Moving Opportunity
Local Connectivity and Spatial Inequality

Luke Heath Milsom*

November 2022

Click here for the most recent version

Abstract

Within-country inequality across space is large. How does the connectivity of place determine underlying spatial inequality of opportunity? To answer this question, I derive a theoretical 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. Policymakers also face an equity-efficiency trade-off: On average only 6 of the top 10 aggregate opportunity-increasing roads also decrease inequality of opportunity. A back-of-the-envelope calculation shows that between 13% and 44% of people living in low-opportunity locations would need to migrate to high-opportunity areas to achieve reductions in inequality similar to those I estimate to be possible through road upgrades.

JEL: R12, R42, O18, H54
Keywords: Inequality, Opportunity, Roads, Spatial, Market Access

*Department of Economics, University of Oxford, luke.milsom@economics.ox.ac.uk. With thanks to: Ferdinand Rauch, Gabriel Ulyssea, Niclas Moneke, Douglas Gollin, Hannah Zillessen, Marco Gonzalez-Navarro, Cecile Gaubert, Vernon Henderson, Dave Donaldson, Christian Dustmann, Jessica Pan, Tony Venables, Elizabeth Sadoulet, Allan Hsiao, Thibault Fally, Nina Guyon, Johannes Abeler, Abigail Adams-Prassl, Barbara Petrongolo, Simon Quinn, Ian Crawford, Isabelle Roland, Hamish Low, Sarah Clifford, Binta Zahra Diop, Shihang Hou, Evan Soltas, Giulio Schinaia, Verena Wiedemann, Sam Altmann, Itzhak Rasooly, Sanghamitra Mukherjee, and Vatsal Khandelwal, for useful comments and conversations. I would also like to thank seminar participants at LEAP Bocconi, NEUDC Yale, UEA Conference Washington DC, OxDev Oxford, CEP Junior Trade Workshop LSE, the CSAE conference, applied micro and development economics workshops at Oxford, development workshop at UC Berkeley ARE, European Urban Economics Association conference LSE, Warwick PhD conference, African Meeting of the Econometric Society, Royal Economics Society Young Economist Symposium, Newcastle PhD conference, Infra4Dev World bank, UEA Urban Economics Workshop, and National University of Singapore.

There is vast regional inequality within countries [UN, 2020]. This inequality is often rooted in locations being cut off from markets and the economic opportunities they foster [Allen et al., 2020c]. Better connecting such areas is a common policy solution. Indeed, in 2021 alone over 10% of the World Bank's lending was to finance transportation infrastructure investment. Although connecting any two locations positively affects the degree of local market integration and ease of migration, it 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 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, 2009]. To quantify the impacts of changing connectivity on inequality of opportunity I develop a sufficient statistic result which is estimated using a novel identification strategy, and used to parameterise a structural spatial general equilibrium model for counterfactual analysis. I use information on the causal effect growing up in a given location has on individuals from Heath Milsom [2021], and data on 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 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.[1]

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

[1]A 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 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.

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 location[2] on their probability of completing primary education.[3] Although there are many dimensions of local opportunity, primary schooling is particularly salient[4] 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 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

[2]Regions 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.

[3]This 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 A for details regarding the estimation strategy used in Heath Milsom [2021].

[4]Primary-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.

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.[5]

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 upgrading[6] 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

[5] An 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.

[6] In 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.

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.[7] 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 oportunity. 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 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 achieve by many roads, between 71,000 and 466,000 individuals would have to

[7] I 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.

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 et al. [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.

This paper proceeds as follows. Section 1 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 then describes my setting and data before developing a novel identification strategy to 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 3 estimates a structural spatial economics model and performs counterfactual analysis. Section 4 then concludes.

1 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 effectively describes how to measure the impact of road building on each location, taking into account spillover and general spatial equilibrium concerns 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 corresponding 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.

1.1 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 based models of economic geography with economies of scale in production (e.g. Bartelme [2015]); Melitz-type frameworks with heterogenious 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 framwork (e.g. Redding and Sturm [2008]); recent quantitative spatial economics models[8] (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. However, the sufficient statistic result is in addition 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 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 based 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.

[8] In appendix section 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.

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 we tackle 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 which Shephard's lemma give the aggregate demand equation for i's goods in j of the familiar gravity form.

$$X_{ij} = \frac{(\tau_{ij} p_i)^{-\phi}}{\sum_k p_{kj}^{-\phi}} E_j \tag{1}$$

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 costs to that of symmetric iceberg trade costs[9], 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 trade[10] $\pi_{ij} = \frac{u_{ij}^{\lambda}}{\sum_{k \in \mathcal{L}} u_{kj}^{\lambda}}$ 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

[9] In 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.

[10] This can be micro-founded form example by assuming type two extreme value preferences over locations as is common.

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 [Koşar et al., 2021]. Together this implies that movement across space, obeys the following equation (in aggregate):

$$M_{ij} = \frac{\kappa_{ij}^{-\lambda} u_i^{\lambda}}{\sum_k \kappa_{kj}^{-\lambda} u_k^{\lambda}} L_j \tag{2}$$

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 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 have wages increase in goods market access, but decrease in labor market access:

$$w_i = \frac{p_i^{-\phi} MA_i}{u_i^{\lambda} LMA_i} \tag{3}$$

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 above[11]: $u_i = A_i \left(\frac{w_i}{P_i}\right)^a$ 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

[11]Note 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 D for details.

find:

$$w_i = \Omega_i \cdot MA_i^{\frac{1-\frac{a\lambda}{\phi}}{x}} \cdot LMA_i^{-\frac{1}{x}} \tag{4}$$

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 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 production[12]. 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 follows[13].

$$MA_i^s = \sum_j (\tau_{ij})^{-\phi^s} \frac{Y_j^s}{MA_j^s} \qquad\qquad LMA_i^s = \sum_j \left(\kappa_{ij}^s\right)^{\lambda^s} \frac{L_j^s}{LMA_j^s} \tag{5}$$

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

$$\mathbb{P}\left[\text{complete primary school}_i\right] = \mu_i = \beta_1 \cdot \ln(r_i) + \varepsilon_i \tag{6}$$

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, $\mathbb{P}\left[\text{complete primary school}_i\right]$ to it's empirical observed counterpart, μ_i, local educational opportunity.

[12]The sufficient statistic result is robust to this rigid production set up, shown formally in appendix D.

[13]The implied assumption on consumption patterns is generalized in the appendix section D, and results are found to not be qualitatively different

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've omitted an important dimension of variation both over time and space. Here I postulate a simple way of endogenizing education supply which doesn't impact the sufficient statistic result. In appendix section D.2.1 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 schools[14] 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 $\tilde{c} = cd^{\chi}$ 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

[14]Perhaps 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.

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.

1.2 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 D.1 I show how the equations above can be combined to arrive at a log linear relationship given in equation 7 which relates local educational opportunity to the four market access terms, and can be directly taken to the data.

$$\mu_i = \gamma_1 \cdot \ln(MA_i^E) + \gamma_2 \cdot \ln(MA_i^N) + \gamma_3 \cdot \ln(LMA_i^E) + \gamma_4 \cdot \ln(LMA_i^N) + v_i \tag{7}$$

As noted above this result is robust to a broad set of modeling assumptions and micro-foundations and this is formally shown in appendix 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 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 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_i^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 that 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_i^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 Taking the sufficient statistic result to the data

In this section I describe how I take the sufficient statistic result developed to the data, focusing 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.

2.1 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, 2009] 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 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.

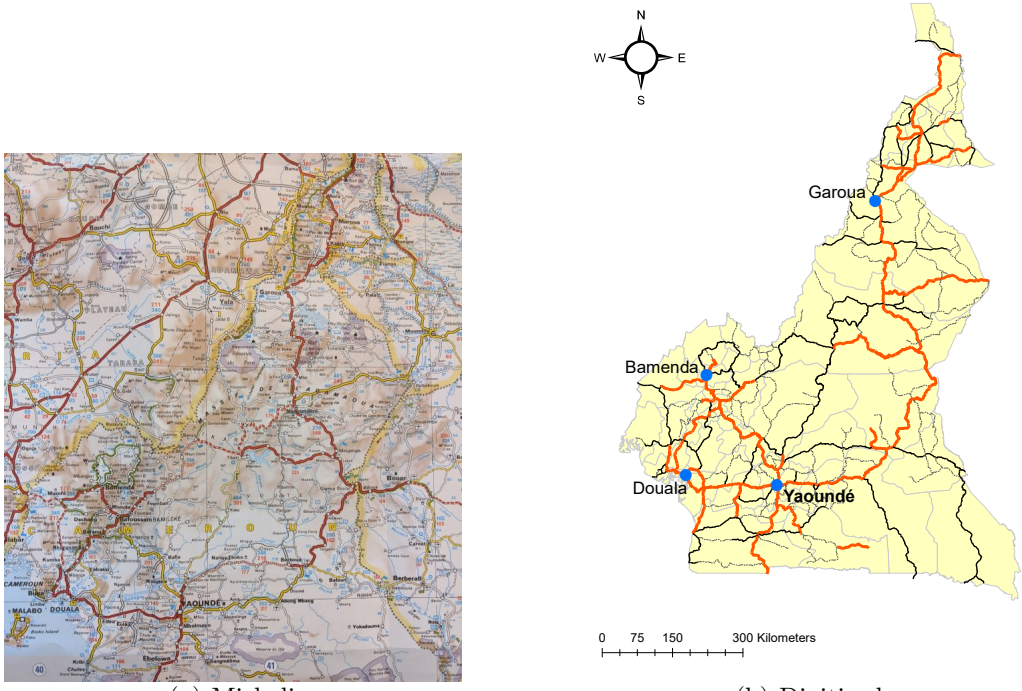
In this section, I describe the main data sources and institutional context, summarize key dimensions of variation, and present some descriptive correlational evidence. This evidence shows that places that became more connected as the result of road building saw larger increases in local returns to education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

2.1.1 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 1 gives an example of this process. Panel 1a shows an image of the raw Michelin map of Cameroon in 2019 and panel 1b shows the digitized version.

Figure 1 Digitizing Michelin road maps — Cameroon 2019



(a) Michelin map (b) Digitized map

Notes: This figure shows in panel 1a the original Michelin road map of Cameroon in 2019 and in panel 1b the digitized version. In panel 1b thick red lines are paved roads, dark black lines are improved roads and gray lines are dirt tracks. Note that in panel 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 C.1.1. 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 2013[15] 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 samples[16]. 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 cycle[17]. This does not include Koranic schools but does include Medersas in Mali as they follow the national curriculum [Boyle, 2014], in appendix B.9 I show that both of these non-traditional schools play a minor role.

[15] In 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.

[16] 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.

[17] In 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.

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, 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 C.1.2.

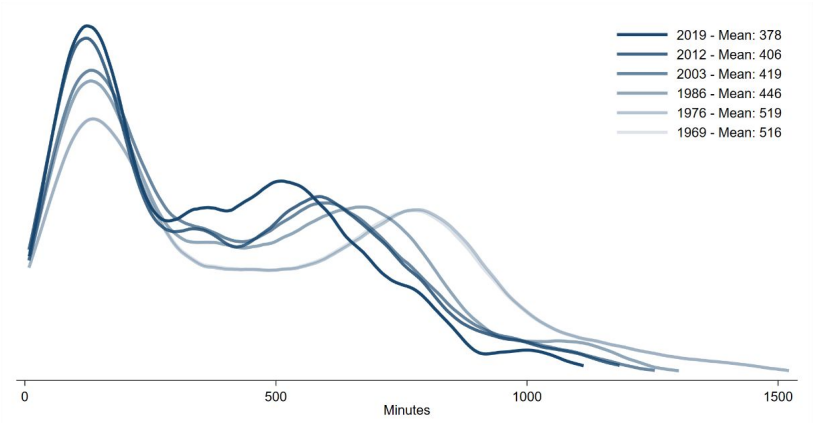
Variation in connectivity over time and space

In figure 2, I plot the distribution of expected travel times for each location, in each year maps are available over the digitized road network[18]. 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.

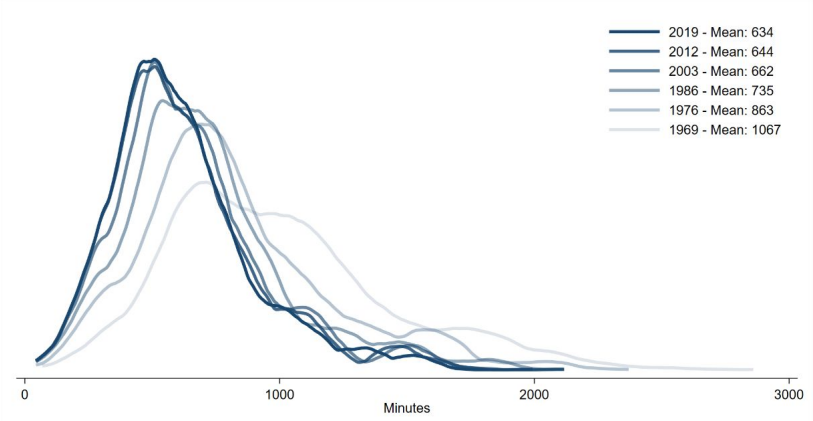
[18]Over my study period, in Benin, Cameroon, and Mali, other forms of transport such as railways or waterways exhibited little variation and are not modeled.

Figure 2 Distribution of expected pairwise travel times between localities

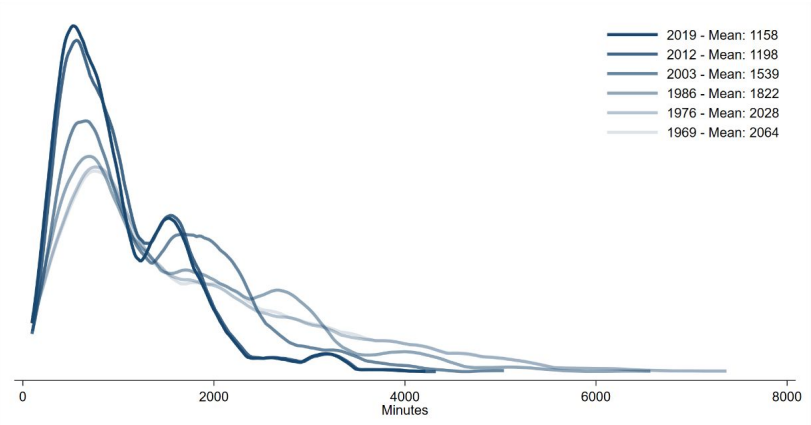
(a) Benin



(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 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 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 Djikstra'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 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 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.

2.1.2 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.

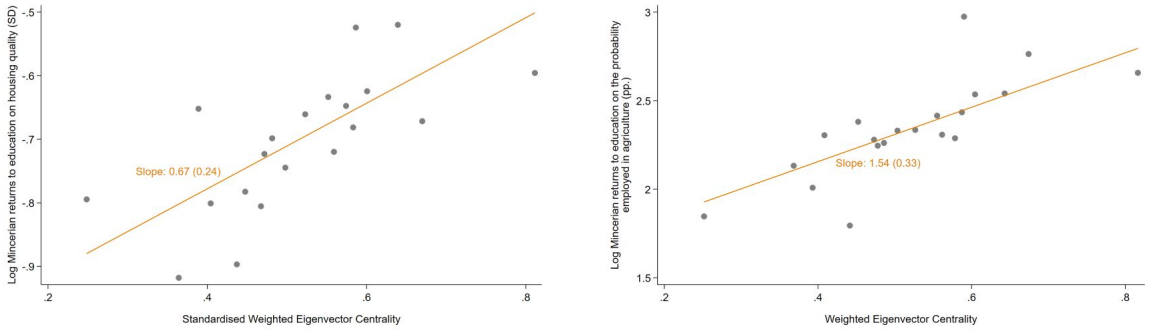
In the binscatter plots and corresponding OLS regressions which follow I include locality and year fixed effects as well as weighting by locality population and clustering standard errors at the locality level. Because of this, coefficients can be interpreted as the association between changes in centrality (due to changes in the road network) and changes in local returns to education in figure 3 and changes in local returns to education and causal place effects in figure 4.

First, I show that there is a strong correlation between local connectivity and the local benefits of completing primary education. Figure 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 (3a) or the probability of not working in agriculture (3b).

