THE CURSE OF GOOD SOIL? LAND FERTILITY, ROADS, AND RURAL POVERTY IN AFRICA*

Leonard Wantchekon† Piero Stanig‡

April 6, 2015

Abstract

Using a global poverty map and standard soil productivity measures, we find that the poorest districts in Africa are more likely to have better (not worse) soil quality and that land fertility is higher in districts with worse roads. Our results are robust to a battery of controls and alternative measures of poverty and soil quality. The results indicate that transportation costs are the main drivers of poverty in Africa and that isolation might turn soil quality into a curse. More specifically, in districts with poor infrastructure, the poverty rate increases as soil quality gets better. We find that these results could be attributed to relatively low human capital investment in isolated districts with abundant agricultural resources. We provide evidence for causality by using colonial road networks as an instrument for current transportation costs.

JEL: O13, O55, Q15, R42

*The authors would like to thank Nicole Mason, Harouna Kazianga, Jean-Francois Maystadt and participants of the PDL seminar at Princeton, ASE-CRED Workshop at the African School of Economics (Benin), the 2014 Annual Bank Conference on Africa in Paris, and the Political Economy Seminar Series at the Hertie School of Governance for comments and suggestions. Russell Morton and Sanata Sy-Sahande provided excellent research assistance. Financial support from Open Society Foundations is gratefully acknowledged. The usual caveat applies.

†Department of Politics, Princeton University.

‡Department of Policy Analysis and Public Management, Bocconi University.

1 INTRODUCTION

Soil quality and land fertility are key determinants of agricultural production and economic growth. As a result, soil degradation and drought are widely perceived to be associated with food insecurity and rural poverty (Barbier 2010; Nkonya et al. 2011). It is therefore not surprising that policy-makers and academics have recently focused on the adoption of modern inputs, including fertilizer use and improved seeds as a possible solution for rural poverty (Demery and Christiaensen 2007; World Bank 2008; Morris et al. 2007; Dercon and Christiansen 2011). For instance, Sanchez (2002) cites soil quality as a key driver of low agricultural production, while Scherr (1999) and Woomer et al. (1994) emphasize the role of soil degradation as a major concern for food security in Africa (See also Nkonya et al. 2011). In other words, the conventional wisdom is that soil quality is negatively correlated with rural poverty and that improving land fertility is crucial to poverty alleviation. We find the opposite to be true.

Using data on soil quality, poverty, transportation costs, and other potential determinants of poverty covering more than five thousand subnational units from forty-six sub-Saharan African countries we establish four important sets of empirical regularities. First of all, there exists a positive correlation between soil quality and poverty in Africa, meaning that regions where land is most fertile are on average more likely to be impoverished than regions where soil is poorer. In addition, transportation costs and isolation are the main drivers of rural poverty. Furthermore, there exists a mismatch between soil quality and infrastructure. Roads tend to be bad in areas with good soil, such as in hills and valleys, and good where the soil is of worse quality, such as in flat terrain close to the coast. Finally, when infrastructure is poorly maintained or non-existent, households are poorer in areas where the soil is quite fertile than in areas where the land is barren. We also argue that these results may be attributed in part to insufficient human capital investment. In the absence of proper roads and transportation, residents may see little point in investing in education and other human capital drivers given their isolation and their abundant agricultural resources. Conversely, in soil-poor districts with bad roads, households have lower opportunity costs of education than those living in soil-rich districts with bad roads.

Several papers have already pointed to a weak association between soil quality and income (Drechsel et al. 2001, Ehui and Pender 2005, Okwi et al 2007). In particular, Okwi et al. (2007)

use data from Kenya and show that if all soil were at the highest quality, it would only lead to a 1 percentage point decrease in poverty. While soil quality is found in some studies to have a limited effect on income, the literature finds a strong association between rural roads and poverty (Jacoby, 2000; Gibson and Rozelle 2003; Mu and van de Walle 2007; Jacoby and Minten, 2009; Khandker et al., 2009). From the theoretical point of view, Gollin and Rogerson (2014) propose a model that links transportation costs and the fraction of the rural workforce engaged in subsistence agriculture. Calibrating the model to match the features of sub-Saharan African countries, they make predictions about the role of transportation productivity in economic development and in the allocation of resources between manufacturing, subsistence agriculture and modern agriculture. In particular, they find that there are large interaction effects between agricultural productivity and improvements in transportation. Stifel and Minten (2008) find a strong relationship between poverty and isolation in Madagascar, highlighting four mechanisms at work: transportation costs, plot size and productivity, price variability and extensification onto less fertile lands, and insecurity. In other words, rural households choose to use larger plots with significantly fewer inputs, and invest in crops for their own consumption rather than for income because they are not sure they will be able to sell surplus crops to the market. Diversification of crops into more types and across more land also acts as insurance against price variability in cash crops, which is stronger in rural areas, as families do not need to purchase food. Meanwhile, extensification provides insurance for insecurity. Minten et al. (2013) also find that transaction and transportation costs together add about 50 percent for the most remote farmer to the fertilizer prices charged at the input distribution center in their case study of the inpout distribution system in Northwestern Ethiopia. However, as Stifel and Minten acknowledge, endogeneity of isolation may be a serious issue in their study, with superior land quality being associated with less isolation. We build upon these results by studying the relationship between isolation and soil quality as determinants of poverty.

The literature also highlights mechanisms through which rural roads decrease poverty and increase food security. Ali (2010) finds that households in Bangladesh invest in better technology once a road is built, replacing traditional local rice with a high-yield variety that requires more inputs. Households that only grow traditional rice increase their acreage instead. Additionally, the treatment effect is stronger for wealthier households, presumably better able to bear the costs of the inputs associated with the high-yield rice. The paper shows that the result only applies to

households close to the road, with distant households not changing their practices. Additionally, the preference for high-yield rice did not appear to change over time significantly outside of the treatment population, with the price for the seed remaining stable across the time period studied. This result is consistent with the general notion that rural households use more inputs once roads are built. Bell and Dillen (2012) find that the introduction of rural roads in India leads households to invest more in education and health, with fewer days of school missed and much higher use of medical services. Yamano and Kijima (2010) find that soil quality is correlated with farm income only after controlling for isolation. They also find that both isolation and soil quality are associated with crop choice. The result suggests that isolated communities with poor soil quality may have stronger incentives to use non-farm income to supplement their agricultural output, while isolated agricultural households with higher quality soil are less likely to use non-farm income for consumption.

Some of the more recent research highlights informational frictions and market aspects of agriculture in isolated areas. In a study of India, internet kiosks displaying current prices have been shown to raise prices in the regulated markets, and decrease price dispersion (Goyal, 2010). Additionally, the internet kiosks are associated with 19% higher production without increasing acreage, signaling either more intensive use of inputs and/or substitution to more profitable crops. A second study applies such logic to a case study of Sierra Leone (Casaburi et al., 2010). Applying a regression discontinuity design to a program where roads were built based on scores (based on population and economic value per kilometer, length and other considerations), the authors find heterogeneity in the treatment effect of roads on net returns for agriculture. In other words, the price impact of road construction depends upon other factors, specifically productivity and linkages with urban consumers, with higher prices observed in productive areas and vice-versa.

Our paper builds upon the notion of heterogeneity in the impact of roads, noting that roads have a much more positive benefit in areas with high soil quality, which is in line with the result that more productive areas receive higher prices, and, thereby, higher net returns due to roads.

