_i$ which using land market clearing can be simplified to the following.

$$Y_i = \left(\frac{w_i}{B_i}\right)^{-\frac{b\phi}{1+\phi(1-b)}} \left(\frac{1-\alpha}{H_i}\right)^{-\frac{(1-\alpha)\phi}{1+\phi(1-b)}} M A_i^{\frac{1}{1+\phi(1-b)}} \quad (2.41)$$

From here it is clear that the sufficient statistic result will be recovered.

Land in consumption and production (endogenous supply)

If land is in fixed supply we can combine sub-sections 2.D.2 and 2.D.2 to derive the model with land both in consumption and production and see that the sufficient statistic result will once again be recovered. Here I extend this set up further by allowing land to be endogenously supplied following [Diamond \[2016\]](#).

With land used in both consumption and production the total demand for housing (land) is given by $HD_i = (1 - a)bY_i + (1 - b)Y_i = Y_i(1 - ab)$. As in [Diamond \[2016\]](#), the price of land is determined by developers who are price takers and sell a homogeneous good at marginal cost $q_i = MC(CC_i, LC_i)$ where CC_i are local construction costs and LC_i are local land costs. In the asset market steady state equilibrium there is no uncertainty and prices equal the discounted value of local rents $q_i = r_i = i_t MC(CC_i, LC_i)$ where i_t is the interest rate in period t . The cost of land LC is a function of aggregate demand for local goods, following [Diamond \[2016\]](#) this is parametrized as a log-linear function: $q_i = i_t CC_i HD_i^\rho$, where ρ is the housing supply elasticity. Local construction costs and interest rates are taken as exogenously given.

Equating demand and supply we then have $Y_i(1 - ab) = H_i q_i = i_t CC_i (Y_i(1 - ab))^\rho H_i$ and therefore $H_i = Y_i^{1-\rho} (1 - ab)^{1-\rho} (CC_i i_t)^{-1}$. Note that $\rho = 1$ collapses to our previous “fixed supply” case. Subbing this equation for H_i into the previous derivations with $\bar{H}_i$ recovers a set up that will clearly result in the sufficient statistic result.

Intermediate goods

Following [Caliendo et al. \[2018\]](#) I can also allow for intermediate goods in production. Allow differentiated goods in each location of each type to use both labor and other goods in production in a Cobb-Douglas fashion such that γ_s and $1 - \gamma_s$ are the labor and intermediate goods factor intensities for sector s . Additionally allow γ_{sk} to denote the factor intensity of intermediate goods from sector k in production by sector s . Thus the marginal cost of production, and so price, is given by the following.

$$p_{is} = B_{is} (w_{is})^{\gamma_s} (P_{iE}^{\gamma_{sE}} P_{iN}^{\gamma_{sN}})^{1-\gamma_s} \quad (2.42)$$

Thus, we have now that output in a given sector-location cell is given by the following.

$$Y_{is} = B_{is} \left((w_{is})^{\gamma_s} (P_{iE}^{\gamma_{sE}} P_{iN}^{1-\gamma_{iN}})^{1-\gamma_s} \right)^{-\phi_s} MA_{is} \quad (2.43)$$

As all other equilibrium equations remain the same it's clear that the sufficient statistic result will be recovered.

Non-homotheticities and structural transformation

In the main text, and in the above extensions, I retain CES preferences and so shut down potential effect acting through structural transformation. It's possible that structural transformation towards sectors that require more educated workers, puts further upward pressure on wages and to the extent to which this force acts in response to changes in connectivity in some locations but not others, could impact spatial inequality of opportunity. In this subsection I show how non-homotheticities can be introduced into this framework. I show that market access terms remain a sufficient statistic under Stone-Geary preferences, although the log-linear relationship is broken.

To do this, I consider perhaps the simplest possible form of non-homotheticity, and generalise CES preferences into the Stone-Geary form. That is, I allow some constants γ_s to denote “subsistence” levels of consumption of goods of type s . Assuming that $\gamma_N > 0$ but

$\gamma_E = 0$ allows me to recover the desired result that the proportion of income spend on N type goods will fall as income levels rise. Generalizing preferences such that both types of consumer consume both types of goods but in the same proportion, we recover a new set of equations which define equilibrium.

$$\begin{aligned}
 u_{it}^s &= A_{it}^s \left(\left(\left(\frac{w_{it}^s - P_{it}^N \gamma_N}{P_{it}^E} \right)^\alpha \left(\frac{w_{it}^s - P_{it}^N \gamma_N}{P_{it}^N} \right)^{1-\alpha} \right)^{1-\beta} E_{it}^\beta \right)^{\lambda_s} \\
 E_{it} &= \frac{w_{it}^E}{w_{it}^N} \\
 w_{it}^s &= Y_{it}^s / L_{it}^s \\
 Exp_{it}^s &= B_{it}^s (w_{it}^s)^{-\phi_s} MA_{it}^s \\
 Exp_{it}^s &= \gamma_s + \alpha_s \left((w_{it}^E - P_{it}^N \gamma_N) L_{it}^E + (w_{it}^N - P_{it}^N \gamma_N) L_{it}^N \right) \\
 L_{it}^s &= u_{it}^s LMNA_{it}^s \\
 MA_{it}^s &= (P_{it}^s)^{-\phi_s} = \sum_j \tau_{ijt}^{-\phi_s} \frac{Exp_{jt}^s}{MA_{it}^s} \\
 LMA_{it}^s &= \sum_j \kappa_{ijt}^{-\lambda_s} \frac{L_{jt}^s}{LMA_{jt}^s}
 \end{aligned}$$

This new set up requires two additional parameters. α determines the proportion of residual income spend on E -type goods, and γ_N is the subsistence level for N -type goods. I set $\alpha = 0.4$ and $\gamma_N = 2.5$ following [Comin et al. \[2021\]](#). Solving this new system in differences using exact-hat algebra results in the following system which can be used to analyze

counterfactuals.

$$\begin{aligned}
(\hat{w}_{it}^s) \widehat{MA}_{it}^s &= \frac{\gamma_s}{X_{it}^s} + \frac{\alpha_s w_{it}^E L_{it}^E}{X_{it}^s} \hat{w}_{it}^E \hat{L}_{it}^E - \frac{\alpha_s \gamma_N (MA_{it}^N)^{-1/\phi_N} L_{it}^E}{X_{it}^s} (\widehat{MA}_{it}^N)^{-1/\phi_N} \hat{L}_{it}^E \\
&\quad + \frac{\alpha_s w_{it}^N L_{it}^N}{X_{it}^s} \hat{w}_{it}^N \hat{L}_{it}^N - \frac{\alpha_s \gamma_N (MA_{it}^N)^{-1/\phi_N} L_{it}^N}{X_{it}^s} (\widehat{MA}_{it}^N)^{-1/\phi_N} \hat{L}_{it}^N \\
\hat{L}_{it}^s &= \left(w_{it}^s (MA_{it}^E)^{1/\phi_E} \hat{w}_{it}^s (\widehat{MA}_{it}^E)^{1/\phi_E} \right. \\
&\quad \left. - \gamma_N (MA_{it}^N)^{-1/\phi_N} (MA_{it}^E)^{1/\phi_E} (\widehat{MA}_{it}^N)^{-1/\phi_N} (\widehat{MA}_{it}^E)^{1/\phi_E} \right)^{\alpha(1-\beta)\lambda_s} \\
&\quad \times \left(w_{it}^s (MA_{it}^N)^{1/\phi_N} \hat{w}_{it}^s (\widehat{MA}_{it}^N)^{1/\phi_N} - \gamma_N \right)^{(1-\alpha)(1-\beta)\lambda_s} \left(\frac{\hat{w}_{it}^E}{\hat{w}_{it}^N} \right)^{\beta\lambda_s} \widehat{LMA}_{it}^s (L_{it}^s)^{-1} \\
\widehat{MA}_{it}^s &= \sum_j \lambda_{ijt}^{MA_s} \hat{\tau}_{ijt}^{-\phi_s} \frac{(\hat{w}_{jt}^s) \widehat{MA}_{jt}^s}{\widehat{MA}_{jt}^s} \\
\widehat{LMA}_{it}^s &= \sum_j \lambda_{ijt}^{LMA_s} \hat{\kappa}_{ijt}^{-\lambda_s} \frac{\hat{L}_{jt}^s}{\widehat{LMA}_{jt}^s}
\end{aligned}$$

Where $X_{it}^s = \gamma_s + \alpha_s ((w_{it}^E - (MA_{it}^N)^{-1/\phi_N} \gamma_N) L_{it}^E + (w_{it}^N - (MA_{it}^N)^{-1/\phi_N} \gamma_N) L_{it}^N)$. Endogenous variables are thus (implicitly) functions of market access terms only, preserving the sufficient statistic result although in a non-linear manner.

Generalizing the education choice problem

In the main text I focus on local returns to education as the driving force behind household decisions to educate their children or not. Above I also consider generalizing this to allowing the local cost of education to vary and potentially do so in response to changes in local connectivity. Here I show formally how this process can further be generalized to include other dimensions relevant for a households education choice problem in a very straight forward manner that remains consistent with the sufficient statistic result. Suppose the utility value of education depends on (i) returns to education, (ii) household income, (iii) the opportunity cost of education, and (iv) the actual cost of education, in the following manner.

$$E_{it}^s = \underbrace{\left(\frac{w_{it}^E}{w_{it}^N} \right)^{\beta_1}}_{\text{Returns to Education}} \cdot \underbrace{(w_{it}^s)^{\beta_2}}_{\text{Income}} \cdot \underbrace{\left(w_{it}^N \right)^{\beta_3}}_{\text{Opportunity cost}} \cdot \underbrace{(c_{it})^{\beta_4}}_{\text{Cost}} \quad (2.44)$$

Returns to education are discussed at length in the main text, and are captured succinctly

by the relative wages those with vs without primary education can demand in a given location. Income effects are captured by β_2 which scales the income of the parents of type s . The opportunity cost of schooling is increasing in the (shadow) wages children could receive if they were not in school. This is assumed to be proportionate to the local wages earned by those without primary schooling. Finally, the actual cost of education, c_{it} , can also be included as a determinant, and it's relevance is discussed at some length in section 2.D.2 where I endogenise education supply and cost.

As equation 2.44 is a function of wages only (with the exception of costs), and as wages are themselves an endogenous variable depending only on market access terms — it's clear that these generalizations will also result in the same sufficient statistic result.

2.D.3 Quantitative Spatial Economics model example

In this section, I write down a quantitative spatial economics model, taken from the broad class of data-generating processes consistent with the sufficient statistic result described above. This model can therefore be seen as a particular case of the more general setup described in section 2.2. I follow Morten and Oliveira [2021], Tsivanidis [2019], Ahlfeldt et al. [2015] and others in writing down a model where individuals have idiosyncratic Fréchet preferences over location amenities and firms draw Fréchet productivity parameters.

First, I parameterize the household utility function. Recall that households gain utility from consuming, local amenities and sending their child to complete primary school, and face negative iceberg transport costs. Households differ ex-ante in their origin location, location preferences, and type. They have utility over an aggregate consumption good C_{jt}^E , this utility is scaled by movement cost weighted local amenities b_{jt}/κ_{ijt} which consist of an exogenous component and a component determined by education opportunities in the region. E and N type good bundles are CES aggregates of a continuum of differentiated consumption goods where firms in each location produce differentiated varieties. Recall households are of type s , are from location, i , move to location, j , and are operating in period, t .

$$U_{ijt}^s = \underbrace{\frac{b_{jt}}{\kappa_{ijt}}}_{\text{Location}} \underbrace{\left(C_{jt}^s\right)^{1-\beta}}_{\text{Consumption}} \quad (2.45)$$

Households seek to maximize 2.45 subject to the budget constraint given below.

$$P_{jt}^s C_{jt}^s = w_{jt}^s \quad (2.46)$$

Where P_{jt}^s is the Dixit-Stiglitz price aggregator and household wealth varies by household type. Each household type commands wages from it's own industry.

In the above equation $b_{jt} = \bar{b}_{jt}E_{jt}$ denotes an individuals draw of local amenities where $E_{jt} = (r_{jt})^\beta$ is the value of educating your child. κ_{ijt} is an iceberg-style movement cost, and α is the proportion of income spent on E-type goods. r_{jt} denotes the value of their child's returns to education to the household. This could be specified in many ways, a completely forward-looking household would denote this as the actual utility the child earns. I take a more naive, and realistic, view that parents value their child's education as proportional to the current local return to education and write $r_{jt} = \frac{w_{jt}^E}{w_{jt}^N}$.

I assume that b_{jt} is type two extreme value (Fréchet) distributed as follows:

$$b_{jt} \sim G_{jt}^B, \quad G_{jt}^B(b) = e^{-T_{jt}^{B_s} b^{-\lambda_s}} \quad (2.47)$$

$\{T_{jt}^{B_s}\}$ governs the centrality of the distribution, that is the absolute advantage of place and λ_s governs the dispersion of each distribution. This setup produces a gravity equation for migration across space of exactly the same form as given in the sufficient statistic result. Given the above distributional assumption and the timing of decisions, we can solve recursively first by considering the consumption problem and then the location problem.

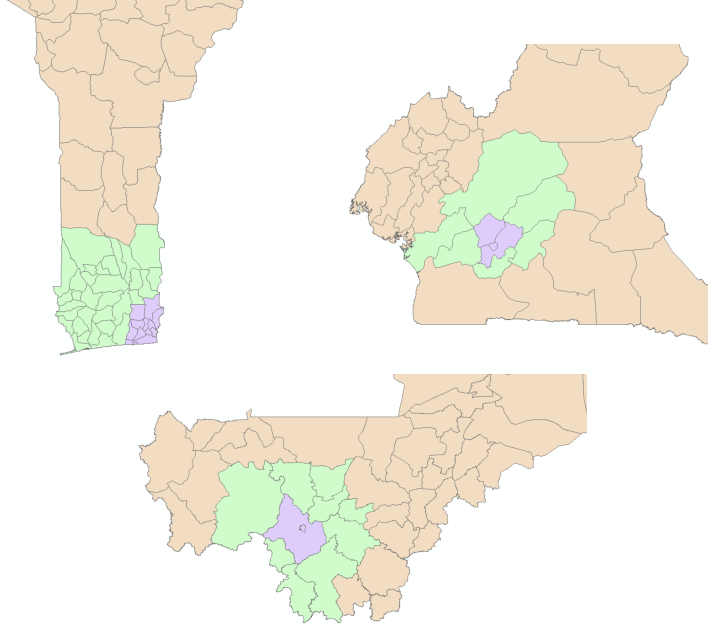
Firms produce a differentiated consumption variety q and production occurs under perfect competition in each location. Labor is the only factor of production. As discussed above there are two industries/ sectors, firms that produce E type goods and those that produce N type goods. Sectors are differentiated in two ways. First, sector s firms only employ individuals with s type education. Second, although firms in both sectors draw idiosyncratic productivities from type two extreme value distributions, these distributions are based on different parameters. Type s firms draw productivities z from $G_{jt}^{z^s}(z) = \exp\{-T_{jt}^{z^s} z^{-\phi_s}\}$. Firms have the technology:

$$Y_{jt}^s(q) = Z_{jt}^s(q)L_{jt}^s \quad (2.48)$$

The price in j of a variety q good from sector s produced in i is given as a constant markup over marginal cost $p_{ijt}^s(q) = \frac{\tau_{ijt}}{z_{it}^s(q)} w_{it}^s$. This set up emits the standard gravity equation as in section 2.2.

2.D.4 Future road upgrading counterfactuals details

FIGURE 2.D.2 Defining road types



Notes: This figure shows graphically the approach taken to categorized roads into *primate* or *hinterland* roads. Primate roads connect hinterland (beige) localities to the primate (purple) locality. Hinterland roads connect two hinterland localities. The *other* category subsumes all roads which don't fall into the previous two categories.

2.D.5 Road-locality level analysis

I run regressions at the locality (i) road (r) level where $\hat{\mu}_{it}$ is the relative change in local educational opportunity in locality i due to upgrading road r . Equation 2.49 summarizes the regression specification.

$$\hat{\mu}_{ir} = \beta_{tk} \cdot \text{RoadType}_{t(r)} \cdot X_{ki} + \alpha_r + \gamma_{c(i)} + \varepsilon_{ri} \quad (2.49)$$

where X_{ki} is a vector of locality level characteristics consisting of: primary completion rate, agricultural employment rate, housing quality index, and expected travel time (remoteness). α_r and $\gamma_{c(i)}$ denote road and country level fixed effects. β_{rk} are the coefficients of interest and allow locality characteristic k to differentially impact the effect road building has on local opportunity by road type t . In table 2.D.3 I show the results from running equations of the form given in 2.49, column (1) looks at locality characteristics without allowing coefficients to vary by road type, column (2) pools all data and looks at interaction effects and columns (3), (4), and (5) consider impacts separately by country.

Column (1) of table 2.D.3 suggests that localities with lower primary completion rates, higher agricultural employment rates, higher housing quality, and higher expect travel times (more remote) are more positively effected by road upgrading on average over all possible road upgrades. In column (2) I brake these average effects down by road type where roads are categorized as connecting hinterland locations to the primate city (primate), hinterland locations to other hinterland locations, or “other” — reported effects are relative to the “other” category. I find that hinterland connections more negatively affected high primary completion areas, but primate connections had a more positive effect. Both types of road had relatively more negative impacts on localities with high agricultural employment rates. Primate connections were worse for localities with high housing quality, and hinterland connections were mildly better. Finally, primate connections were worse for more remote locations.

Finally, columns (3), (4), and (5) break down the impact of types of road and locality characteristics by country, that is by *network*. In general, the large network-specific differences highlighted above can also be seen here. I find that the large negative impact higher primary completion rate locations experience with hinterland connections, and positive impacts with primate connections are driven by Cameroon — with impacts in Mali not varying. On the other hand although locations with higher agricultural employment rates in both Benin and Cameroon are more negatively affected by primate locations, in Cameroon such locations are more positively effected by hinterland connections and in Benin the result remains negative. Again, there is no effect in Mali. Again the result linking higher housing quality with larger effects for both types of roads is driven by Cameroon, with Benin seeing more muted impacts and Mali little action. Finally, considering remoteness, I find considerable heterogeneity across locations. In Benin primate roads had little impact on more remote locations relative to less, whereas in both Cameroon and Mali this impact was significantly negative. On the other hand in Benin hinterland connections hurt more remote locations considerably more, but in Cameroon the effect is positive and in Mali there is little impact at all.

TABLE 2.D.3 The impact of locality-road level characteristics on μ_{ir}

	(1)	(2)	(3)	(4)	(5)
	Overall	By Road Type	By Road Type Benin	By Road Type Cameroon	By Road Type Mali
Primary completion rate	-0.0397*** (0.00253)	0.0153** (0.00632)	0.00270 (0.0108)	0.0188*** (0.00601)	-0.115*** (0.0242)
Primate $\times$ Primary completion rate		0.0230** (0.00950)	0.00551 (0.0283)	0.0369*** (0.00834)	-0.0101 (0.0522)
Hinterland $\times$ Primary completion rate		-0.0739*** (0.00691)	-0.0214* (0.0114)	-0.0666*** (0.00683)	0.0257 (0.0297)
Agricultural employment rate	0.0341*** (0.00287)	0.0708*** (0.00698)	0.0917*** (0.00944)	-0.0941*** (0.0114)	-0.00986 (0.0125)
Primate $\times$ Agricultural employment rate		-0.0396*** (0.0125)	-0.0780*** (0.0239)	-0.0376** (0.0168)	-0.00984 (0.0242)
Hinterland $\times$ Agricultural employment rate		-0.0466*** (0.00769)	-0.0728*** (0.00996)	0.0473*** (0.0124)	0.00692 (0.0155)
Housing quality index	1.296*** (0.0967)	0.986*** (0.237)	0.771*** (0.295)	-5.898*** (0.489)	3.125*** (0.367)
Primate $\times$ Housing quality index		-1.674*** (0.450)	-1.202 (0.750)	-2.796*** (0.752)	1.038 (0.709)
Hinterland $\times$ Housing quality index		0.500* (0.261)	-0.622** (0.312)	5.321*** (0.527)	-0.134 (0.470)
Log(Expected travel time)	1.907*** (0.120)	2.318*** (0.267)	14.30*** (0.505)	-5.150*** (0.521)	-0.899*** (0.304)
Primate $\times$ Log(Expected travel time)		-4.242*** (0.485)	0.828 (1.349)	-3.002*** (0.721)	-3.758*** (0.564)
Hinterland $\times$ Log(Expected travel time)		-0.179 (0.302)	-6.457*** (0.538)	6.015*** (0.596)	-0.613 (0.385)
Observations	25744	25744	7144	10140	8460
R^2	0.069	0.092	0.584	0.180	0.054

Notes: This table shows the road-locality level impacts of upgrading each road on the relative local educational opportunity in a given locality. Coefficients are recovered from estimating the following equation:

$\hat{\mu}_{ir} = \beta_{ik} \cdot \text{RoadType}_{i(r)} \cdot X_{ki} + \alpha_r + \gamma_{c(i)} + \varepsilon_{ri}$ and in columns (2) to (5) road-type specific effects are given relative to the left out category *other*. Column (1) estimates effects on locality level characteristics only, column (2) interacts such effects with the category of road being upgraded. In columns (3), (4), and (5) I then restrict the sample to each country in turn. Standard errors are clustered at the road level and are given in parenthesis below estimates.