Figure 3 Correlation between connectivity and local returns to education

(a) Housing quality

(b) Not working in agriculture



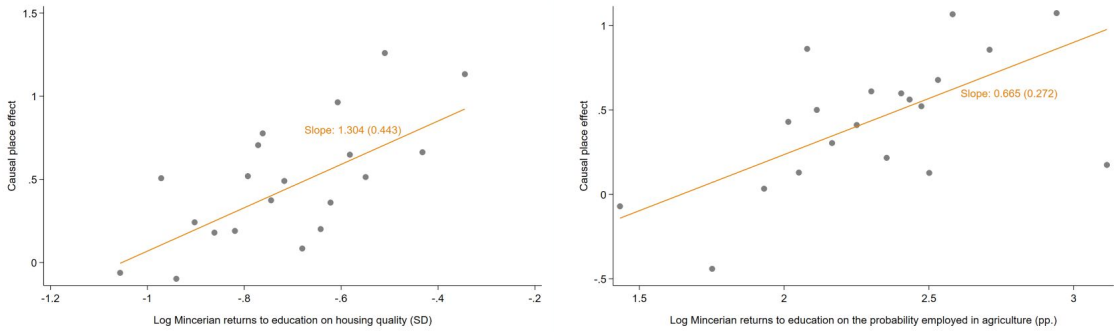
Notes: This figure shows in panel 3a the correlative relationship between the log Mincerian returns to education in terms of housing quality on the y-axis and eigenvector centrality on the x-axis. Panel 3b shows the correlative relationship between the log Mincerian returns to education in terms of the 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 $y_i = \beta_l^y Primary_i + \beta_{1l} age_i + \beta_{2l} age_i^2 + \varepsilon_i$ 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 $Cent_l$ 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 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 4 Correlation between local returns to education and opportunity

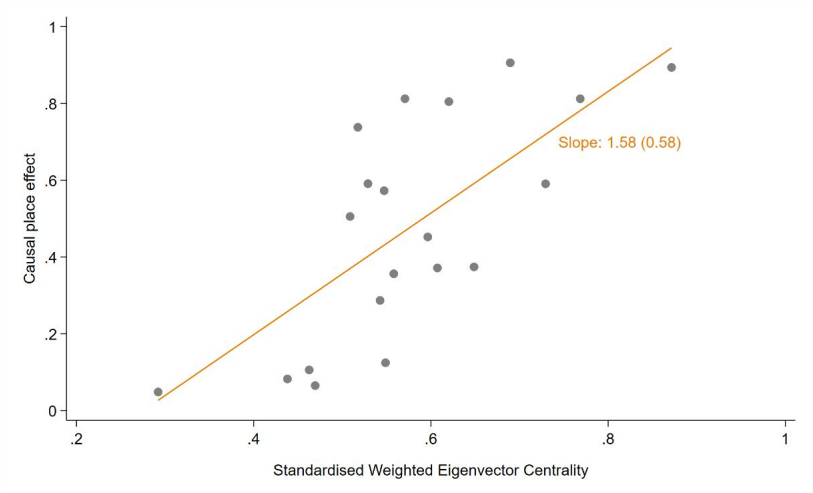
(a) Housing quality (b) Not working in agriculture



Notes: This figure shows in panel 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 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 25 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 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 are calculated from the analogous linear regression. Associated standard errors are given in parenthesis clustering at the locality level.

2.2 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 instruments[19] 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 the variation

[19]Locations 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.

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 locations[20]. 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 whether at some point coefficients stabilize giving evidence to suggest that sources of endogeneity have

[20]This 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.

been removed[21].

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 B.5 for a graphical explaination 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 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.

2.2.1 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 local variation in observed primary completion rates which is due to causal place effects

[21] As 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're reached the limits of policy makers sophistication.

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 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 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.

$$\mu_{it} = \gamma_1 \cdot \ln(MA_{it}^E) + \gamma_2 \cdot \ln(MA_{it}^N) + \gamma_3 \cdot \ln(LMA_{it}^E) + \gamma_4 \cdot \ln(LMA_{it}^N) + \tau_i + \alpha_t + v_{it} \quad (8)$$

I estimate equation 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 1 and figure 23 display the main results from this section. In each column of table 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 variation[22]. Removing higher than 4th order least-cost path variation results in weak instruments. For each regression (column) the Kleibergen-Paap under-identification 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

[22]In 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.

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 B.10.

Table 1 Results from estimating the sufficient statistic relationship

	(1) OLS	(2) No including own market	(3) Constant market size	(4) Not including least cost path variation	(5) 1st order least cost + only far away variation	(6) 2nd order least cost + only far away variation	(7) 3rd order least cost + only far away variation	(8) 4th order least cost + only far away variation
Log(LMA Educ)	-0.0598	-0.0569	-0.170	-0.0312	-0.264	-0.192*	-0.179*	-0.175
	(0.0509)	(0.0564)	(0.107)	(0.125)	(0.356)	(0.108)	(0.106)	(0.110)
Log(MA Educ)	0.0434**	0.0961***	0.137***	0.0545	0.0904	0.0795**	0.0699	0.0696
	(0.0201)	(0.0271)	(0.0290)	(0.0598)	(0.237)	(0.0396)	(0.0425)	(0.0453)
Log(LMA No Educ)	0.151***	0.0767	0.265**	0.156*	0.619	0.252*	0.215**	0.218**
	(0.0249)	(0.0850)	(0.124)	(0.0879)	(0.475)	(0.131)	(0.0997)	(0.106)
Log(MA No Educ)	-0.0814***	-0.132**	-0.185***	-0.136**	-0.111	-0.151**	-0.123**	-0.118*
	(0.0212)	(0.0529)	(0.0403)	(0.0563)	(0.293)	(0.0682)	(0.0555)	(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 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 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 23 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 23 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 10 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 10 in the appendix. In column (2) of table 10 I show results controlling for expected market access (in all four variables) calculated as described in subsection 2.3.1. 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 13 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 15 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.

2.3.1 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.

2.3.2 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 18 in the appendix section 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 4 and 5 in the appendix section B.7 show the results. In table 4 I regress changes for each variable on levels of all variables and in table 5 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 B.8 by including lagged market access terms in the main sufficient statistic regression — and find no evidence that this is a relevant chanel for concern in this setting.

2.3.3 Inference

Inference surrounding the coefficients recovered from estimating equation 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 natural 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 B.10. I find that standard errors remain similar to those reported in this section, never significantly inflating.

2.3.4 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 maybe 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.[23] 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.

3 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 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 1 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 $\{\tau_{ijt}, \kappa_{ijt}\}$ (transporta-

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

tion costs), and parameters $\{\phi_E, \phi_N, \lambda_E, \lambda_N, \beta\}$, I can solve for the endogenous quantities $\{u^s_{it}, E_{it}, w^s_{it}, Y^s_{it}, L^s_{it}, MA^s_{it}, LMA^s_{it}\}$.

1. $u^s_{it} = A^s_{it}\left(\left(\frac{w^s_{it}}{(P^s_{it})}\right)^{1-\beta} E^{\beta}_{it}\right)^{\lambda_s}$
2. $E^{\beta}_{it} = \left(\frac{w^E_{it}}{w^N_{it}}\right)^{\beta}$
3. $w^s_{it} = \frac{Y^s_{it}}{L^s_{it}}$
4. $Y^s_{it} = B^s_{it}\left(w^s_{it}\right)^{-\phi_k} MA^s_{it}$
5. $L^s_{it} = u^s_{it} LMA^s_{it}$
6. $MA^s_{it} = (P^s_{it})^{-\phi_s} = \sum_j \tau^{-\phi_s}_{ijt} \frac{Y^s_{jt}}{MA^s_{jt}}$
7. $LMA^s_{it} = \sum_j \kappa^{-\lambda_s}_{ijt} \frac{L^s_{jt}}{LMA^s_{jt}}$

Appendix section D.1 shows that these equations can be simplified into a series of simultaneous non-linear equations given by equation 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.

$$MA^r_{it} = \sum_j K^r_{ijt} \prod_{h=1}^{4} \left(MA^h_{jt}\right)^{b_{rh}} \tag{9}$$

Where MA^r_{it} denotes market access of type $r \in \{1,2,3,4\}$ such that $MA^1_{it} = MA^E_{it}$, $MA^2_{it} = MA^N_{it}$, $MA^3_{it} = LMA^E_{it}$, $MA^4_{it} = LMA^N_{it}$. K^r_{ijt} denotes the kernel associated with market access of type r and is a bundle of exogenous shifters and iceberg costs, for example $K^1_{ijt} = \tau^{-\phi_E}_{ijt}(B^E_{jt})^{a_{11}}(B^N_{jt})^{a_{12}}(A^E_{jt})^{a_{13}}(A^N_{jt})^{a_{14}}$. The scalars $\{a_{rh}\}$ and $\{b_{rh}\}$ are known functions of structural parameters.

Equation 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 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 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 $\widehat{\rho}^r_{ijt}$ 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 1. Then we can write the system given in equation 9 (see appendix section D.1 for a formal derivation) in changes where the only unknown objects are the parameters $\{b_{rh}\}_{r,h=1,2,3,4}$.

$$\widehat{MA}^r_{it} = \sum_j \widehat{\rho}^r_{ijt} \lambda^r_{ijt} \prod_{h=1}^{4} \left(\widehat{MA}^h_{jt}\right)^{b_{rh}} \tag{10}$$

λ^r_{ijt} 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 $\Delta\mu_l = \hat{\gamma} \ln(\widehat{MA}_l)$.

3.1 Finding parameter estimates

I first set β (the utility value of education relative to consumption) to equal to 0.2[24], although quantitative results presented in section 3.2 are not sensitive to other reasonable values of β as shown in appendix section D.6.2. There are four remaining parameters, $\phi_E, \phi_N, \lambda_E, \lambda_N$. In appendix 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. Finally, I also show in appendix 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 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.

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

3.2 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 3 to estimate the counterfactual vector of market access terms in 2020 in the absence of any changes to the road network since 1970. Figure 6 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

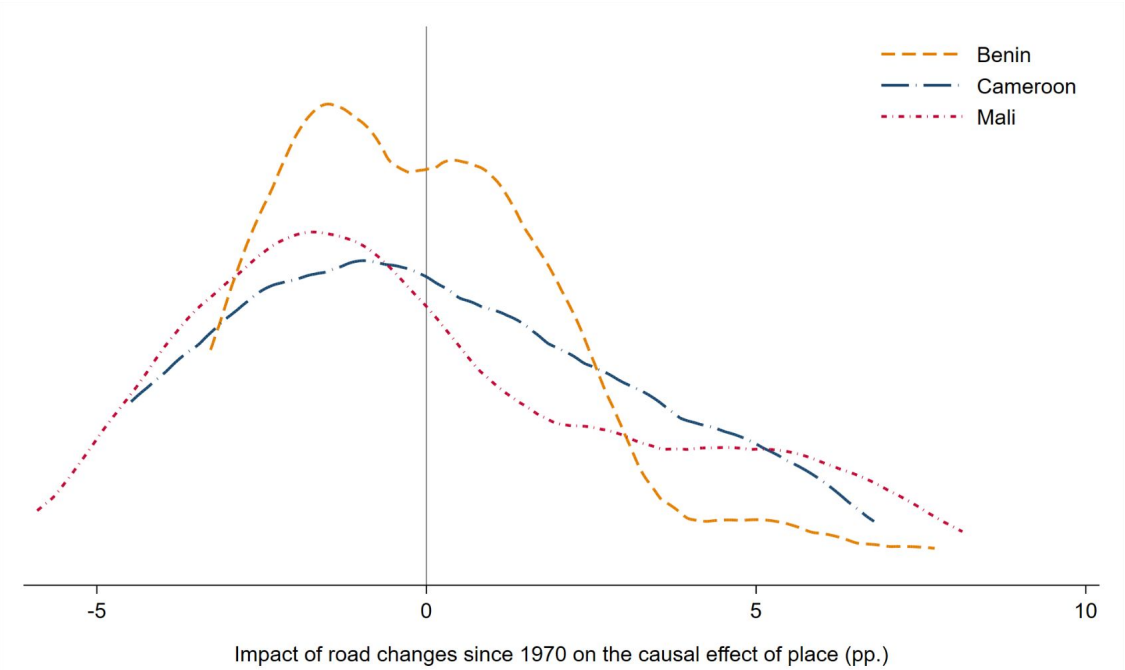
vector of estimated coefficients from section 1, and the log-operator is taken element-wise.[25]

Figure 6 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 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%.

[25] In 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 B.11.

Figure 6 The distributional impact of roads built since 1970 on the causal effect of place on primary education completion

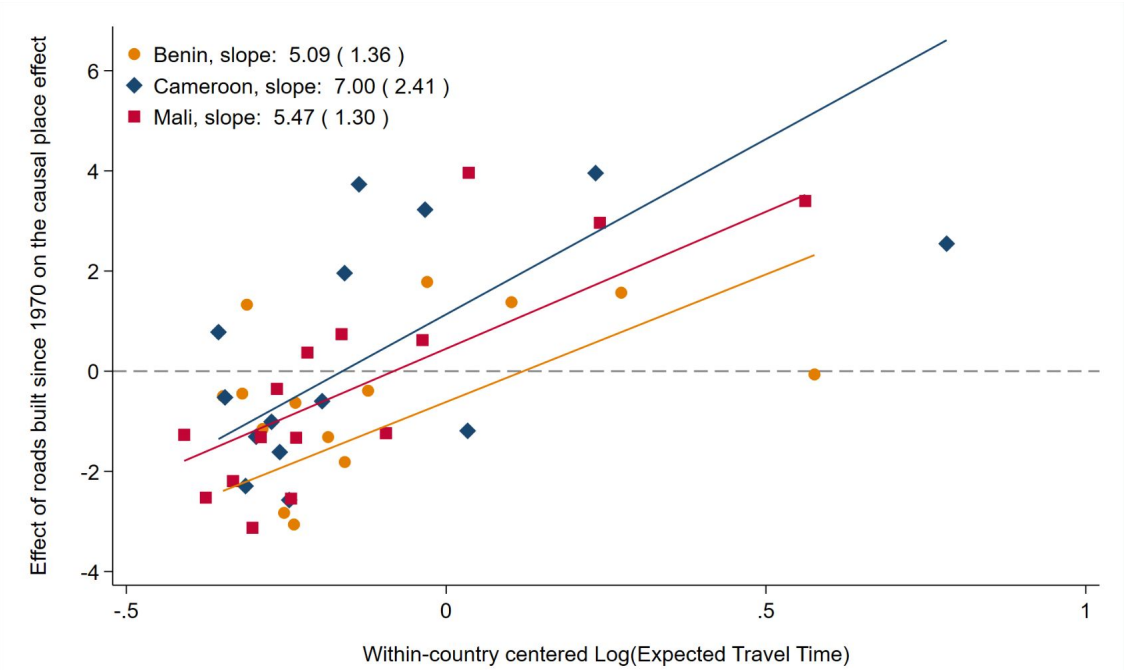


Notes: This figure shows the distribution of the total relative effects of road building from 1970 to 2020, on the causal effect of place on completing primary education in each location in each of Benin, Cameroon, and Mali.

Figure 6 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 7 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 relationship[26] 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.

[26]In figure 7 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 7 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.

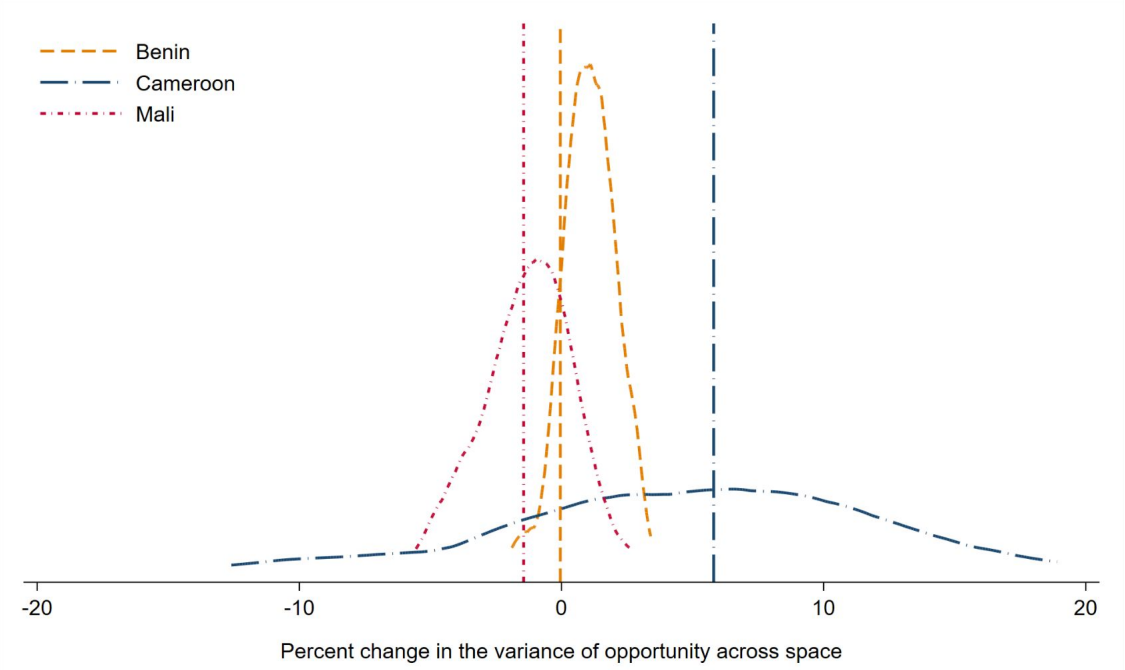
Second, I turn to investigating cross-country heterogeneity. But studying three countries I am uniquely able to investigate the extent to which network-level characteristics are important, a key ingredient in consider the external validity of any single-country study. In general, there is no particular 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 ran-

domly 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 8 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 8 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 8 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 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.