On the methodological front, two recent papers, both studying China, have advanced a new identification strategy to tackle the issue of endogeneity of infrastructure. Banerjee et al. (2012) use colonial-era railroads as an instrument for current transportation networks to show that proximity to transportation infrastructure promotes long-term growth, while Faber (2012) treats a large

infrastructure-building project as a natural experiment. We build on this type of intuition, and exploit variation in infrastructure induced by decisions made during the colonial period to assess the role that transportation plays in determining rural poverty in contemporary Africa.

Our paper makes several important contributions to the literature. Foremost, in contrast to the conventional wisdom, we find a negative correlation between soil quality and poverty. This implies that poverty cannot be explained simply by invoking land degradation. In fact, as a stylized fact, a majority of poor households in Africa dwell on relatively high-quality soil.

Further, we find that the interaction between infrastructure and soil quality is positive and statistically significant. This can be interpreted in different ways: (1) soil quality matters most when there is good transportation infrastructure, (2) infrastructure is the most useful when soil is of good quality. But it might also point to the existence of a "curse of good soil." We estimate flexible models to assess whether good soil without infrastructure may lead to worse economic outcomes. The data show that among bad-infrastructure districts the expected poverty rate is higher when soil quality is better.

We provide some additional evidence about the correlates both of infrastructure and soil quality. We find an element of "bad luck" in how infrastructure and soil are matched. In particular, the best soil is in districts with medium-high level slopes (e.g., hilly terrain) and, all else equal, at an altitude of 200 meters, while the best infrastructure is found in flatter lands and low altitude. We also find that districts containing rivers and those close to the 1900 colonial borders have better land quality and worse infrastructure.

We adopt several strategies to deal with plausible endogeneity of infrastructure provision. First, we use the presence of roads in the colonial period as an instrument for current infrastructure provision or transportation costs. We show that the location of colonial infrastructure can be predicted based on geographic characteristics and, importantly, by the presence of extractive resources (mines and quarries) but not by soil quality. This is in line with conventional wisdom in economic history: colonial powers were not after farmland but minerals. More specifically, we find that the probability of having a colonial road is around 22 percentage points higher in mining districts than in non-mining ones.[1]

[1]The results implicitly link food security to historical factors. Given that contemporary infrastructure availability is associated quite closely to the road network in colonial times, decisions made by colonizers for other reasons affect how land is used today, and therefore the current patterns of poverty.

One possible concern might be that mining could have an independent effect on poverty; if colonial roads also capture some of the direct effect of mining on poverty, then the exclusion restriction assumption might be violated. But conditional on mining, this assumption should hold. We estimate a model in which we condition on mining and find that mining districts have somewhat lower poverty, but our basic results still hold: the interaction between soil and infrastructure, instrumenting for current infrastructure with colonial roads, is still negatively correlated with poverty. We perform various robustness and sensitivity checks to assess the extent to which our results might be driven by violations of the exclusion restriction. Our main result for the interaction between soil quality and infrastructure survives even if the exclusion restriction is assumed to hold only conditionally, or to hold only approximately (so that the direct effect of the instruments on the poverty rate is not zero).

Overall we find a negative association between soil quality and income in isolated areas. The results are robust to the inclusion of measures of urbanization and the exclusion of urban districts. In other words, our results are not simply an indication that rural areas are poorer than urban areas due to agricultural productivity (see Lewis 1955, McMillan and Rodrik 2011). Instead, we relate the income differentials across rural areas to market access and high transportation costs and the dominance of subsistence agriculture, which is strongly associated with rural poverty. Our results suggest that the most important factor driving rural poverty and food security is the provision of rural infrastructure and access to markets. In other words, to understand the relationship between factor endowment (e.g., soil) and rural development or rural poverty, it is crucial to take into account the complementarity between soil quality and infrastructure availability and market access. If one were to overlook this interaction, one would underestimate the return to infrastructure in poverty reduction. [2]

Our result implies that the return to infrastructure depends on productive asset endowments: in other words, infrastructure has a larger poverty-reduction effect when the productive potential of the land is high. In addition, by estimating separate coefficients for all combinations of soil quality and infrastructure provision, we show that districts with good soil and bad roads are worse off than districts with lower-quality soil and equally bad roads.

[2] In fact, the estimated effect of infrastructure on poverty would be an average of the effect it has in places where roads would not matter much (because of low soil quality) and roads matter a lot (because the soil quality is high and the productive potential at high investment and high effort is large).

We discuss potential mechanisms that account for this perverse effect of soil quality on rural poverty. We use a case study of Kenya to explore the role of human capital as a possible mechanism mediating the relationship between soil quality and isolation on one hand, and rural poverty on the other. Investment in human capital might be low in isolated districts with good soil because of higher opportunity cost of education and poor service delivery in these areas. As expected, soil quality is associated with worse education outcomes, both at primary and secondary school levels.

2 DATA

Our dataset covers 5334 subnational units in 46 sub-Saharan African countries. The number of districts per country varies from a minimum of two districts (Sao Tome and Principe) to over 550 (in Nigeria) and its median is 80. The dataset covers a wide array of phenomena, ranging from poverty to infrastructure to soil quality to population density.

The backbone for constructing the data set is the GIS map of African sub-national entities at level 3, published by the UN's Food and Agriculture Organization (FAO). We overlay the shapefile of the subnational districts to the geocoded sources of soil quality, poverty, as well as other relevant variables and compute district-level summaries, that we treat as the value of that given variable for the district. In other words, when the resolution of the original measure is higher than that of the districts (so that one district encompasses many cells in the grid) we average the values of the cells. This is the case for the vast majority of variables, and is always the case when the original data are released as gridded datasets or shapefiles (e.g., of rivers). In the case of the few measures that have lower resolution than the districts borders, we attribute to the district the value of the larger unit in which it is contained (or the mean of the values of the larger units it spans, if the district belongs to more than one larger unit).

2.1 MAIN VARIABLES

Soil quality For soil quality we rely on several measures, from different sources. First of all, we collected the scores on seven dimensions, published by the International Institute for Applied Systems Analysis (IIASA) and part of the Harmonized World Soil Database (HWSD). These evaluate soil quality according to the following criteria: nutrient availability; nutrient retention capacity;

rooting conditions; oxygen availability to roots; excess salts; toxicity; and workability. The data is released for small cells (30 arcseconds, or approximately less than one square kilometer at the equator) and the values in each cell can range from 0 to 7, with higher values meaning worse soil quality. For each subnational entity, we compute the average soil quality for each of the seven dimensions (as the simple average of the cells contained in the perimeter of the subnational entity). We then perform principal components analysis on the data for the whole continent (the unit of observation, here, is the subnational district). From this, we extract the first principal component. This loads negatively on all the seven variables, hence it is an index of soil quality where higher values indicate better quality.

We also use the index of soil production published by FAO. This index "considers the suitability of the best adapted crop to each soil's condition in an area and makes a weighted average for all soils present in a pixel." (FAO 2007) We calculate, for each subnational district, the average of this index. Finally, we use the classification of problem land published by FAO. This classifies each cell, at high resolution, in one of several categories. We calculate the proportion of cells in the district that fall in a category characterized as "No problem soils > 30% of the mapping unit", to construct the variable "goodsoil". For the robustness checks, we also create analogous variables with proportions for cells classified as "steep" or "infertile" by the problem land data set.