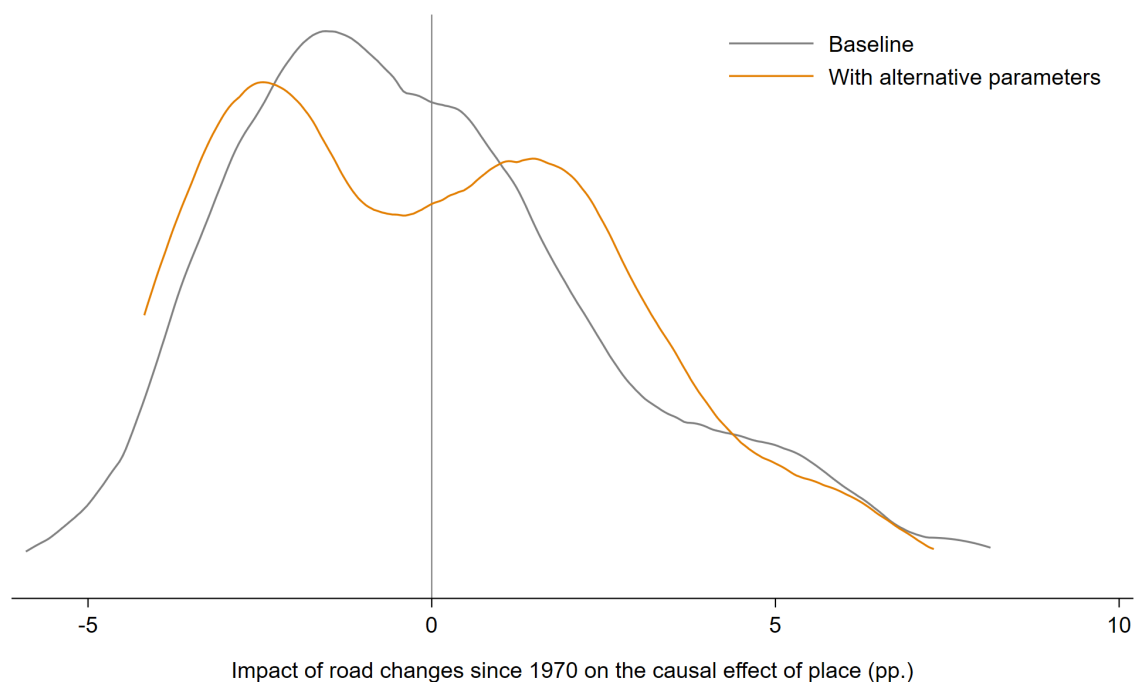
2.D.6 No-roads counterfactual results with alternative parameters

Using parameter values taken from the literature

In this section I compare my baseline results to those under the alternative parameterization taken from the literature and presented in the main text. I make this comparison in the counterfactual where I consider the total impact of roads built since 1970. The bottom line is that there is little qualitative (or quantitative) difference between results under either set of parameters — that is my main findings are robust to my structural parameter identification procedure.

The correlation between baseline results and results under the alternative parameterization is .87. Figure 2.D.3 compares the distribution of effects using the baseline and alternative parameters. Table 2.D.4 then compares the relationship between total affects and baseline 1970 remoteness in both parameterisations. Although differences are apparent in both table and figure, they are not sufficient to alter the qualitative conclusions from either presented in the text.

FIGURE 2.D.3 Compare the distribution effects of road building since 1970 under the baseline and alternative parameterization



Notes: This figure shows the distribution over locations of the causal effect of road building since 1970 on local opportunity. In gray I plot baseline results, replicating figure 2.4.1 pooling over all three countries. In orange I plot the analogous results with the model estimated using the alternative parameters.

TABLE 2.D.4 Relationship between the total relative effect of roads built since 1970 and locality remoteness in 1970. Comparing the baseline and alternative parameterization

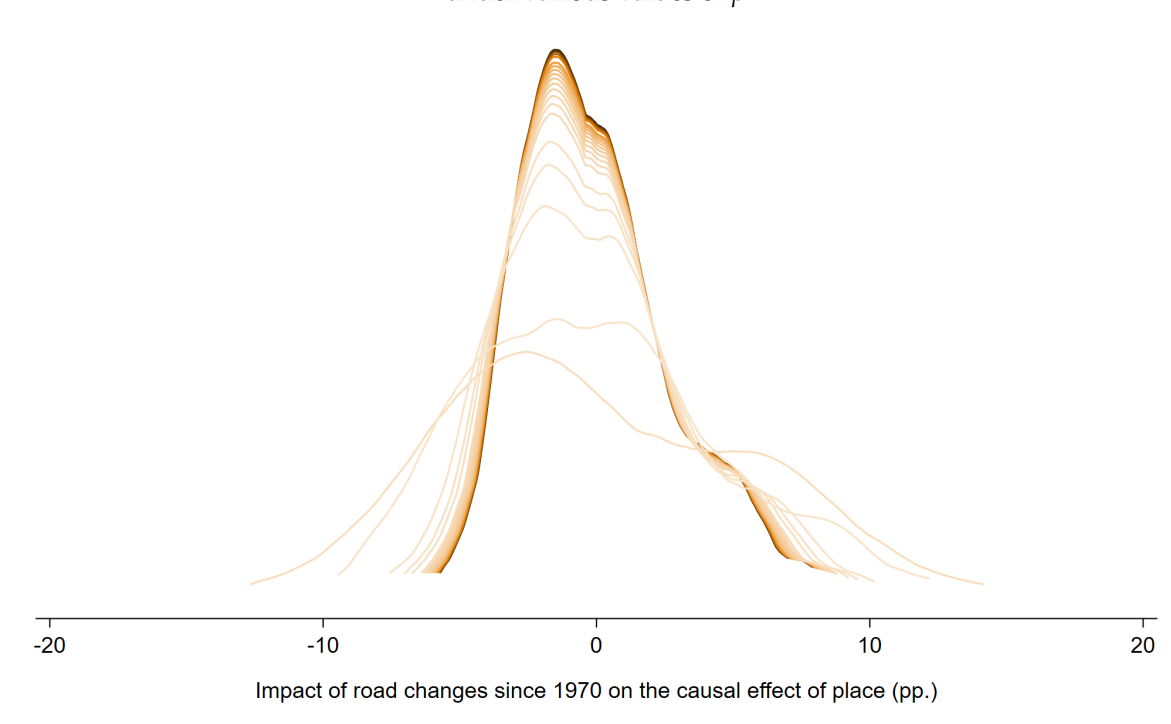
	(1) Baseline	(2) With alternative parameters
Benin $\times$ 1970 Remoteness	5.092*** (1.357)	9.472*** (1.300)
Cameroon $\times$ 1970 Remoteness	6.995*** (2.409)	9.596*** (1.664)
Mali $\times$ 1970 Remoteness	5.468*** (1.299)	7.491*** (0.876)
Observations	156	156
R^2	0.334	0.642

Notes: This table correlates 1970 remoteness with the total impact of road building since 1970 on local opportunity. Column one replicates results from figure 2.4.2 in the main text. Column two presents analogous results under the alternative parameterization.

Alternative values of β

Figure 2.D.4 shows the results from estimating the effects of road building since 1970 on the spatial distribution of opportunity, varying the β parameter. Recall from the main text that β governs the relative importance of education to consumption in household utility — higher β means education is valued more. To produce figure 2.D.4 I pick a value of β , re-estimate the remaining structural parameters using the sufficient statistic result and then using the new set of parameters use the model to estimate the total effect of road building. Figure 2.D.4 shows lower values of beta in darker orange (starting at 0.01) and higher values in lighter orange (ending at 0.95). Although high values of β are clearly associated with a more dispersed distribution, it's clear from figure 2.D.4 that for any reasonable values of β , results are not significantly changed.

FIGURE 2.D.4 Comparing the distributional effects of road building since 1970 under various values of β



Notes: This figure shows the distribution over locations of the causal effect of road building since 1970 on local opportunity. Each line gives the results for a different value of β , darker orange lines correspond to lower values (starting at 0.01) and higher lines to higher values (ending at 0.95). Lines move in increments of 0.05.

Chapter 3

The Butcher, the Brewer, or the Baker: The Role of Occupations in Explaining Wage Inequality

Abstract

Google does not pay its canteen chefs the same premium as it does its software engineers; there is heterogeneity in pay premia within a firm. To study the implications of this observation, we estimate a two-way, worker-job, fixed effect model that builds upon the canonical model by [Abowd et al. \[1999\]](#) by allowing for within-firm heterogeneity. Estimating our model on UK administrative data, we find that even if workers, and firms, were identical, 25% of current log-wage variance would remain, due solely to heterogeneity between occupations. We further document that worker heterogeneity accounts for a relatively small proportion of log-wage variance, and that between-firm pay heterogeneity is more important in higher-wage occupations and larger labour markets.

3.1 Introduction

Google does not pay its canteen chefs the same premium as it does its software engineers. As a result, the best engineers go to Google, but the best chefs work at Michelin star restaurants: workers match to jobs, not firms. This simple observation implies that the canonical model of wage decomposition pioneered by [Abowd et al. \[1999\]](#) (hereafter AKM) – which proposes a statistical model that explains log-wages as the sum of a set of worker fixed effects, firm fixed effects, and idiosyncratic errors – is misspecified because it implies that firms pay the same wage premium to all their workers¹.

In this paper, we argue that AKM models lead to the omission of an important dimension of inequality – inequality of wages across occupations. This omission leads to the overstatement of the degree of wage heterogeneity explained by individual differences, and the understatement of significant heterogeneity of earnings across occupations within firms. This implies that even if we could eliminate all wage dispersion between firms or all wage dispersion between individuals, there would still be substantial wage dispersion due to differences in occupations.

We extend the canonical framework to allow different firms to pay occupations different wage premia, by estimating a model with individual and *job* fixed effects, where jobs are defined as occupation-firm pairs. Our model embeds the AKM model as a particular case but is more flexible by allowing firms to pay different wage premia to workers in different occupations. To estimate such a model we use a two-stage approach suggested by [Bonhomme et al. \[2019\]](#) to overcome limited mobility bias ([Andrews et al. \[2012\]](#), [Jochmans and Weidner \[2019\]](#)), first clustering jobs, and secondly employing worker job-cluster fixed effect regressions.

Estimating worker-job wage variance decompositions using matched employee-employer data from the UK², we show that the share of overall log-wage variance due to individual heterogeneity is about 18%. This is significantly lower than what is typically found in AKM

¹The line of research pioneered by [Card et al. \[2013\]](#) has largely emphasised the role played by firms in explaining wage inequality [[Song et al., 2019](#), [Card et al., 2018](#), [Bonhomme et al., 2019](#), [Alvarez et al., 2018](#)]. The point that the model used is misspecified is taken up by recent theoretical work by [Haanwinckel \[2020\]](#) and [Bloesch et al. \[2021\]](#), which both argue that workers who are in some sense more important to a firm are paid more within it.

²We use the Annual Survey of Hours and Earnings (ASHE) data in this paper. The UK offers a particularly interesting setting for several reasons. First, although significant prior work has studied UK wage inequality [[Faggio et al., 2010](#), [Mueller et al., 2017](#), [Machin, 2011](#), [Lee et al., 2016](#), [Rice et al., 2006](#), [Schaefer and Singleton, 2020](#)] to the authors' knowledge this is the first time large-scale employee-employer data has been brought to bear on the question using AKM style wage decompositions. Second, the UK typifies many debates surrounding wage inequality in western economies with spatial and occupational wage inequality squarely positioned in everyday political and social discourse.

regressions. [Card et al. \[2013, 2016\]](#), [Song et al. \[2019\]](#), [Alvarez et al. \[2018\]](#), [Bonhomme et al. \[2020\]](#) all find that between 50% and 60% of log wage variance is explained by worker differences. We show this reduction relative to worker-firm decompositions is to be expected theoretically if more productive workers match to more productive occupations within firms. If the true model has varying job wage premia within a firm, then estimating an AKM model would cause this varying occupation premia to become an omitted variable, which will be loaded onto the estimated worker FE and firm FE. This results in AKM estimates of log-wage variance decompositions to attribute job premia heterogeneity to worker heterogeneity.

Consequently, our estimates of the role of job heterogeneity are larger than is typically found in the AKM context for firm heterogeneity. 38% of the log-wage variance can be explained by job heterogeneity, and 29% can be explained by positive assortative matching between workers and jobs. Both of these components are considerably larger than those found in firm-worker decompositions. In bias-corrected specifications, [Bonhomme et al. \[2020\]](#) find that firm heterogeneity accounts for between 5% and 13% across various European countries and the US, and that sorting between workers and firms accounts for at most 20%. Thus, previous work somewhat puzzlingly finds that sorting is strong even when the associated differentiation between firms is relatively less significant. Our results resolve this puzzle by showing both considerable job heterogeneity and considerable positive assortative matching.

We attempt to reconcile our results with those of the AKM literature by decomposing the variance in job fixed effects and in sorting into that which can be attributed to between-occupation variance and within-occupation variance. Variation within occupation across firms captures heterogeneity in firm-level pay premia, and a weighted average of this variation is conceptually comparable to the variance of firm fixed effects estimated in a standard AKM decomposition. When considering the variation in job fixed effects, we find that the between-occupation component is twice as important as the within-occupation component and that they account for 25% and 13% of overall log wage inequality respectively. Therefore, most variation is a between- rather than within-occupation phenomenon. We find similar effects when decomposing sorting with 20% explained by sorting between occupations and 8% by sorting within occupations. These results highlight the importance of occupations in explaining log wage inequality but do not completely discount the role of firms.

Armed with worker and job fixed effects, we test and reject the assumption in the canonical model that firms pay the same wage premia over occupations. We reject the null hypothesis at traditional levels of statistical significant, but also show that differences are of a

magnitude that is economically significant. We then study the distribution of match effects between occupations and firms by computing the differences between our framework and a naive framework that uses linear worker, firm, and occupation fixed effects. We find that there are substantial match effects between occupations and firms; low-skill occupations have a negative match effects at most firms, while high-skill occupations have a positive match effect at the most productive firms. This result supports the results by [Haanwinkel \[2020\]](#) and [Bloesch et al. \[2021\]](#) that firms pay disproportionate large wage premia to certain, typically high-skill, workers, and disproportionately small wage premia to typically low-skill workers.

This result naturally points to the hypothesis that since wage premia are especially large for high-wage occupations, it is likely that occupational labour markets in those occupations ought to be characterised by more heterogeneity and sorting. To investigate this, we analyse the sources of wage variance in each occupation. Looking at differences across occupations we find that sorting between workers and firms is only a substantial component of wage variance within higher-wage occupations. In relatively lower-wage occupations, the sorting component is negligible. This evidence suggests that models which emphasise positive matching between workers and firms are more appropriate for characterising high-wage occupations. In lower-wage occupations where it appears workers and jobs are not so heterogeneous, search is likely to be less important.

Turning to heterogeneity across local labor markets and using a methodology first developed in [Dauth et al. \[2018\]](#), we show that high-paying jobs and workers co-locate in larger labour markets. This finding lends evidence to a large literature in spatial and urban economics which highlights agglomeration economies and/or sorting across space leading to a city wage premium (see [D'Costa and Overman \[2014\]](#), [Glaeser and Mare \[2001\]](#), and [Rice et al. \[2006\]](#) for example). We also find that the variance of wages is increasing in local labour market size and sorting, especially within-occupation, explains an increasingly large proportion of this variation with size. In the smallest labour markets, almost none of the wage variance is explained by within-occupation sorting (between individuals and firms within an occupation) whereas in the largest labour markets over 10% is. Differences in pay between firms conditional on occupation and worker are only a feature of large labour markets. To our knowledge, this result is new to the literature.

Our work relates to the large and active literature exploring the factors underlying wage inequality using the two-way fixed effects approach. This literature was initiated by [Abowd et al. \[1999\]](#) and [Abowd et al. \[2002\]](#), using French data. More recently, [Card et al. \[2013\]](#)

used this framework to argue that increasing wage inequality in recent years was driven by increasing dispersion in firm wage premia. Their result was replicated in the American context by [Song et al. \[2019\]](#). [Alvarez et al. \[2018\]](#) find that the implementation of the minimum wage drove reductions in wage inequality in the Brazilian context. Recent econometric developments in addressing the limited mobility bias faced by two-way fixed effect regressions in relatively sparsely connected networks ([Andrews et al. \[2012\]](#), [Jochmans and Weidner \[2019\]](#), [Kline et al. \[2020\]](#), [Bonhomme et al. \[2019\]](#) and [Bonhomme et al. \[2020\]](#)) have suggested that the true share of log-wage variance explained by firm effects is potentially much smaller than the 20-25% found in the earlier literature. [Dauth et al. \[2018\]](#) use the AKM framework to study geographical income inequality in the German context.

We contribute to this literature by emphasising that if we allow firms to pay workers in different occupations varying premia, as suggested by [Haanwinckel \[2020\]](#) and [Bloesch et al. \[2021\]](#), we obtain qualitatively very different results from the AKM approach. In particular, while results from most AKM regressions find that wage inequality is largely due to individual heterogeneity, we find that this proportion is substantially smaller when we consider between-occupation moves. Instead, a lot of wage inequality is due to disparate pay between occupations, and this is particularly important in explaining wage inequality in small geographical labour markets. While there is less within-occupation between-firm variation in pay for smaller labour markets and low-wage occupations, there are large match effects between high-wage occupations and high-premium firms, and thus within-occupation variation is more important for higher-paying occupations. Partly due to these occupations typically being located in larger labour markets, and partly due to a market size effect, we also find that within-occupation heterogeneity between firms only matters substantially in larger geographical labor markets.

The rest of our paper proceeds as follows. Section 3.2 describes the specification and identifying assumptions underlying our model. Section 3.3 discusses the UK data that we use. We present our main results relating to the log wage variance decomposition in section 3.4. Section 3.5 considers differences across occupations. We discuss the implications of our model for wage inequality across space in section 3.6. Section 3.7 concludes.

3.2 Econometric model

In our econometric framework, we aim to disentangle log-wage variation into the components due to worker specific heterogeneity, job specific heterogeneity, and that due to the

sorting of workers into jobs. For each individual $i \in \mathcal{I}$ in each period $t \in \mathcal{T}$ we observe the firm $F(i, t) \in \mathcal{F}$ and occupation $O(i, t) \in \mathcal{O}$ they work in as well as the wage w_{it} received. We denote by $J(i, t)$ the job an individual i is employed doing, in period t where j is a firm-occupation pair i.e. $J(i, t) \in \mathcal{J} = \mathcal{F} \times \mathcal{O}$. We assume that log wages $\ln(w_{it})$ can be described by the sum of worker, α_i , and job, $\psi_{J(i,t)}$, components³ as well as an idiosyncratic error as shown in equation 3.1.

$$\ln(w_{it}) = \alpha_i + \psi_{J(i,t)} + \varepsilon_{it} \quad (3.1)$$

This setup follows the literature started by Abowd et al. [1999] and continued by others such as Card et al. [2013] and Song et al. [2019]. In this model person fixed effects α_i can be interpreted as a combination of skills and other factors that are awarded equally across jobs j . Similarly, ψ_j can be interpreted as a pay premium offered to all workers in job j irrespective of unobserved worker-specific characteristics. Note that the usual worker-firm models (as opposed to worker-job) use ψ_f instead of ψ_j and interpret it as a pay premium offered to all workers at a specific firm. This means that, for example, software developers, administrators, managers, cleaners, and chefs at Google all receive the same log-additive pay premium.

Our model nests this as a special case where all occupations within a firm have the same ψ_j , but in general is significantly more flexible by allowing this premium to vary within a firm by occupation. For example, this would allow Google to pay its software developers a higher premium than its chefs and the Ritz to pay a higher premium to its chefs than its software developers. Given the nested structure we can test formally whether firms do indeed offer the same wage premium across all occupations and do so in section 3.4.3 finding evidence of statistically and economically significant varying premia.

The fixed effects in the AKM model are identified by *movers* between firms. In the same way, the job fixed effects are identified by the same individual moving from one job to another and noting how their wages changed due to the move, and individual fixed effects were identified by observing workers' wages between jobs relative to other workers in those jobs. Denote by y the stacked vector of individual wages, $A = [a_1, \dots]$ the design matrix of individual indicators, $J = [b_1, \dots]$ the design matrix of job indicators and ε the stacked vector of error terms. Then we can write equation 3.1 as $y = A\alpha + J\psi + \varepsilon$ or more succinctly as $y = Z\beta + \varepsilon$ where $Z = [A, J]$ and $\beta = [\alpha, \psi]$. We will estimate equation 3.1 by OLS to find $\hat{\beta} = \beta + (Z'Z)^{-1}Z'\varepsilon$. This result highlights the two main hurdles facing identification. First

³Throughout we additionally condition on a set of time varying worker covariates X_{it} such as age, these have been omitted in this discussion for notational brevity.

note that implicit in the above derivation is the invertibility of $Z'Z$. This is not the usual innocuous assumption; in general $Z'Z$ is not invertible in this set up if there are *unconnected* individuals or firms. That is, $Z'Z$ is only invertible in the subset of the data representing the largest connected set. In our data the largest connected set covers over 90% of observations and will be the sub-sample that we use for estimation (this is summarised in table 3.1).

The specification in equation 3.1 allows a simple decomposition of the observed variance of log wages as given in equation 3.2 below.

$$\mathbb{V}[\ln(w_{ijt})] = \underbrace{\mathbb{V}[\alpha_i]}_{\text{Variance due to individual heterogeneity}} + \underbrace{\mathbb{V}[\psi_j]}_{\text{Variance due to job heterogeneity}} + \underbrace{2 \cdot \text{Cov}[\alpha_i, \psi_j]}_{\text{Variance due to workers sorting into jobs}} + v \quad (3.2)$$

This shows how, for each period, wage variance can be decomposed into four components. $\mathbb{V}[\alpha_i]$ captures the variance due to individual characteristics, and $\mathbb{V}[\psi_j]$ similarly captures the variance due to job characteristics. Finally, $2\text{Cov}(\alpha_i, \psi_j)$ captures the component of the variance which is due to sorting or matching of individuals to jobs, and v captures unexplained variance.

3.2.1 Limited mobility bias

AKM type decompositions suffer from limited mobility bias which was recently found to be qualitatively and quantitatively of importance (Bonhomme et al. [2020]). In our dataset for example, although we have almost 1.4 million observations, equation 3.1 requires estimating 303,000 worker fixed effects and 282,000 job fixed effects. However, the problem is significantly worse than it at first appears. This is because intuitively job fixed effects are only identified from *movers* between jobs — of which there are even fewer. While this doesn't lead to biased estimates of worker or job fixed effects, it does lead to biased second order (or higher) moments calculated from the resulting estimated fixed effects⁴.

In the main body of this paper, we follow the two-step approach suggested in Bonhomme et al. [2019] and Bonhomme et al. [2020] to overcome the limited mobility bias, where we first reduce the dimensions of the problem by classifying jobs into L job classes using a k-means algorithm⁵. This gives us a very connected network of workers and job classes, on which we can estimate a two-way linear regression on workers and job classes and related variance terms with little bias. In the first step, we cluster jobs into L classes

⁴The econometrics of this bias are reviewed in the appendix in section 3.4.