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

Section 1 showed that road building alters the spatial distribution of opportunity and section 3.2 showed that since 1970 changes to the transport network have had large, but het-

erogeneous, 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 network[27] 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.

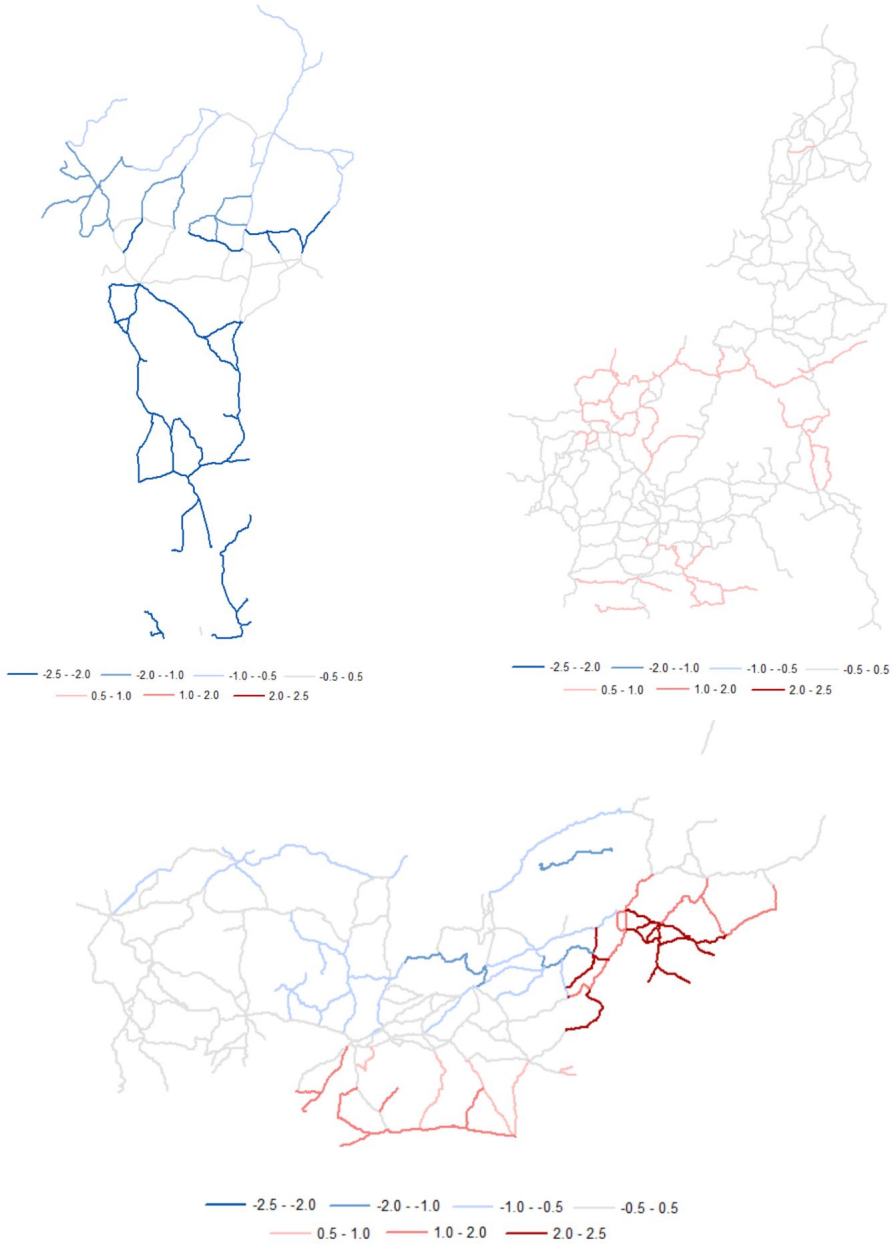
This approach can also be used to approximate in a computationally feasible manner the optimal network problem as discussed in Fajgelbaum and Schaal [2020], and is similar to the procedure taken in Balboni et al. [2020]. That is, although I don't solve over the entire space of potential new road locations, 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

[27] A 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.

— is a particularly policy-relevant margin, whilst remaining computationally 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 9 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 9 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 9 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 9. 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 9 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 9 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 31 in appendix 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 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 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 Impact of different types of road on spatial inequality of opportunity

	(1) Overall	(2) Benin	(3) Cameroon	(4) 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: $\Delta SD(\mu_l)_r = \beta_t \text{RoadType}_r + \varepsilon_r$ 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 bellow point estimates.

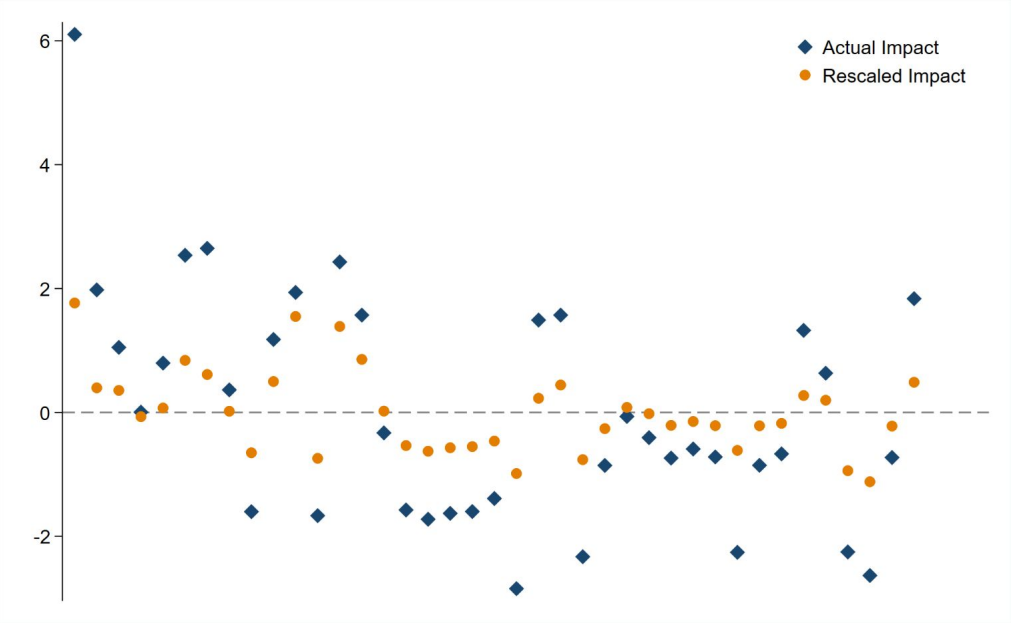
The intuition behind the result presented in table 2 is simple. A periphery location stands to gain more by trading with a relatively richer main location than vice-versa. The first order effects of connecting any two locations is to double down on their current trading and migration patterns — as periphery 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 B.13.

Within-country cross-road-type heterogeneity is important, but figure 9 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 10 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.

Figure 10 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 Cameron 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 11 in appendix 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 12 in appendix B.14 shows the impact of re-scaling on the road-type level effects reported in table 2. 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.

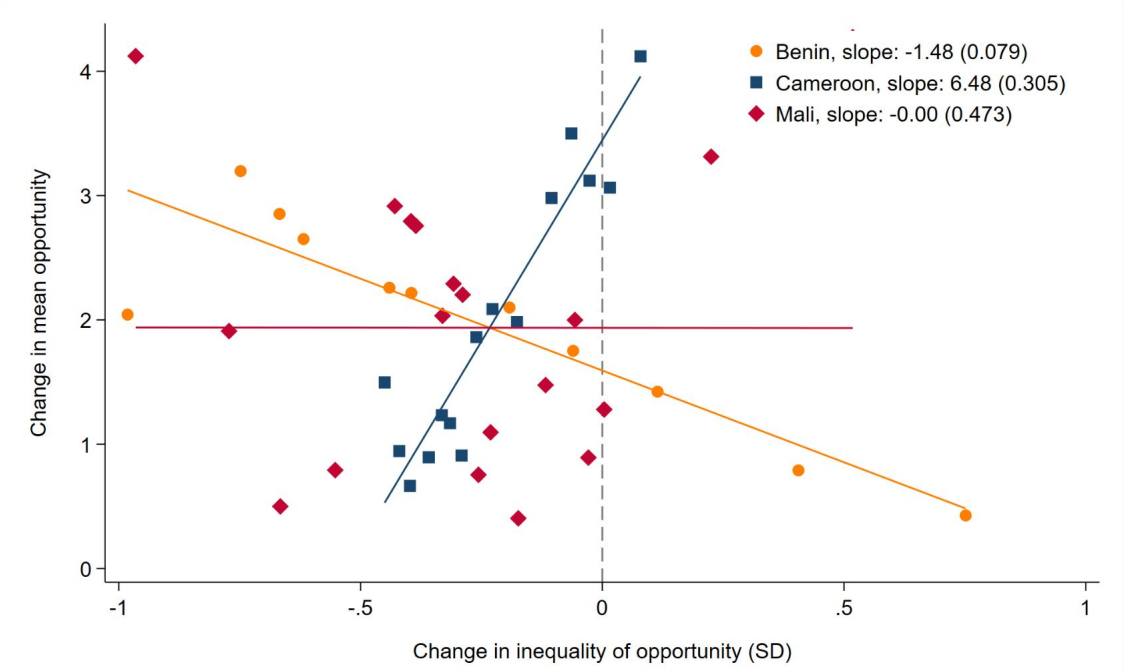
These results highlight that country (network) as well as road-type level heterogeneity matters. These exercises however, don't shed light on how different locations are deferentially effected by different roads in appendix D.4 I conside how effects vary by locality and road characteristics.

3.3.1 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 11 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 11 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 slop 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 27 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.

4 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.

References

Tasso Adamopoulos. Spatial integration, agricultural productivity, and development: A quantitative analysis of ethiopia's road expansion program. In *2019 Meeting Papers*, number 86. 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.

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.

Alberto Alesina, Sebastian Hohmann, Stelios Michalopoulos, and Elias Papaioannou. Intergenerational 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. 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.

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.

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? 2020.

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.

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.

Martin Battle and Jennifer C Seely. It's all relative: Modeling candidate support in benin. *Nationalism and Ethnic Politics*, 16(1):42–66, 2010.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Thiemo Fetzer. Can workfare programs moderate conflict? evidence from india. *Journal of the European Economic Association*, 18(6):3337–3375, 2020.

Vivien Foster and Cecilia M Briceño-Garmendia. Africa's infrastructure: a time for transformation. 2009.

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, 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, David Laibson, and Bruce Sacerdote. An economic approach to social capital. *The economic journal*, 112(483):F437–F458, 2002.

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.

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. 2022.

IPUMS. Integrated public use microdata series, international: Version 7.2 [dataset]. 2020.

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.

Hundanol A Kebede et al. The gains from market integration the welfare effects of new rural roads in ethiopia. 2020.

Gaurav Khanna. Large-scale education reform in general equilibrium: Regression discontinuity evidence from india. *The American Economic Review*, 2022.

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.

Jean-William P Laliberté. Long-term contextual effects in education: Schools and neighborhoods. *American Economic Journal: Economic Policy*, 2021.

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.

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.

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.

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.

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.

Sebastian Sotelo. Domestic trade frictions and agriculture. *Journal of Political Economy*, 128(7):2690–2738, 2020.

Nick Tsivanidis. Evaluating the impact of urban transit infrastructure: Evidence from bogota's transmilenio. *Unpublished manuscript*, 2019.

UN. 2018 revision of world urbanization prospects, 2018.

UN. *World social report 2020: Inequality in a rapidly changing world*. Department of Economic and Social Affairs, UN, 2020.

Raoul van Maarseveen. The effect of urban migration on educational attainment: evidence from africa. *Available at SSRN 3836097*, 2021.

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.

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.

Appendix

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.

$$y_i = \alpha_{odt} + \mu_{lt} \cdot e_{ilt} + \varepsilon_i \tag{11}$$

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 12.

$$e_{ilt} = \begin{cases} 13 - m_i & \text{if } l = d(i), t = t(i) \\ m_i & \text{if } l = o(i), t = t(i) \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Where, y_i is an indicator variable taking the value one if individual i has competed primary school, and m_i is the year i moved. This equation is estimated on a sample of 14-18 year old one-time movers[28] 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.

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

[28] 13% of 14-18 year olds in my sample can be classified as one-time movers.

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 population[29]. 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 $\mathbb{E}[e_{il} \cdot \varepsilon_i | \alpha_{odt}] = 0 \ \ \forall l \in \mathbb{L}$. 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].

B Empirics appendix

B.1 Descriptive statistics

B.1.1 Spatial variation in primary completion

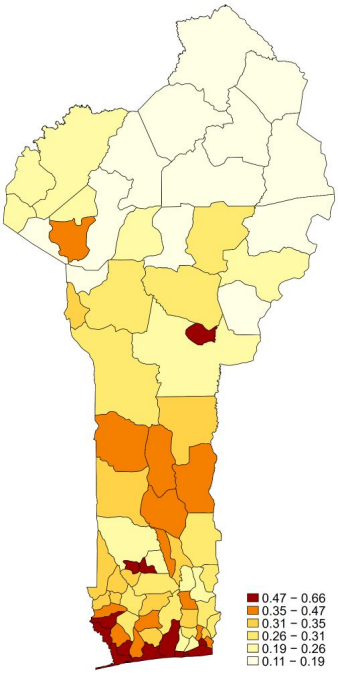
There is substantial variation in primary education completion within country across localities as can be seen by figures 12a, 12b and 12c. In Benin in 2013 the proportion of individuals who had completed primary education in an area varied from 11% in the north to as high as 66% on the coast close to 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

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

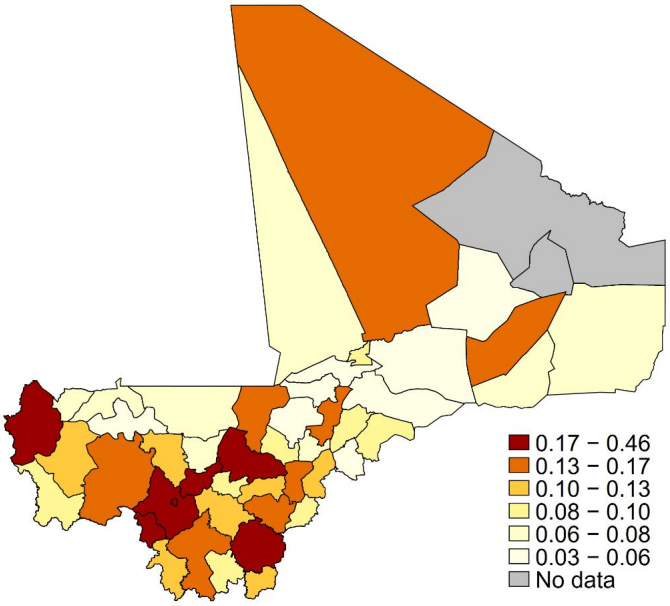
compared to Benin or Cameroon. Figure 12b also displays significant cross-locality variation with some areas completion rates as low as 3% and some, especially round the capital, closer to 50%. Cameroon shows a similar pattern, the most educated areas are around the capital, or close to the large coastal city of Douala. Cameroon also has the highest completion rates over all varying from 20% to almost 90%.