Poverty For poverty, we rely mainly on the Global Poverty Map Derived From Satellite Data published by the National Geophysical Data Center (NGDC). The map is released at the 30 arcseconds resolution (again, approximately one square kilometer at the equator). This dataset matches night lights visible from satellites with population density estimates from the LandScan dataset. These are then benchmarked with available poverty data at the national and subnational levels. See Elvidge et al. (2006) for a detailed description. This dataset has a very high resolution, which allows us to calculate a distinct value of the poverty rate for each subnational district. This means that we are able to perform all of our analysis at the subnational district level including country fixed effects. In this estimation approach, all the coefficients on the explanatory and control variables are identified exclusively by within-country variation. The fixed effects account for all the features of the country that do not vary across districts (e.g., political regime, legal provisions regarding freedom of association, quality of governance at the national level, etc.) as well as the

mean of all omitted variables that do vary across districts.

Alternative measures of poverty that are geo-coded and are based on direct evidence (survey estimates, census data, etc.) are available. In particular, the Poverty Mapping Project: Global Subnational Prevalence of Child Malnutrition published by the Center for International Earth Science Information Network (CIESIN), Columbia University, provides geocoded poverty estimates based on "hard" data. It reports the percentage of children with weight-for-age z-scores that are more than two standard deviations below the median of the NCHS/CDC/WHO International Reference Population and aims at providing "a global subnational map of the prevalence of underweight children that can be used by a wide user community in interdisciplinary studies of health, poverty and the environment." The data refers, depending on the location, to the most recent available year between 1990 and 2002.

In an analogous effort, HarvestChoice/International Food Policy Research Institute (IFPRI) publishes a geocoded map, referring to the year 2005, of sub-national poverty headcount ratios, derived from 23 nationally representative household surveys and population censuses. Poverty is defined at the \$2/day level, expressed in 2005 international equivalent purchasing power parity (PPP) dollars. Rates are in percentages of total population.

The resolution of these alternative measures is, unsurprisingly, much coarser than the satellite-imaging data. This makes them unsuitable for district-level estimation with country fixed effects.[3] In fact, the CIESIN and the IFPRI data provide few distinct values per country, and, especially in smaller countries, this makes their use in our analysis problematic. First of all, the variation across districts within country is smaller for these two measures than for the satellite data, as reflected by coefficients of variation for the satellite data that are in the overwhelming majority of cases quite higher. [4] In addition, the district-level values have within-country coefficient of variation exactly equal to zero (implying that the index takes the same value in all districts) in a quarter of the countries in the IFPRI data, and in two countries in the CIESIN data. See Appendix for full reports. These countries would drop out of the analysis if a measure like this were used as the dependent variable in country-fixed-effects estimation.

[3]CIESIN releases the data in raster format at the nominal resolution of a quarter degree; yet, the data just replicates in the grid the few polygons per country available in the shapefile.

[4]The coefficient of variation discussed here is calculated by dividing the *within-country* standard deviation of the poverty value by the country-wise mean.

Yet, it is still worth comparing the main measure we use to these alternative estimates of poverty, in order to validate the measure we chose as dependent variable in the analysis below. Both alternative measures are positively correlated with the satellite data (with correlations respectively .24 and .18) and linear regressions of the satellite values on the CIESIN and IFPRI data yield (positive, and substantively large) coefficients, with t-values respectively 24 and 28.8. These associations survive the inclusion of country fixed effects (so that variation across countries is not accounted for when estimating the regression coefficient, but only variation within country). In fixed effects models, in particular, the coefficient on the CIESIN measure is not statistically distinguishable from one. Given that both variables are measured as a percentage of total population (and hence on the same scale) this establishes an almost-perfect coincidence between changes in the expected value of the satellite measure and changes in the CIESIN measure.

Infrastructure For infrastructure, we rely on one main variable, road cost, published by FAO. This measure is calculated as follows. First, the road network in Africa is classified according to the accompanying road type classification system. Then, the cost to travel from one cell to the next is estimated, assuming that "the time required to travel from one cell to another in absence of main roads is 5 times longer than the time needed on the main road." The information on the road network was derived from ArcWorld (ESRI, 1992). In addition, we also employ the average of the class of the roads found in the district. This is based on the Roads of Africa dataset (also published by FAO).

2.2 CONTROL VARIABLES AND INSTRUMENTS

As for the control variables and the instruments, these can be divided into pure geographic/geological variables, demographic variables, and historical variables.

Geographic variables The geographic controls are elevation and slope class of the terrain, distance from the coast, the presence of water bodies, the presence (and type) of rivers in the district, the length of the growing period, rainfall, and classification into an agro-ecological zone.

Elevation data comes from the IIASA-LUC Global Terrain Slopes and Aspect Database, and is originally reported at the 30 arcseconds resolution. We compute the average by district. Elevation

is reported in meters, and we rescaled it in hundreds of meters. The slope data comes from the FAO Geonetwork, originally from the FAO-UNESCO Soil Map of Africa: each cell is classified in one of three classes, and we compute the median, the mean and the standard deviation of the slope class within the district.

We compute the distance of the district from the closest coast, based on the data on coastlines published by naturalearthdata.com. This takes the value of zero for a coastal district. For rivers, we rely on the World Rivers GIS file (from http://worldmap.harvard.edu). We compute several summaries: whether there is a river in the district; the distance of the district from the closest river (equal to 0 if a river flows in the district); the average rank class of the rivers present in the district; the rank class of the largest river in the district. We also calculate the proportion of cells in the district classified as being water bodies. The data on water bodies also comes from the FAO Geonetwork.

To build our instrument for soil quality, we overlay the map of subnational districts on the Digital Soil Map of the World. This classifies the dominant soil for relatively fine-grained areas (FAO, 1974). In general, a district belongs to more than one of the soil areas. We get the list of the dominant soil in each of the geological areas that span the district, and we record all the classes of the dominant soil present in a district. We then create a set of dummies, one for each major soil grouping. We have a total of 22 dummies, for 22 soil groupings. The dummy for a given class of soil takes the value of one in a given district if part of that district is spanned by a geological area whose dominant soil belongs to that class.

Demographic variables We collect several demographic variables. First of all, we compute the average and maximum level of urbanization in the district. The data on urbanization comes from IIASA and is released at the 30 arcseconds resolution; for each cell, the percentage of urban population is reported. Similarly, we compute the percentage of cultivated land, also reported by IIASA.

We also collect data on travel times to towns of at least 20,000 thousand inhabitants, and to cities of more than half a million inhabitants. The data on which our measures are based are published by HarvestChoice/IFPRI in raster format. We compute the average value of the cell-level scores in the district.

In addition, we also rely on the data on settlements published by CIESIN and now part of the Global Rural-Urban Mapping Project (CIESIN, IFRPRI, and CIAT 2011). This dataset provides detailed information about human settlements based on a variety of sources. For every settlement, the location, and the population as of 1995 (among other pieces of information) is provided. We restrict the analysis to settlements of at least 5000 inhabitants. We compute the number of settlements with more than 5000 and 10,000 inhabitants (both in absolute terms and relative to the area of the district) and the mean, the median, and the maximum size of the settlements in the district. If a district has no recorded settlement with a population of at least 5000 as of 1995, all of these take the value of zero.

We also compute the average of the rural population density figures published by FAO at the resolution of 5 arc-minutes. Each pixel classified as "rural" by the urban area boundaries map has information about the number of persons per square kilometer, aggregated from the 30 arcsecond data layer. Again, we compute the average value by district. The measure of rural population density is missing by construction in cells (pixels) classified as urban by FAO. This fact has two consequences: first of all, the variable we compute (the mean by district) is unaffected by the presence of a high-density urban area in the district. Urban areas are basically considered non existent when the district-wise average is taken. In addition, districts that encompass only urban areas have a missing value on this variable. This also means that when we include rural population density in regression models, we are excluding from the analysis urban-only districts.