⁵K-means is an unsupervised machine learning algorithm that classifies data into K groups by minimising each within-group sum of square differences from the corresponding group mean.

using an unsupervised machine learning algorithm, k-means clustering based on 20 points (percentiles) of the job wage distribution. We find that these thirty clusters capture some 85% of the underlying variation in job wages; this is encouraging evidence of the success of the dimensionality reduction with minimal information loss⁶. In the second step we then estimate equations of the form given in 3.1 using job-class rather than job fixed effects.

3.2.2 Endogeneity and specification tests

To uncover unbiased estimates of the fixed effects, we require that $\mathbb{E}[Z'\varepsilon|Z] = 0$, the classic exogeneity assumption in OLS. The validity of this assumption for the classic AKM framework was discussed in detail in Card et al. [2013], and in our paper, we focus on what changes in our context relative to the standard AKM model. As in Card et al. [2013], we focus on whether the assignment of workers to jobs satisfies exogeneity, i.e. $\mathbb{E}[J'\varepsilon|Z] = 0$, and consider the sufficient condition that $\forall j, \Pr(J(i, t) = j|\varepsilon) = \Pr(J(i, t) = j)$.

The assumption made is not violated if the labour market is characterised by sorting on individual fixed effects and job fixed effects, i.e. if workers move to higher paying jobs or firms sort to more productive workers. Rather, there are two ways in which the assumption could be violated. First, the assumption is violated if workers sort to jobs on the basis of a worker-job match-specific characteristic not captured by α_i and $\psi_{J(i, t)}$. Relative to the AKM specification, this is more of a concern, as the skills necessary for different occupations are likely to be non-unidimensional (as would be implied in our model) (see e.g. Lindénlaub [2017] and Lise and Postel-Vinay [2020] for work arguing for the importance of multi-dimensional skills). Thus, there may be dimensions of individual skill not captured by the fixed effect that generate match-specific surplus at particular occupations which are pushed to the error term.

A second possible problem is that temporary variation in wages may be correlated to the job that workers perform. A concern of this type in Card et al. [2013] is that the statistical model is incompatible with models of the labour market where workers move to jobs due to high transitory wage offers not related to the firm fixed effect⁷. In our model, an additional problem of this kind is if temporary occupation-specific productivity shocks lead to substantial movement between occupations in the period.

An upshot of both concerns is that wage changes due to moves in one direction should be larger than wage changes due to moves in the other. If workers sort to jobs on the basis of

⁶All results presented here are not sensitive to the number of clusters used. We have replicated very similar findings using 10, 20, 50, and 100 clusters.

⁷A prominent model with such a mechanism is Postel-Vinay and Robin [2002].

elements not captured in either the worker fixed effect nor the firm fixed effect, we should expect the relative wage change of a worker moving from a low fixed effect job to a higher fixed effect job to exceed in magnitude the relative wage change of workers moving the other way. On the other hand, if the exogeneity assumptions holds, the average wage change of movers from a less productive job to a more productive job should be equal to the negative of movers from the more productive job to the less productive job.

We check for this endogeneity in appendix 3.2, using methods developed in Card et al. (2018) in the context of the AKM model. The main idea behind these tests is to check if wage changes in moves of workers between job clusters is symmetric. We find that wage changes between estimated job clusters are approximately symmetric.

Finally, another possible criticism of this approach is that the two-way fixed effect regression imposes an inappropriate log-additive functional form on firm and job fixed effects. The log-additive structure we impose on the data could mask important heterogeneity, mechanisms or disallow potentially relevant theoretical channels. However, we feel relatively assured that, even when taking a coarse job clustering approach⁸, most of the variation of interest is captured by our approach. This is due to two reasons. First, a relatively small part is played by unexplained variation in wage decompositions (roughly 8%). Second, we follow the approach in Card et al. (2013) and show that the mean of the residuals of the two-way FE regression are near zero on average, and by job and worker cells. This exercise and its results are summarised in appendix 3.3. We find that the specification performs well, except for very low-skill workers at very low wage premia jobs. This could be due to institutional factors like minimum wage that are imperfectly captured by this model. This finding is qualitatively similar to what other authors find when analysing the suitability of linearity in the context of more standard AKM models (e.g. Card et al. [2013], Bonhomme et al. [2019]).

3.3 Data and empirical setting

We apply the empirical specification discussed in section 3.2 to matched employee-employer data from the United Kingdom. We study the UK for two main reasons. First, to the best of our knowledge, this is one of the first studies (other than Schaefer and Singleton [2020]) to bring employee-employer data and AKM-like decompositions to bear on the question

⁸We check the robustness of the results to various numbers of clusters ($K = 30, 50, 100$) and clustering algorithms and find no difference.

of UK wage inequality. Second, the UK typifies contemporary debates surrounding wage inequality in aggregate as well as across space, common in developed economies.

Our main data source for this paper is the UK Annual Survey of Hours and Earnings (hereafter ASHE). The ASHE is a matched employee-employer data set collected annually by surveying employers and comprises of a 1% sample of the government tax agency's job records. The survey, which is administered in April each year, records information about the identity of the establishment, its location and the industry classification of its primary business, the worker's identity, sex, and home address, and details about the job itself, including earnings, hours, and occupational classifications. The survey is administered by the UK Office for National Statistics (ONS) and response to the ASHE is compulsory. The survey provides a panel data set as the same individuals are tracked over time, and as filling in the ASHE survey is a legal requirement of the employer attrition is very low. The ASHE can be compared to other data sets used in the AKM literature, such as the Integrated Employment Biographies data in Germany (used by [Card et al. \[2013\]](#)), the Master Earnings File in the USA used in [Song et al. \[2019\]](#), and a mix of microdata sets including the ANST in Sweden used in [Bonhomme et al. \[2019\]](#).

Details of our data cleaning process can be found in appendix 3.1. Table 3.3.1 below provides summary statistics on the entire data set and the largest-connected set used in our main clustered worker-job fixed effect specification. We lose data on about 50,000 workers and 15,000 firms in the largest connected set, or about 16% of all unique workers and 10% of all unique firms in the data set. This amounts to 3.5% of the sample in the second period in total. Reassuringly, the hourly wage moments in the two periods do not change significantly between the connected set and the full sample. Table 3.1 in the appendix summarises how much of our data is retained in the largest connected set for various specifications.

TABLE 3.3.1 Summary statistics

	02-10	11-19
All observations		
Total observations	1398098	1598919
No. unique workers	303,122	339,697
No. unique firms	127,014	156,178
No. unique firm-occ pairs	282,251	335,298
Average log hourly wage	2.21	2.18
SD log hourly wage	0.53	0.51
Largest connected set of main specification		
Total observations	1346163	1542383
No. unique workers	251,187	283,161
No. unique firms	112,817	139,533
No. unique firm-occ pairs	258,676	308,095
Average log hourly wage	2.22	2.19
SD log hourly wage	0.53	0.51

Notes: This table shows summary statistics from the ASHE data for two time periods: 2002-2010 and 2011-2019 as well as for the full data and the largest connected set.

3.4 Log wage decomposition

3.4.1 Overall wage variance decomposition

We present the main results of our decomposition in table 3.4.1 below, as well as several alternative specifications that have been produced in the literature⁹. In column (1), we produce estimates from the classic AKM two-way regression with worker and firm fixed effects. We find results that are qualitatively similar to those reported in other countries: a small and negative covariance between firm and worker fixed effects and 24% of log-wage variance due to firms' fixed effects, with most of the variance in log-wages being due to worker fixed effects. Column two shows the two-way fixed effect regression with worker and firm-cluster fixed effects, where firms are clustered into 30 firm clusters using a k-means algorithm on the basis of their wage distributions. In this specification, sorting becomes positive and increases at the expense of the contribution of firms and to a lesser extent workers. This again serves to highlight the importance of accounting for limited mobility bias.

⁹We compare these estimates against estimates in the literature in table 3.2 in appendix section 3.6

TABLE 3.4.1 Comparison of different specifications (2011-2019)

Proportion of overall log-wage variance	Basic-firm	Cluster-firm	Cluster-Occ	Cluster-Job
Worker component	0.64	0.54	0.56	0.18
Firm component	0.24	0.10		
Occupation component			0.06	
Job component				0.38
Sorting component	-0.02	0.19	0.20	0.29
Residual component	0.08	0.10	0.11	0.08
Adjusted R2	0.89	0.88	0.87	0.90

Notes: This table presents log-wage variance decompositions for four specifications: (i) simple AKM, (ii) bias-corrected AKM, (iii) worker-occ twfe, and (iv) worker-job twfe (main specification). We use the ASHE data over the period 2011-19 for these results.

In column (3)¹⁰, we consider a specification where log earnings are decomposed into worker and occupation-specific effects. This regression serves two purposes. First, it is of interest in and of itself and shows a remarkable similarity to the cluster firm regression in column (2). A large share of the log-wage variance appears to be due to worker heterogeneity, with between-occupation heterogeneity accounting for only 6% of the total variance. Second, it serves to highlight that the differences between (low mobility adjusted) traditional decompositions in column (2) and those we suggest in column (4) are not just driven by the inclusion of occupation fixed effects but rather by specifically the inclusion of firm-occupation fixed effects.

Our preferred job-level analysis on the other hand in column (4) produces different results in several ways¹¹. There are three qualitative differences in these results relative to the results using the traditional AKM approach. First, the share of log-wage variance attributed to differences between workers falls to only about 18% of overall wage variance, from 64% in the standard AKM decomposition. As discussed in section 3.2, this result arises due to the misspecification of the wage process in traditional AKM models. Intuitively, where the job fixed effect is not included in the regression in column (2), it shows up as part of the omitted variable bias in the estimate of α_i in the misspecified model if there is residual covariance between α_i and $\psi_{J(i,t)}$ after accounting for the firm workers work at. This bias component is

¹⁰As there are only roughly 367 occupations, there is no limited mobility issue in individual-occupation fixed effect regressions. The results do not differ when clustering is used.

¹¹We check that the job-level estimates are related to firms' revenue in appendix 3.8, suggesting that these estimates can plausibly be interpreted as job-level wage premia.

counted as part of the individual fixed effect in the variance decomposition of the misspecified model and must be larger than the actual variance of α_i if the covariance between the individual fixed effect and the residualised, weighted sum of job fixed effects across an individual's career is positive. This explanation implies that, in practice, workers who work at high-wage occupations tend to work at high-wage occupations for most of their career, and vice versa¹². We discuss this analysis of the differences between our results and the canonical AKM results in appendix 3.5, and show that in general, if there is positive sorting of workers to jobs conditional on firms, then in the misspecified model, the variance of worker fixed effects will be over estimated and that additionally the variance of firm fixed effects in worker-firm specifications will be incorrectly estimated (although the bias is of ambiguous sign).

These results paint a fundamentally different picture of wage inequality than those from the AKM regressions. In the AKM story, the substantial amounts of wage inequality are primarily because workers are very different, due to different degrees of ability or human capital accumulation. In our case, we find that when we make use of information from moves by workers across occupations, we see that differences between workers are much smaller than previously suggested. As a caveat, there is substantial heterogeneity in the importance of the individual worker differences in explaining wage dispersion between different occupations. As we discuss in section 3.4.3, wage dispersion in high-wage occupations is more readily explained by individual differences than in low-wage occupations. But, our results suggest that policies which reduce job-level differences in pay can be effective in reducing overall income inequality.

Second, we find that job heterogeneity and sorting both contribute significantly to wage inequality, specifically 38% and 29% respectively. Therefore, whereas previous specifications find low firm-level complementarities (and indeed we find low occupation level complementarities in column (3)) we find large job level complementarities in wages. We also find a larger impact of sorting of workers into jobs, which explains 29% of the variation, validating the original intuition that workers match to jobs, not firms.

A strange qualitative feature of results from AKM regressions, corrected for low-mobility bias, is that they have typically found low firm heterogeneity (10% in our sample) but a relatively larger share due to sorting between firms and workers, implying high degrees of sorting to firms when there is much smaller heterogeneity between firms. This feature is not

¹²We provide some direct evidence of this in appendix 3.7.

present in our findings; we find that there is substantial job-level heterogeneity, and consequently, substantial degrees of sorting between jobs and workers. In the next section, we present evidence that suggests that a possible explanation of the differences between our specification and the standard AKM specification is that the latter approach only captures within-occupation job heterogeneity while we also capture heterogeneity between occupations.

3.4.2 Between and within-occupation log-wage decompositions

A natural question in our framework is whether heterogeneity in the job fixed effect is due primarily to differences between firms or between occupations. To answer this, we can further decompose the variation of job fixed effects, and the covariance between job and worker fixed effects, into between- and within- occupation components, using the law of total variance. The index of occupations partitions the data, and let $\mathbb{E}_o(\cdot)$, $\mathbb{V}_o(\cdot)$, and $Cov_o(\alpha_i, \cdot)$ denote the mean, variance and covariance with α_i of the random variable within the partition indexed by o . Equation 3.3 presents the decomposition of job variance into between and within occupation components, while equation 3.4 does the same for the decomposition of covariance.

$$\mathbb{V}[\psi_j] = \underbrace{\mathbb{V}[\mathbb{E}_o[\psi_j]]}_{\text{Between } o} + \underbrace{\mathbb{E}[\mathbb{V}_o[\psi_j]]}_{\text{Within } o} \quad (3.3)$$

$$Cov[\alpha_i, \psi_j] = \underbrace{Cov[\mathbb{E}_o[\alpha_i], \mathbb{E}_o[\psi_j]]}_{\text{Between } o} + \underbrace{\mathbb{E}[Cov_o[\alpha_i, \psi_j]]}_{\text{Within } o} \quad (3.4)$$

Results of this decomposition are reported in table 3.4.2. Table 3.4.2 also splits the sample into two periods and compares the decomposition across both. We find no significant differences in the drivers of inequality in the UK over time. This is in contrast to Card et al. [2013] who find, in Germany, that increases in firm heterogeneity and sorting explain changes in wage inequality.

We find that about two-thirds of the variation in job effects can be explained by between-occupation variation and a third can be explained by within-occupation variation. The latter part, an occupation-weighted summary of wage variation between firms that is not explained by the occupation title of the worker, is about 13% in our data, which is close to the 10% estimated for the variance of firm fixed effect in the main AKM specification. This implies that the standard AKM approach misses most of the variability in job effects, which

is due to differences in occupations across jobs rather than firm heterogeneity within occupations.

Similarly sorting between workers and jobs can be decomposed into within and between occupation components. The between occupation component is given by the covariance between mean worker type within an occupation and mean job fixed effect within that occupation. Whereas the within occupation component is given by the expected covariance between workers and jobs in a given occupation, across all occupations. We find that over two-thirds of the variation due to sorting can be explained by sorting between occupations, which reflects the sorting of higher-paid workers into higher-paying occupations, whereas only 9% of the log-wage variation can be explained by sorting of workers within occupations to higher-paying jobs.

TABLE 3.4.2 Variance decomposition from jobs model
over time

Component of total log-wage variance	Period 2002-2010	Period 2011-2019
Variance of worker FE	0.20	0.18
Variance of job FE	0.34	0.38
Between-occ variation	0.23	0.25
Within-occ variation	0.11	0.13
Correlation between worker and job FE	0.32	0.29
Sorting between occupations	0.23	0.20
Sorting within occupations	0.09	0.08
Adjusted R2	0.90	0.90
RMSE	0.17	0.16

Notes: This table decomposes the log wage variance according to our job-worker two-way fixed effects model, for two period, 2002-10, and 2011-19. We further decompose the job variance and covariance components into between and within-occupation parts using the law of total variance.

Our results thus offer a different perspective on log wage inequality than most of the literature. Results from standard AKM regressions have tended to primarily emphasise the role of individual heterogeneity and secondarily the role of firms. In contrast, our results show that the occupational structure plays a large role in the story of wage inequality, and most of the variance due to job heterogeneity and sorting of workers to jobs can be attributed to between-occupation heterogeneity and sorting of workers to occupations. Wage heterogeneity between firms within occupations is quantitatively less important, and we show in

a subsequent section that it is relatively more important in high-wage occupations than in low-wage occupations.

3.4.3 Evidence on varying wage premia across occupations within firms

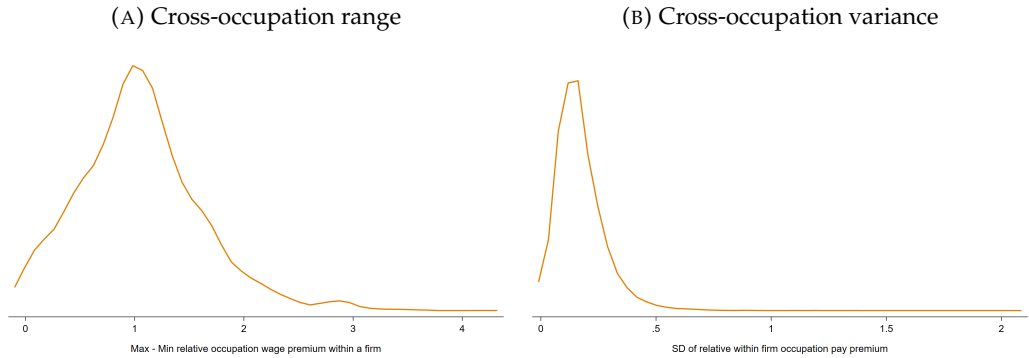
Armed with fixed effects estimates, we can turn to test the idea that motivated this paper: that firms offer different wage premia to workers in different occupations¹³. One reason for this may be that workers in some occupations are in some sense more central or important to a firm's profit¹⁴. In this section, we present empirical evidence to suggest that firms offer varying wage premia across the occupations they employ. We take two approaches, first we analyse the estimated job fixed effects ψ_j and show that these vary within-firm across occupations. Second, we consider to what extent a naive additively linear specification in firm and occupation effects would be misspecified relative to our more general specification, which is equivalent to testing for occupation-firm specific match effects.

First, we plot the within-firm cross-occupation job fixed effect variance and range having removed aggregate occupation variation, in figure 3.4.1. This figure shows significant dispersion in wage premia across occupations within firms, even after accounting for aggregate occupation variation. In the left-hand side panel, the within-firm cross-occupation range is shown, and in the right-hand panel the within-firm, cross occupation variance. These statistics are calculated from the residuals of a regression of estimated job fixed effects on occupation and firm effects and therefore reflect varying wage premia within a firm accounting for aggregate occupational differences.

¹³In appendix 3.8, we regress the estimated job fixed effects on firm characteristics to check whether the rent-sharing interpretation is appropriate and conclude that it is. We further check that the relation between the job FE and firm revenue is weaker for public firms than for private firms without collective bargaining agreements, and document heterogeneity across occupations. We leave a fuller analysis with firm profit data, which is sadly not available in the Business Structure Dataset (BSD) data set, for future work.

¹⁴In the model of Haanwinkel [2020], firms require tasks in different intensities for production, and workers in different occupations produce units of different tasks. In Bloesch et al. [2021], workers in different occupations can receive different wage premia because they have different hold-up power in the production of the firm.

FIGURE 3.4.1 Distribution of the variance of job fixed effects across occupations within firms



Notes: This figure shows the within firm, cross occupation dispersion in wage premia removing aggregate cross-occupation effects. Both figures show density plots over firms, in the left-hand panel the cross occupation range (max-min) is displayed and in the right-hand panel, the cross occupation variance is displayed. The sample is restricted to firms with 10 or more employees. To calculate the statistics graphed estimated job fixed effects are residualised on firm and occupation effects.

Figure 3.4.1 shows significant variation, and therefore gives evidence to suggest that job wage premia do vary across occupations within firms. However, it is unclear from the figure alone whether this variation is statistically significant. Therefore, we formally test the differences graphically observed, by performing a test of the hypothesis that the variance of job fixed effects within the firm is no greater than five percent of the overall variance in job effects¹⁵. Performing this test we find that we can reject the null hypothesis for 49% of firms and therefore conclude that there is significant within-firm, cross-occupation, wage variance.

An alternative hypothesis is that while firms may offer an occupational premium to workers in different occupations, they also impose a constant firm premium on all workers within a firm. That is, we can approximate the job fixed effect estimated in our model by an additive specification given by equation 3.5, where $occ(i, t)$ denotes the occupation worker i is in at time t and $F(i, t)$ denotes the firm the worker is in at time t .

$$\psi_{J(i,t)} = \phi_{occ(i,t)} + \chi_{f(i,t)} \quad (3.5)$$

To test this interpretation, we estimate regressions of the form $\ln(w_{it}) = \alpha_i + \tau_{O(i,t)} + \zeta_{F(i,t)} + \varepsilon_{it}$, where τ_O are occupation fixed effects and ζ_F are firm fixed effects¹⁶. Then we can study the difference between these estimated fixed effects and our previously estimated job

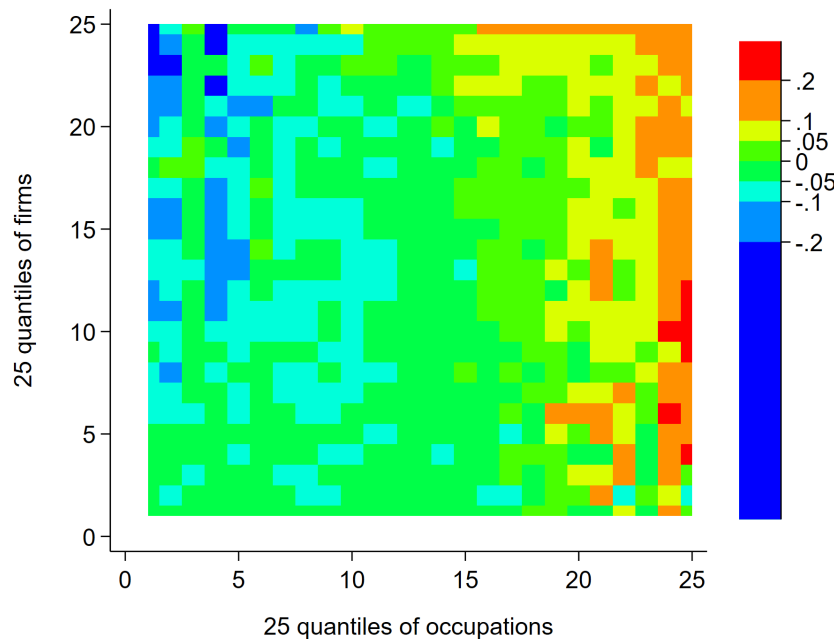
¹⁵To be precise we test whether within-firm, cross-occupation variation in job fixed effects is statistically significantly greater than 5% of the overall variance in job fixed effects across all firms and occupations using a Chi-Squared test [Snedecor and Cochran, 1989]. We find that we cannot reject the null hypothesis at the 5% level in 49% of firms with more than 10 employees in our data.