Figure 12 Locality level primary completion rates

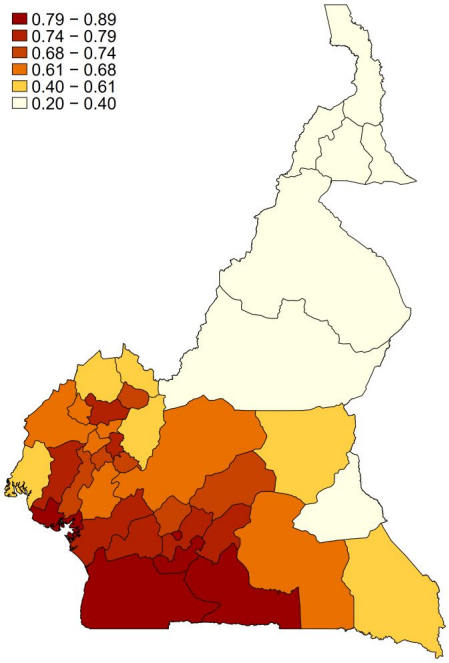
(a) Benin (2013)



(b) Mali (2009)



(c) Cameroon (2005)

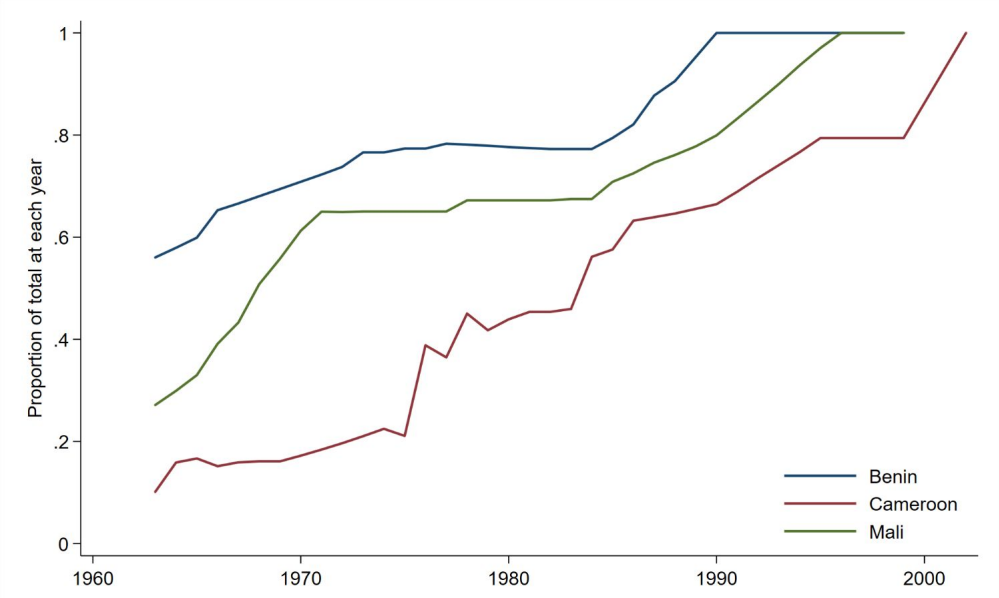


Notes: this figure shows the spatial distribution of primary education completion rates for all those above the age of 12 in each of Benin, Cameroon and Mali. Each figure has it's own scale and corresponding legend where darker orange/red indicates higher completion rates. The data for Benin comes from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

B.1.2 Secondary source showing variation in road building

Figure 13 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 this 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 13 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.

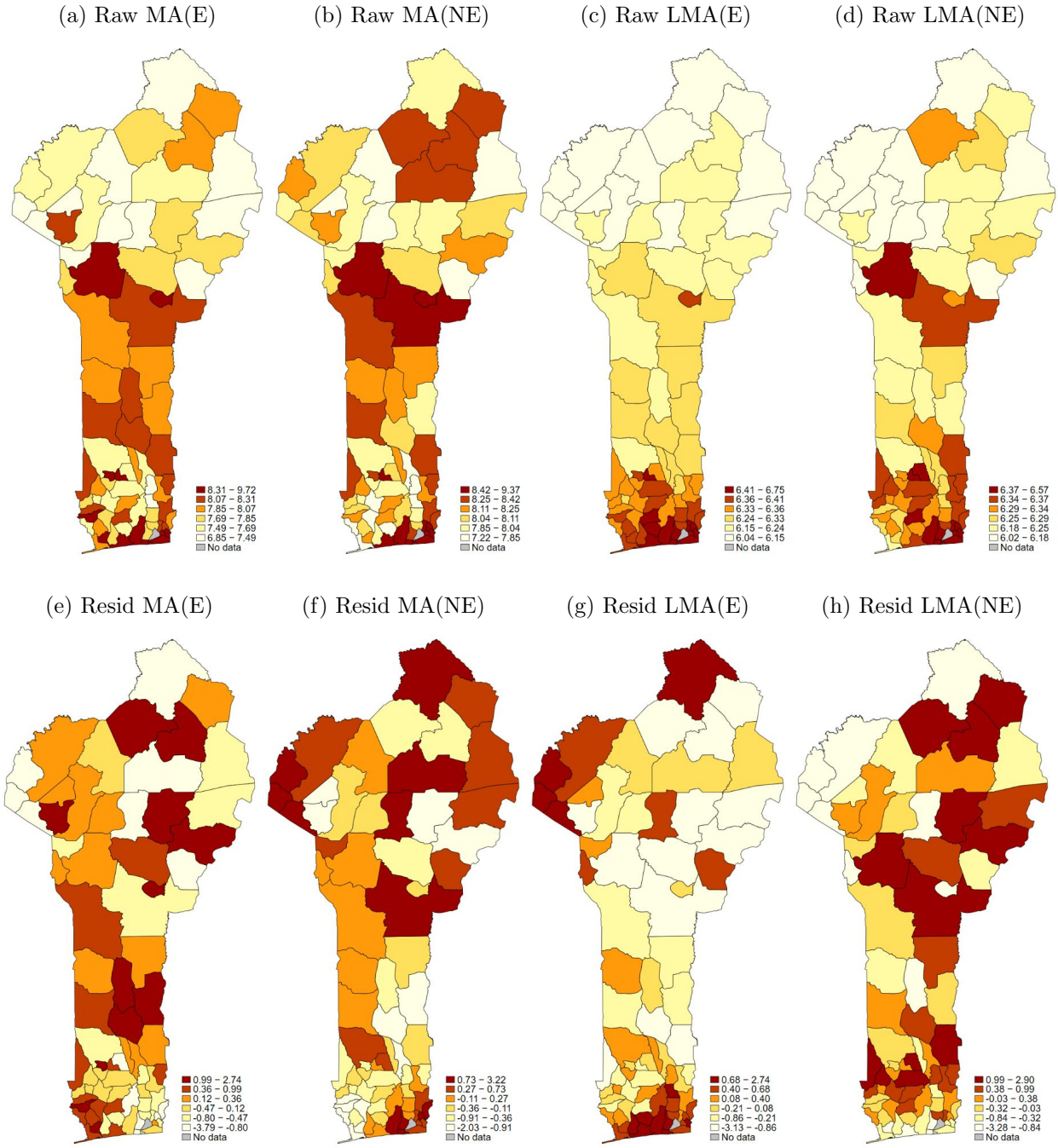
Figure 13 Variation in paved roads



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].

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

Figure 14 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.

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 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 15 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 15 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 locality[30] of the form given in equation 13 where y_ω denotes either housing quality or whether individual ω works in agriculture. Equation 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 15.

$$y_\omega = \beta_i^y \cdot \text{Primary}_\omega + \beta_{1i} \cdot \text{age}_\omega + \beta_{2i} \cdot \text{age}_\omega^2 + \varepsilon_\omega \tag{13}$$

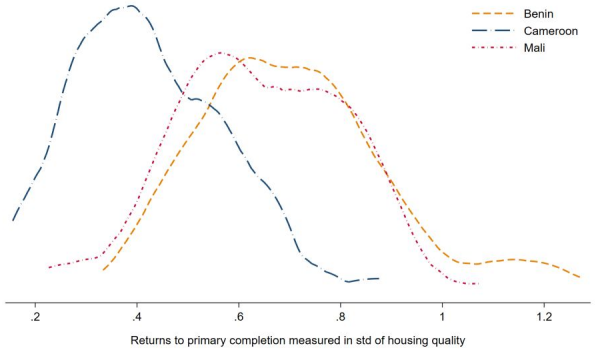
Figure 15 shows considerable variation in returns: in some birth locations those who complete primary education have over 1SD better housing quality, whereas in others quality of housing is very similar to those who don't complete primary school. In addition in appendix C.1 I show that moving across space is costly, lending credence to the non-equalization of

[30] Although 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 B.4 I show that in this limited sample observed wages are strongly correlated with my proxies at both the individual and regional level.

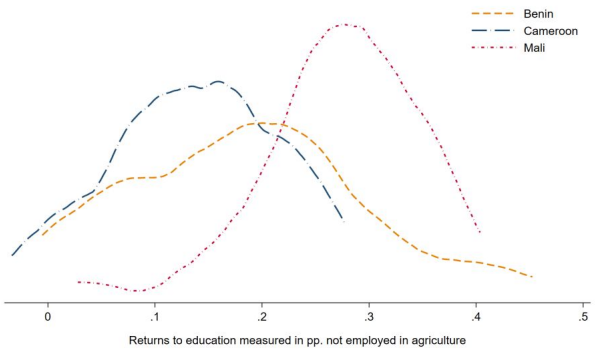
returns.

Figure 15 Spatial variation in the local benefits of education over birth location

(a) Housing quality



(b) Not working in agriculture



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 15a measures returns to education as the Mincerian return to education in terms of housing quality. Panel 15b 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 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. The distribution of the recovered β_l^y are then plotted.

The spatial distribution of local educational opportunity

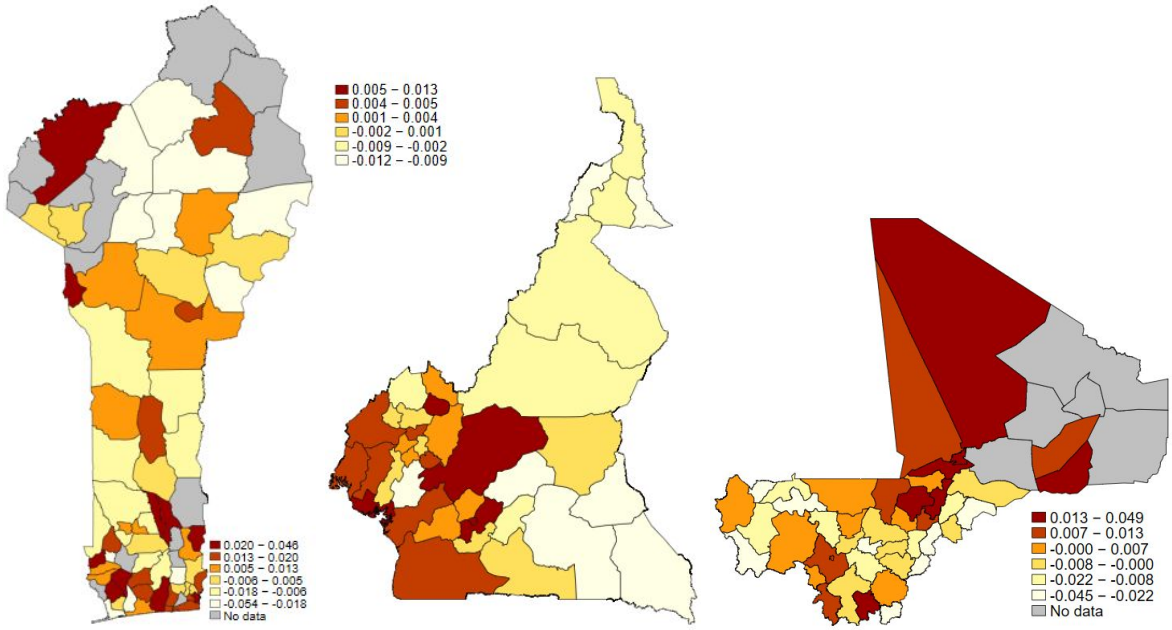
The causal effect of place on primary education completion is distinct from locality level information on observed primary completion rates as the latter conflates two forces, causal place effects and the differing characteristics of individuals over space. Heath Milsom [2021] disentangles these two 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 A.

Figure 16 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 16 Estimates of local educational opportunity from Heath Milsom [2021]



(a) Benin 2013 (b) Cameroon 2005 (c) Mali 2009

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.

B.4 Correlating wage proxies with available wage data

In the main text I use two variables to proxy for skill permia 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

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

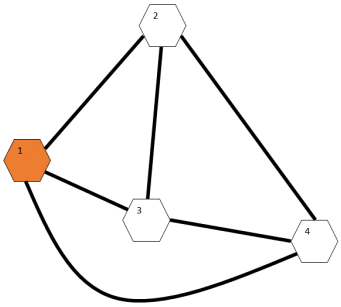
Figure 17 explains the novel not-on-least-cost-path identification strategy developed and used in the main text. Sub-figure 17a shows a stylized network with three locations represented by hexagons and the least-cost-path from each denoted by solid black lines. Through this exposition we shall consider sources of variation from the view point of location one which is colored orange. Sub-figure 17a shows the entire network without any restrictions on the source of variation used to instrument for location one's market access. Sub-figure 17b 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 17b 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 17c 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 17c, 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

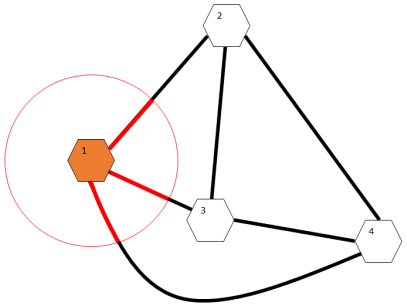
one's market access to vary. Sub-figure 17d goes one step further by depicting second-order not on least cost path variation, again from location one's perspective. Here, I additionally freeze changes in the transport network when calculating every other locations market access which then in turn feeds into calculating location one's market access. In sub-figure 17d 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 fours perspective. The instrument for location one's market access represented in 17d therefore uses indirect variation in the indirect variation due to how changes in the road network effect market access terms which indirectly effects location fours market access (and indeed every location other than location one) which indirectly effects location ones market access.

Figure 17 Not-on-least-cost-path identification strategy in figures

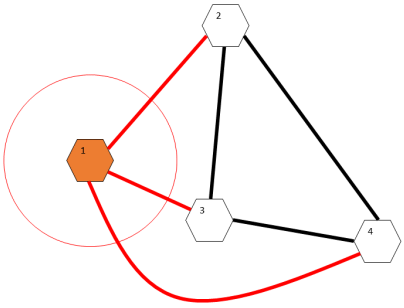
(a) Using all variation



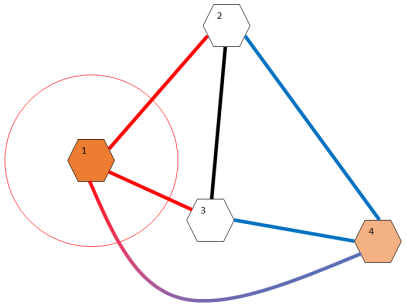
(b) Restricting to far-away variation



(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 17a depicts all variation being used. Panel 17b visually restricts variation in a circle around the focal location, indicating the use of only *far-away* variation. Panel 17c additionally restricts variation on the least cost path from the focal location to all other locations. Finally panel 17d illustrates restricting second order variation from the perspective of location 4.

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 groups[31]. 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 3 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.

[31] Broadly defined as: Ewe, Yoruba, Somba or Barba.

Table 3 Controlling for ethnicity by leader

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

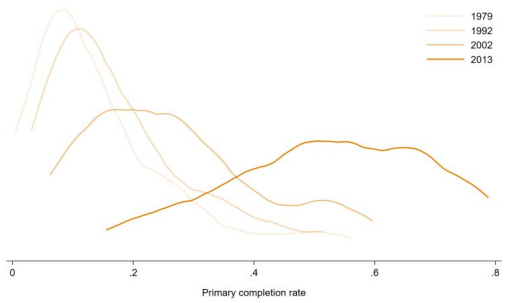
Notes: This table compares the baseline results to those including controls for clientelism. Column one replicates the main results from table 10 column one. Column two includes a dummy variable equal to one if the majority ethnicity of the locality is equal to that of the leader in Benin and zero otherwise.

B.7 Top-coding and non-linearities

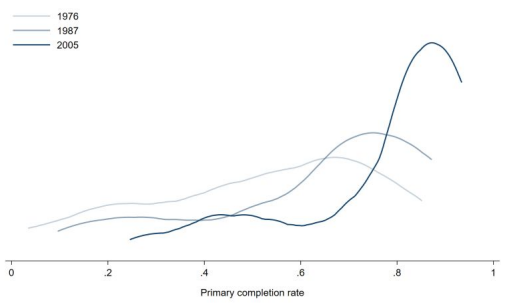
In this section I provide figures and tables referred to in sub-section 2.3.2 of the main text. Figure 18 shows how the distribution over localities of primary completion rates for those aged between 15 and 20 has changed in each country over the sample period. These figures show considerable rightward 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 4 and 5 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 18 Changes in the distribution of primary completion rates

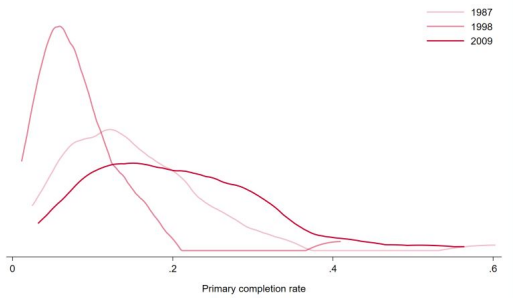
(a) Benin



(b) Cameroon



(c) Mali



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 4 Relation between base level and changes (combined regressions)

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

	(1) Long Diff $Log(LMA_E)$	(2) Log Diff $Log(LMA_N)$	(3) Long Diff $Log(MA_E)$	(4) Long Diff $Log(MA_N)$	(5) Long Diff Prim Educ	(6) Long Diff μ
$Log(LMA_E)$	-0.0581					
	(0.0471)					
$Log(LMA_N)$		0.199***				
		(0.0728)				
$Log(MA_E)$			-0.0741			
			(0.0510)			
$Log(MA_N)$				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.