Historical/political variables Finally, we collect data on transportation infrastructure and political boundaries in the colonial era. For colonial borders, we rely on the historical political maps published in geocoded format by the Harvard Geospatial Library. We compute the distance to the closest colonial-era border. We also record to which mapping unit (e.g., large colonial administrative unit) the district belonged as of 1900 and as of 1950.

To construct measures of colonial-era infrastructure, which we use for instrumental-variable estimation in Section 5, we georeferenced a German road map of Africa as of 1941, found in the collection of the library at Princeton University. For each road featured in the map, we also code whether it is a "primary," "secondary," or lower-rank road. Then, from the shapefile with the information about the roads, we create some district-level summaries about transportation infras-

tructure in the colonial era. The first set of summaries are simply dummies (indicator variables) that take the value of one if the district was crossed by a road in the colonial era. We rely on *current* district borders, as defined in our main data source for administrative districts. We create two dummy variables, Primary colonial and Secondary colonial, equal to one for those districts that were crossed, at least in part, by colonial transportation infrastructure of either type. We also create a third dummy, equal to one if there was any colonial road infrastructure in the district.

We also compute the (point to set) distance between the geometric center of the district, and the closest road of each class. The road does not necessarily pass through the district and it is not required to lie within the country to which the district belongs in the post-colonial period. Roads in neighboring districts and, for that matter, in different countries, are also included in the computation. Very high values mean that not only infrastructure was underprovided in that district, but also in the overall region in which the district is located.[5]

3 BASIC ANALYSIS

The plot in Figure 1 displays the level of poverty (percentage of the population living on less than two dollars a day) according to the IFPRI data, averaged by soil category, after controlling for country fixed effects. Higher values of the soil production index reflect better soil. From the raw data, then, it appears that, without conditioning on other district characteristics, poverty is more of an issue in districts with better land.

To explore the pattern rigorously, we regress the measure of poverty on the measures of soil quality. All the models include country fixed effects (omitted from the tables to improve readability). The fixed effects capture the effect of all features of the districts that do not vary within country. In addition, the standard errors are clustered at the country level: doing so accounts for the fact that there might be remaining correlation in the errors at the district level even after the inclusion of fixed effects to account for the nesting of the districts within countries. (Angrist and Pischke 2009, chapter 8; see also Arellano 1987)

Model 1 measures soil quality with the first principal component of the 7 IIASA measures;

[5] As long as contemporary infrastructure is affected by the long-term patterns set in the colonial era, the fact that the distance measures are calculated also based on colonial roads located very far from the district does not affect our identification strategy directly. Above a certain level of remoteness, distance might not matter much. But this does not violate the exclusion restriction.

Model 2 includes the Soil production index; Model 3 uses the proportion of the district that is classified as having good soil according to FAO's problem soil classification. In all the models, the coefficient on the measures of soil quality is estimated as positive, and substantively large; in models 1 and 3, the association is statistically significant at conventional levels. According to the estimate of Model 1, if we compare a district with soil quality at the first quartile with one at the third quartile, the poverty rate in the latter is expected to be three percentage points higher than in the former. This points to the fact that, on average, districts that have better land tend to have higher poverty rates than districts with worse soil. This result is highly counterintuitive.

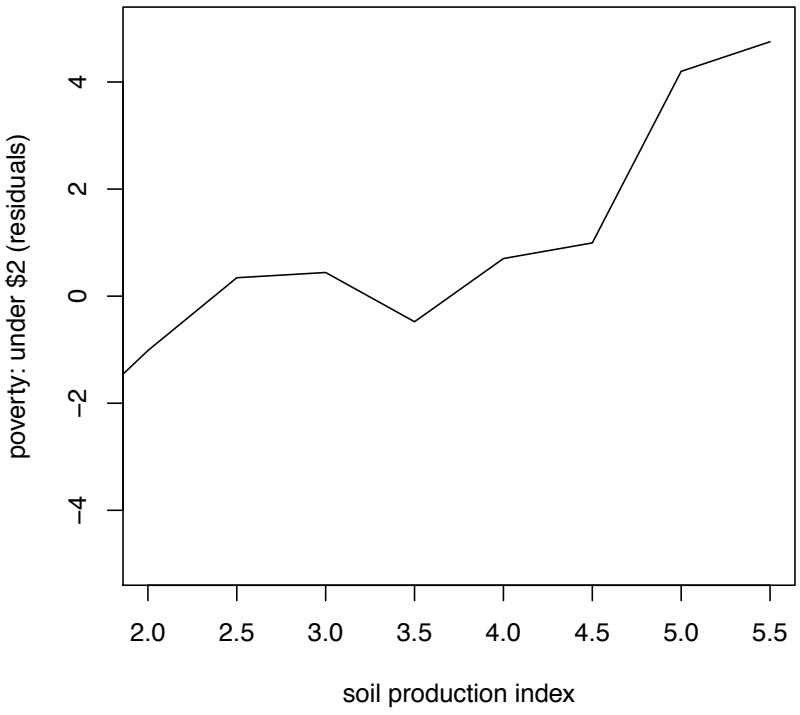

Figure 1: Average poverty by category of soil productivity. The poverty figures are residuals from a regression on country fixed effects.

It is interesting to note that, working on smaller datasets at different levels of aggregation and on smaller samples of countries, the literature in agricultural economics has not been able to establish a clear correlation between soil quality and development outcomes such as poverty alleviation. If anything, the literature has found that the two tend not to be correlated in clear and

statistically detectable ways, and when systematic relationships are detectable, they are at times counterintuitive. For instance, Yamano and Kijima (2010a), studying a sample of households in rural Uganda, find that soil quality (measured at the the household level) is associated with higher crop income but *lower* non-crop income; Yamano and Kijima (2010b) find that soil quality has a positive association with income in Kenya, but not in Uganda and Ethiopia. In a study of rural Kenya, Okwi et al. (2007) estimate that improving soil fertility (from poor to good soil) would "reduce poverty by up to one percentage point."

Table 1: Basic models for poverty, soil quality, and transportation infrastructure

DV: Poverty rate	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	78.97**	79.98**	79.37**	74.24**	59.6**	41.1**
	(0.01)	(0)	(0)	(5.55)	(6.27)	(5.35)
Soil quality index	4.02**			4.13**		
	(0.29)			(0.29)		
Soil production index		2			3.11	
		(1.54)			(2.1)	
Goodsoil			2.22*			2.29**
			(1.01)			(0.8)
Urbanization				-1.15**	-1.11**	-1.07**
				(0.11)	(0.11)	(0.11)
Cultivation				0.19**	0.25**	0.27**
				(0.02)	(0.02)	(0.02)
Elevation				1.82**	1.64**	1.52**
				(0.44)	(0.49)	(0.48)
Elevation^2				-0.07**	-0.06*	-0.05$^+$
				(0.02)	(0.02)	(0.02)
Median slope				-1	15.02**	33.63**
				(5.17)	(5.47)	(5.29)
Median slope^2				-0.23	-4.1**	-8.42**
				(1.12)	(1.2)	(1.23)

Basic models with poverty rate as dependent variable. Standard errors in parentheses. $*$: statistically significant at the 5% level.$**$: statistically significant at the 1% level.

Using a much larger dataset, we are able to detect this counterintuitive positive correlation. The result, as we show in the following subsection, survives the inclusion of controls for many of the possible determinants of poverty. It is also worth bearing in mind that the model includes fixed effects at the country level, so that all the features of the country that do not vary across districts are accounted for. Hence, this correlation (or lack thereof) is not driven by the fact that poor

countries have better soil than rich countries. The only variation used to estimate the coefficients comes from variation in soil quality and in poverty across districts within each country.