¹⁶This specification is similar to the one examined in Torres et al. [2018].

fixed effects $\psi_{J(i,t)}$. We interpret this difference as an occupation-job specific match effects on the offered wage premium. If firms do not pay differential wage premia across occupations, $\Delta = \psi_J - (\tau_O + \xi_F)$ will be small and not systematically related to any firm or occupation characteristics. A positive match effect of e.g. 0.1 implies that the premium offered by the particular job is roughly 10% greater than the premium implied by an additive occupation and firm fixed effect (and vice versa).

In figure 3.4.2, we produce a heat map of the difference between the job fixed effect in our main model and the sum of the occupation and firm fixed effects in the auxiliary model, that is of match effects. We plot this for 25 quantiles of occupations ranked by the average log hourly pay in the occupation and 25 quantiles of firms ranked by the average log hourly pay in the firm. Thus, a positive number (e.g. 0.05) can be interpreted as follows: if we naively replace the job fixed effect by the sum of the two fixed effects in our predicted values in our main model, we would underestimate the hourly wage by 5%. Green tiles represent quantile combinations where the error is less than 5% in either direction.

FIGURE 3.4.2 Contribution of worker-job match effects by firm and occupation mean pay quantiles



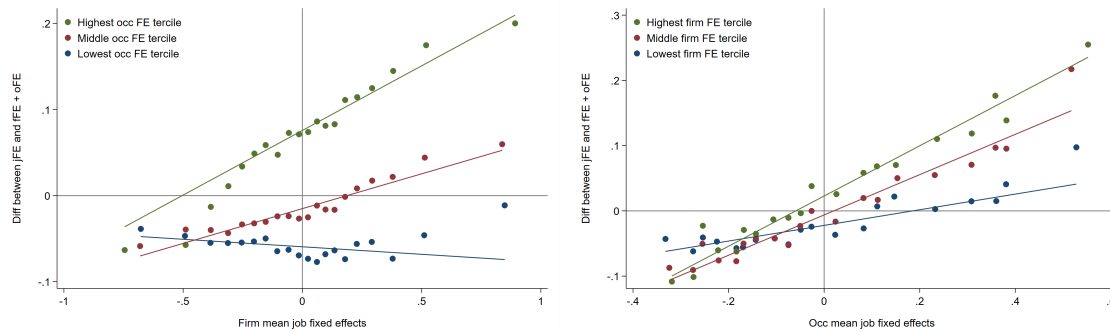
Notes: The figure shows a heatmap of the difference between the estimated job fixed effect in our main specification, and the sum of a firm fixed effect and an occupation fixed effect in an auxiliary specification, which we interpret as match effects between occupation and firms. Green squares imply that if the job fixed effect were replaced by the additive firm and occ fixed effects, the resulting predicted log wage would be within 5% of the true predicted log wage with job FE. Red implies that using the additive specification would underpredict the true wage by over 10% and blue implies that overprediction by over 10%. The x-axis indexes 25 quantiles of occupations, sorted by the mean wage within the occupation, and the y-axis indexes 25 quantiles of firms, sorted by the mean wage within the firm.

This figure shows that the linear form performs quite badly, especially at the extremes for each occupation. The linear combination would severely overestimate the wages of workers in low type occupations at all firms (as these have a negative match effect), and severely underestimate the wage of high type occupations at all types of firms. The problem seems especially acute for high wage occupations at low wage firms and low wage occupations at high wage firms, leading to mistaken predictions by over 20%. This leads us to conclude that the proportionate firm fixed effect is inappropriate for capturing the relation of pay premia to occupation and firm characteristics.

To analyse this dynamic further, we produce two complementary figures, first plotting the relation between the estimated residual against the mean log hourly pay at the firm for three terciles of occupations sorted by mean occupation wage, and the relation between the estimated residual against the mean log hourly pay at the occupation for three terciles of firms sorted by mean firm wage.

FIGURE 3.4.3 Plotting estimated match effects

(A) Match effect by firm average job fixed effects for three occupation mean wage terciles (B) Match effect by occupation job fixed effects for three firm mean wage terciles



Notes: This figure plots the distribution of the difference between the estimated job fixed effect in our main model, and the sum of estimated firm and occupation fixed effects in an auxiliary regression, that is, match effects. Panel 3.4.3a plots the differences by firms, sorted by the mean job fixed effect within the firm for three terciles of occupations, while panel 3.4.3b plots the differences by occupation, sorted by mean occupation job fixed effect, for three terciles of firms. par

Figure 3.4.3 shows binned scatterplots of match effects which emphasise the importance of match effects for job-level pay premia¹⁷. Panel 3.4.3a shows that match effects increase from negative to positive as mean job fixed effects within firms increase. This relationship, however, is mediated by occupation terciles. Low-wage occupations have a negative match effect at all firms, regardless of the firm-level pay premium. On the other hand, high-wage occupations have a positive match effects at most firms, even those who offer on average below-average wages. Panel 3.4.3b shows the analogous results when plotting match effects over occupation average job fixed effects and cutting by firm occupation fixed effect terciles. We find that low occupation fixed effects occupations consistently have negative match effects and that although these increase with occupation average job fixed effects it does so faster for higher firm fixed effect firms.

Finally, we can formally test for the difference between the naive model: $\ln(w_{it}) = \alpha_i + \tau_{O(i,t)} + \zeta_{F(i,t)} + \varepsilon_{it}$ and our more flexible specification $\ln(w_{it}) = \alpha_i + \psi_{J(i,t)} + \varepsilon_{it}$ using an F-test. This gives a F statistic of 3.93 and 3.24 when tested against the naive model with and without including occupation fixed effects, in both cases the critical value at the 5% level is very close to 1, so we resoundingly reject equality between models.

Our results support papers that argue that the standard AKM approach is misspecified, including Haanwinckel [2020], and Bloesch et al. [2021]. Haanwinckel [2020] argues that firms should pay larger premia to workers who perform tasks which the firm use more

¹⁷The scatterplots can be interpreted as similar to the data in figure 3.4.2, ‘sliced’ into a 2-D graph by reducing the dimensions of one of the axes.

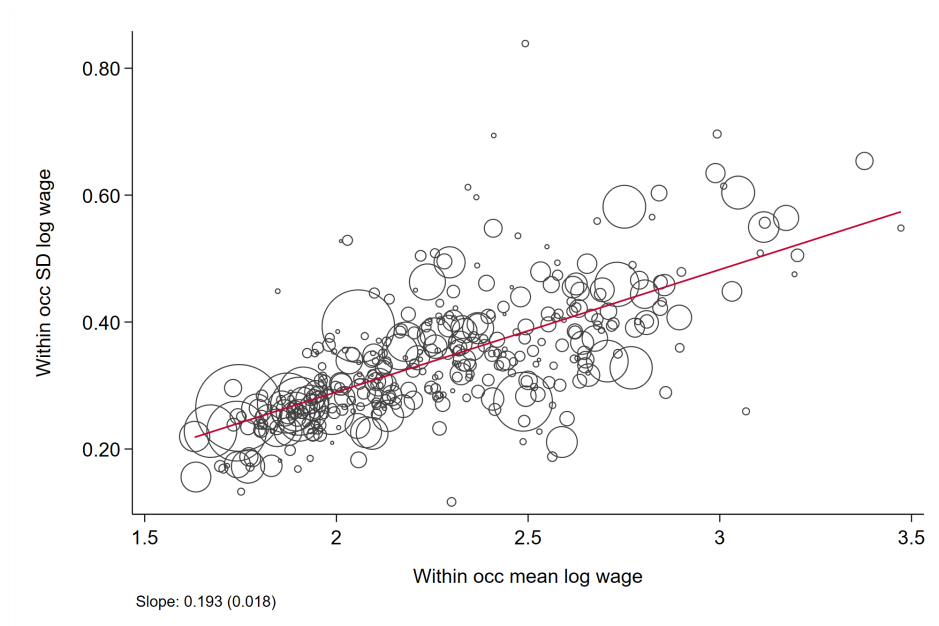
intensively, and in a similar vein, [Bloesch et al. \[2021\]](#) show that workers which have more hold-up power within a firm should be paid more. Panel [3.4.3a](#) shows that these jobs are likely to be more highly paid and skilled occupations. Assuming a proportional firm wage premium leads to the systematic overestimation of the pay of low-wage occupations and the underestimation of the pay of high-wage occupations at high productivity firms.

3.5 Differences in the log-wage decomposition across occupations

An implication of the variation of wage premia across occupations within firms is that the sources of wage inequality within occupations may be quite heterogeneous. An advantage of our more flexible specification is that we can investigate the extent to which the overall wage decomposition presented is representative of the wage variance in individual occupations. Within an occupation, variation in the job fixed effect reflects differences in wage premia offered by firms, and we can more straightforwardly compute decompositions according to equation [3.2](#) within occupations. We find that the overall wage decomposition however disguises substantial heterogeneity in the wage decomposition in each occupation.

We start with the stylised fact that in the UK labour market, the variance of log hourly wages is co-increasing in the mean log hourly wage within an occupation. This is shown in figure [3.5.1](#), which plots a scatter plot of occupations with the mean log hourly wage on the x-axis and the standard deviation of log hourly wage within that occupation on the y-axis. The size of the bubble represents the number of workers in that occupation in the data. There is a clear approximately log-linear relationship between the mean and variance of log earnings, and dramatic outliers to this trend tend to be fairly small occupations.

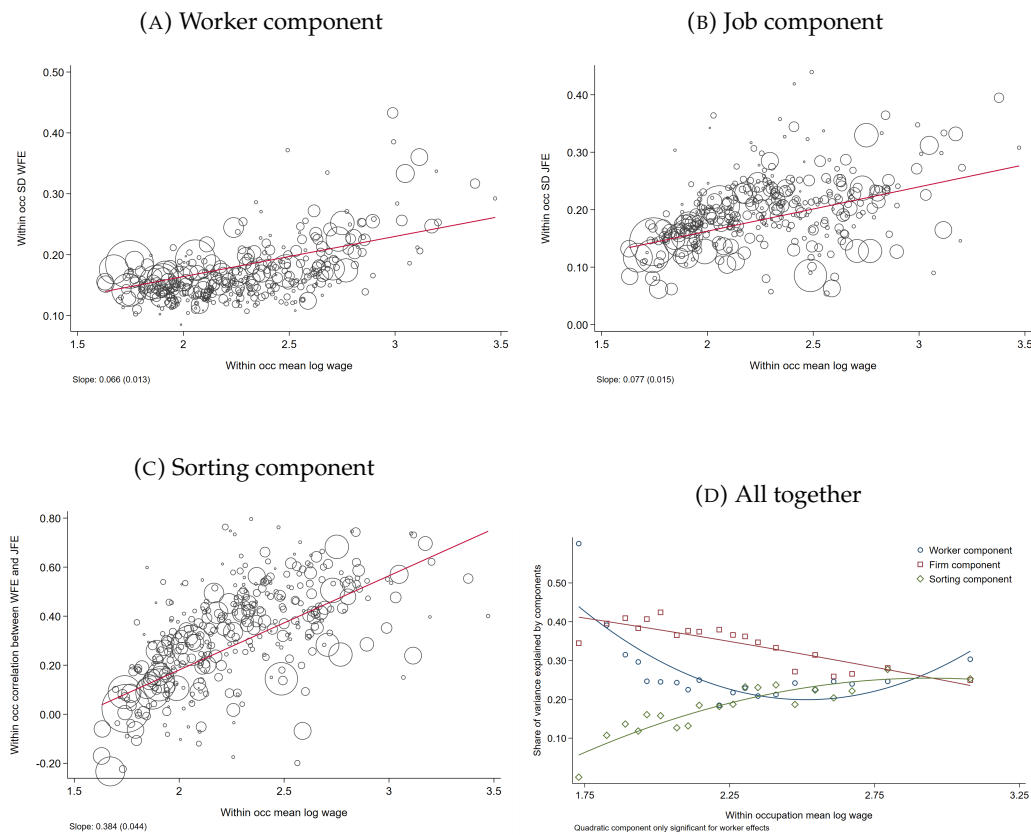
FIGURE 3.5.1 Within occupation mean vs SD of wages



Notes: This figure plots the mean log-wage (x-axis) and the standard deviation of log-wages (y-axis) within each occupation for all occupations in the data. Each bubble represents a particular occupation and the larger the bubble is, the more workers there are in the occupation. A red linear trend line is drawn, with slope 0.193.

What accounts for the greater wage variance in high-wage occupations relative to lower wage occupations? To understand this, we decompose the variance in each occupation into an individual component, a firm component and a sorting component. We plot the log-variance due to each component for each occupation in figure 3.5.2, panels (a) to (c). In panel 3.5.2d, we plot the variance of the components as a share of total log-wage variance within the occupation.

FIGURE 3.5.2 Variance of worker FE, job FE and correlation between the two components



Notes: Panel 3.5.2a plots the standard deviation of worker fixed effects in each of the 368 occupations in period 2 of the data (2011-19). Panel 3.5.2b plots the standard deviation of within-occupation job fixed effects within each occupation; since a job is identified as an occupation-firm pair, the variation is from firms offering different wages for the same occupation. Panel 3.5.2c plots the correlation between worker and firm fixed effects for each occupation. The occupations are ordered by the mean log-wage within the occupation on the x-axis.

The variances of all three components are increasing in the within-occupation mean wage, but at different rates. In particular, sorting is zero or even negative for many low-wage occupations and does not contribute to within-occupation wage variance for those occupations. However, it becomes as important as individual and firm heterogeneity for high-wage occupations. The variance of individual worker fixed effects also increases exponentially; it is relatively low for low and medium wage occupations but is quite high for the highest wage occupations. Thus, the drivers of wage variance appear to be quite different for low, medium and high wage occupations.

Our results suggest that different occupational labour markets require different modelling approaches based on the degree to which firms and individuals are heterogeneous. In particular, the approaches emphasising firm-level wage heterogeneity and the role of

matching (summarised in [Mortensen \[2003\]](#)) seem more applicable to higher-wage occupations, in which matching to a less productive firm or worker leads a larger loss of wages or profits. Lower-wage occupations exhibit almost no sorting and much smaller degrees of both worker and firm heterogeneity, and thus may be better approximated by classical or monopsonistic models of the labour market which do not emphasise matching. This conclusion resembles that reached by [Lise and Postel-Vinay \[2020\]](#), who arrive at it using a structural model of occupational mismatch with heterogeneous skills.

3.6 Understanding spatial wage variation through worker-job decompositions

An important dimension of wage inequality, in the UK¹⁸ but also more broadly in the world, is geographical inequality. It is well-documented that wage inequality is higher in larger labour markets (e.g. [Nord \[1980\]](#), [Baum-Snow and Pavan \[2013\]](#), and specifically in the UK context, [Lee et al. \[2016\]](#))¹⁹. Our novel²⁰ wage variance decomposition allows us to study inequality within and across geographies in the UK with a new lens, and better understand the salient features of inequality across space.

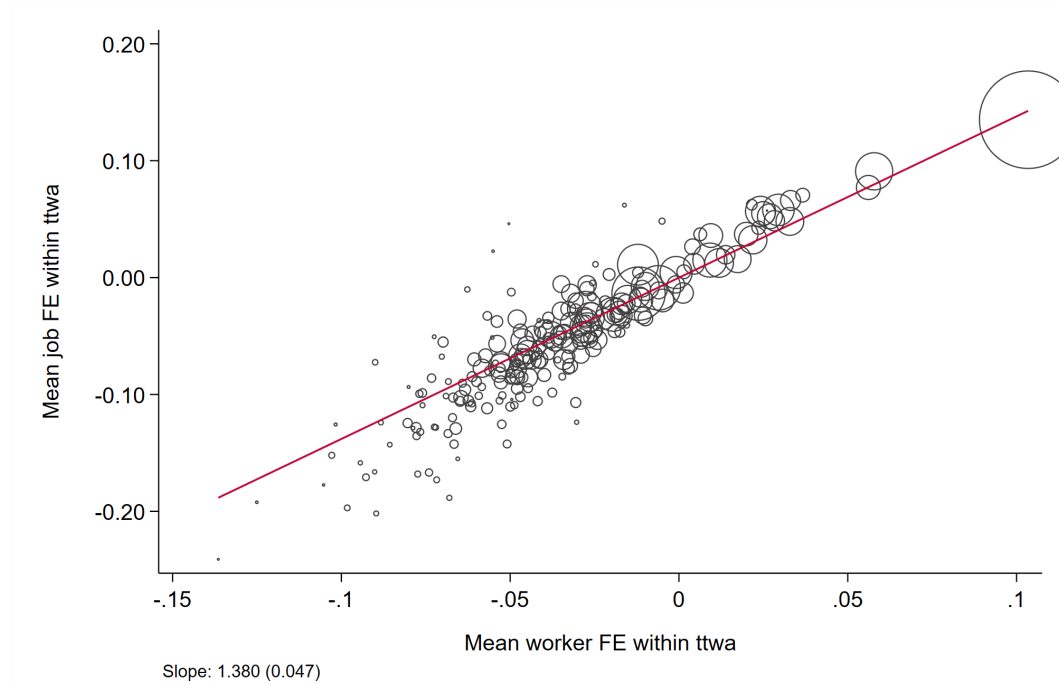
First, we find that higher type workers tend to co-locate with higher type jobs in larger labour markets. Figure 3.6.1 plots the mean worker fixed effect and mean job fixed effect within a travel-to-work area, where the size of the circle reflects the size of the local labour market. This finding supports a similar finding by [Dauth et al. \[2018\]](#) using a worker-firm two-way fixed effect regression, and papers based on the observation that higher productivity firms sort to larger labour markets, like [Gaubert \[2018\]](#) and [Bilal \[2021\]](#). Analogously, it coheres with a large body of literature, including [Venables \[2011\]](#), [Davis and Dingel \[2019\]](#), and [Behrens et al. \[2014\]](#), which has found evidence to suggest and propose explanations for why more productive individuals sort to larger cities.

¹⁸[Agrawal and Phillips \[2020\]](#) show that nominal earnings gaps across space are large with incomes in Wales on average 40% below those in London. However, [Agrawal and Phillips \[2020\]](#) suggest that real differences are smaller and stress that within-city inequalities can be large, especially so in London. Spatial differences in income and opportunity are also increasingly dominating the political discourse in the UK, in September 2021 the department for Housing, Communities and Local Government became the Department for Levelling Up, and the new Secretary of State for Levelling Up, Michael Gove MP, called the Levelling up Agenda “the defining mission of this government”.

¹⁹We replicate this fact in the UK in appendix 3.9, and show that the mean and variance of earnings is increasing in log travel-to-work area population. Our spatial unit is the UK travel-to-work area, as defined using data from the 2011 census.

²⁰For an analogous study that we build on based on using the canonical AKM model to study geographic income inequality, see [Dauth et al. \[2018\]](#).

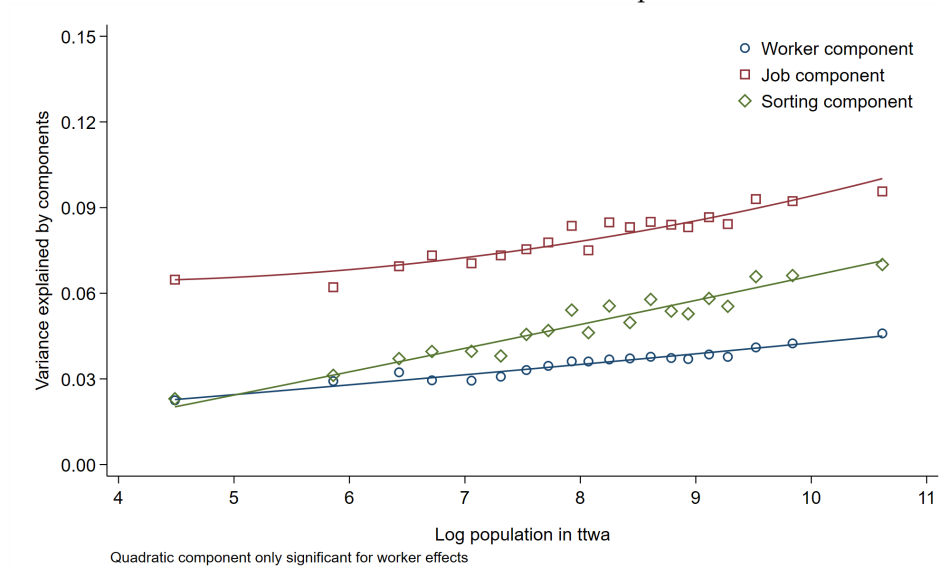
FIGURE 3.6.1 Within TTWA mean of worker FE and job FE (2011-19)



Notes: This figure plots the mean worker fixed effect (x-axis) and the mean job class fixed-effect (y-axis) within each travel to work area for each of 218 travel to work areas for 2011-19. Each bubble represents a particular travel-to-work area and the larger the bubble is, the more workers there are in the travel-to-work area observed in the data. A red linear trend line is drawn.

To study differences across labor markets in the constituent parts of within-labor market wage variation we apply our standard decomposition. Splitting log wage variance into a component due to the variance of individual fixed effects, a component due to the variance of job fixed effects, and a component due to the covariance between the two. We relate each of these components to (log) travel-to-work area size in figure 3.6.2 to explain the increasing within labor-market variation. All three constituent parts are increasing in local labour market size, but at different rates. It's clear from figure 3.6.2 that sorting becomes relatively more important in larger labor markets. This result coheres with a previous result in [Dauth et al. \[2018\]](#) which finds that sorting to firms accounts for more variance in larger labour markets in Germany. This finding could be explained by the value of search being greater in larger cities due to scale effects, as was found in [Petrongolo and Pissarides \[2006\]](#).