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 $Cov(MA_{it}, \varepsilon_{it}) = 0$. This will not be the case if lagged market access enters the equation i.e. $\varepsilon_{it} = MA_{it-1} + \xi_{it}$, 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 6 shows the results. In columns (1) and (3) I replicate column (1) in table 1 in the main text and column (1) of table 10 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 6 Including lagged market access terms

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

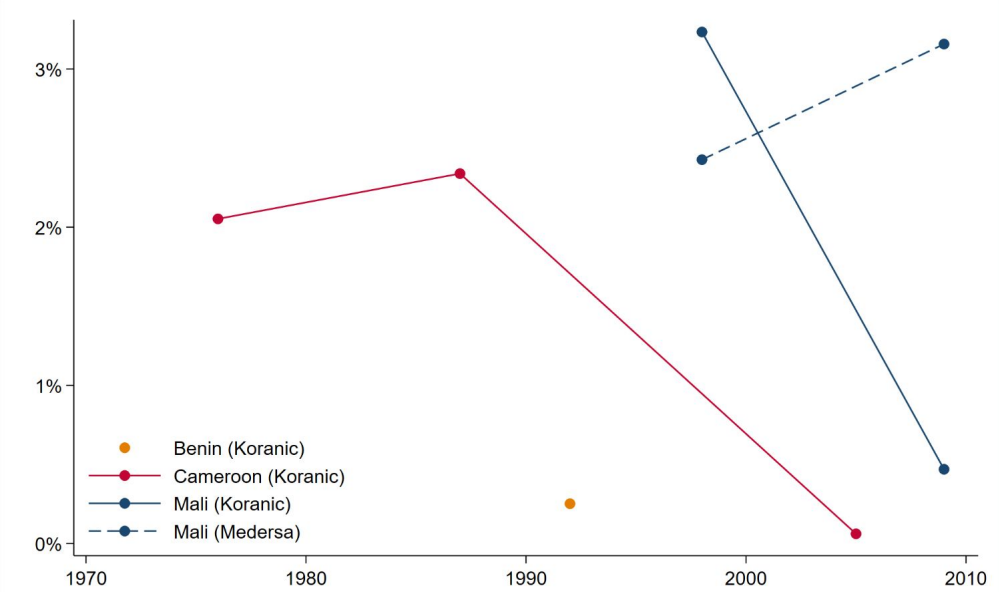
Notes: This table shows the results from estimating the sufficient statistic result described in the main text including lagged values of market access terms. In columns (1) and (3) I replicate column (1) in table 1 in the main text and column (1) of table 10 in the appendix. In column (2) I add lagged 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.

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 19 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 19 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 19 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 19 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.

B.10 Inference

As discussed in the main text performing inference with regressions of the form given in equation 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 maybe artificially attenuating standard errors I perform Conley inference [Conley, 1999] using various distance

bands and lag lengths[32]. Table 7 shows the results from each procedure on the OLS standard errors. In column one I report the coefficients from estimating equation 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 7 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.

Table 7 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

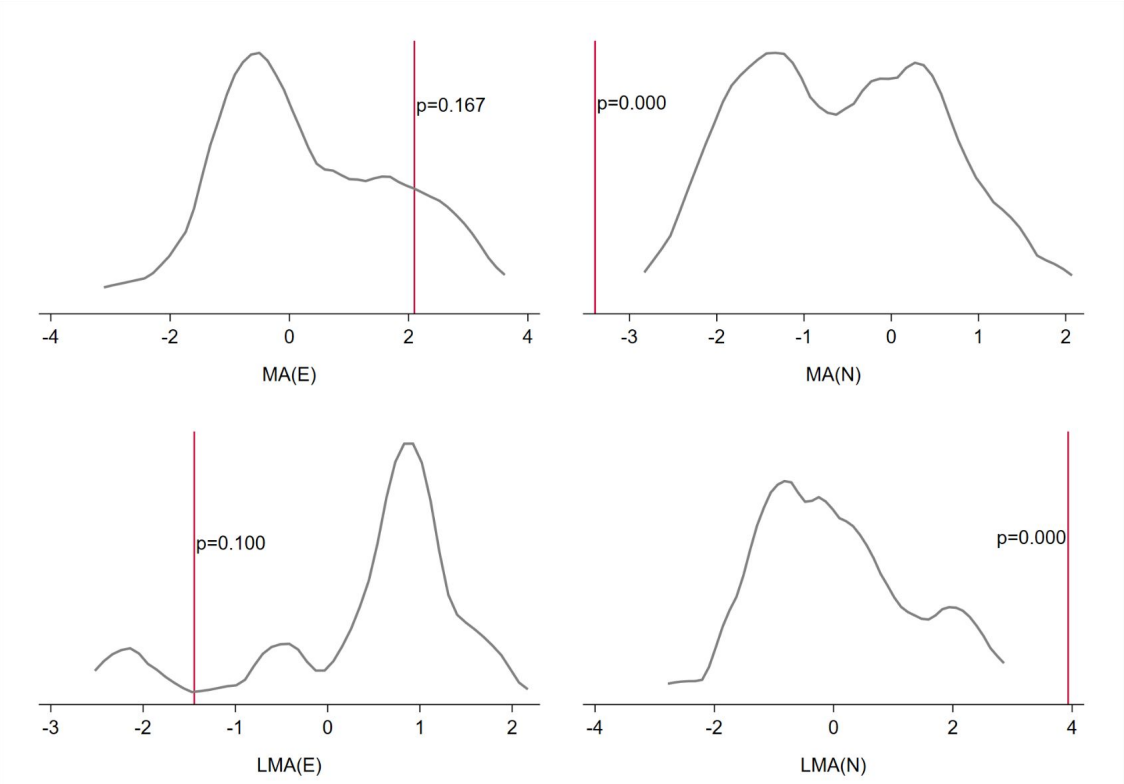
Notes: This table shows the results from performing various inference procedures. In column one I report the coefficients from estimating equation 8 by OLS. Column two reports the unadjusted standard errors associated with these coefficients. Column three adjusts standard errors by clustering at the 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 spatial 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

[32]Implementation of Conley standard errors is performed using the Stata program *reg2hdfespatial* Fetzer [2020], Hsiang [2010].

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 20 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 20 Randomization inference on OLS results



Notes: This figure shows the result from performing randomization inference on the coefficients in equation 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.

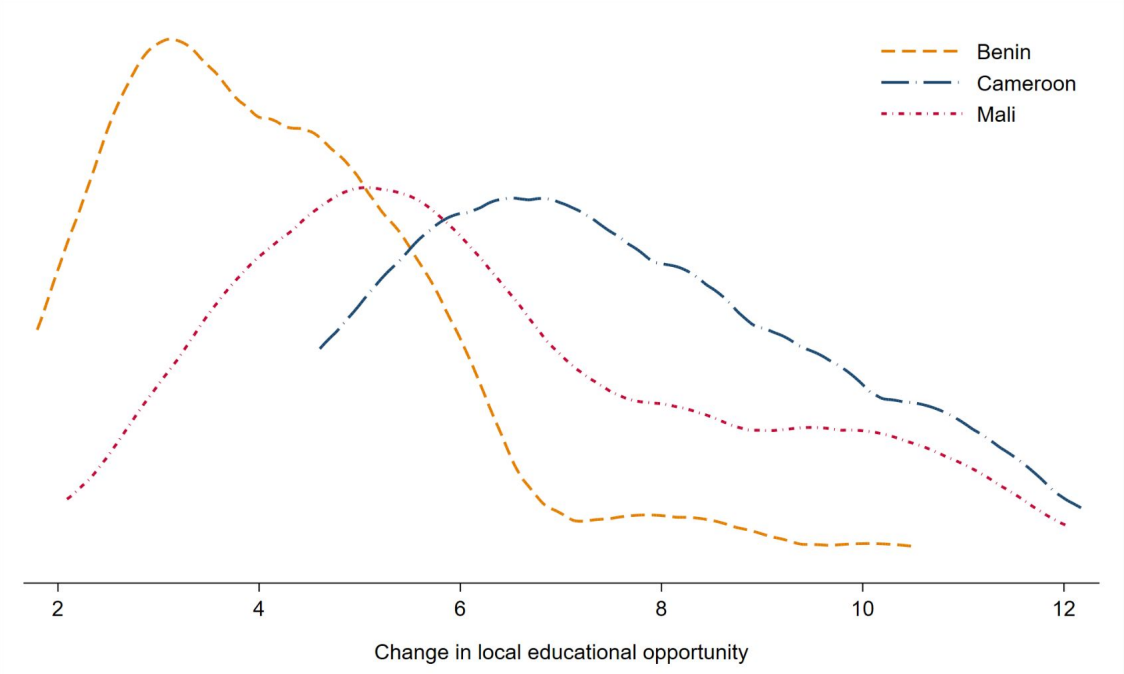
The final complication is that the main outcome variable of interest μ_{it}, is itself an estimated quantity from an earlier paper Heath Milsom [2021]. As a result of this, it will exhibit measurement error. However, there is little reason to suppose that this measurement error is not classical in nature,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

stricter requirements on rejecting the null of 0 effect.

B.11 Mean shifts in the spatial distribution of opportunity

Figure 21 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 21 Aggregate distributional shifts due to road building since 1970



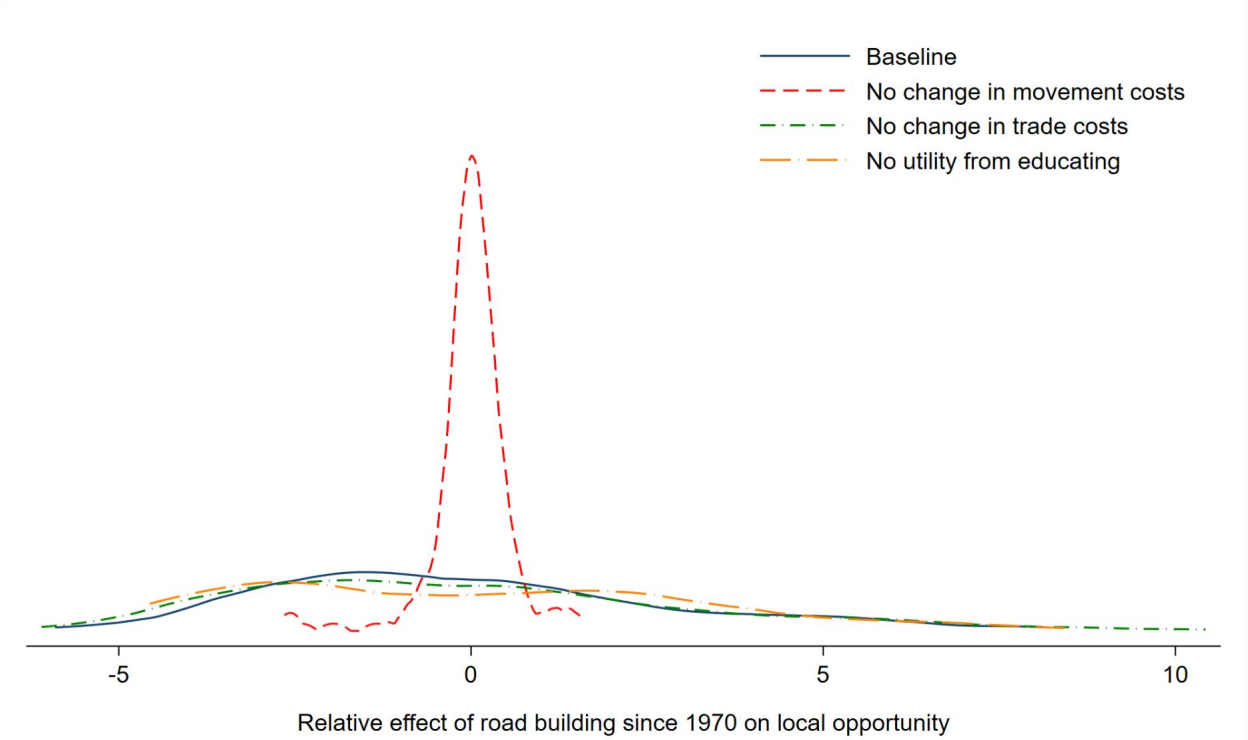
Notes: This figure shows the aggregate effects of changes to the road network since 1970 on spatial inequality of opportunity in each of Benin (yellow dashed), Cameroon (blue dash-dot), and Mali (red small dash-dot), including mean shifts.

B.12 Decomposing effects

I consider the same counterfactual exercise "what if no roads had been built since 1970" and in figure 22 show the distributional effect from shutting down each of the main channels. In blue I replicate the results from figure 6 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 22 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 26 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 22 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 8 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 8 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.

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.

$$\hat{\mu}_{ir} = \sum_{h=1}^{4}\gamma_h\left(\ln\left(\hat{\rho}_{ikr}^{h}\right) + \ln\left(\lambda_{ik}^{h}\right) + \ln\left(\prod_{q=1}^{4}\left(\widehat{MA}_{kr}^{q}\right)^{b_{qh}}\right)\right) \tag{14}$$

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 D.1 we have that $\prod_{h=1}^{4}\left(\widehat{MA}_{k}^{h}\right)^{b_{rh}} = \hat{J}_k^r/\widehat{MA}_k^r$ 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.

$$\hat{\mu}_{ir} = \beta_1\cdot\ln\left(\lambda_{i,k(i,r)}^{1}\right) + \beta_2\cdot\ln\left(\lambda_{i,k(i,r)}^{2}\right) + \beta_3\cdot\ln\left(\lambda_{i,k(i,r)}^{3}\right) + \beta_4\cdot\ln\left(\lambda_{i,k(i,r)}^{4}\right) + \varepsilon_{ir} \tag{15}$$

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 9 shows the results from estimating equation 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 9 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 9 Predicting the Impact of Roads on Local Opportunity

	(1)	(2)	(3)	(4)	(5)
$\ln(\lambda_{ik}^{MA(E)})$	3.175***	1.837*	2.858***	2.475**	2.771**
	(0.914)	(1.070)	(0.928)	(1.149)	(1.186)
$\ln(\lambda_{ik}^{MA(N)})$	-2.694**	-0.858	-2.321	-2.718	-3.219*
	(1.298)	(1.657)	(1.432)	(1.775)	(1.811)
$\ln(\lambda_{ik}^{LMA(E)})$	-10.52***	-8.784***	-5.927***	-5.595**	-4.507*
	(2.306)	(2.936)	(2.190)	(2.662)	(2.658)
$\ln(\lambda_{ik}^{LMA(N)})$	4.189*	1.947	1.368	3.012	2.581
	(2.403)	(2.865)	(2.959)	(2.979)	(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

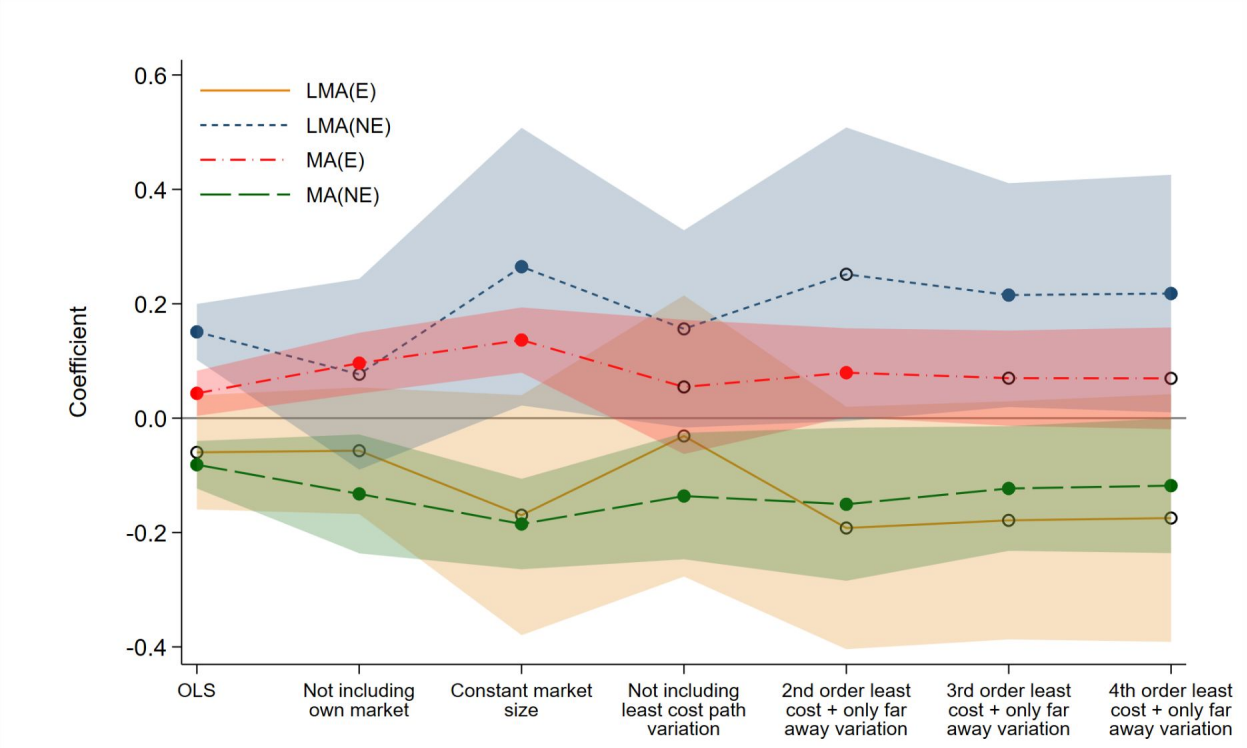
Notes: This table shows the results from estimating equations of the form given in 15. The left-hand-side variable is the change in local educational opportunity in locality i due to upgrading road r. λ_{ik}^{h} is the proportion of i's market access of type h due to locality k. Column one includes country 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.