The second set of models in Table 1 adjusts for some basic characteristics of the district: urbanization, cultivation (average percentage of the district that is cultivated), and quadratic terms for elevation and for slope class. Poverty tends to be lower in more urbanized districts, and higher in more cultivated districts [6]; in addition, poverty increases (at a decreasing rate) with elevation and (in Models 6 and 7) with steepness. Most importantly, the magnitude (and statistical significance) of the coefficients on the soil quality measures are left unchanged when we include these covariates.

The coefficient on soil quality should not, obviously, be interpreted as an estimate of a causal effect. All we establish with these models is that, on average, districts with better soil tend to be poorer, statistically significantly so in most cases. This is a snapshot of the existing situation, which as we will show depends, in turn, on several observable factors. In the remainder, we explore this counterintuitive correlation.

3.1 The role of transportation infrastructure and access to markets

The literature on rural infrastructure has shown the importance of roads for development (See Ayogu (2007) for a review). This literature has not, however, explored the possible complementarities between agricultural factor endowments and infrastructure provision. Measures of transportation infrastructure or market access are included in models of rural poverty in additive fashion, overlooking the fact that the role of soil quality for development and poverty reduction depends itself on the availability of infrastructure and the accessibility of markets. For instance, Radeny and Bulte (2012) find that distance from the nearest market and the nearest town are negatively associated with per capita income in Kenya; similarly, Okwi et al.(2007) include measures of soil quality and of access to markets (distance from towns). Yet, these papers overlook the possible complementarity between soil quality and market access.

Our empirical models take seriously the complementarity between factor endowment (and soil quality in particular) and market access. We estimate econometrically the variation in the

[6] That urban districts are less poor is not surprising but an explanation of the positive correlation is left for future research.

soil/poverty association that is driven by the availability of transportation infrastructure. Here and in the next section, we establish the following basic results: (1) the effect of soil on poverty depends on transportation infrastructure, and (2) bad infrastructure might turn high soil quality into a curse.

In the models reported in Table 2, we first include soil quality and the measure of transportation cost in isolation; then, we also include their interaction. The variable "road cost" measures the transportation cost (averaged over the district), with higher values reflecting worse roads. Positive values of the coefficient mean that worse roads are associated with more poverty. Transportation cost is centered to have mean zero (and scaled so it has standard deviation one half), hence the main effect of soil quality captures the effect of soil quality on poverty in a district with average road quality, and symmetrically the main effect of road quality captures the effect of road quality on poverty for a district with average soil quality. The interaction term captures how the association between soil and poverty varies across different levels of road quality. In these models, we do not include any controls (other than the country fixed effects). The models with a fuller set of control variables are reported in the next subsection, where we probe the robustness of the results to various specifications. In Subsection 3.3, we also model non-parametrically the interaction between infrastructure provision and soil quality.

In line with the conventional wisdom and results in the literature, lack of transportation infrastructure is systematically associated with poverty. The effect of infrastructure on poverty outcomes has large economic significance too. According to the estimates of Model 1, if one compares two districts in a given country, with the same soil quality, and respectively one standard deviation below and one above the mean of road quality, they are expected to differ by 13 percentage points in poverty.

The variables that capture soil quality are coded so that higher values mean better land. A positive coefficient on the soil quality measure means that better soil is associated with more poverty. In a district that is average in terms of road cost, higher soil quality (as measured by the Soil production index) is associated with higher poverty.

The interaction between the measure of soil quality and the measure of infrastructure captures how the association between soil quality and poverty varies depending on the quality of transportation infrastructure. This interaction between infrastructure and soil quality is a very strong

predictor of poverty.

Table 2: Models with interaction between soil quality and transportation infrastructure

DV: Poverty rate	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	75.82**	77.26**	77.33**	75.95**	77.12**	77.38**
	(0.02)	(0.01)	(0.02)	(0.09)	(0.02)	(0.02)
Soil quality index	4.45**			4.17**		
	(0.28)			(0.37)		
Road cost	12.81**	11.33**	10.27**	12.53**	11.54**	10.28**
	(2.13)	(2.26)	(2.29)	(2.13)	(2.19)	(2.3)
Soil production index		1.8			1.59	
		(1.57)			(1.39)	
Goodsoil			2.09*			2.12*
			(0.9)			(0.87)
Soil quality index by road cost				1.19+		
				(0.68)		
Soil production index by road cost					5.09*	
					(2.48)	
Goodsoil by Road cost						1.98
						(2.22)

Models with poverty rate as dependent variable and interaction between measures of soil quality and measures of infrastructure. Standard errors in parentheses. +: statistically significant at the 10% level.∗: statistically significant at the 5% level. ∗∗: statistically significant at the 1% level.

The coefficient on the interaction is positive in all models, and statistically significant (with the exception of the model that uses the "goodsoil" measure). This points to the following implications: on the one hand, infrastructure is most beneficial when the quality of the soil is higher. This is far from counterintuitive. In other words, providing infrastructure in places with low soil quality should have a negligible effect on poverty (unless the infrastructure can be used to exploit some other type of natural resource or lead to industrialization). From another perspective, good soil helps reduce poverty only if there is infrastructure of sufficiently high quality. The evidence points to the fact that in a place with average infrastructure, soil quality does not have a poverty-reducing role. The coefficient on the main effect of soil quality (that captures the effect of soil quality in a district with average quality infrastructure) is positive (and statistically significant in Models 4 and 6). According to the estimates of Model 4, for instance, only if the road cost is more than 1.5 standard deviations below its mean, soil quality starts having a poverty-reducing effect.

In addition, the results for the interaction might potentially provide evidence (whose robustness we probe below) that good soil can be a *curse* in the absence of infrastructure.

Again, these are not causal regressions (or estimates of structural parameters, for that matter) but in terms of conditional expectations (and therefore, as descriptive summaries) they point to the fact that infrastructure in most of Africa is insufficient for the available resources, in terms of soil, to be used to significantly reduce rural poverty.

3.2 ROBUSTNESS CHECKS

There are several variables that have been shown in the literature to be associated with rural poverty and which could drive the relationship we detect. While we deal directly with potential endogeneity issues in Sections 4 and 5, here we show that the relationship we detect is robust to the inclusion of several "obvious" controls. These can be thought as belonging to one of three categories: geographic factors (e.g., elevation, terrain slope, and distance from the coast); demographic and economic factors (e.g., population density, proportion of land allocated to cultivation, distance from large cities); and finally, long-term historical legacy factors (for instance, location in the context of colonial political borders).

Accounting for other variables The models reported in Table 3 include measures of soil quality, measures of infrastructure, and controls for geographic characteristics of the district. In particular, we include average altitude of the district, median terrain slope of the district, and their squares (to capture potential non-monotonicities in the associations), as well as measures about rivers: the dummy for river districts, the rank class of the largest river in the district, and the average rank class of rivers in the district. These controls turn out to be closely associated with poverty: in particular, river districts seem to be systematically poorer than districts without rivers. The models also estimate that both elevation and terrain slope have a non linear effect on poverty, with poverty increasing faster when moving from sea-level flatland to somewhat elevated and sloping terrain, and then tapering off at very high altitude. In fact, the highest poverty rate is found at around 1200 meters, where poverty starts decreasing (as areas become, plausibly, less populated) and similarly, in areas with median slope between class one and two.

In any case, the strong association between soil quality, infrastructure, and rural poverty that

we detect and discuss in the previous Subsection, is robust to the inclusion of these controls. The next set of controls we include have to do with demographic (broadly meant) and economic characteristics of the district.