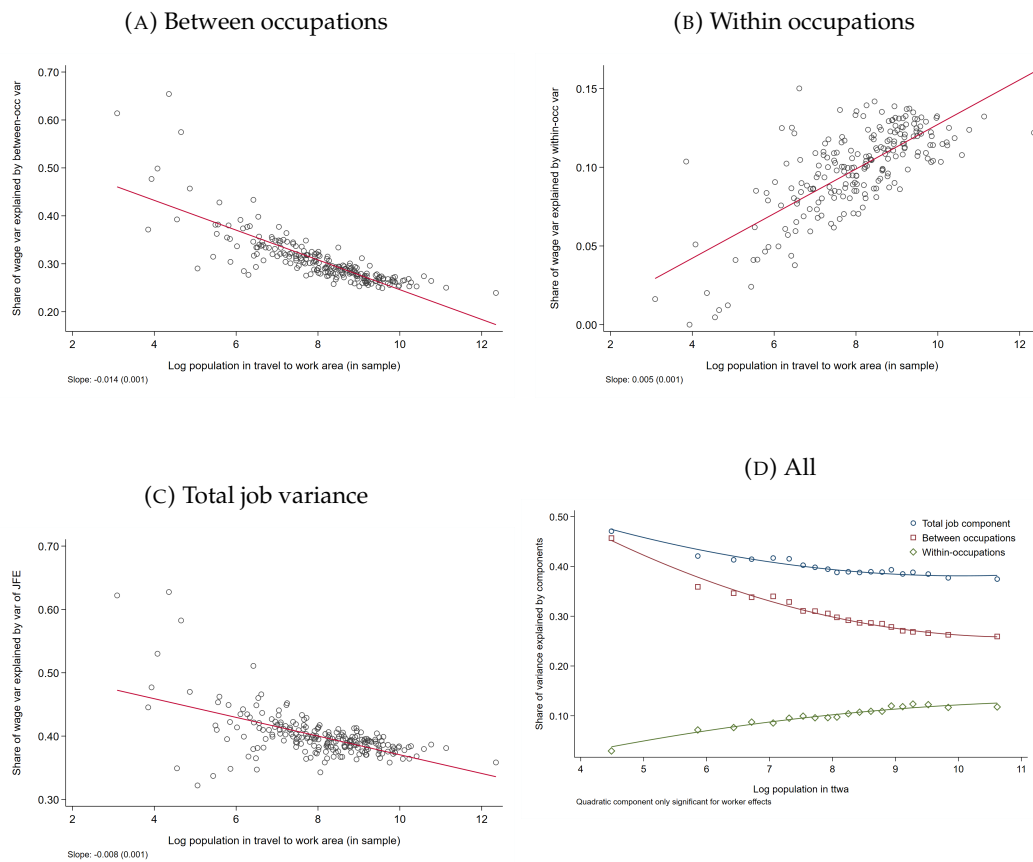
FIGURE 3.6.2 Within TTWA variance components vs local LM size



Notes: We compute the variance of individual fixed effects (blue line), the variance of job fixed effects (red line), and the covariance between the two (green) line for each travel-to-work area. We then plot a binned scatterplot against log-population within the ttwa and draw quadratic lines of best fit for all three components.

To delve deeper into what explains the patterns in figure 3.6.2, we decompose the local labor market job-class variance and covariance components into a between-occupations component and a within-occupations component using equations 3.3 and 3.4. First, we examine whether the decline in the relative importance of job heterogeneity is due to within or between occupations variance. We plot the relative importance of each factor in each TTWA in figure 3.6.3 below. Panels (a) and (b) plot the between- and within- occupation part of total log-wage variance within the TTWA as a function of log population in the TTWA. Panel (c) plots the total job variance as a share of total log-wage variance, and panel (d) combines all the other three panels in a binned scatterplot for clarity of comparison. The fall in the relative importance of job heterogeneity is primarily due to the falling relevance of between-occupation job variance as the local labour market gets larger. In larger local labour markets, wage dispersion between firms for the same occupation becomes relatively more important from a low base, eventually accounting for about 10% of the job component of log-wage variance. Absolute, rather than relative, trends are plotted in figure 3.8 in appendix 3.9.

FIGURE 3.6.3 Decomposing the within TTWA job variance component into within and between occupations

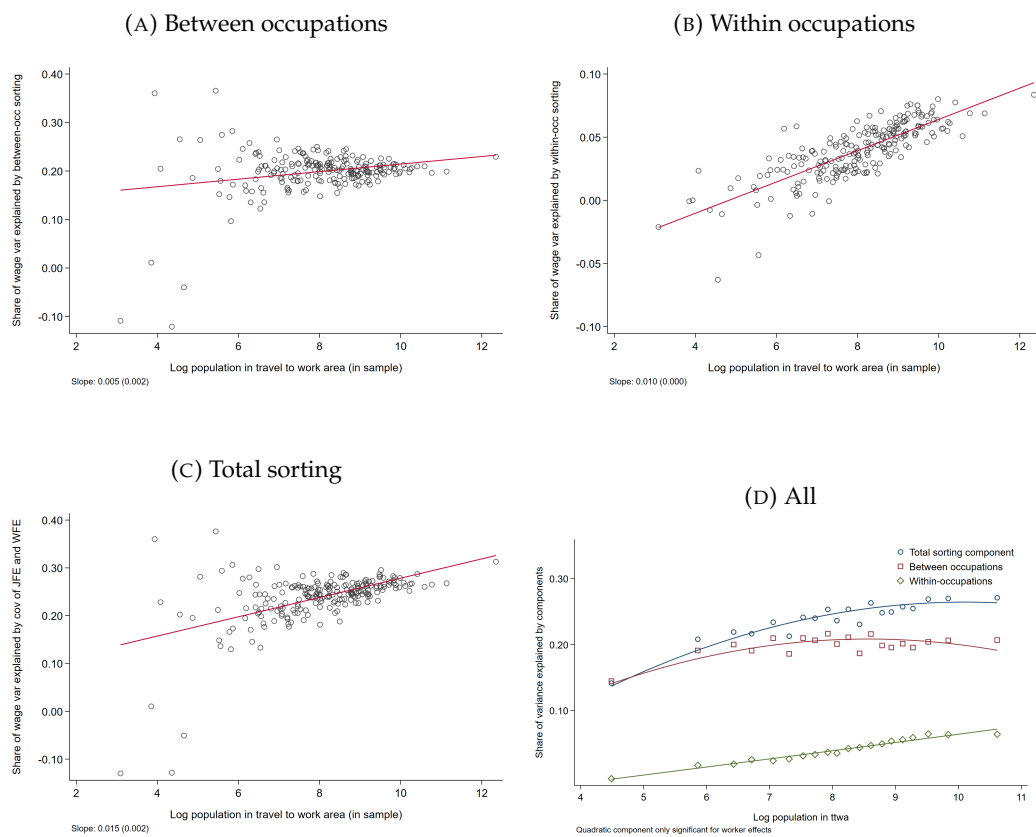


Notes: This figure presents scatterplots with log-population within the TTWA (x-axis) against between-occupation job-class variance in panel (a), within-occupation job-class variance in panel (b), total job-class variance in panel (c), and all components in bin-scatter plots in panel (d).

Similarly, figure 3.6.4 decomposes the share of within local labour market wage variance due to sorting between workers and jobs into the component between occupations and the component within occupations. (The between- and within- occupation covariances as raw values are included in figure 3.9 in appendix 3.9.) What is immediately clear from this figure is that the majority of the previously documented increased importance of sorting in larger labour markets can be attributed to increased occupation sorting. That is, in larger labour markets better workers sort into better firms within occupation classes to a much greater degree than they do in smaller labour markets²¹.

²¹The increase in within-occupation sorting could reflect the tendency for jobs in high-wage occupations to locate in larger labour markets, given that we know that the degree of sorting is higher than in higher wage occupations. In appendix 3.10, we show that while a share of this relation is explained by occupational composition effects, not all of it is explainable by composition.

FIGURE 3.6.4 Decomposing the within TTWA wage variance sorting component into within and between occupations



Notes: This figure presents scatterplots with log-population within the TTWA (x-axis) against between-occupation sorting between worker and job-class in panel (a), within-occupation sorting between worker and job-class in panel (b), total worker and job-class sorting in panel (c), and all components in bin-scatter plots in panel (d).

In sum, although overall wage variation in the UK can mainly be described by a between-occupation story, within-occupation variation becomes a relevant feature of larger labor markets. Colloquially, wage inequality in villages and small towns is largely due to occupations, because of the differences in pay of the plumber and the shopkeeper rather than to differences between pay of plumbers at particular firms. As the local labour market gets larger, wage dispersion between firms within an occupation becomes a quantitatively more important element of wage variance. The fact that better software engineers work at Google instead of smaller software firms is more relevant to wage inequality in London than in a provincial town.

3.7 Conclusion

In this paper, we propose that the classic two-way fixed effect model with worker and firm fixed effects is misspecified and that this misspecification produces a misleading decomposition of the components of log-wage variance. Instead of matching to firms, workers match to jobs which may differ within firms — e.g. a software engineer and a chef at Google are different positions. Further, it is likely that the firm would want to offer differing wage premia to workers in different occupations at the firm.

We estimate a two-way fixed effects model with individual and job fixed effects. We find that the component of variance due to individual heterogeneity, usually found to be 50-60%, accounts for only 18% of log wage inequality in our model. Conversely, we find that job heterogeneity and the sorting of workers into jobs account for 38% and 29% of the log-wage variance respectively. The majority of variation in these components can be explained by cross rather than within-occupation differences, highlighting the importance of occupations and jobs rather than firms, in explaining UK wage inequality.

Our approach also allows us to analyse how wage inequality within occupations differs across occupations. We find that wage variance in higher mean wage occupations is explained by sorting to a greater degree. We thus conclude that occupational labour markets should be analysed using different approaches, depending on the degree to which individual and firm heterogeneity affect productivity. Occupations in which heterogeneity is more important are better explained by search and matching models which produce more wage dispersion, while occupations, where workers and firms are more interchangeable, are better modelled by spot models of the labour market.

Finally, we consider the implications of analysis for understanding wage inequality in local labour markets. We find that sorting is increasingly important in larger labour markets, which could be due to scale effects in search markets. We also find that sorting to firms within occupations are relatively much more important in larger labour markets than in smaller ones; while wage inequality in small towns is likely to be due to differences in occupation, wage inequality in large cities are as much due to workers being at different firms within the same occupation as being in differing occupations.

Appendix

3.1 Details on data cleaning

We focus our decomposition analysis on the most recent period from 2011-19, during which the UK standard occupation classification 2010 is used. We observe the most fine-grained 4-digit occupational classifications, giving us 368 occupations in total²². Our preferred wage variable, used throughout as the focus of our analysis, is the hourly wage, which is computed as the gross pay divided by the total number of hours. We trim the top and bottom 1% of the sample to remove outliers. Our main unit of geographical analysis is the travel to work area (hereafter TTWA), a non-overlapping geographical unit provided by the Office for National Statistics. The TTWA is defined as an area where at least 75% of the resident workforce work in the area and 75% of the people who work in the area also live in the area and is thought of as a local labour market. TTWAs are the UK equivalent of commuter zones in the US which have become the standard unit for measuring local labour markets. TTWAs differ in size, and urban areas such as London can be very large. We use the 2011 classification of the TTWA in our analysis, giving us 218 travel-to-work areas in the data in total. Observations with incomplete information about their home address are used in the estimation of the fixed effects model but are not included in the descriptive analysis on geographical dimensions.

In the main decomposition analysis we can only use the largest connected set of workers and jobs to estimate fixed effects. Table 3.1 shows the overall size of the data set and the proportion thereof that is retained when considering the largest connected set over various decompositions reported in the main text. As one can see, the majority of the data is retained in most specifications with over 96% of observations used in our main decomposition reported in the final row.

²²One of these, [2321], has fewer than 9 observations in the entire period, and is excluded for statistical disclosure purposes.

TABLE 3..1 Largest connected set of common specifications (2011-19)

	No. obs	No. workers	No. firms	No. occs	No. jobs
All	1598919	339697	156178	368	335298
LCS for firm FE	93.57	81.01	63.57	100.00	79.30
LCS for clustered firm FE	96.46	83.36	89.34	100.00	91.89
LCS for job FE	88.50	77.98	59.14	100.00	56.93
LCS for clustered job FE	96.46	83.36	89.34	100.00	91.89

Notes: This table documents the proportion of the data retained in the largest connected set over various decompositions performed in the main text.

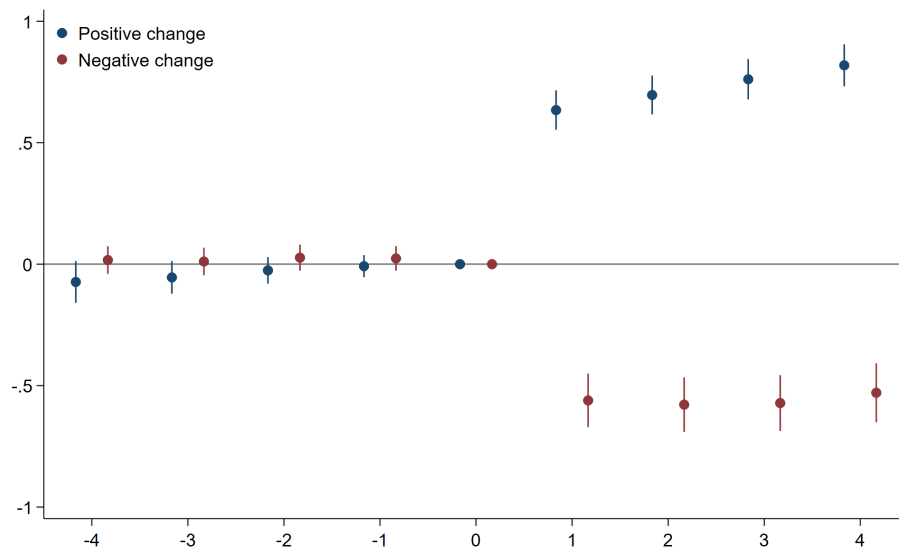
3..2 Testing the identifying assumption

In section 3.2, we described the exogeneity assumption we need for our estimated fixed effects to not be biased, and listed some examples under which exogeneity would fail. In this section, we perform a number of tests to show that this assumption is likely to hold. We follow analogous strategies in our context to the approaches that [Card et al. \[2018\]](#) takes to establish the suitability of the base AKM model.

The two-way fixed effect framework implies that when workers change jobs, they would experience a log wage change equal to the difference in the job fixed effect at the old job and at the new job. One implication of this is that for similar workers, the log wage change of moving from a former job to a latter job should be equal to the negative of log wage change of moving from that latter job to the former job. This may not be the case if for example job transitions occur due to workers matching to more suitable jobs over time. In that case, the mean wage changes associated with moves from less to more productive jobs should exceed that of moves in the other direction.

This idea is the basis of both tests presented in this section. First, we adopt the event study approach in [Card et al. \(2013\)](#) when we consider the wage effects of a worker moving from a less productive job to a more productive job and vice versa. [Figure 3.1](#) considers the wage rate in years relative to the year before a job switch occurs of workers who will (or have) switched jobs from the 1st to 4th (or 4th to 1st) quartile of jobs in terms of job fixed effects. This figure estimates wage trajectories using only within job switch pair variation and taking into account individual fixed effects. Wage trajectories are normalised to the year before the move (0'th year).

FIGURE 3..1 1 to 4 (and 4 to 1) job switches

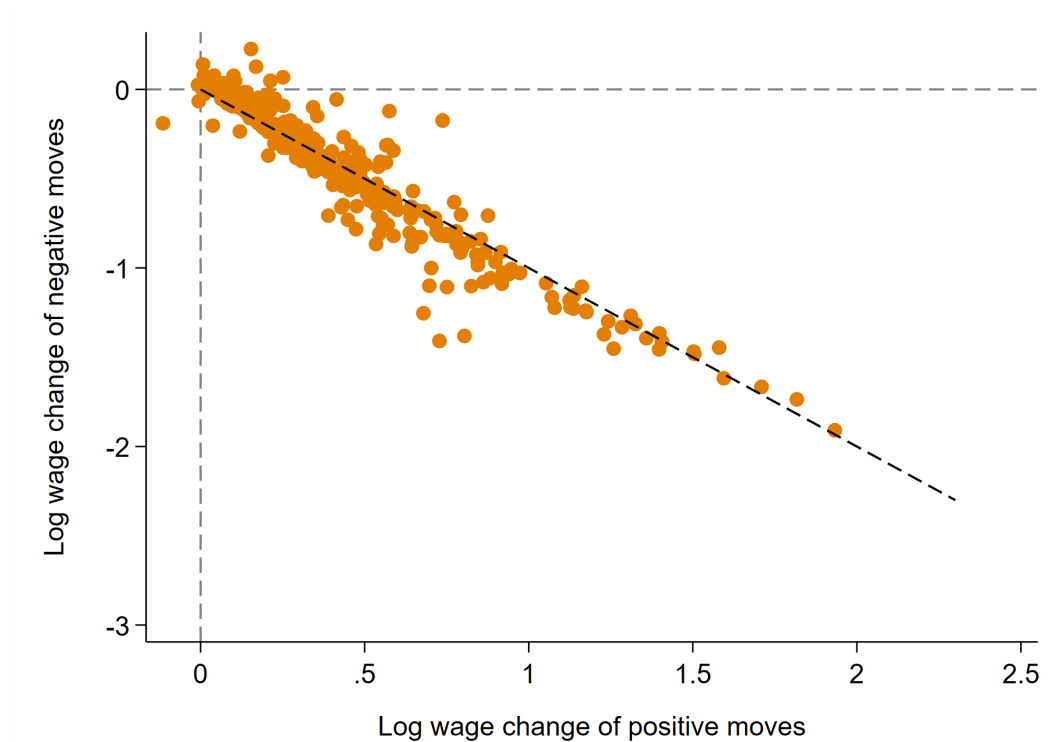


Notes: This figure shows the estimated log wage over time for workers who switch between jobs in the first and fourth quartile of the job fixed effect distribution. By including job pair fixed effects and individual fixed effects this figure shows the relative wages of workers who make a positive vs negative switch where wages are normalised to the year before the switch (0th year).

Two things are apparent from figure 3..1. First, there are no clear increases or decreases in pay in the year prior to, or proceeding, job changes, supporting what [Card et al. \[2018\]](#) describe as the step-like pattern of wage changes following job changes. Second, the degree to which wages change is symmetric; positive switches are associated with a similar increase in wages as negative switches are associated with a decrease. This gives evidence to suggest that there is not a large worker-job match-specific component of wages.

Second, the implied symmetry of wage changes in between job pair moves can be tested more directly by considering the average log wage change when workers move from a job to another. Because the set of moves between jobs is fairly sparse, we consider the average wage changes of moves between the wage clusters which we use to correct for limited mobility bias. We plot the changes in figure 3..2, restricting our analysis to firm-cluster pairs where we have observed at least 10 moves. If the exogeneity assumption is satisfied, then all the points should lie on the 45 degree line. We find that wage changes are roughly asymmetric as the model predicts, although there are a few outliers. The severity of the problem seems more acute for small moves, while larger moves between clusters seem largely symmetric as the model predicts. Many of the points also show that the wage loss associated with a move to a lower productivity cluster exceeds the wage gain of moving to a more productive cluster, which similarly contradicts prominent narratives of between job transitions that do not cohere with the implications of the two-way fixed effect framework.

FIGURE 3..2 Average log wage changes of moves between firm clusters

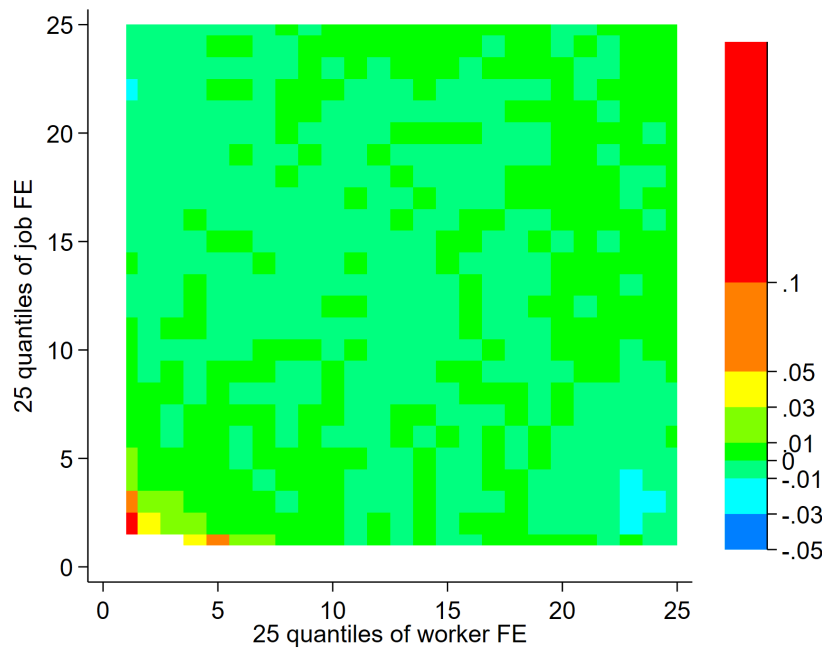


Notes: This figure plots the within-cluster average wage changes of moves between firm clusters. We use 30 firm-clusters which you use to correct for limited mobility bias. On the x-axis, we plot the wage change associated with moves from less productive to more productive clusters, while on the y-axis we plot the average wage change of moves from a more productive cluster to a less productive cluster. Each point represents a firm cluster pair, and there are 435 such pairs. We drop 94 such pairs where there are fewer than 10 moves in either direction. The dashed line is the 45 degree line which represents perfect symmetry.

3.3 Is the log-additive functional form a good specification for wages?

In this section, we check whether the log additive functional form in fixed effects is a good approximation to the data. To do this, we follow the test in [Card et al. \[2013\]](#), and aggregate our estimated fixed effects into 50 quantile groups. For each worker quantile and job quantile, we take the mean of the residual of the two-way fixed effect regression, reasoning that if the specification is a good fit to the data, then the mean of the residuals should be 0 or close to zero. We plot a heat map in figure 3.3 with the quantiles of worker FE on the x-axis, the quantiles of job FE on the x-axis, and the size of the mean residuals as the colour within the cell. Warmer colours (yellow, orange, red) mean that the model underpredicts the true wage in the data while cooler colours (cyan, blue) means that the model overpredicts the actual wage data. Green implies that the mean residual in the cell is within 1% on each side of 0.

FIGURE 3..3 Mean of residuals by worker FE and job FE quantiles
(25 groups each)



Notes: This figure plots the mean of the two-way fixed effect residuals for 25 worker FE quantiles by 25 job FE quantile cells for the two-way fixed effect regression with worker and job fixed effects without clustering.

The figure shows that for the vast majority of cells the difference between the predicted value and the true data is within 1% of the actual log wage. The model is particularly bad for a small number of low-skill workers at low-wage premium jobs, severely underestimating the pay of such workers by more than 10% and up to 30% at worst. This could be because of institutional factors, like minimum wage which are not explicitly considered in the model specification.