B.14 Additional results and robustness

Figure 23 Results from estimating the sufficient statistic relationship



Notes: This figure graphically displays results from running regressions of the form as given in 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 10 Results combining instruments and implementing the Borusyak and Hull [2020] correction.

	(1) Baseline	(2) Including average MA variables
Log(LMA Educ)	-0.198**	-0.228***
	(0.0767)	(0.0844)
Log(MA Educ)	0.118***	0.136***
	(0.0281)	(0.0355)
Log(LMA No Educ)	0.284***	0.334***
	(0.0581)	(0.0953)
Log(MA No Educ)	-0.171***	-0.211***
	(0.0329)	(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 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 2.3.1.

Table 11 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 12 Impact of different types of road on spatial inequality of opportunity in the re-scaled networks

	(1) Overall	(2) Benin	(3) Cameroon	(4) Mali
Primate	0.0828	0.00368	-0.0768***	0.247
	(0.0690)	(0.0159)	(0.0225)	(0.166)
Hinterland	0.243***	0.234***	0.0698***	0.421***
	(0.0557)	(0.0222)	(0.0214)	(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 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 bellow point estimates.

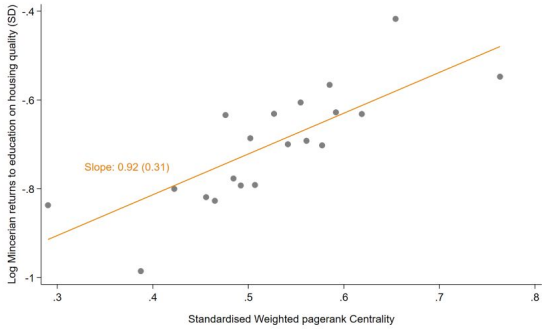
Table 13 Sufficient statistic result by country

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

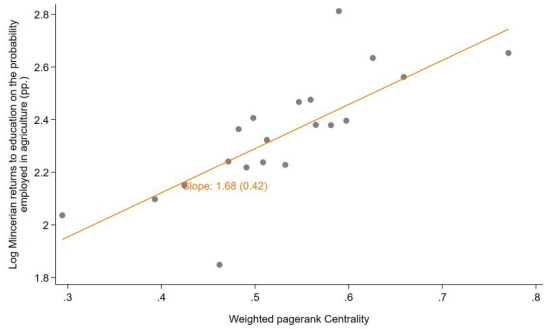
Notes: This table shows the results from re-estimating the sufficient statistic equation on samples from each country individually. The first column replicates the main (pooled) results for comparison. For each country I present a column of results for each market access term, and a column displaying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

Figure 24 Correlation between connectivity and local returns to education: pagerank centrality

(a) Housing quality

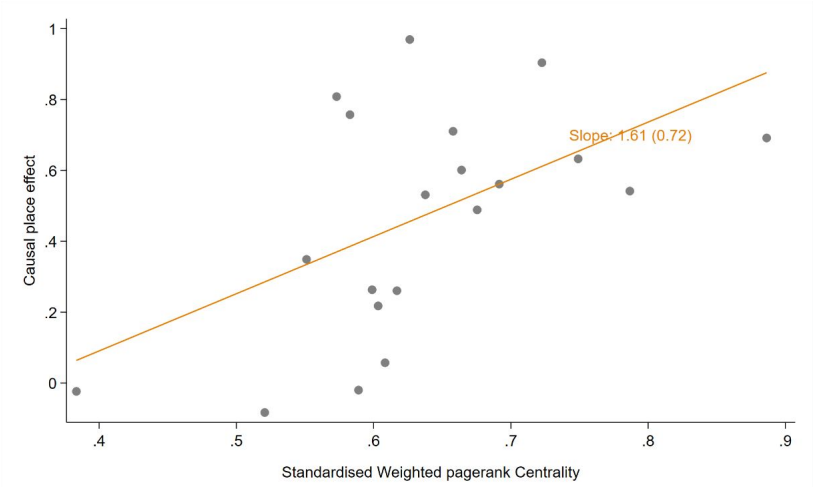


(b) Not working in agriculture



Notes: This figure shows in panel 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 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 $y_i = \beta_l^y Primary_i + \beta_{1l} age_i + \beta_{2l} age_i^2 + \varepsilon_i$ 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 $Cent_l$ 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 25 Correlation between connectivity and local opportunity: pagerank centrality



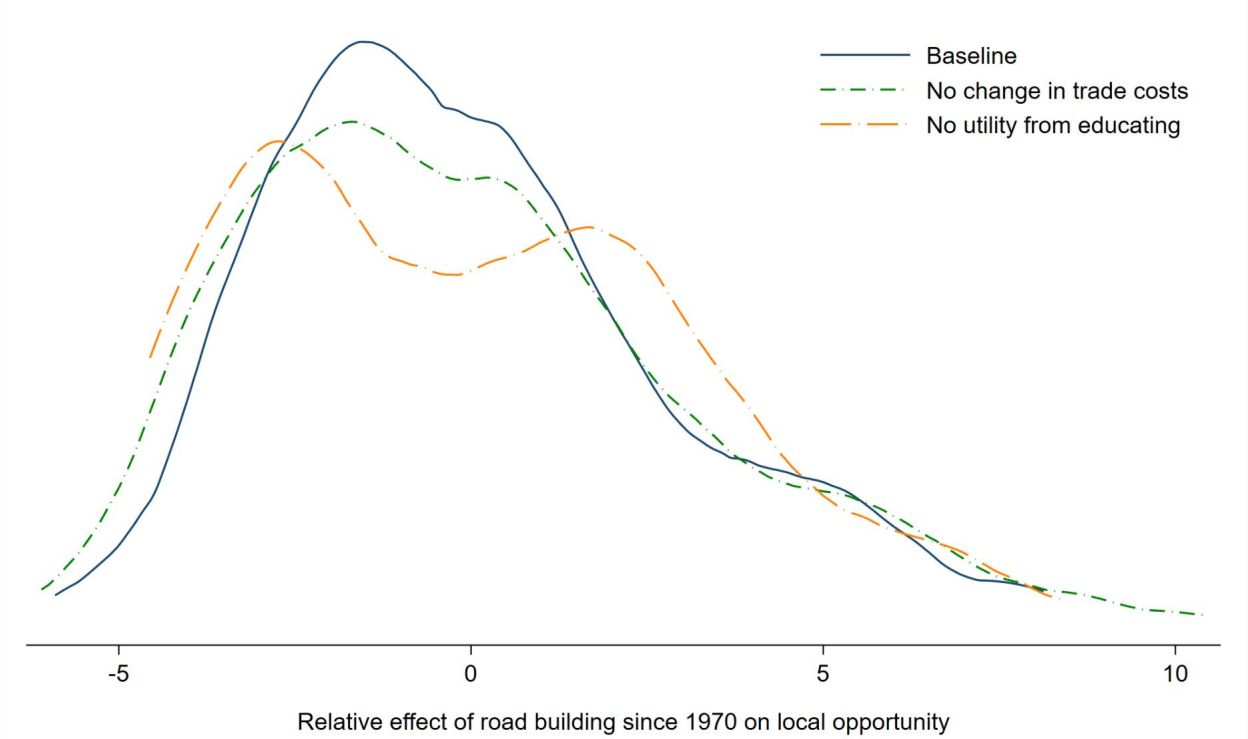
Notes: This figure shows the correlative relationship between locality pagerank centrality on the x-axis and causal place effects on the y-axis. The binscatter plots are constructed by comparing μ_l with $cent_l$ controlling for locality and year fixed effects. The slope coefficient is indicated in orange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

Table 14 Robustness of the main sufficient statistic result

	(1) Baseline	(2) No weighting	(3) Trimmed	(4) Trimmed and no weighting	(5) Weight by inverse SE
Log(LMA Educ)	-0.198***	-0.424*	-0.201***	-0.333*	-0.228
	(0.0656)	(0.241)	(0.0540)	(0.169)	(0.166)
Log(MA Educ)	0.118***	0.159**	0.115***	0.110*	0.145**
	(0.0287)	(0.0797)	(0.0235)	(0.0570)	(0.0627)
Log(LMA No Educ)	0.284***	0.575**	0.239***	0.342**	0.444**
	(0.0577)	(0.228)	(0.0480)	(0.165)	(0.175)
Log(MA No Educ)	-0.171***	-0.261***	-0.150***	-0.158**	-0.232***
	(0.0346)	(0.0962)	(0.0286)	(0.0681)	(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 10. 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 26 Distribution of effects



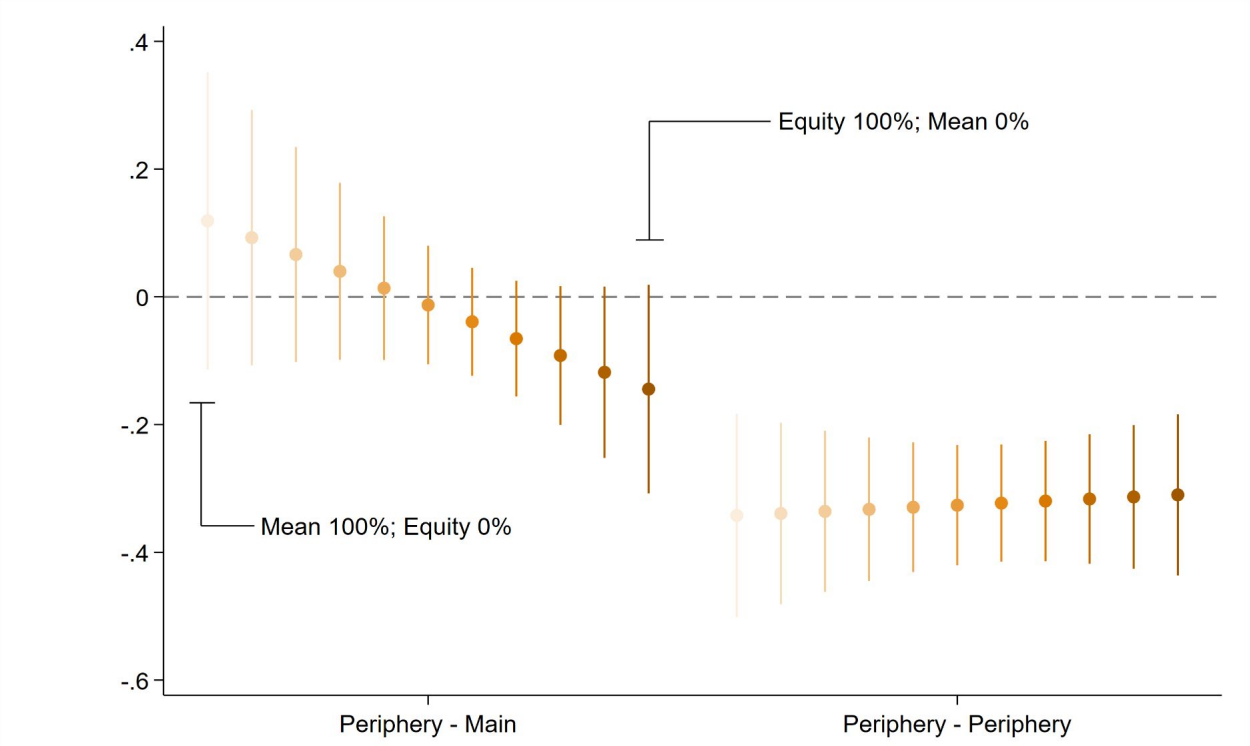
Notes: This figure shows the (centered) distribution over locations of effects of road building since 1970 on local education opportunity, pooled over Benin, Cameroon, and Mali. In blue (solid line) baseline results are plotted. In green (dash-dot line) changes in goods trade costs are shut down. In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

Table 15 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	-1.373**	0.124
	(0.295)	(0.654)	(0.350)
Log(MA Educ)	0.392**	0.699**	0.172
	(0.162)	(0.328)	(0.183)
Log(LMA No Educ)	-0.761***	2.042***	0.442
	(0.275)	(0.675)	(0.312)
Log(MA No Educ)	-0.0385	-1.322***	-0.463**
	(0.164)	(0.356)	(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 27 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".

C Data construction

C.1 Market Access data construction

C.1.1 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.

$$t_{ijt} = \frac{1}{|i|} \sum_{p \in i} \frac{1}{|j|} \sum_{q \in j} d_{pqt} \tag{16}$$

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.

$$t_{ijt} = \sum_{p \in i} \frac{P_{p,1970}}{P_{i,1970}} \sum_{q \in j} \frac{P_{q,1970}}{P_{j,1970}} d_{pqt} \tag{17}$$

This can be seen as a coarse discretisation of equation 16, the best that can be done with the data available.

C.1.2 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.

$$\ln(Q_{ahit}) = \alpha_a + \eta_a \ln(C^N_{hit}) + \xi_a \ln(P_{it}) + \beta X_{hit} + \varepsilon_{ahit} \tag{18}$$

Where α_a are product constants, η_a is the (quasi) income elasticity of demand, C^N_{hit} 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 18 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 16.

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 16 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 $\tilde{Y}^e_{it} = 1/N_{ite} \sum_{h\in\{i,t,e\}} \sum_a \frac{1}{\hat{\eta}_a} Q_{hait} + base$ 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 auxillary regressions in the first stage allow me to approximate how much more income

Table 16 Asset demand equations using DHS data

	(1)	(2)	(3)	(4)
Conrete floor	0.00550***	0.00557***	0.00479***	0.00496***
	(0.000914)	(0.000923)	(0.000820)	(0.000915)
Electricity	0.00212***	0.00218***	0.00205***	0.00158**
	(0.000653)	(0.000668)	(0.000593)	(0.000665)
Sanitation	0.00312***	0.00319***	0.00313***	0.00258***
	(0.000890)	(0.000898)	(0.000814)	(0.000893)
Asset × Region FE	X	X	X	X
Age polynomial		X	X	X
Asset × Region × 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 18 using DHS data.

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.

C.1.3 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-poission-maximum-likelihood (PPML) as in Yotov, Piermartini, Monteiro, and Larch [2016] and others. The estimating equation is given in 19.

$$M_{ijt}^s = \alpha_{jt}^s \cdot \kappa_{ijt}^{-\lambda_s} \cdot \rho_{it}^s \cdot \varepsilon_{ijt}^s \tag{19}$$

i is the origin location, j is the destination location, and s denotes individual type. Where α^s_{jt} are destination-time fixed effects, ρ^s_{it} are origin-time fixed effects and ε^s_{ijt} is an idiosyncratic error term. I approximate bilateral migration costs by the following specifications: $\kappa^{-\lambda_s}_{ijt} = \ln(t_{ijt})^{\tilde{\lambda}_s}$. As M^s_{ijt} and t_{ijt} are observed I can estimate this equation by PPML to recover $\hat{\tilde{\lambda}}_s$. Table 17 shows the results from estimating equations of the form given in equation 19. Columns (1) and (3) estimate 19 by OLS after first taking logs. Columns (2) and (4) estimate equation 19 by PPML. Column (4) represents my main specification and thus I use values $\hat{\tilde{\lambda}}_s = 1.48$ and $\hat{\tilde{\lambda}}_s = 0.985$ to derive labour market access terms.

Table 17 Gravity regressions to recover thetas

	No primary education		Primary education	
	(1) Log-linear	(2) PPML	(3) Log-Linear	(4) PPML
Log(Travel time)	-1.270***	-1.477***	-0.933***	-0.985***
	(0.0193)	(0.0356)	(0.0171)	(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 19 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 $\hat{\tilde{\phi}} = -3.8 \times 0.088$ from Donaldson and Hornbeck [2016], Donaldson [2018].

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 28 gives examples of the original maps used for Mali, reproduced with permission from Michelin, and figure 29 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 28 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.