Excluding urban districts from the analysis In all of the analysis above, we include all the districts in each country, but we account for level of urbanization in the regressions. Notice that districts are often large enough to contain both urban and rural areas, and excluding all the districts that contain also some urban areas would not be appropriate. We now show that the basic results are unaffected if we drop from the analysis districts that can be classified as "urban" according to some criteria. The first criterion is that the district lies in the top 2.5 percent of most urbanized[7]; the second, that it lies in the bottom 2.5 percent of distance from the capital; the third is that every cell in the district is classified as urban according to the FAO data on rural population density. This leads to the exclusion of around 1100 districts from the analysis. Notice that these exclusion criteria are quite stringent: the median excluded district has around 15 percent of its surface classified as cultivated, and is only 6 percent urban overall. Importantly, the patterns we detect cannot be considered to be driven by comparisons we are carrying out between rich and urban areas (for which soil quality does not matter) and poor and rural areas. The results are reported in the first two columns of Table 3.

Alternative measures of poverty In the section devoted to the dataset, we justify why we rely mostly on the satellite-based estimate of poverty in our models. In spite of the limitations of the alternative measures we collected, we now probe the robustness of our results if these alternative measures of poverty and underdevelopment are used as dependent variables. We also create an index as a simple average of the (rescaled) values of the poverty measures we have, and use it as an alternative measure of poverty. The results using these alternative measures of poverty are reported in columns 3–5 in Table 3.

[7]Specifically, we consider as urban a district that contains at least one cell with urbanization rate higher than 41 percent.

Table 3: Robustness checks

DV: Poverty rate	(1)	(2)	(3)	(4)	(5)
Intercept	79.04**	80.08**	57.62**	31.33**	-2.04**
	(0.24)	(3.39)	(0.01)	(0.01)	(0.02)
Soil quality index	4.09**	3.01**			
	(0.5)	(0.36)			
Road cost	3.45	5.28**	3.47**	1.79**	12.43**
	(2.32)	(1.61)	(0.84)	(0.39)	(2.17)
Soil quality index by road cost	4.12**	4.53**			
	(0.97)	(0.69)			
Distance coast		0.87**			
		(0.11)			
River district		0.88			
		(0.76)			
Elevation		1.87**			
		(0.37)			
Elevation^2		-0.08**			
		(0.02)			
Median slope		4.39			
		(3.17)			
Median slope^2		-1.41*			
		(0.68)			
Distance town		-1.26**			
		(0.25)			
distance.border		-0.87			
		(1.52)			
distance.capital		-0.24			
		(0.17)			
Soil production index			1.36	-0.04	2.16
			(1.06)	(0.4)	(1.88)
Soil production index by Road cost			1.86	0.78	5.98$^+$
			(1.47)	(0.67)	(3.15)

In models 1 and 2, more urbanized districts are excluded from the analysis. In models 3-5, the alternative measures of poverty described in subsection 2.1 and an index that combine the three available measures are used as dependent variables. Standard errors in parentheses. $+$: statistically significant at the 10% level. $*$: statistically significant at the 5% level. $**$: statistically significant at the 1% level.

3.3 CAN GOOD SOIL REALLY BE A CURSE?

The estimates reported above show that infrastructure has a stronger association with reduced poverty in areas with good soil, and, symmetrically, soil quality has a stronger association with reduced poverty in areas with good infrastructure. At the same time, the result could also be interpreted as saying that when the quality of infrastructure is sufficiently bad, soil quality is associated with *increased* poverty. This would point to the possible existence of a "curse" of good soil: if we were to compare two districts, with equally insufficient infrastructure, the one with the better soil would be expected to have a higher poverty rate than the one with the worse soil.

The linear multiplicative interaction is, by construction, symmetric; hence it cannot tell apart a proper "curse" from the simpler claim that the return to infrastructure, in terms of poverty alleviation, is higher in districts with higher soil quality. In addition, the multiplicative model assumes linearity over the entire support of soil quality. In other words, the result we present above for the interaction between soil and infrastructure might reflect exclusively the increase in returns to soil quality when infrastructure is better (and symmetrically, the increase in returns to infrastructure connecting to locations with good soil), rather than a potential curse that affects locations that are poorly served in terms of infrastructure but "sit" on good soil. In sum, the interaction term might just capture the complementarity between soil endowments and transportation infrastructure availability. Hence, the claim that good soil is associated with increased poverty when infrastructure is insufficient and requires further evidence.

We address this issue by turning the soil quality measure and the infrastructure availability measure into categorical variables and then estimating a model fully saturated in these categorical variables. In practice, to estimate to what extent an actual curse is at work, we create a full set of dummies for all the possible combinations of soil quality (turned into a six-category variable by rounding to the nearest integer the value for the district) and road cost (turned into a three-category variable, grouping in turn the bottom two levels and the middle two levels of road cost). There is a total of 18 possible combinations in which both road cost and soil quality are observed, plus those cases in which one of the two variables is not observed. We exclude from the analysis the districts for which one of the two (or both) measures is not observed.

We then estimate models of the form

$$y_i = \alpha_{j(i)} + \beta X_i + \sum_r \sum_s \gamma_{r,s} 1(r_i = r \& s_i = s) + \epsilon_i \quad (1)$$

where the α terms are just country fixed effects and the $\gamma_{r,s}$ is the intercept for districts that belong to the r road cost category and the s soil quality category. In other words, we include one intercept for each combination of soil and road cost. We estimate two variants of the model. In one, we simply include the dummies for the combinations of soil and infrastructure. In econometric terms, these are fixed effects for the groups defined by a given soil and infrastructure combination. In the second variant, we model the γ as drawn from a normal distribution with variance estimated from the data: the γ coefficients are random effects for given soil-infrastructure combinations. The advantage of this approach, as opposed to estimating them as fixed effects (or category dummies) is that, whenever (or if) they are estimated imprecisely for a given category (for instance, because there are few districts that belong to the category) they are shrunk towards zero, and the expected level of poverty is shrunk towards the grand mean for the country (see Gelman et al. 2008; Ghitza and Gelman 2013).

The plots in Figure 2 display the estimates from these models. For each category of road cost and soil quality, we display the expected level of poverty based on the estimates of the saturated model. Each line is for a given level of road cost (with darker lines indicating higher road cost –hence worse infrastructure). On the horizontal axis is the Soil quality index, and on the vertical axis the poverty level (re-centered). The plot on the top left is for a model that only includes fixed effects for soil-road combination and country fixed effects, while the plot on the top right is for a model that also includes controls: urbanization, the cultivation measure, the dummy for the presence of a river in the district, distance from the capital, distance from towns of at least 20,000 inhabitants, distance from the coast, distance from the border, and quadratic polynomials for elevation and terrain slope.

From the inspection of the two plots, one can infer that while poverty increases (all else equal) with road cost, the association between soil quality and poverty is far from straightforward. In the top two categories of road cost (the most remote districts with the worst transportation infrastructure) poverty turns out to be higher in districts with relatively high soil quality. While

poverty is lower on average in the districts with the very best soil (category 6), categories 3 to 5 of soil quality have on average higher poverty than districts with soil in the lowest categories. The evidence suggests that poverty is at its worst when we have a combination of good soil and very poor infrastructure.

The bottom two plots in Figure 2 probe this relationship further, by displaying the results from estimation that models soil/road combination categories as random effects. The left plot only includes the random effects for soil and road combination and the country fixed effects, while the right plot includes the controls listed above.