3.4 Review of limited mobility bias

Limited mobility bias occurs when second (or higher) order moments are estimated for fixed effects, which are estimated off few observations. In this case, the estimated fixed effects themselves are unbiased, but the high-order moments are biased. [Andrews et al. \[2012\]](#) pointed out this potential threat to identification and [Jochmans and Weidner \[2019\]](#) show how to understand the likely size of the bias as a function of the number of moves between workers and firms, or of the connectedness of the network of workers and firms. To see this more formally consider the variance of the fixed effects β , denoted as $V_Q = \beta' Q \beta$ for some matrix Q . Then our estimate of the variance would be given by $\hat{V}_Q = \hat{\beta}' Q \hat{\beta}$ the bias of which

can be characterized as follows:

$$\mathbb{E}[\hat{V}_Q] = V_Q + \mathbb{E}[\varepsilon'Z(Z'Z)^{-1}Q(Z'Z)^{-1}Z'\varepsilon] \quad (3.6)$$

This bias arises from an insufficient number of movers between jobs (or firms in the classic AKM literature) and is therefore termed the *limited mobility bias*. As a result, the variance of job (or firm) fixed effects is overstated at the expense of understating the co-variance between workers and jobs. This is a problem the whole literature attempting AKM decompositions faces but the scale of the issue has only recently become apparent with [Bonhomme et al. \[2020\]](#) for example showing that the limited mobility bias is severe and that bias correction is very important. In our case, relative to more standard worker-firm decompositions, by introducing jobs we increase mobility intuitively by capturing within-firm, cross-occupation moves, but simultaneously increase the number of parameters to be estimated. Therefore it's unclear ex-ante whether limited mobility bias represents a smaller or larger problem in our framework.

We solve the limited mobility bias by following [Bonhomme et al. \[2019\]](#) in first clustering job using a k-means algorithm, into K clusters. By performing this first step we greatly increase movers between clusters whilst retaining much of the underlying variation (30 clusters capture over 85% of the underlying job level wage variation).

3.5 Analysing the result of misspecifying the model with worker-firm FE when the true DGP is with worker-job FE

In the paper, we noted that when the true data-generating process underlying the data involves worker-job (firm-occupation pair) fixed effects but a worker-firm FE two-way fixed effect regression is specified, the model leads to the overestimation of variance attributed to individuals under certain conditions. We show this in more detail in this section. This result is conceptually similar to the work in [Abowd et al. \[1999\]](#) to show the effects of misspecifying the model with worker-industry FE when the true model is worker-firm.

We can rewrite the true data-generating process as follows.

$$\begin{aligned} \ln(w_{it}) &= \alpha_i + \phi_{F(i,t)} + (\psi_{J(i,t)} - \phi_{F(i,t)}) + \varepsilon_{it} \\ \phi_f &= \frac{1}{n_f} \sum_{j \in \bar{f}} \psi_j \end{aligned}$$

It seems from this rewriting that what misspecification does is to add a term to the part that varies across individuals and periods which is now an unobserved omitted variable, and which will therefore cause the regression to estimate the individual fixed effects and the mean firm fixed effects with bias. Let W denote a $NT \times 1$ vector of wages, L a $NT \times N$ matrix of indicators for individuals $i = 1, \dots, N$, and J a $NT \times H$ matrix of indicators for jobs $j = 1, \dots, H$. Let F denote a $NT \times G$ matrix of indicators for the firm the job is at (for $f = 1, \dots, G$). Let ε denote a $NT \times 1$ vector of ε_{it} errors. Finally, let K be a $H \times G$ matrix where each row $j = 1, \dots, H$ indexes which firm the job is at. Let α denote a $N \times 1$ vector of personal fixed effects, Ψ a $H \times 1$ vector of job fixed effects and Φ a $G \times 1$ vector of firm-specific mean job fixed effects.

$$\begin{aligned} W &= L\alpha + J\Psi + \varepsilon \\ &= L\alpha + F\Phi + J(\Psi - K\Phi) + \varepsilon \end{aligned}$$

When J is not observed, estimating an OLS regression on L and F leads to omitted variable bias. Suppose we estimate the misspecified model. Then the parameters estimated have omitted variable bias as follows, where $M_A = I - A(A'A)^{-1}A'$ for an arbitrary matrix A .

$$\alpha^* = \alpha + (L'M_F L)^{-1} L'M_F J(\Psi - K\Phi) \quad (3.7)$$

$$= \alpha + (L'M_F L)^{-1} L'M_F J\Psi \quad (3.8)$$

$$\Phi^* = \Phi + (F'M_L F)^{-1} F'M_L J(\Psi - K\Phi) \quad (3.9)$$

$$= (F'M_L F)^{-1} F'M_L J\Psi \quad (3.10)$$

The M_F and M_L matrices are residualising matrices, where $M_F L$ refers to the residual of a linear projection of F on L as in the Frisch-Waugh-Lovell theorem. It is instructive to re-write the expressions above as follows.

$$\alpha^* = \alpha + \beta \quad (3.11)$$

$$\Phi^* = \iota \quad (3.12)$$

Where β is the regression coefficient found from estimating the following by OLS: $J\Psi =$

$(M_F L)\beta + \xi$ and ι is similarly the regression coefficient from estimating $J\Psi = (M_L F)\iota + \xi$. Thus, the bias in variance estimation is easily characterisable.

$$\mathbb{V}[\alpha^*] - \mathbb{V}[\alpha] = \mathbb{V}[\beta] + 2Cov(\alpha, \beta) \quad (3.13)$$

$$\mathbb{V}[\Phi^*] - \mathbb{V}[\Phi] = \mathbb{V}[\iota] - \mathbb{V}[\Phi] \quad (3.14)$$

From this formula we can see that the individual fixed effects will be over estimated if $Cov(\alpha, \beta)$ is not sufficiently negative. To gain intuition as to the signs and magnitudes of $\mathbb{V}[\beta]$ and $2Cov(\alpha, \beta)$, we must understand β . β is an $N \times 1$ vector of coefficients, an individual element of β is large if for that individual the correlation between the firm-residualised individual and job effect is large. If the strength of this relationship varies over individuals $\mathbb{V}[\beta]$ will be large. If this relationship is stronger for higher-FE individuals $Cov(\alpha, \beta)$ will be positive, this is the most likely scenario.

It's harder to sign the difference in firm fixed effect variance. ι is the firm-analogue version of β , it is large if the firm conditional on worker is correlated with job FE. However, it's unclear whether the strength of variation in this coefficient will be less than or greater than the variation in firm effects itself and so signing the difference is difficult.

In general, from the above we can say that we expect the variance of the estimated worker and firm fixed effects in the mis-specified model to be wrong, and for it to be increasingly wrong the more residual variation there is in the relationship over workers between workers and jobs (conditional on firms) and over firms between firms and jobs (conditional on workers).

3.6 Comparing log-wage decomposition results against other results in the literature

In this section we compare results from estimating the usual worker-firm AKM model decomposition on our UK data to that from previous studies using data in other countries. Table 3.2 displays the main results. Panel [A] compares results using the basic AKM decomposition which does not correct for limited mobility bias whereas panel [B] shows results correcting for limited mobility bias. The main take away from this table is that in these specifications the UK data appears to behave similarly to that from other contexts. One possible exception is in the estimated negative effect found due to sorting in panel [A]. However, this is not as unusual a finding as it would appear from this table; in a recent survey,

Bonhomme et al. [2020] show that in 10 of the 37 AKM estimates from previous studies, a negative sorting effect is found.

TABLE 3..2 Comparison of basic AKM and clustered correlated fixed effect decomposition results with select estimates from previous studies.

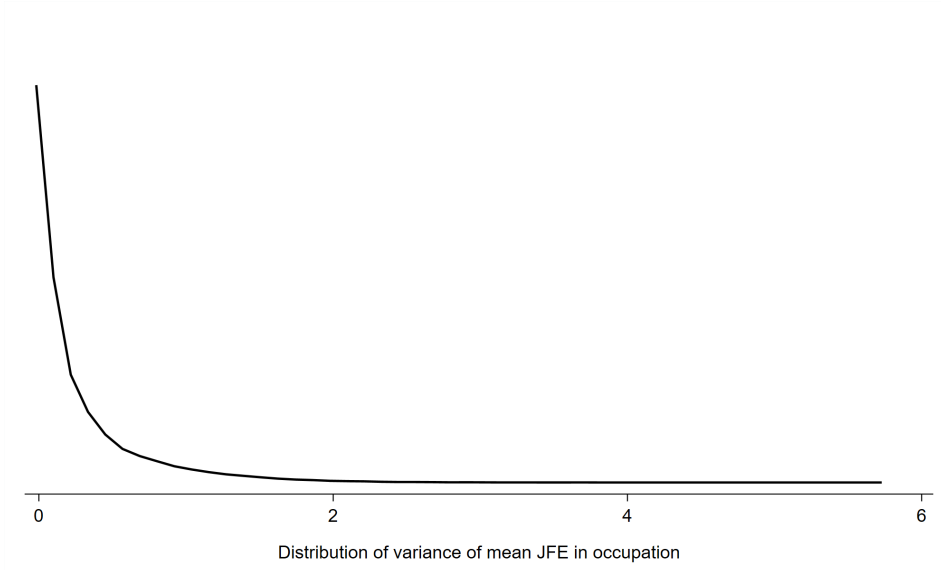
[A] Basic AKM Decomposition					
Paper	Period	Country	WFE	FFE	Sort
Card et al. [2013]	2002-09	Germany	51%	21%	16%
Card et al. [2016]	2002-09	Portugal	58%	20%	11%
Song et al. [2019]	2007-13	USA	52%	9%	12%
Alvarez et al. [2018]	2008-12	Brazil	58%	16%	18%
Our data	2011-19	UK	64%	24%	-2%
[B] Clustered Random Effects Bias Correction					
Paper	Period	Country	WFE	FFE	Sort
Bonhomme et al. [2019]	2002-04	Sweden	60%	3%	12%
Our data	2011-19	UK	54%	10%	19%

Notes: This table compares results using the UK data used in this paper when estimating classic and bias-corrected AKM models, to those found in selected papers from the previous literature.

3..7 Direct evidence that occupational changes are relatively unlikely

In section 3.4.1, we argued that the misspecification of the AKM leads to over estimation of the variance of the individual fixed effect especially if an individual's occupation is persistent over time. In this section we provide some direct evidence of the persistence of occupation choice. We use mean job fixed effect within an occupation as a proxy for the relative productivity of the job. We then calculate the variance of the mean job fixed effect across occupations for each individual, and normalise it by the value of $\mathbb{V}[\mathbb{E}_o[\psi_j]]$ for the whole population. This measures an individual's relative variance in occupational quality for their time in the panel relative to the overall population between-occupation job fixed effect variance. We plot the values for workers who have been in the panel for at least 5 years in figure 3.4 below. We find that this variance is very low for most individuals and the distribution is skewed towards 0. This suggests that most workers do not substantially change the quality of the occupation they are employed at over time.

FIGURE 3..4 Distribution of the variance of mean JFE within occupations over individuals



Notes: This figure plots the variance of the occupation mean fixed effect for the occupations an individual was observed in over the length of the panel, normalised over the overall variance in occupational mean job fixed effects.

3..8 Additional results on the relation between job FE as wage premia related to rent-sharing

As discussed in the introduction, estimates of the firm fixed effects in implementations of the AKM model have been interpreted as wage premia arising from firms rent-sharing with their workers (see [Card et al. \[2018\]](#) in particular for an attempt to synthesise the rent-sharing and AKM literatures). We check this story in our setting by investigating whether the job fixed effect $\psi_{J(i,t)}$ correlate with the firm's average productivity. Specifically, we consider a model where the job fixed effects can be modelled as follows, where $\phi_{F(i,t)}$ denotes industry fixed effects and $\chi_{Occ(i,t)}$ denotes occupation fixed effects.

$$\psi_{J(i,t)} = \alpha_0 + \beta \cdot \ln(\text{prod}_{F(i,t)}) + \phi_{F(i,t)} + \chi_{Occ(i,t)} + u_i \quad (3.15)$$

Since $\psi_{J(i,t)}$ is an estimated coefficient, we have to take into account its covariance matrix when estimating the parameters of equation 3.15 including β . We do this by substituting equation 3.15 into equation 3.1, and estimating the whole model²³. We use data on firm characteristics from the Business Structure Database (BSD), which contains detailed panel

²³An alternative approach is to use Generalised Least Squares after estimating the covariance matrix of $\psi_{J(i,t)}$.

information on firm formation, as well as some basic characteristics like industry, employment, tax status and revenue. A major disadvantage of the BSD is that it does not contain information on firm profits, and we are unable as a result to test directly whether the estimated coefficients relate to firm productivity. We proxy for productivity using revenue per worker.

We run our analysis on three samples. Besides the full sample, we also consider splitting the sample into two groups (1) private workers not covered by a collective bargaining agreement and (2) private workers who are covered by a collective bargaining agreement and public/non-profit sector workers²⁴. If the estimated fixed effects were driven by rent-sharing, then we would expect that workers in private firms not covered by a collective bargaining agreement to have a higher value of β_o over all occupation groups. We thus estimate equation 3.15 on sub-samples of the two groups. Results for private, non-CA workers are presented in column (2) while results for public sector or workers under CA are presented in column (3)²⁵.

We present the results in table 3.3. We find that the job fixed effect is increasing in the revenue per worker of the firm, and that this correlation is greater for the sub-sample of jobs at private firms that are not covered by collective agreements. This coheres with the interpretation of these job fixed effects as wage premia paid to workers.

²⁴We use indicators in the ASHE dataset on whether the worker works at a public, private or non-profit institution. We also use indicators in the dataset on whether the worker is covered by a collective agreement. We lump workers under all kinds of collective agreements in the same category.

²⁵We also calculated the regression on public/private, and CA/non-CA sub-samples separately. Results for public sector workers are similar to those for CA workers and results for private workers are similar to those for workers not under CAs. Thus, we present only results for the two combined groups, noting that the results are not driven in particular by either the public/private or the CA/no-CA dimension.

TABLE 3..3 Estimates of β in equation 3.15 (2011-19)

	All workers	Private, no CA workers	Public or CA workers
Variable	(1)	(2)	(3)
Firm revenue per worker	0.018 (0.002)	0.027 (0.002)	0.005 (0.002)
Obs	1537793	740,090	715,920
R^2	0.897	0.911	0.906

Notes: This table shows the results from estimating $\psi_{j(i,t)} = \alpha_0 + \beta \cdot \ln(\text{prod}_{F(i,t)}) + \phi_{F(i,t)} + \chi_{Occ(i,t)} + u_i$ on the full sample in column one, a sample of private workers not covered by a collective bargaining agreement in column two, and on public workers or those covered by a collective bargaining agreement in column three.

We run a similar regression with different specifications to check if the relation to firm characteristics differs across occupations. Table 3.4 summarises the coefficients of a similar regression to equation 3.15, but where we interact the log worker productivity variable with the 1 digit occupation group; we also plot the coefficients in figure 3.5. We find that for all occupation groups, there is a positive relation between revenue per worker and the job fixed effect. We find statistically significant relationships for all occupation categories for the sample of jobs at private firms and jobs not covered by collective agreements. On the other hand, we find that the relations between the wage premium for managers, professionals, customer service and machine operators and log revenue per worker are not statistically significant. This again seems to be evidence in support of the wage premium interpretation of the job fixed effects.

An interesting aspect of the results in figure 3.5 is how the coefficient varies across occupations. The coefficient is broadly higher for more skilled workers than for less skilled workers for jobs in private firms, with workers in care and leisure having a particularly low correlation and customer service having a high correlation. In the public sector however, the premium paid to managers and professionals have no relation to firm revenue per worker, but the premium paid to mid and low-skill occupations seem to still exhibit some relation. This could reflect workers in high-skill occupations in these public institutions being better able to insure for wage variation.

For robustness, we check whether this pattern persists if we use a measure of the skill

requirement of the occupation²⁶. The results for these regressions are summarised in table 3.5 and figure 3.6. We similarly find that overall the relation between revenue per worker and the job fixed effect is greater for lower-skilled workers than for higher-skilled workers, but this is primarily driven by a composition effect. The relation is statistically insignificant and near zero for high skill workers at public firms and at jobs covered by collective agreements, but very large for similarly skilled workers in jobs at private firms not covered by a collective agreement. In contrast, the correlation of job fixed effect with revenue per worker is positive and statistically significant in all samples for lower skilled workers. We leave a full explication of these puzzles to future work.

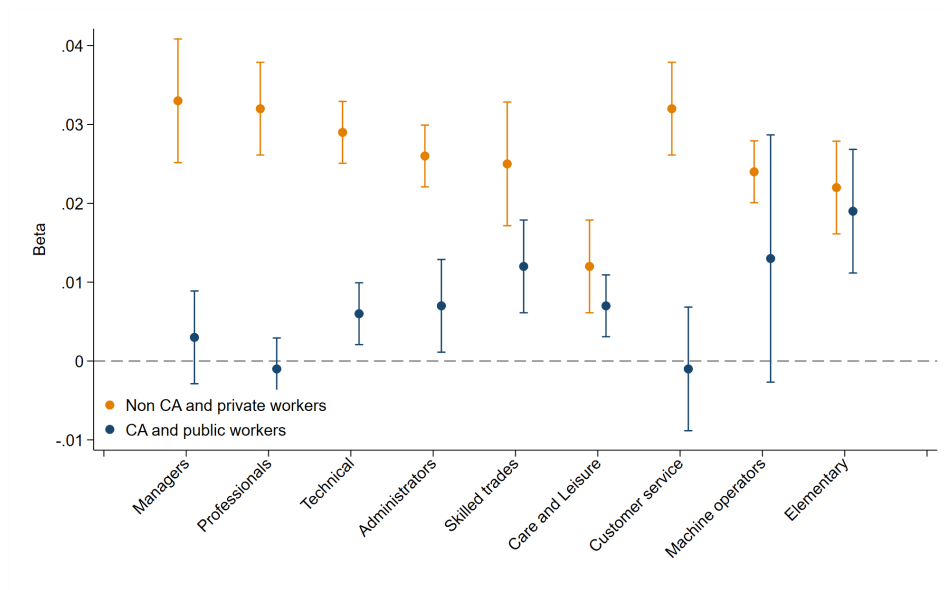
²⁶We use the skill requirements for occupations used by the UK Home Office, and set out in their Appendix J tables. RQF 3 corresponds to a high school degree, while RQF 4 corresponds to 1 year of a bachelors degree. RQF 6 corresponds to a bachelor degree.

TABLE 3..4 Coefficients on log revenue per worker interacted with
1 digit occupation group

Occupation Group	All workers (1)	Private, no CA workers (2)	Public or CA workers (3)
Managers, directors and senior officials	0.023 (0.003)	0.033 (0.004)	0.003 (0.003)
Professional occupations	0.010 (0.004)	0.032 (0.003)	-0.001 (0.002)
Associate professional and technical occupations	0.019 (0.002)	0.029 (0.002)	0.006 (0.002)
Administrative and secretarial occupations	0.021 (0.002)	0.026 (0.002)	0.007 (0.003)
Skilled trades occupations	0.024 (0.003)	0.025 (0.004)	0.012 (0.003)
Caring, leisure and other service occupations	0.010 (0.002)	0.012 (0.003)	0.007 (0.002)
Sales and customer service occupations	0.022 (0.004)	0.032 (0.003)	-0.001 (0.004)
Process, plant and machine operatives	0.025 (0.003)	0.024 (0.002)	0.013 (0.008)
Elementary occupations	0.022 (0.002)	0.022 (0.003)	0.019 (0.004)
Obs	1537793	740,090	715,920
R ²	0.897	0.911	0.906

Notes: This table shows the coefficients and corresponding standard errors found when estimating equation 3.15, interacting log revenue per worker with occupation group.

FIGURE 3..5 Relation between job fixed effect and firm revenue per worker, by occupation group (2011-19)



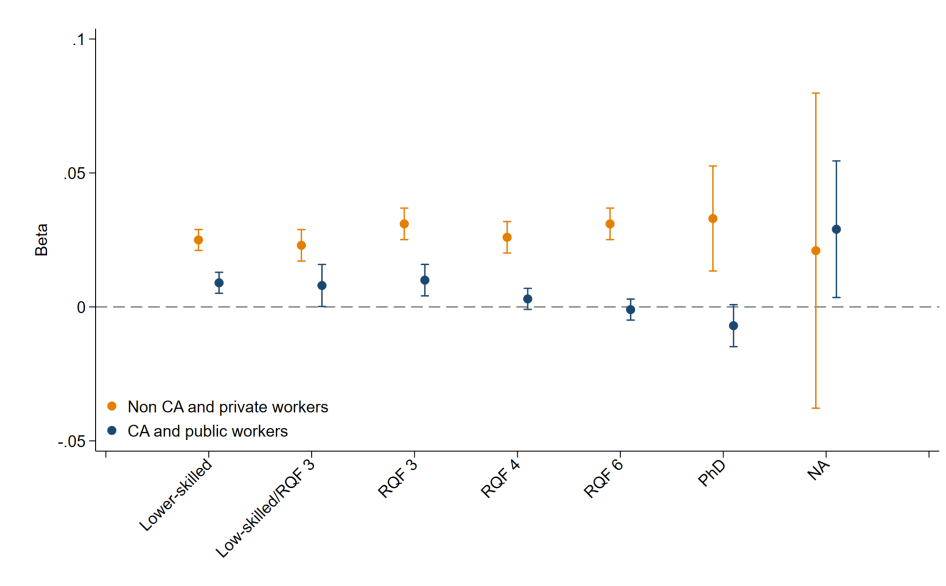
Notes: This figure shows the coefficients and corresponding 95% confidence intervals found when estimating equation 3.15, interacting log revenue per worker with occupation group.

TABLE 3..5 Coefficients on log revenue per worker interacted with skill groups

	All workers	Private, no CA workers	Public or CA workers
Occupation Group	(1)	(2)	(3)
Lower-skilled	0.019 (0.002)	0.025 (0.002)	0.009 (0.002)
RQF 3/Lower-skilled	0.016 (0.003)	0.023 (0.003)	0.008 (0.004)
RQF 3	0.023 (0.002)	0.031 (0.003)	0.010 (0.003)
RQF 4	0.015 (0.002)	0.026 (0.003)	0.003 (0.002)
RQF 6	0.014 (0.004)	0.031 (0.003)	-0.001 (0.002)
PhD	0.014 (0.006)	0.033 (0.010)	-0.007 (0.004)
Unclassified	0.016 (0.010)	0.021 (0.030)	0.029 (0.013)
Obs	1534631	738,048	714,953
R^2	0.897	0.911	0.906

Notes: This table shows the coefficients and corresponding standard errors found when estimating equation 3.15, interacting log revenue per worker with the skill level of the occupation, as determined by the UK Home Office.