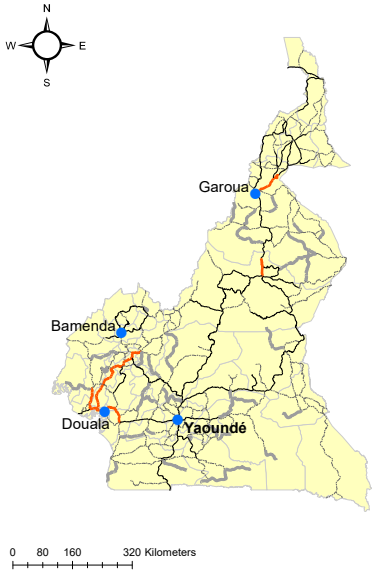
Figure 29 Digitized road networks from historical Michelin maps

(a) Benin 1969

(b) Benin 2019

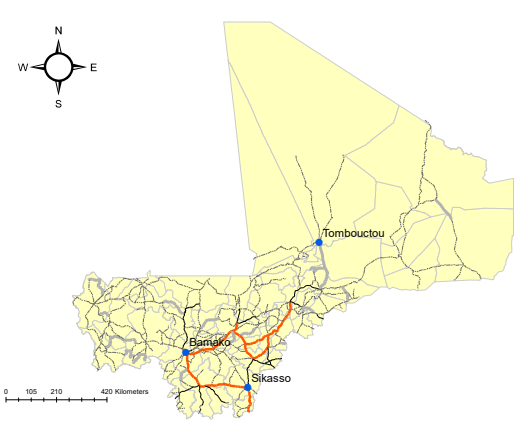
(c) Cameroon 1969

(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.

D Theory appendix

D.1 Solving the spatial model

Theorem 1. *The equations given below:*

$$u_{lt}^{s} = T_{lt}^{B_s}\left(\left(\frac{w_{lt}^{s}}{P_{lt}^{s}}\right)^{1-\beta} E_{lt}^{\beta}\right)^{\lambda_s} \tag{20}$$

$$E_{lt} = \left(\frac{w_{lt}^{E}}{w_{lt}^{N}}\right) \tag{21}$$

$$w_{lt}^{s} = \frac{Y_{lt}^{s}}{L_{lt}^{s}} \tag{22}$$

$$Y_{lt}^{s} = T_{lt}^{Z_s}\left(w_{lt}^{s}\right)^{-\phi} MA_{lt}^{s} \tag{23}$$

$$L_{lt}^{s} = u_{lt}^{s} LMA_{lt}^{s} \tag{24}$$

$$\left(P_{lt}^{s}\right)^{-\phi} = MA_{lt}^{s} = \sum_{k} \tau_{lkt}^{-\phi} \frac{Y_{kt}^{s}}{MA_{kt}^{s}} \tag{25}$$

$$LMA_{lt}^{s} = \sum_{k} \kappa_{lkt}^{-\lambda_s} \frac{L_{kt}^{s}}{LMA_{kt}^{s}} \tag{26}$$

can be rewritten to be of the form:

$$MA_{it}^{r} = \sum_{j} K_{ijt}^{r} \prod_{h=1}^{4} \left(MA_{jt}^{h}\right)^{b_{rh}} \tag{27}$$

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 suppress unless strictly necessary to ease notation. Plug 20 into 24 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 26 and 21 into this we find

$$L_s = \left(A_s\left(\frac{w_E}{w_N}\right)^{\beta}\left(\frac{w_s}{MA_s^{-1/\phi_s}}\right)^{1-\beta}\right)^{\lambda_s} LMA_s \tag{28}$$

. Then noting that by plugging 22 into 23 we have $w_s = B_s^{\frac{\phi_s}{1+\phi_s}} L_s^{-\frac{1}{1+\phi_s}} MA_s^{\frac{1}{1+\phi_s}}$. This can

then be plugged into 28 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 $x = x_E x_N - \beta(x_E\lambda_E - x_N\lambda_N)$).

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

This equation gives the structural interpretation of the coefficients from estimating the reduced form equation in equation 7 in the main text. Plugging this into the above derived equation for L_s gives the following.

$$\begin{aligned}
L_E =& A_E^{\frac{\lambda_E(1+\phi_E)}{x_E} - \frac{\beta\lambda_E(1+\phi_E)}{x_E}\frac{\lambda_E}{x_E}\frac{x_E x_N}{x}} A_N^{\frac{\beta\lambda_E(1+\phi_E)}{x_E}\frac{\lambda_N}{x_N}\frac{x_E x_N}{x}} \\
&\times B_E^{\frac{\phi_E\lambda_E(1-\beta)}{x_E} + \frac{\beta\lambda_E(1+\phi_E)}{x_E}\frac{x_E x_N}{x}\left(\frac{\phi_E}{1+\phi_E} - \frac{\phi_E}{1+\phi_E}\frac{\lambda_E(1-\beta)}{x_E}\right)} \\
&\times B_N^{\frac{\beta\lambda_E(1+\phi_E)}{x_E}\frac{x_E x_N}{x}\left(\frac{\phi_N}{1+\phi_N}\frac{\lambda_N(1-\beta)}{x_N} - \frac{\phi_N}{1+\phi_N}\right)} \\
&\times M A_E^{\frac{(1-\beta)\lambda_E}{x_E} + \frac{(1+\phi_N)}{\phi_N}\frac{(1-\beta)\lambda_E}{x_E} + \frac{\beta\lambda_E(1+\phi_E)}{x_E}\frac{x_E x_N}{x}\left(\frac{1}{1+\phi_E} - \frac{(1-\beta)\lambda_E}{x_E(1+\phi_E)} - \frac{(1-\beta)\lambda_E}{\phi_E x_E}\right)} \\
&\times M A_N^{\frac{\beta\lambda_E(1+\phi_E)}{x_E}\frac{x_E x_N}{x}\left(-\frac{1}{1+\phi_N} + \frac{(1-\beta)\lambda_N}{x_N(1+\phi_N)} + \frac{(1-\beta)\lambda_N}{\phi_N x_N}\right)} \\
&\times L M A_E^{\frac{1+\phi_E}{x_E} - \frac{\beta\lambda_E(1+\phi_E)}{x_E}\frac{x_E x_N}{x}\frac{1}{x_E}} \\
&\times L M A_N^{\frac{\beta\lambda_E(1+\phi_E)}{x_E}\frac{x_E x_N}{x}\frac{1}{x_N}}
\end{aligned}$$

Therefore, using L_E as derived above I can write $LMA_{iE} = \sum_j \kappa_{ij}^{-\lambda_E} L_{iE} LMA_{iE}^{-1}$. Noting that 22 and 23 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.

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

Note the above system does not include fundamentals. $\lambda_{ljt}^{MA^E}$ *denotes the fraction of* l*'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:

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

Where $\lambda_{ij}^{MA^E} = (\tau_{ij}^{\phi_E} Y_j^E \left(MA_j^E \right)^{-1}) / MA_i^E$, 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 $\widehat{Y}_j^k$ 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. □

D.2 Model extensions

D.2.1 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.

$$\max_{S_i} \sum_i A_i \quad s.t. \quad c \sum_i S_i \leq I \tag{29}$$

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 minuets 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.

$$A_i = \pi S_i^{\alpha_1} D_i^{\alpha_2} R_i^{\alpha_3} \tag{30}$$

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.

$$S_i = \left(\frac{I}{cA}\right)^{\frac{1}{1-\alpha_1}} D_i^{\frac{\alpha_2}{1-\alpha_1}} R_i^{\frac{\alpha_3}{1-\alpha_1}} \tag{31}$$

Equation 31 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.

$$\mu_i = \beta_0 + \beta_1 \ln(r_i) + \beta_2 \ln(c_i^E) + \varepsilon_i \tag{32}$$

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

$$\mu_i = \bar{\beta}_0 + \beta \ln(r_i) + \frac{\alpha_2 \gamma_1 \beta_2}{1 - \alpha_1} \ln(D_i) + \left(\beta_2 \gamma_1 \frac{\alpha_3}{1 - \alpha_1} + \gamma_2 \beta_2 \right) \ln(R_i) + \varepsilon_i \tag{33}$$

Where $\bar{\beta}_0$ captures all constants. As $D_i = L_i / Area_i$ and market access terms are a sufficient statistic for L_i, equation 33 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 18 shows the results from this. Column one uses the school data (which is only available in the most recent period) to estimate equation 31 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 10 in the body of the paper. Column two runs the same specification but includes local road quality as suggested by equation 33. 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 18 Endogenising Schooling Supply

	(1) $ln(\#Schools)$	(2) μ	(3) μ
Local Road Network	-0.0827	0.000528	
	(0.141)	(0.00620)	
Log Density	0.332***		
	(0.0839)		
$ln(LMA-E)$		-0.211**	-0.198**
		(0.0768)	(0.0767)
$ln(MA-E)$		0.116***	0.118***
		(0.0280)	(0.0281)
$ln(LMA-N)$		0.291***	0.284***
		(0.0610)	(0.0581)
$ln(MA-N)$		-0.171***	-0.171***
		(0.0333)	(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 31 in a log-linear form. In column two I show the results from estimating equation 33 and in column three I give the baseline sufficient statistic results for comparison.

D.2.2 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:

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

, 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.

$$\hat{L}_{is} = \left(\left(\frac{\hat{w}_{is}}{\hat{P}_{is}} \right)^{(1-\beta)} \left(\frac{\frac{\hat{w}_{iE}}{\hat{P}_{iE}}}{\frac{\hat{w}_{iN}}{\hat{P}_{iN}}} \right)^{\beta} \right)^{\lambda_s} \widehat{LMA}_{is} \tag{34}$$

$$\hat{E}_{is} = \hat{w}_{is}^{-\phi_s} \widehat{MA}_{is} \tag{35}$$

$$\hat{E}_{is} = \hat{w}_{iE} \hat{L}_{iE} \underbrace{\left(\frac{\alpha_{Es} Y_{iE}}{\alpha_{Es} Y_{iE} + \alpha_{Ns} Y_{iN}} \right)}_{\xi^E_{iEs}} + \hat{w}_{iN} \hat{L}_{iN} \underbrace{\left(\frac{\alpha_{Ns} Y_{iN}}{\alpha_{Es} Y_{iE} + \alpha_{Ns} Y_{iN}} \right)}_{\xi^E_{iNs}} \tag{36}$$

$$\widehat{MA}_{is} = \sum_j \lambda_{ij}^{MA_s} \hat{\tau}_{ij}^{-\phi_s} \hat{E}_{js} \widehat{MA}_{js}^{-1} \tag{37}$$

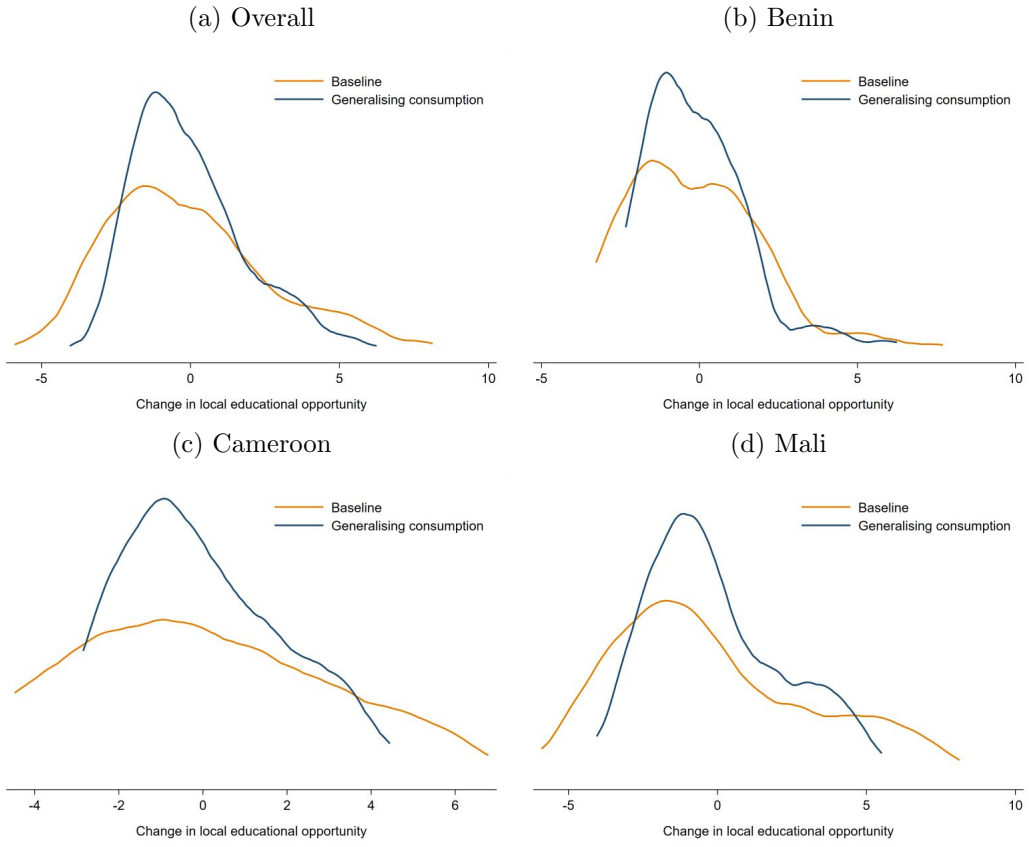
$$\widehat{LMA}_{is} = \sum_j \lambda_{ij}^{LMA_s} \hat{\kappa}_{ij}^{-\lambda_s} \hat{L}_{js} \widehat{LMA}_{js}^{-1} \tag{38}$$

The system of equations given in 34 together with a counterfactual change in the road network $(\hat{\tau}, \hat{\kappa})$, data on $\lambda_{ij}^{MA_s}, \lambda_{ij}^{LMA_s}, \xi^E_{iEs}, \xi^E_{iNs}$, 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. $\lambda_{ij}^{MA_s}, \lambda_{ij}^{LMA_s}, \lambda^E_{iEs}, \lambda^E_{iNs}$ 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 30 then compares the distribution of (relative) effects in the baseline and in the generalizing consumption extension. Sub-figure 30a shows the overall distribution whereas sub-figures 30b, 30c, and 30d 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 30 Changes in the distribution of primary completion rates



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 19 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 19 Relationship with remotness, 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.

D.2.3 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.

$$w_{it} = B_{it} L_{it}^{\alpha} p_{it}^{b} \tag{39}$$

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 $Y_{it} = B_{it} L_{it}^{\alpha} w_{it}^{-\phi} MA_{it}$ which can be rearranged to $Y_{it} = B_{it}^{1/(1-\alpha)} w_{it}^{-\frac{\phi+\alpha}{1-\alpha}} MA_{it}^{1/(1-\alpha)}$. 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.

D.2.4 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: $A^s_{it} = \tilde{A}^s_{it}\left(\frac{L^E_{it}}{L^N_{it}}\right)^{\gamma}$. 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.

D.2.5 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 $u_{it} = A_{it}\left(w_{it}/P_{it}\right)^{a}\left(H_{it}/q_{it}\right)^{(1-a)}$. 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.

$$u_{it} = A_{it}\left(\left(\frac{w_{it}}{P^a_{it}q^{1-a}_{it}}\right)^{1-\beta}E^{\beta}\right)^{\lambda}$$

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

$(1-a)/\bar{H}_{it}$ 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.

D.2.6 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 $Y_i = \left((w_i/B_i)^b q_i^{1-b}\right)^{-\phi} MA_i$ which using land market clearing can be simplified to the following.

$$Y_i = \left(\frac{w_i}{B_i}\right)^{-\frac{b\phi}{1+\phi(1-b}} \left(\frac{1-\alpha}{H_i}\right)^{-\frac{(1-\alpha)\phi}{1+\phi(1-b)}} MA_i^{\frac{1}{1+\phi(1-b)}} \tag{40}$$

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

D.2.7 Land in consumption and production (endogenous supply)

If land is in fixed supply we can combine sub-sections D.2.5 and D.2.6 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 uncertainity 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 parametized 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.