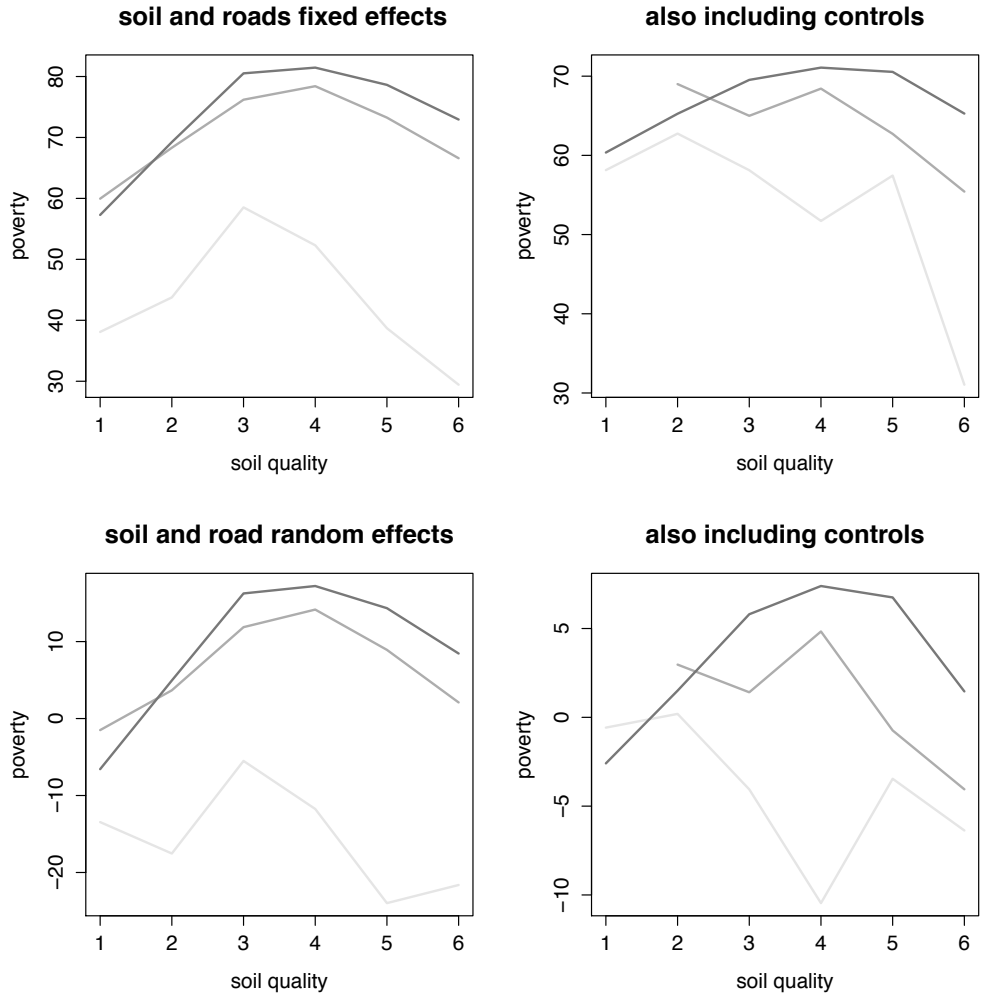

Figure 2: Saturated models for poverty. Darker lines represent districts with worse transportation infrastructure.

Table 4 reports in tabular form the same information. For each combination of soil quality

and road cost, we report the expected value of poverty according to the model estimates. The coefficients on the controls are, unsurprisingly, not different in any substantial way from those in the models reported in the main tables, hence we omit them to save space. Each row reports all the estimates and, for the fixed effects estimations, also the standard error (in the columns labeled "se"). So for instance, the first row in the table reports the estimates for road category one (districts with very low road cost) and soil category one (districts with very bad soil).

Table 4: Non-parametric interactive models for poverty rate

road cost	soil	fe	se	fe+controls	se	re	re+controls
1	1	38.09	13.96	58.14	17.01	-13.45	-0.58
1	2	43.75	6.05	62.75	12.17	-17.53	0.19
1	3	58.53	3.06	58.12	4.65	-5.51	-4.06
1	4	52.31	2.30	51.72	4.19	-11.75	-10.46
1	5	38.70	3.97	57.45	5.60	-23.98	-3.46
1	6	29.43	11.26	31.04	12.08	-21.63	-6.37
2	1	59.96	19.52			-1.50	
2	2	68.31	5.26	69	6.43	3.67	2.97
2	3	76.19	2.10	64.99	3.93	11.87	1.41
2	4	78.41	1.80	68.43	3.86	14.16	4.83
2	5	73.26	2.52	62.72	4.18	8.92	-0.74
2	6	66.59	5.83	55.43	6.83	2.09	-4.05
3	1	57.29	3.91	60.36	4.66	-6.58	-2.59
3	2	69.24	2.36	65.25	3.92	4.95	1.50
3	3	80.52	1.68	69.53	3.75	16.25	5.80
3	4	81.46	1.56	71.08	3.79	17.22	7.39
3	5	78.63	1.82	70.54	3.89	14.34	6.75
3	6	72.94	3.02	65.27	4.42	8.45	1.46

Estimates from the saturated models, with district-level poverty rate as dependent variable. In the columns labeled "fe" the estimates come from the fixed-effects estimates, in the columns labeled "re" from random effects. The columns labeled "se" report the standard errors from the fixed-effects estimation.

3.4 UNDERSTANDING THE DISTRIBUTION/ALLOCATION OF INFRASTRUCTURE

In order to understand the phenomenon we highlight, we must examine which systematic patterns can be detected in the allocation of infrastructure. For this purpose, we regress the measures of road quality on geographic characteristics of the districts. In the first three columns of Table 5, we report models that regress infrastructure on quadratic terms for elevation and terrain slope,

on the dummy for river districts, and on the measures of distance from rivers, distance from the coast, and distance from current borders. The measure of infrastructure we use is such that higher values reflect higher transportation costs, hence worse transportation infrastructure. Positive coefficients therefore mean that a given variable is associated with worse roads. First of all, river districts, and districts closer to rivers, tend to have significantly worse infrastructure. In addition, infrastructure is worse in districts farther from the coastline, in districts closer to borders, and in districts farther from the current capital. Finally, elevation and slope have non linear effects. In the case of elevation, both terms are positive, implying that districts at sea level have better infrastructure than districts at higher altitude.

Table 5: The mismatch between good soil and good infrastructure

DV:	Road cost (1)	Road cost (2)	Road cost (3)	Soil prod. (4)	Soil prod.(5)	Soil prod. (6)
Intercept	4.73**	4.73**	4.79**	-8.6**	2.29**	2.35**
	(0.11)	(0.1)	(0.11)	(0.99)	(0.36)	(0.35)
River district	0.12**	0.12**		0.13	0.08	
	(0.03)	(0.03)		(0.09)	(0.05)	
Elevation	0.01	0.01	0.02	-0.04	-0.04*	-0.03*
	(0.02)	(0.02)	(0.02)	(0.04)	(0.02)	(0.01)
Elevation^2	0	0	0	0.01*	0+	0+
	(0)	(0)	(0)	(0)	(0)	(0)
Median slope	-0.14	-0.13	-0.13	8.75**	1.38**	1.37**
	(0.11)	(0.1)	(0.1)	(0.92)	(0.34)	(0.34)
Median slope^2	0.03	0.03	0.04+	-2.05**	-0.25**	-0.24**
	(0.02)	(0.02)	(0.02)	(0.21)	(0.07)	(0.07)
Distance coast	0.07*	0.07*	0.05	0.12	0.08+	0.07
	(0.03)	(0.03)	(0.04)	(0.08)	(0.05)	(0.05)
Distance capital	0.42**	0.41**	0.44**	-0.52**	-0.15**	-0.12**
	(0.05)	(0.05)	(0.05)	(0.13)	(0.05)	(0.04)
Distance border		-0.04+	-0.07**			-0.01
		(0.02)	(0.02)			(0.03)
Distance river			-0.1**			-0.09
			(0.03)			(0.07)
Distance 1950 border			0.04**			
			(0.01)			

In columns 1-3 the dependent variable is road cost; in columns 4-6 the dependent variable is soil quality (the soil production index). Standard errors in parentheses. +: statistically significant at the 10% level. ∗: statistically significant at the 5% level. ∗∗: statistically significant at the 1% level.