FIGURE 3.6 Relation between job fixed effect and firm revenue per worker, by skill group (2011-19)

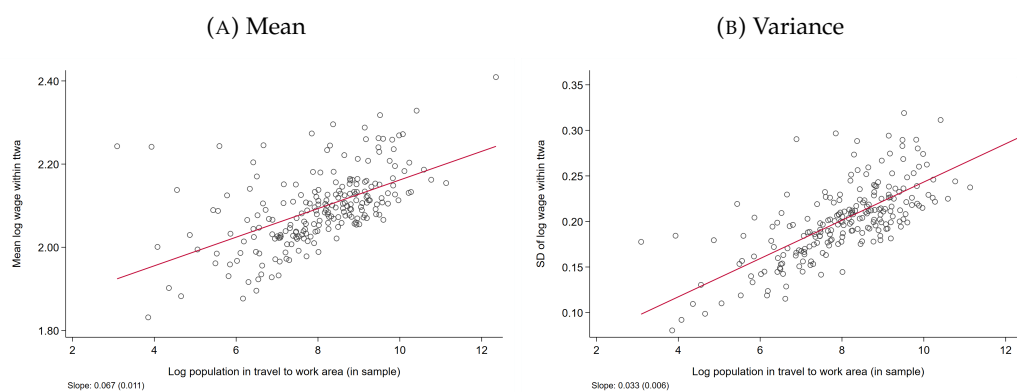


Notes: This figure shows the coefficients and corresponding 95% confidence intervals found when estimating equation 3.15, interacting log revenue per worker with the skill level of the occupation, as determined by the UK Home Office.

3.9 Appendix figures and tables relating to wage inequality across geography

Panel 3.7a reproduces the well-known fact that mean wages are higher in larger labour markets, while panel 3.7b shows that the variance of log-wages is also increasing log-linearly with log-population within a TTWA.

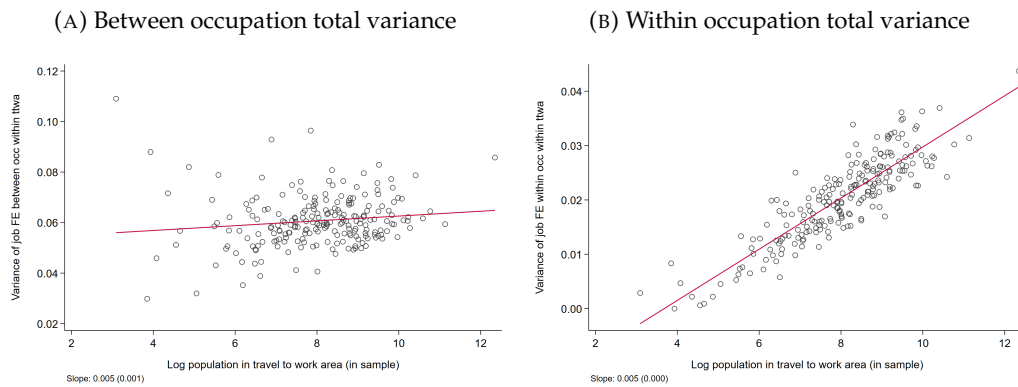
FIGURE 3.7 Within TTWA log population vs moments of log pay (2011-19)



Notes: Panel 3.7a plots the mean of log-hourly wage (y-axis) against log-population (x-axis) in the travel-to-work area. Panel 3.7b plots the variance of log hourly wage (y-axis) against log-population (x-axis) in the travel-to-work area.

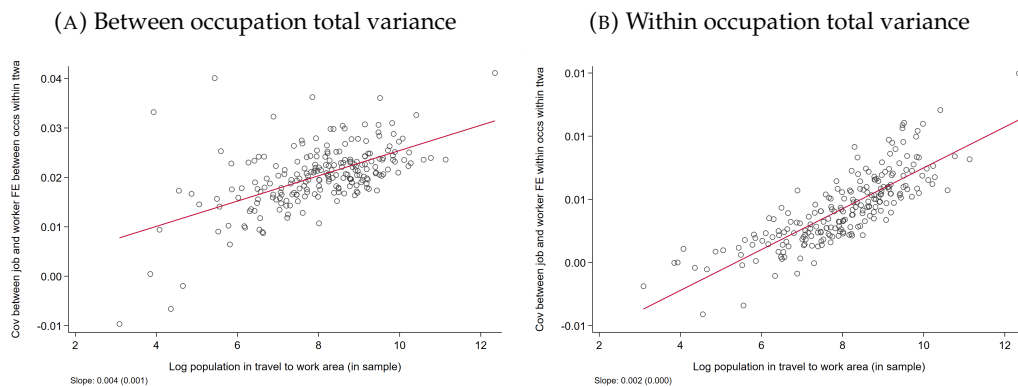
In figure 3.8, we plot scatterplots of the absolute values of the part of log-wage variance explainable by between-occupation job fixed effect variance (panel 3.8a), and the part explainable by within-occupation job fixed effect variance (panel 3.8b). Analogously, figure 3.9 plots the absolute values of the covariance of job fixed effects and individual fixed effects attributable to between-occupation sorting (panel 3.9a) and within-occupation sorting (panel 3.9b).

FIGURE 3.8 Within TTWA log population vs variance of job FE between and within occupation



Notes: Panel (a) plots a scatterplot of TTWAs with log-population on the x-axis and the between-occupation component of job-class fixed effect variance on the y-axis. Panel (b) plots a scatterplot of TTWAs with log-population on the x-axis and the within-occupation component of job-class fixed effect variance on the y-axis.

FIGURE 3.9 Within TTWA log population vs covariance components decomposed into between and within occupation parts



Notes: Panel (a) plots a scatterplot of TTWAs with log-population on the x-axis and the between-occupation component of worker-job-class covariance on the y-axis. Panel (b) plots a scatterplot of TTWAs with log-population on the x-axis and the within-occupation component of worker-job-class covariance on the y-axis.

3.10 Is the relation between size of travel to work area and variance due to occupational composition?

The increase in within-occupation sorting documented in figure 3.6.4d could reflect the tendency for jobs in high-wage occupations to locate in larger labour markets, given that we know that the degree of sorting is higher in higher wage occupations. Thus, the relation documented here between TTWA size and positive assortative matching could reflect a composition effect rather than a true correlation with size. To obtain direct evidence on this, we calculate the standard deviation of job fixed effects within each occupation for each travel-to-work area in each period ($\sigma_{kl\tau}$) as well as the correlation between worker fixed effects and job fixed effects within each travel-to-work area, occupation and period ($\rho_{kl\tau}$) (where τ indexes the period). We then perform simple regressions of these moments on log population within the travel-to-work area, controlling for occupation and period fixed effects (denoted by γ_k , and γ_τ respectively for the standard deviation regression, and by δ_k and δ_τ for the correlation regression).

$$\sigma_{kl\tau} = \gamma_0 + \gamma_1 \cdot \ln(pop_l) + \gamma_k + \gamma_\tau + u_{kl\tau} \quad (3.16)$$

$$\rho_{kl\tau} = \delta_0 + \delta_1 \cdot \ln(pop_l) + \delta_k + \delta_\tau + v_{kl\tau} \quad (3.17)$$

The results of this regression for γ_1 and δ_1 are reported in table 3.6 for both the full set of TTWA-occupation-period cells and for the set of cells with at least 30 observations. We find that even after controlling for occupational differences, there is still a residual correlation between travel-to-work area size and both the standard deviation of job fixed effects and the correlation between worker and job fixed effects. This qualitative result is robust to dropping smaller cells for which these moments might not be as reliable. Thus, we conclude that being in a larger local labour market is associated with larger variance of job fixed effects, larger variance of worker fixed effects, and a greater correlation between worker and job fixed effects, even conditional on the occupational composition of the travel-to-work area.

TABLE 3..6 Coefficients on log population from regression
of occ-ttwa moments

Coefficient on log pop	All > 10	All > 30
Standard deviation of job FE	0.018 (0.001)	0.013 (0.001)
Standard deviation of worker FE	0.015 (0.001)	0.008 (0.001)
Correlation of worker and job fixed effects	0.002 (0.000)	0.001 (0.000)
N	42,328	18,749

Notes: This table shows the results from estimating equations of the form given in 3.16 of the relationship between the TTWA-occupation-period standard deviation of job and worker FE as well as the correlation between the two and TTWA population. Column one gives results for all cells with more than 10 observations and column two for all cells with more than thirty observations. Standard errors are clustered at the TTWA level.

Closing Remarks

This thesis aims, in some small way, to further our understanding of both the degree of and determinants of, inequality over two dimensions. First inequality of opportunity over space, and second wage inequality. This analysis, however, leaves many stones unturned and (as is so often the case) raises many more questions. In these closing remarks, I will first summarize each chapter and the possible future directions for research that arise from each chapter, before concluding with more general “unturned stones”.

In the first chapter, I quantified and explored the impact of spatial inequality of opportunity relating to primary school completion in Benin, Cameroon, and Mali using a movers design. I found that moving to a better location at birth has a significant impact on a child’s likelihood of completing primary school. Specifically, moving to a one standard deviation better location at birth increases the probability of completing primary school by 15 percentage points. Gender is an essential factor in this analysis, with place having a more significant effect on girls than boys. Furthermore, local characteristics such as urbanization, access to electricity, and employment in agriculture are associated with place effects.

Religion also plays a role in the educational outcomes of children in these countries. Although it is weakly correlated with average causal place effects, it is strongly correlated with the gender gap in the effect of place on educational attainment. These findings highlight the importance of considering both gender and local characteristics in addressing spatial inequality of opportunity and improving access to education. This work illustrates the need for, and potential scope of, place-based policies in reducing inequality of opportunity. In future work, it will be crucial to understand the underlying causes of this variability, and arm policymakers with the tools to mediate spatial inequality of opportunity most effectively. Additionally, although across-space inequality of opportunity was found to be large, future work should also aim to quantify within-space, cross-group inequalities of opportunity and consider dimensions of inequality outside of those captured by education.

The second chapter of this thesis builds on the first chapter’s questions and explores how policymakers can reduce spatial inequality of opportunity in Benin, Cameroon, and Mali through road building. The chapter develops an approach to measure the impact of

any given road on locations across the entire network, using a sufficient statistic approach that endogenizes skill premia across localities. I find that changes in the road network significantly affected the spatial distribution of opportunity, with road investments raising the probability of primary school completion by 7.5p.p. more in the most-affected locations relative to the least-affected ones. However, the impact of possible future roads varies across roads within each country, with roads that connect two periphery locations more likely to increase inequality than those that connect a main and a periphery location.

To answer counterfactual questions, I write down a structural spatial general equilibrium model consistent with the sufficient statistic result, and parameterizes it using the estimated coefficients. By considering cross-network variation in counterfactuals, I find that network-level density plays an important role in the magnitude of effects. There is also an equity-efficiency trade-off in road building — if a social planner values equity this will change where roads should be built in the future. This work opens the door to a potentially exciting area of future research considering how local and global network-level characteristics determine the impact of any place-based policy. Another natural next step for analysis would be to estimate the impact of other types of public policies or infrastructure on local opportunities.

In the third chapter, we propose that the classic two-way fixed effect model with worker and firm fixed effects is misspecified and that this misspecification produces a misleading decomposition of the components of log-wage variance. Instead of matching to firms, workers match to jobs that may differ within firms — e.g. a software engineer and a chef at Google are different positions. We estimate a two-way fixed effects model with individual and job fixed effects. We find that the component of variance due to individual heterogeneity, usually found to be 50-60%, accounts for only 18% of log wage inequality in our model. Conversely, we find that job heterogeneity and the sorting of workers into jobs account for 38% and 29% of the log-wage variance respectively. The majority of variation in these components can be explained by cross rather than within-occupation differences, highlighting the importance of occupations and jobs rather than firms, in explaining UK wage inequality.

We also find that sorting plays a greater role in explaining wage variance in higher mean-wage occupations. As a result, different approaches should be used to analyze occupational labor markets depending on the level of heterogeneity between individuals and firms. Occupations with greater heterogeneity are better explained by search and matching models that produce more wage dispersion, while those with more interchangeable workers and firms are better explained by spot models. Finally, we find that sorting becomes increasingly

important in larger labor markets, with sorting to firms within occupations being more significant in larger cities than in smaller towns. Wage inequality in small towns is likely due to differences in occupation, whereas wage inequality in large cities is due to both workers being at different firms within the same occupation and being in differing occupations. A natural next step for this, mainly empirical, analysis, would be to nest it within a more structural model of wage setting and inequality to better understand how government policies can influence such observed inequality.

In general, the themes discussed in this thesis, although hopefully providing some answers, leave many questions and areas for future work. One larger cross-cutting area for future work is the intersectional analyses of inequality: exploring the intersectional nature of inequality across multiple dimensions such as gender, location, occupation, and socioeconomic status. The first chapter highlights the importance of gender and local characteristics in addressing spatial inequality of opportunity, while the second chapter discusses the equity-efficiency trade-off in road building. The third chapter emphasizes the role of job heterogeneity and sorting in explaining wage inequality. An intersectional analysis that combines these insights could provide a more nuanced understanding of how different dimensions of inequality interact and influence each other.

To understand the complex, nuanced, and intersection nature of inequalities, and more importantly, what policymakers can do about them, we also need methodological advancements in analyzing inequality. The first chapter uses a movers design to explore the impact of spatial inequality of opportunity, while the second chapter develops an approach to measure the impact of roads on the spatial distribution of opportunity. The third chapter challenges the standard two-way fixed effects model and proposes a new approach to decomposing wage variance. Future research could build on these methodological advancements and develop new approaches to analyzing inequality that accounts for the intersectional nature of inequality and its multidimensional aspects.

Finally, once we have a better grasp of the degree of inequalities, and how policies might affect this, we then of course need to know the objective function. What level of equity, and on what dimensions, should we aim for? In this thesis, I have mainly sidestepped this question, but when expanding this work this is an important question one will encounter quickly.

Bibliography

Daniel Aaronson. Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. *Journal of Human Resources*, pages 915–946, 1998.

John M Abowd, Francis Kramarz, and David N Margolis. High wage workers and high wage firms. *Econometrica*, 67(2):251–333, 1999.

John M. Abowd, Robert H. Creecy, and Francis Kramarz. Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data. Longitudinal Employer-Household Dynamics Technical Papers 2002-06, Center for Economic Studies, U.S. Census Bureau, March 2002.

Tasso Adamopoulos. Spatial integration, agricultural productivity, and development: A quantitative analysis of ethiopia's road expansion program. In *2019 Meeting Papers*. Society for Economic Dynamics, 2019.

Anjali Adukia, Sam Asher, and Paul Novosad. Educational investment responses to economic opportunity: evidence from indian road construction. *American Economic Journal: Applied Economics*, 12(1):348–76, 2020.

Sarthak Agrawal and David Phillips. Catching up or falling behind? geographical inequalities in the uk and how they have changed in recent years. *The Institute for Fiscal Studies*, page p2, 2020.

Gabriel M Ahlfeldt, Stephen J Redding, Daniel M Sturm, and Nikolaus Wolf. The economics of density: Evidence from the berlin wall. *Econometrica*, 83(6):2127–2189, 2015.

Sheila Aikman and Elaine Unterhalter. *Beyond access: Transforming policy and practice for gender equality in education*. Oxfam, 2005.

Alberto Alesina, Sebastian Hohmann, Stelios Michalopoulos, and Elias Papaioannou. Religion and educational mobility in africa. Technical report, National Bureau of Economic Research, 2020.

- Alberto Alesina, Sebastian Hohmann, Stelios Michalopoulos, and Elias Papaioannou. Inter-generational mobility in africa. *Econometrica*, 89(1):1–35, 2021.
- Treb Allen and Costas Arkolakis. Trade and the topography of the spatial economy. *The Quarterly Journal of Economics*, 129(3):1085–1140, 2014.
- Treb Allen and Costas Arkolakis. Supply and demand in space. *NBER Working Paper*, 2022.
- Treb Allen and Dave Donaldson. Persistence and path dependence in the spatial economy. Technical report, National Bureau of Economic Research, 2020.
- Treb Allen, Costas Arkolakis, and Xiangliang Li. On the equilibrium properties of network models with heterogeneous agents. Technical report, National Bureau of Economic Research, 2020a.
- Treb Allen, Costas Arkolakis, and Yuta Takahashi. Universal gravity. *Journal of Political Economy*, 128(2):393–433, 2020b.
- Treb Allen, David Atkin, Santiago Cantillo, and Carlos Hernandez. Trucks. *presentation slides*, 2020c.
- Jorge Alvarez, Felipe Benguria, Niklas Engbom, and Christian Moser. Firms and the decline in earnings inequality in brazil. *American Economic Journal: Macroeconomics*, 10(1):149–89, 2018.
- Sonya Anderson. *Girls’ education and the “pedagogy of difference”: A study of teacher attitudes and gender-equitable pedagogy in Benin*. Harvard University, 2005.
- Martyn J Andrews, Leonard Gill, Thorsten Schank, and Richard Upward. High wage workers match with high wage firms: Clear evidence of the effects of limited mobility bias. *Economics Letters*, 117(3):824–827, 2012.
- Sam Asher and Paul Novosad. Rural roads and local economic development. *American economic review*, 110(3):797–823, 2020.
- David Atkin. Endogenous skill acquisition and export manufacturing in mexico. *American Economic Review*, 106(8):2046–85, 2016.
- Emmanuelle Auriol, Julie Lassebie, Amma Panin, Eva Raiber, and Paul Seabright. God insures those who pay? formal insurance and religious offerings in ghana. *The Quarterly Journal of Economics*, 135(4):1799–1848, 2020.

- Michael Bailey, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3):259–80, 2018.
- Clare Balboni, Oriana Bandiera, Robin Burgess, Maitreesh Ghatak, and Anton Heil. Why do people stay poor? *The Quarterly Journal of Economics*, 137(2):785–844, 2022.
- Abhijit Banerjee, Esther Duflo, and Nancy Qian. On the road: Access to transportation infrastructure and economic growth in china. *Journal of Development Economics*, 145:102442, 2020.
- Mahdi Barouni and Stijn Broecke. The returns to education in africa: some new estimates. *The Journal of Development Studies*, 50(12):1593–1613, 2014.
- Dominick Bartelme. Trade costs and economic geography: evidence from the us. *Work. Pap., Univ. Calif., Berkeley*, 2015.
- Matthias Basedau and Alexander Stroh. How ethnic are african parties really? evidence from four francophone countries. *International Political Science Review*, 33(1):5–24, 2012.
- Matthias Basedau, Gero Erdmann, Jann Lay, and Alexander Stroh. Ethnicity and party preference in sub-saharan africa. *Democratization*, 18(2):462–489, 2011.
- Joerg Baten, Michiel de Haas, Elisabeth Kempter, and Felix Meier zu Selhausen. Educational gender inequality in sub-saharan africa: A long-term perspective. *African Economic History Network Working Paper Series*, 54, 2020.
- Martin Battle and Jennifer C Seely. It’s all relative: Modeling candidate support in benin. *Nationalism and Ethnic Politics*, 16(1):42–66, 2010.
- Nathaniel Baum-Snow and Ronni Pavan. Inequality and city size. *Review of Economics and Statistics*, 95(5):1535–1548, 2013.
- Nathaniel Baum-Snow, Loren Brandt, J Vernon Henderson, Matthew A Turner, and Qinghua Zhang. Roads, railroads, and decentralization of chinese cities. *Review of Economics and Statistics*, 99(3):435–448, 2017.
- Nathaniel Baum-Snow, J Vernon Henderson, Matthew A Turner, Qinghua Zhang, and Loren Brandt. Does investment in national highways help or hurt hinterland city growth? *Journal of Urban Economics*, 115:103124, 2020.

- Kristian Behrens, Gilles Duranton, and Frédéric Robert-Nicoud. Productive cities: Sorting, selection, and agglomeration. *Journal of Political Economy*, 122(3):507–553, 2014.
- Mattia C Bertazzini. The long-term impact of italian colonial roads in the horn of africa, 1935–2015. *Journal of Economic Geography*, 22(1):181–214, 2022.
- Eleonora Bertoni, Michele Di Maio, Vasco Molini, and Roberto Nistico. Education is forbidden: The effect of the boko haram conflict on education in north-east nigeria. *Journal of Development Economics*, 141:102249, 2019.
- Adrien Bilal. The geography of unemployment. Technical report, National Bureau of Economic Research, 2021.
- Hoyt Bleakley and Jeffrey Lin. Portage and path dependence. *The quarterly journal of economics*, 127(2):587–644, 2012.
- Moussa P Blimpo, Robin Harding, and Leonard Wantchekon. Public investment in rural infrastructure: Some political economy considerations. *Journal of African Economies*, 22 (suppl_2):ii57–ii83, 2013.
- Justin Bloesch, Birthe Larsen, and Bledi Taska. Which workers earn more at productive firms? position specific skills and individual worker hold-up power. *SSRN Working paper*, 2021.
- Stéphane Bonhomme, Thibaut Lamadon, and Elena Manresa. A distributional framework for matched employer employee data. *Econometrica*, 87(3):699–739, 2019.
- Stéphane Bonhomme, Kerstin Holzheu, Thibaut Lamadon, Elena Manresa, Magne Mogstad, and Bradley Setzler. How much should we trust estimates of firm effects and worker sorting? Technical report, National Bureau of Economic Research, 2020.
- Kirill Borusyak and Peter Hull. Non-random exposure to exogenous shocks: Theory and applications. Technical report, National Bureau of Economic Research, 2020.
- Kirill Borusyak, Peter Hull, and Xavier Jaravel. Quasi-experimental shift-share research designs. Technical report, National Bureau of Economic Research, 2018.
- Helen N Boyle. Between secular public schools and qur’anic private schools: The growing educational presence of malian medersas. *The Review of Faith & International Affairs*, 12(2): 16–26, 2014.