D.2.8 Intermediate goods

Following Caliendo et al. [2018] I can also allow for intermediate goods in production. Allow differentiatred 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.

$$p_{is} = B_{is} \left(w_{is}\right)^{\gamma_s} \left(P_{iE}^{\gamma_{sE}} P_{iN}^{\gamma_{sN}}\right)^{1-\gamma_s} \tag{41}$$

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

$$Y_{is} = B_{is} \left(\left(w_{is}\right)^{\gamma_s} \left(P_{iE}^{\gamma_{sE}} P_{iN}^{1-\gamma_{iN}}\right)^{1-\gamma_s} \right)^{-\phi_s} MA_{is} \tag{42}$$

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

D.2.9 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.

$$
\begin{aligned}
u_{it}^{s} &= A_{it}^{s}\left(\left(\left(\frac{w_{it}^{s}-P_{it}^{N}\gamma_{N}}{P_{it}^{E}}\right)^{\alpha}\left(\frac{w_{it}^{s}-P_{it}^{N}\gamma_{N}}{P_{it}^{N}}\right)^{1-\alpha}\right)^{1-\beta}E_{it}^{\beta}\right)^{\lambda_{s}}\\
E_{it} &= \frac{w_{it}^{E}}{w_{it}^{N}}\\
w_{it}^{s} &= Y_{it}^{s}/L_{it}^{s}\\
Exp_{it}^{s} &= B_{it}^{s}\left(w_{it}^{s}\right)^{-\phi_{s}}MA_{it}^{s}\\
Exp_{it}^{s} &= \gamma_{s}+\alpha_{s}\left(\left(w_{it}^{E}-P_{it}^{N}\gamma_{N}\right)L_{it}^{E}+\left(w_{it}^{N}-P_{it}^{N}\gamma_{N}\right)L_{it}^{N}\right)\\
L_{it}^{s} &= u_{it}^{s}LMNA_{it}^{s}\\
MA_{it}^{s} &= \left(P_{it}^{s}\right)^{-\phi_{s}} = \sum_{j}\tau_{ijt}^{-\phi_{s}}\frac{Exp_{jt}^{s}}{MA_{it}^{s}}\\
LMA_{it}^{s} &= \sum_{j}\kappa_{ijt}^{-\lambda_{s}}\frac{L_{jt}^{s}}{LMA_{jt}^{s}}
\end{aligned}
$$

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.

$$
\begin{aligned}
\left(\hat{w}_{it}^{s}\right)\widehat{MA}_{it}^{s} &= \frac{\gamma_{s}}{X_{it}^{s}}+\frac{\alpha_{s}w_{it}^{E}L_{it}^{E}}{X_{it}^{s}}\hat{w}_{it}^{E}\hat{L_{it}^{E}}-\frac{\alpha_{s}\gamma_{N}(MA_{it}^{N})^{-1/\phi_{N}}L_{it}^{E}}{X_{it}^{s}}(\widehat{MA}_{it}^{N})^{-1/\phi_{N}}\hat{L}_{it}^{E}\\
&+\frac{\alpha_{s}w_{it}^{N}L_{it}^{N}}{X_{it}^{s}}\hat{w}_{it}^{N}\hat{L}_{it}^{N}-\frac{\alpha_{s}\gamma_{N}(MA_{it}^{N})^{-1/\phi_{N}}L_{it}^{N}}{X_{it}^{s}}(\widehat{MA}_{it}^{N})^{-1/\phi_{N}}\hat{L}_{it}^{N}\\
\hat{L}_{it}^{s} &= \Bigg(w_{it}^{s}\left(MA_{it}^{E}\right)^{1/\phi_{E}}\hat{w}_{it}^{s}\left(\widehat{MA}_{it}^{E}\right)^{1/\phi_{E}}\\
&-\gamma_{N}\left(MA_{it}^{N}\right)^{-1/\phi_{N}}\left(MA_{it}^{E}\right)^{1/\phi_{E}}\left(\widehat{MA}_{it}^{N}\right)^{-1/\phi_{N}}\left(\widehat{MA}_{it}^{E}\right)^{1/\phi_{E}}\Bigg)^{\alpha(1-\beta)\lambda_{s}}\\
&\times\left(w_{it}^{s}\left(MA_{it}^{N}\right)^{1/\phi_{N}}\hat{w}_{it}^{s}\left(\widehat{MA}_{it}^{N}\right)^{1/\phi_{N}}-\gamma_{N}\right)^{(1-\alpha)(1-\beta)\lambda_{s}}\left(\frac{\hat{w}_{it}^{E}}{\hat{w}_{it}^{N}}\right)^{\beta\lambda_{s}}\widehat{LMA}_{it}^{s}(L_{it}^{s})^{-1}\\
\widehat{MA}_{it}^{s} &= \sum_{j}\lambda_{ijt}^{MA_{s}}\hat{\tau}_{ijt}^{-\phi_{s}}\frac{\left(\hat{w}_{jt}^{s}\right)\widehat{MA}_{jt}^{s}}{\widehat{MA}_{jt}^{s}}\\
\widehat{LMA}_{it}^{s} &= \sum_{j}\lambda_{ijt}^{LMA_{s}}\hat{\kappa}_{ijt}^{-\lambda_{s}}\frac{\hat{L}_{jt}^{s}}{\widehat{LMA}_{jt}^{s}}
\end{aligned}
$$

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

D.2.10 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.

$$E_{it}^s = \underbrace{\left(\frac{w_{it}^E}{w_{it}^N} \right)^{\beta_1}}_{\substack{\text{Returns to} \\ \text{Education}}} \cdot \underbrace{\left(w_{it}^s \right)^{\beta_2}}_{\text{Income}} \cdot \underbrace{\left(w_{it}^N \right)^{\beta_3}}_{\substack{\text{Opportunity} \\ \text{cost}}} \cdot \underbrace{\left(c_{it} \right)^{\beta_4}}_{\text{Cost}} \tag{43}$$

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 D.2.1 where I endogenise education supply and cost.

As equation 43 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.

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 1. 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.

$$U_{ijt}^{s} = \underbrace{\frac{b_{jt}}{\kappa_{ijt}}}_{\text{Location}} \underbrace{\left(C_{jt}^{s}\right)^{1-\beta}}_{\text{Consumption}} \tag{44}$$

Households seek to maximize 44 subject to the budget constraint given below.

$$P_{jt}^{s} C_{jt}^{s} = w_{jt}^{s} \tag{45}$$

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:

$$b_{jt} \sim G_{jt}^{B}, \quad G_{jt}^{B}(b) = e^{-T_{jt}^{Bs} b^{-\lambda_s}} \tag{46}$$

$\{T_{jt}^{B_s}\}$ 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.

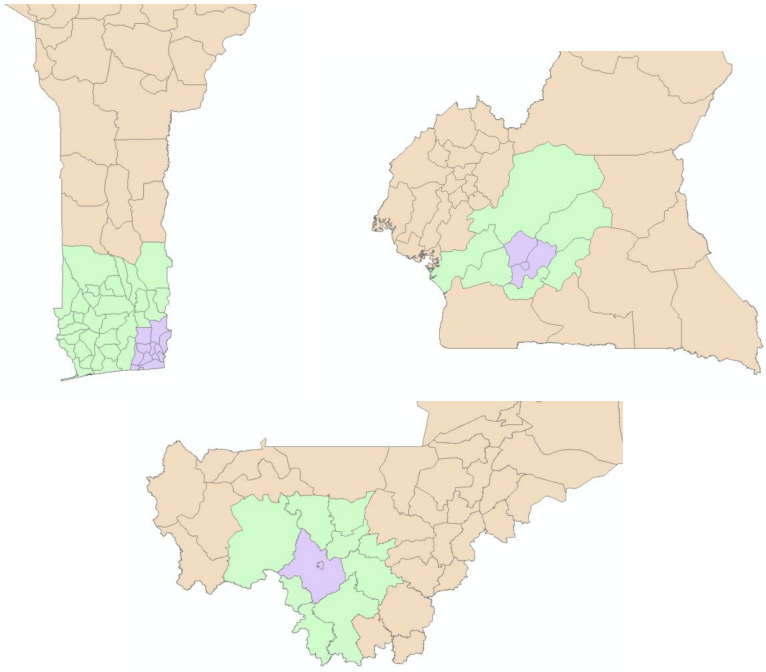
Firms produce a differentiated consumption variety q and production occurs under perfect competition in each location. Labor is the only factor of production. As discussed above there are two industries/ sectors, firms that produce E type goods and those that produce N type goods. Sectors are differentiated 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:

$$Y_{jt}^s(q) = Z_{jt}^s(q) L_{jt}^s \tag{47}$$

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 1.

D.4 Future road upgrading counterfactuals details

Figure 31 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.

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 48 summarizes the regression specification.

$$\hat{\mu}_{ir} = \beta_{tk} \cdot \text{RoadType}_{t(r)} \cdot X_{ki} + \alpha_r + \gamma_{c(i)} + \varepsilon_{ri} \tag{48}$$

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 deferentially impact the effect road building has on local opportunity by road type t. In table 20 I show the results from running equations of the form given in 48, 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 20 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) breakdown 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 20 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***	0.0153**	0.00270	0.0188***	-0.115***
	(0.00253)	(0.00632)	(0.0108)	(0.00601)	(0.0242)
Primate × Primary completion rate		0.0230**	0.00551	0.0369***	-0.0101
		(0.00950)	(0.0283)	(0.00834)	(0.0522)
Hinterland × Primary completion rate		-0.0739***	-0.0214*	-0.0666***	0.0257
		(0.00691)	(0.0114)	(0.00683)	(0.0297)
Agricultural employment rate	0.0341***	0.0708***	0.0917***	-0.0941***	-0.00986
	(0.00287)	(0.00698)	(0.00944)	(0.0114)	(0.0125)
Primate × Agricultural employment rate		-0.0396***	-0.0780***	-0.0376**	-0.00984
		(0.0125)	(0.0239)	(0.0168)	(0.0242)
Hinterland × Agricultural employment rate		-0.0466***	-0.0728***	0.0473***	0.00692
		(0.00769)	(0.00996)	(0.0124)	(0.0155)
Housing quality index	1.296***	0.986***	0.771***	-5.898***	3.125***
	(0.0967)	(0.237)	(0.295)	(0.489)	(0.367)
Primate × Housing quality index		-1.674***	-1.202	-2.796***	1.038
		(0.450)	(0.750)	(0.752)	(0.709)
Hinterland × Housing quality index		0.500*	-0.622**	5.321***	-0.134
		(0.261)	(0.312)	(0.527)	(0.470)
Log(Expected travel time)	1.907***	2.318***	14.30***	-5.150***	-0.899***
	(0.120)	(0.267)	(0.505)	(0.521)	(0.304)
Primate × Log(Expected travel time)		-4.242***	0.828	-3.002***	-3.758***
		(0.485)	(1.349)	(0.721)	(0.564)
Hinterland × Log(Expected travel time)		-0.179	-6.457***	6.015***	-0.613
		(0.302)	(0.538)	(0.596)	(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:
$\hat{\mu}_{ir} = \beta_{tk} \cdot \text{RoadType}_{t(r)} \cdot X_{ki} + \alpha_r + \gamma_{c(i)} + \varepsilon_{ri}$ 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.

D.6 No-roads counterfactual results with alternative parameters

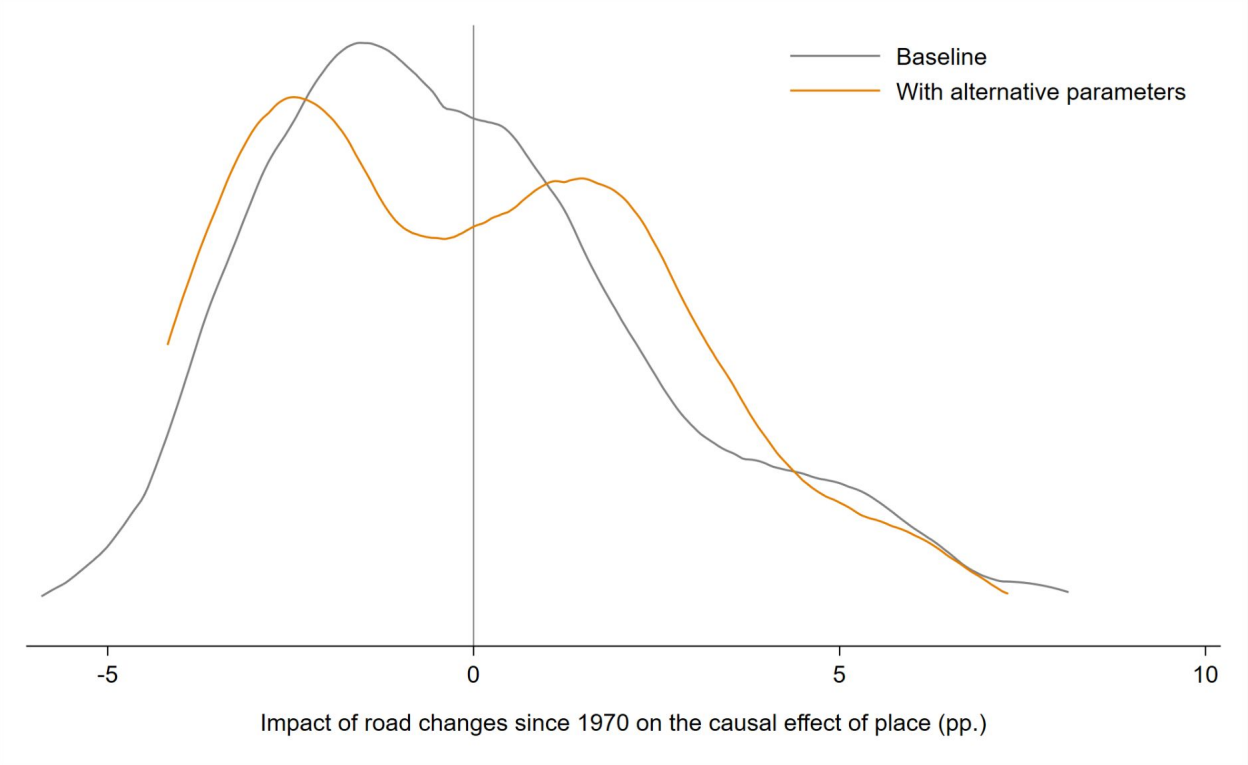
D.6.1 Using parameter values taken from the literature

In this section I compare my baseline results to those under the alternative parameterization taken from the liturature 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 32 compares the distribution of effects using the baseline and alternative parameters. Table 21 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 32 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 6 pooling over all three countries. In orange I plot the analagous results with the model estimated using the alternative parameters.

Table 21 Relationship between the total relative effect of roads built since 1970 and locality remoteness in 1970. Comparing the baseline and alternative parameterization

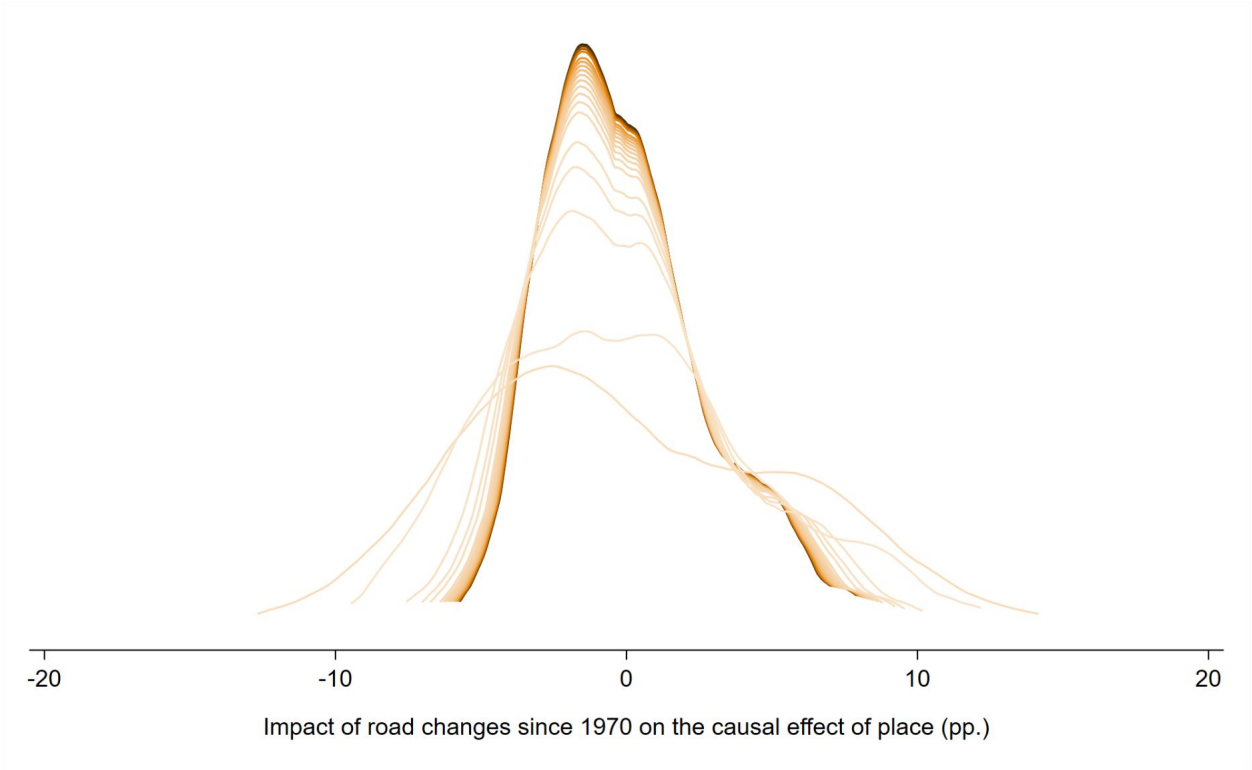
	(1) Baseline	(2) With alternative parameters
Benin × 1970 Remoteness	5.092*** (1.357)	9.472*** (1.300)
Cameroon × 1970 Remoteness	6.995*** (2.409)	9.596*** (1.664)
Mali × 1970 Remoteness	5.468*** (1.299)	7.491*** (0.876)
Observations	156	156
R^2	0.334	0.642

Notes: This table correlates 1970 remoteness with the total impact of road building since 1970 on local opportunity. Column one replicates results from figure 7 in the main text. Column two presents analogous results under the alternative parameterization.

D.6.2 Alternative values of β

Figure 33 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 33 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 33 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 33 that for any reasonable values of β, results are not significantly changed.

Figure 33 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.