- Bart J Bronnenberg, Jean-Pierre H Dubé, and Matthew Gentzkow. The evolution of brand preferences: Evidence from consumer migration. *American Economic Review*, 102(6):2472–2508, 2012.
- Wyatt Brooks and Kevin Donovan. Eliminating uncertainty in market access: The impact of new bridges in rural nicaragua. *Econometrica*, 88(5):1965–1997, 2020.
- Gharad Bryan, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak. Underinvestment in a profitable technology: The case of seasonal migration in bangladesh. *Econometrica*, 82(5):1671–1748, 2014.
- Robin Burgess, Remi Jedwab, Edward Miguel, Ameet Morjaria, and Gerard Padró i Miquel. The value of democracy: evidence from road building in kenya. *American Economic Review*, 105(6):1817–51, 2015.
- Piet Buys, Uwe Deichmann, and David Wheeler. *Road network upgrading and overland trade expansion in Sub-Saharan Africa*, volume 4097. World Bank Publications, 2006.
- Lorenzo Caliendo, Fernando Parro, Esteban Rossi-Hansberg, and Pierre-Daniel Sarte. The impact of regional and sectoral productivity changes on the us economy. *The Review of economic studies*, 85(4):2042–2096, 2018.
- David Canning and Peter Pedroni. Infrastructure, long-run economic growth and causality tests for cointegrated panels. *The Manchester School*, 76(5):504–527, 2008.
- David Card, Jörg Heining, and Patrick Kline. Workplace heterogeneity and the rise of west german wage inequality. *The Quarterly journal of economics*, 128(3):967–1015, 2013.
- David Card, Ana Rute Cardoso, and Patrick Kline. Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *The Quarterly Journal of Economics*, 131(2):633–686, 2016.
- David Card, Ana Rute Cardoso, Joerg Heining, and Patrick Kline. Firms and labor market inequality: Evidence and some theory. *Journal of Labor Economics*, 36(S1):S13–S70, 2018.
- David Card, Jesse Rothstein, and Moises Yi. Industry wage differentials: A firm-based approach. *Unpublished draft, University of California, Berkeley*, 2022.
- Marie Castaing Gachassin. Should i stay or should i go? the role of roads in migration decisions. *Journal of African Economies*, 22(5):796–826, 2013.

- Raj Chetty and Nathaniel Hendren. The impacts of neighborhoods on intergenerational mobility i: Childhood exposure effects. *The Quarterly Journal of Economics*, 133(3):1107–1162, 2018a.
- Raj Chetty and Nathaniel Hendren. The impacts of neighborhoods on intergenerational mobility ii: County-level estimates. *The Quarterly Journal of Economics*, 133(3):1163–1228, 2018b.
- Raj Chetty, Nathaniel Hendren, and Lawrence F Katz. The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review*, 106(4):855–902, 2016.
- Eric Chyn. Moved to opportunity: The long-run effects of public housing demolition on children. *American Economic Review*, 108(10):3028–56, 2018.
- Eric Chyn and Diego Daruich. An equilibrium analysis of the effects of neighborhood-based interventions on children. Technical report, National Bureau of Economic Research, 2022.
- Diego Comin, Danial Lashkari, and Martí Mestieri. Structural change with long-run income and price effects. *Econometrica*, 89(1):311–374, 2021.
- Timothy G Conley. Gmm estimation with cross sectional dependence. *Journal of econometrics*, 92(1):1–45, 1999.
- Arusha Cooray and Niklas Potrafke. Gender inequality in education: Political institutions or culture and religion? *European Journal of Political Economy*, 27(2):268–280, 2011.
- Flavio Cunha and James J Heckman. Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of human resources*, 43(4):738–782, 2008.
- Anna Piil Damm and Christian Dustmann. Does growing up in a high crime neighborhood affect youth criminal behavior? *American Economic Review*, 104(6):1806–32, 2014.
- Wolfgang Dauth, Sebastian Findeisen, Enrico Moretti, and Jens Suedekum. Matching in cities. Technical report, National Bureau of Economic Research, 2018.
- Donald R Davis and Jonathan I Dingel. A spatial knowledge economy. *American Economic Review*, 109(1):153–70, 2019.
- Donald R Davis and David E Weinstein. Bones, bombs, and break points: the geography of economic activity. *American economic review*, 92(5):1269–1289, 2002.

- Donald R Davis and David E Weinstein. A search for multiple equilibria in urban industrial structure. *Journal of Regional Science*, 48(1):29–65, 2008.
- Sabine D’Costa and Henry G Overman. The urban wage growth premium: Sorting or learning? *Regional Science and Urban Economics*, 48:168–179, 2014.
- Robert Dekle, Jonathan Eaton, and Samuel Kortum. Global rebalancing with gravity: Measuring the burden of adjustment. *IMF Staff Papers*, 55(3):511–540, 2008.
- Joseph DeStefano, Audrey-Marie Schuh Moore, David Balwanz, and Ash Hartwell. Meeting efa: Reaching the underserved through complementary models of effective schooling. working paper. *Academy for Educational Development*, 2007.
- Nathan Deutscher. Place, peers, and the teenage years: Long-run neighborhood effects in australia. *American Economic Journal: Applied Economics*, 12(2):220–49, 2020.
- Julian Di Giovanni and Andrei A Levchenko. Firm entry, trade, and welfare in zipf’s world. *Journal of International Economics*, 89(2):283–296, 2013.
- Rebecca Diamond. The determinants and welfare implications of us workers’ diverging location choices by skill: 1980–2000. *American Economic Review*, 106(3):479–524, 2016.
- Dave Donaldson. Railroads of the raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4-5):899–934, 2018.
- Dave Donaldson. Blending theory and data: A space odyssey. *Journal of Economic Perspectives*, 36(3):185–210, 2022.
- Dave Donaldson and Richard Hornbeck. Railroads and american economic growth: A “market access” approach. *The Quarterly Journal of Economics*, 131(2):799–858, 2016.
- Thad Dunning and Lauren Harrison. Cross-cutting cleavages and ethnic voting: An experimental study of cousinage in mali. *American Political Science Review*, 104(1):21–39, 2010.
- Fabian Eckert, Tatjana Karina Kleineberg, et al. Saving the american dream? education policies in spatial general equilibrium. Technical report, The World Bank, 2021.
- Eric V Edmonds, Nina Pavcnik, and Petia Topalova. Trade adjustment and human capital investments: Evidence from indian tariff reform. *American Economic Journal: Applied Economics*, 2(4):42–75, 2010.

- Parfait M Eloundou-Enyegue. Pregnancy-related dropouts and gender inequality in education: A life-table approach and application to cameroon. *Demography*, 41(3):509–528, 2004.
- David K Evans and Amina Mendez Acosta. Education in africa: What are we learning? *Journal of African Economies*, 30(1):13–54, 2021.
- Benjamin Faber. Trade integration, market size, and industrialization: evidence from china’s national trunk highway system. *Review of Economic Studies*, 81(3):1046–1070, 2014.
- Giulia Faggio, Kjell G Salvanes, and John Van Reenen. The evolution of inequality in productivity and wages: panel data evidence. *Industrial and Corporate Change*, 19(6):1919–1951, 2010.
- Pablo D Fajgelbaum and Edouard Schaal. Optimal transport networks in spatial equilibrium. *Econometrica*, 88(4):1411–1452, 2020.
- Oliver Falck, Stephan Heblich, Alfred Lameli, and Jens Südekum. Dialects, cultural identity, and economic exchange. *Journal of urban economics*, 72(2-3):225–239, 2012.
- Robert C Feenstra, Robert Inklaar, and Marcel P Timmer. The next generation of the penn world table. *American economic review*, 105(10):3150–82, 2015.
- Thiemo Fetzer. Can workfare programs moderate conflict? evidence from india. *Journal of the European Economic Association*, 18(6):3337–3375, 2020.
- Amy Finkelstein, Matthew Gentzkow, and Heidi Williams. Sources of geographic variation in health care: Evidence from patient migration. *The quarterly journal of economics*, 131(4):1681–1726, 2016.
- Vivien Foster and Cecilia Briceño-Garmendia. *Africa’s infrastructure: a time for transformation*. World Bank, 2010.
- Raphael Franck and Ilia Rainer. Does the leader’s ethnicity matter? ethnic favoritism, education, and health in sub-saharan africa. *American Political Science Review*, 106(2):294–325, 2012.
- Matthew Freedman, Shantanu Khanna, and David Neumark. Jue insight: The impacts of opportunity zones on zone residents. *Journal of Urban Economics*, page 103407, 2021.

- Junichi Fujimoto, David Lagakos, and Mitchell Vanvuren. Aggregate and distributional effects of 'free' secondary schooling in the developing world. *manuscript, University of California at San Diego*, 2019.
- Thomas Fujiwara and Leonard Wantchekon. Can informed public deliberation overcome clientelism? experimental evidence from benin. *American Economic Journal: Applied Economics*, 5(4):241–55, 2013.
- Cecile Gaubert. Firm sorting and agglomeration. *American Economic Review*, 108(11):3117–53, 2018.
- Cecile Gaubert, Patrick M Kline, and Danny Yagan. Place-based redistribution. Technical report, National Bureau of Economic Research, 2021.
- Ejaz Ghani, Arti Grover Goswami, and William R Kerr. Highway to success: The impact of the golden quadrilateral project for the location and performance of indian manufacturing. *The Economic Journal*, 126(591):317–357, 2016.
- Edward L Glaeser and David C Mare. Cities and skills. *Journal of labor economics*, 19(2): 316–342, 2001.
- Edward L Glaeser, David Laibson, and Bruce Sacerdote. An economic approach to social capital. *The economic journal*, 112(483):F437–F458, 2002.
- Claudia Goldin. The u-shaped female labor force function in economic development and economic history, 1994.
- Douglas Gollin, Remi Jedwab, and Dietrich Vollrath. Urbanization with and without industrialization. *Journal of Economic Growth*, 21(1):35–70, 2016.
- Douglas Gollin, Casper Worm Hansen, and Asger Mose Wingender. Two blades of grass: The impact of the green revolution. *Journal of Political Economy*, 129(8):2344–2384, 2021.
- Ken Gwilliam. *Africa's transport infrastructure: Mainstreaming maintenance and management*. The World Bank, 2011.
- Daniel Haanwinckel. Supply, demand, institutions, and firms: A theory of labor market sorting and the wage distribution. *Unpublished manuscript*, 2020.
- Joan Hamory, Marieke Kleemans, Nicholas Y Li, and Edward Miguel. Reevaluating agricultural productivity gaps with longitudinal microdata. *Journal of the European Economic Association*, 19(3):1522–1555, 2021.

- Luke Heath Milsom. The changing spatial inequality of opportunity in west africa. Technical report, University of Oxford, Department of Economics, 2021.
- Richard Hornbeck and Daniel Keniston. Creative destruction: Barriers to urban growth and the great boston fire of 1872. *American Economic Review*, 107(6):1365–98, 2017.
- Solomon M Hsiang. Temperatures and cyclones strongly associated with economic production in the caribbean and central america. *Proceedings of the National Academy of sciences*, 107(35):15367–15372, 2010.
- Allan Hsiao. Educational investment in spatial equilibrium: Evidence from indonesia. *Working paper*, 2022.
- IPUMS. Integrated public use microdata series, international: Version 7.2 [dataset]. 2020.
- Melina Platas Izama. Muslim education in africa: Trends and attitudes toward faith-based schools. *The Review of Faith & International Affairs*, 12(2):38–50, 2014.
- Rémi Jedwab. Urbanization without structural transformation: Evidence from consumption cities in africa. *George Washington University, Washington, DC. Processed*, 2013.
- Remi Jedwab and Alexander Moradi. The permanent effects of transportation revolutions in poor countries: evidence from africa. *Review of economics and statistics*, 98(2):268–284, 2016.
- Remi Jedwab and Adam Storeygard. The average and heterogeneous effects of transportation investments: Evidence from sub-saharan africa 1960-2010. *Journal of the European Economic Association*, 2021.
- Remi Jedwab, Edward Kerby, and Alexander Moradi. History, path dependence and development: Evidence from colonial railways, settlers and cities in kenya. *The Economic Journal*, 127(603):1467–1494, 2017.
- Remi Jedwab, Elena Ianchovichina, and Federico Haslop. Consumption cities vs. production cities: New considerations and evidence. *Production Cities: New Considerations and Evidence*, 2022.
- Koen Jochmans and Martin Weidner. Fixed-effect regressions on network data. *Econometrica*, 87(5):1543–1560, 2019.

- Tabi Atemnkeng Johannes and Armand Gilbert Noula. Gender and increased access to schooling in cameroon: A marginal benefit incidence analysis. *Journal of International Women's Studies*, 12(1):94–106, 2011.
- Hundanol Kebede. The gains from market integration the welfare effects of new rural roads in ethiopia. *Working paper*, 2020.
- Gaurav Khanna. Large-scale education reform in general equilibrium: Regression discontinuity evidence from india. *The American Economic Review*, 2022.
- Patrick Kline, Raffaele Saggio, and Mikkel Sølvsten. Leave-out estimation of variance components. *Econometrica*, 88(5):1859–1898, 2020.
- Gizem Koşar, Tyler Ransom, and Wilbert Van der Klaauw. Understanding migration aversion using elicited counterfactual choice probabilities. *Journal of Econometrics*, 2021.
- Paul Krugman. Scale economies, product differentiation, and the pattern of trade. *The American Economic Review*, 70(5):950–959, 1980.
- David Lagakos. Urban-rural gaps in the developing world: Does internal migration offer opportunities? *Journal of Economic perspectives*, 34(3):174–92, 2020.
- Jean-William P Laliberté. Long-term contextual effects in education: Schools and neighborhoods. *American Economic Journal: Economic Policy*, 2021.
- Neil Lee, Paul Sissons, and Katy Jones. The geography of wage inequality in british cities. *Regional Studies*, 50(10):1714–1727, 2016.
- Ilse Lindenlaub. Sorting multidimensional types: Theory and application. *The Review of Economic Studies*, 84(2):718–789, 2017.
- Jeremy Lise and Fabien Postel-Vinay. Multidimensional skills, sorting, and human capital accumulation. *American Economic Review*, 110(8):2328–76, 2020.
- Stephen Machin. Changes in uk wage inequality over the last forty years. *The Labour Market in Winter-The State of Working Britain*, 2011.
- Kirsten Majgaard and Alain Mingat. *Education in sub-Saharan Africa: A comparative analysis*. World Bank Publications, 2012.

- Nicolette D Manglos-Weber. Identity, inequality, and legitimacy: religious differences in primary school completion in sub-saharan africa. *Journal for the Scientific Study of Religion*, 56(2):302–322, 2017.
- Leandro Medina, Mr Andrew W Jonelis, and Mehmet Cangul. *The informal economy in Sub-Saharan Africa: Size and determinants*. International Monetary Fund, 2017.
- Guy Michaels. The effect of trade on the demand for skill: Evidence from the interstate highway system. *The Review of Economics and Statistics*, 90(4):683–701, 2008.
- Edward Miguel and Gerard Roland. The long-run impact of bombing vietnam. *Journal of development Economics*, 96(1):1–15, 2011.
- David Molitor. The evolution of physician practice styles: evidence from cardiologist migration. *American Economic Journal: Economic Policy*, 10(1):326–56, 2018.
- Niclas Moneke. Can big push infrastructure unlock development? evidence from ethiopia. Technical report, Working paper, 2020.
- Melanie Morten and Jaqueline Oliveira. The effects of roads on trade and migration: Evidence from a planned capital city. *NBER Working Paper*, 22158:43, 2021.
- Dale Mortensen. *Wage dispersion: why are similar workers paid differently?* MIT press, 2003.
- Holger M Mueller, Paige P Ouimet, and Elena Simintzi. Wage inequality and firm growth. *American Economic Review*, 107(5):379–83, 2017.
- Rachel Ngai, Claudia Olivetti, and Barbara Petrongolo. Structural transformation over 150 years of women’s and men’s work. *Unpublished Working Paper*, 2022.
- Stephen Nord. Income inequality and city size: An examination of alternative hypotheses for large and small cities. *The Review of Economics and Statistics*, pages 502–508, 1980.
- Dozie Okoye, Roland Pongou, and Tite Yokossi. New technology, better economy? the heterogeneous impact of colonial railroads in nigeria. *Journal of Development Economics*, 140:320–354, 2019.
- Victor Ombati and Mokuia Ombati. Gender inequality in education in sub-saharan africa. *JWEE*, 3-4:114–136, 2012.

- Amma Panin et al. Four reasons why analysis of economic policy and religion go hand-in-hand in sub-saharan africa. Technical report, PEGNet-Poverty Reduction, Equity and Growth Network, Kiel Institute for the . . . , 2021.
- Barbara Petrongolo and Christopher Pissarides. Scale effects in markets with search. *The Economic Journal*, 116(508):21–44, 2006.
- Melina R Platas. Culture and the persistence of educational inequality: Lessons from the muslim–christian education gap in africa. *New York University Abu Dhabi Working Paper*, 2018.
- Fabien Postel-Vinay and Jean-Marc Robin. Equilibrium wage dispersion with worker and employer heterogeneity. *Econometrica*, 70(6):2295–2350, 2002.
- Leandro Prados de la Escosura. Augmented human development in the age of globalization. *The Economic History Review*, 2021.
- George Psacharopoulos and Harry Anthony Patrinos. Returns to investment in education: a decennial review of the global literature. *Education Economics*, 26(5):445–458, 2018.
- Diego Puga. The magnitude and causes of agglomeration economies. *Journal of regional science*, 50(1):203–219, 2010.
- Stephen J Redding and Esteban Rossi-Hansberg. Quantitative spatial economics. *Annual Review of Economics*, 9:21–58, 2017.
- Stephen J Redding and Daniel M Sturm. The costs of remoteness: Evidence from german division and reunification. *American Economic Review*, 98(5):1766–97, 2008.
- Stephen J Redding and Matthew A Turner. Transportation costs and the spatial organization of economic activity. *Handbook of regional and urban economics*, 5:1339–1398, 2015.
- Patricia Rice, Anthony J Venables, and Eleonora Patacchini. Spatial determinants of productivity: analysis for the regions of great britain. *Regional science and urban economics*, 36(6): 727–752, 2006.
- Fernanda Catalina Rojas Ampuero. *Sent Away: The Long-Term Effects of Slum Clearance on Children and Families*. PhD thesis, UCLA, 2022.
- Eleanor Sanderson and Frank Windmeijer. A weak instrument f-test in linear iv models with multiple endogenous variables. *Journal of Econometrics*, 190(2):212–221, 2016.

- Marta Santamaria. Reshaping infrastructure: Evidence from the division of germany. Technical report, University of Warwick, Department of Economics, 2020.
- Daniel Schaefer and Carl Singleton. Recent changes in british wage inequality: Evidence from large firms and occupations. *Scottish Journal of Political Economy*, 67(1):100–125, 2020.
- T Paul Schultz. Evidence of returns to schooling in africa from household surveys: Monitoring and restructuring the market for education. *Journal of African Economies*, 13(suppl_2): ii95–ii148, 2004.
- M Najeeb Shafiq, Robert K Toutkoushian, and Alexandria Valerio. Who benefits from higher education in low-and middle-income countries? *The Journal of Development Studies*, 55(11): 2403–2423, 2019.
- Muna Shifa and Murray Leibbrandt. Spatial inequality in sub-saharan africa. *African Geographical Review*, pages 1–17, 2022.
- George W Snedecor and William G Cochran. Statistical methods, eight edition. *Iowa state University press, Ames, Iowa*, 1191, 1989.
- Jae Song, David J Price, Fatih Guvenen, Nicholas Bloom, and Till Von Wachter. Firming up inequality. *The Quarterly journal of economics*, 134(1):1–50, 2019.
- Sebastian Sotelo. Domestic trade frictions and agriculture. *Journal of Political Economy*, 128(7):2690–2738, 2020.
- Nik Stoop, Marijke Verpoorten, and Koen Deconinck. Voodoo, vaccines, and bed nets. *Economic Development and Cultural Change*, 67(3):493–535, 2019.
- Sónia Torres, Pedro Portugal, John T Addison, and Paulo Guimaraes. The sources of wage variation and the direction of assortative matching: Evidence from a three-way high-dimensional fixed effects regression model. *Labour Economics*, 54:47–60, 2018.
- Clarence Tsimpo and Quentin Wodon. Assessing the role of faith-inspired primary and secondary schools in africa: Evidence from multi-purpose surveys. *The Review of Faith & International Affairs*, 12(2):5–15, 2014.
- Nick Tsivanidis. Evaluating the impact of urban transit infrastructure: Evidence from bogota’s transmilenio. *Unpublished manuscript*, 2019.

Theresa Tuwor and Marie-Antoinette Sossou. Gender discrimination and education in west africa: strategies for maintaining girls in school. *International Journal of Inclusive Education*, 12(4):363–379, 2008.

UN. 2018 revision of world urbanization prospects, 2018.

Raoul van Maarseveen. The effect of urban migration on educational attainment: evidence from africa. *Available at SSRN 3836097*, 2021.

Anthony J Venables. Productivity in cities: self-selection and sorting. *Journal of Economic Geography*, 11(2):241–251, 2011.

Leonard Wantchekon. Clientelism and voting behavior: Evidence from a field experiment in benin. *World politics*, 55(3):399–422, 2003.

Nils B Weidmann, Jan Ketil Rød, and Lars-Erik Cederman. Representing ethnic groups in space: A new dataset. *Journal of Peace Research*, 47(4):491–499, 2010.

Felix Weinhardt. Social housing, neighborhood quality and student performance. *Journal of Urban Economics*, 82:12–31, 2014.

Yoto V Yotov, Roberta Piermartini, José-Antonio Monteiro, and Mario Larch. *An advanced guide to trade policy analysis: The structural gravity model*. World Trade Organization Geneva, 2016.

Alwyn Young. The african growth miracle. *Journal of Political Economy*, 120(4):696–739, 2012.

Wouter Zant. Measuring trade cost reductions through a new bridge in mozambique: Who benefits from transport infrastructure? *Journal of African Economies*, 31(4):384–408, 2022.

Román David Zárate. Spatial misallocation, informality, and transit improvements: Evidence from mexico city. *University of California, Berkeley*, 2020.

Office for National Statistics. (2020). Annual Survey of Hours and Earnings, 1997-2020: Secure Access. [data collection]. 17th Edition. UK Data Service. SN: 6689, <http://doi.org/10.5255/UKDA-SN-6689-16>

Office for National Statistics. (2021). Business Structure Database, 1997-2021: Secure Access. [data collection]. 14th Edition. UK Data Service. SN: 6697, <http://doi.org/10.5255/UKDA-SN-6697-14>