3.5 WHERE IS THE GOOD LAND?

Here we describe in general terms, some patterns of association between soil quality and observable geographic characteristics of the districts. We regress the measures of soil quality on the

quadratic terms for elevation and slope, the dummy for river district, another measure of presence of rivers (the class rank of the largest river in the district), and distance from the coast. The best land is located in river districts (even if the evidence is not strong enough to be statistically significant at conventional levels), and in districts farther from the coast (also in this case not statistically significant). In addition, soil quality is not correlated with distance from the border, while it is worse farther from the capital. Again, we estimate quadratic polynomial terms in elevation and terrain slope. According to the estimates, soil quality is maximum (all else equal) at an altitude of 200 meters. In addition, the coefficients on terrain slope also point to the fact that soil quality is highest in "hilly" areas: soil quality is at a maximum when the median slope of the district is in class three. We explore these non-linear effects in Figure 3.

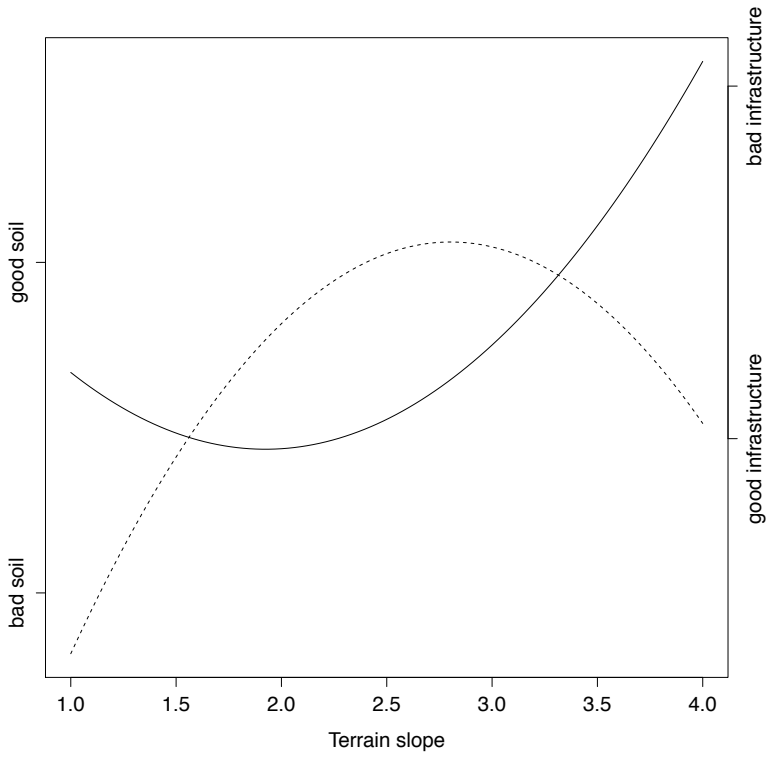

Figure 3: Expected values of soil quality (dashed) and road cost (solid) as a function of median terrain slope in the district, from the models respectively columns 2 and 5 of Table 5.

The disconnect between good land and infrastructure The plot in Figure 3 displays the expected value of soil quality (the dashed line) and of road cost (the solid line) as a function of the median

terrain slope of the district. Higher values of road cost reflect lower transportation infrastructure provision. The best soil is estimated to be located, all else equal, in hilly districts, while the worst is in flat areas and, to a lesser extent, in mountain areas. The worst infrastructure is found in mountainous areas with very steep terrain (i.e., right side of the plot). There is some mismatch between soil quality and infrastructure provision, on average: the best land from the point of view of soil quality is not the tract that attracts the best transportation infrastructure.

Our results implicitly suggest that rugged terrain is associated with poverty in Africa through its effect on infrastructure provision. This seems to be at odds with Nunn and Puga (2012) which indicates that ruggedness has a positive effect on economic development in Africa since more rugged African countries have experienced less slave exports. However, we should note that their measure of ruggedness captures "small-scale terrain irregularities, such as caverns, caves, and cliff walls, that afforded protection to those being raided during the slave trades" (p. 21), while our geographic controls (average elevation and average slope in the district) capture more "macro" features of a given territory. In addition, once quality of institutions (measured by an index of rule of law) is included in their model, Nunn and Puga (2012) find that their measure ruggedness has no association with level of development. Our models include country fixed effects, that capture all the institutional features of the country that do not vary across districts: our results for geographic characteristics then have to be interpreted net of (country-specific) indirect effects through quality of institutions.

4 ENDOGENEITY OF SOIL QUALITY

While the measures of soil quality we use are related to stable features of the soil, a significant amount of literature has focused on the relationship between soil degradation and poverty, and (of particular interest for our argument), on the potential spillovers from poverty to soil quality. Such spillovers might take the form of poorer areas depleting the resources of the land, or failing to use fertilizers and to engage in other practices that enhance soil fertility. (Drechsel et al., 2001)

The basic correlation we detect does not seem to support this type of mechanism directly: in fact, in our data, poverty is associated with *higher* soil quality; if poverty-related soil degradation were the main driver of soil quality in our data, we would expect soil quality and poverty to be

negatively correlated. Nonetheless, we deem it worth investigating the robustness of our result.

In order to address this potential concern, we rely on the set of dummies described in Section 2.2 and based on a purely geological classifications of soil, using them as instruments for soil quality. The soil classes have jointly quite strong predictive power for soil quality: around 11 percent of the variation in soil quality (as measured by the Soil production index) is explained by the soil class dummies alone. When country fixed effects are included along with the soil class dummies, around 23 percent of the variation in soil quality is explained by the model.

The first model in Table 6 reports the 2SLS estimate of the effect of soil quality on poverty in isolation, the second model reports the estimates for soil quality interacted with road quality (using the interactions between road quality and geological classes as instruments for the interaction) and the third model reports the estimates of a specification including all the control variables discussed in the previous section. The unconditional association between soil quality and poverty is again positive, and the interactive model confirms the findings presented in the previous section: soil quality is a blessing in districts with high-quality infrastructure, but might be a curse in districts in which transportation infrastructure is underprovided.

5 Endogeneity of infrastructure

One might be more concerned that the associations we report are driven by possible endogeneity of the amount of infrastructure to rural poverty. In other words, current infrastructure provision might be driven by current rural poverty, or by other (unobserved) contemporary features that also affect poverty levels. For this reason, we want to show that the systematic relationship between infrastructure provision and rural poverty that we document in the previous sections survives if a temporally pre-determined instrument, colonial transportation infrastructure, is used to correct for possible endogeneity of current transportation infrastructure to rural poverty. It is worth noting that endogeneity of infrastructure would not explain why we find that better soil in the absence of infrastructure is associated with more poverty. Also, as documented in the previous section, using purely geological features of the soil as an instrument for soil quality leaves our main results unaffected. In any case, in order to establish with more confidence the role played by infrastructure, we use the presence of transportation infrastructure in the colonial period as an

Table 6: IV models, with soil quality instrumented by geological classification

DV: Poverty rate	(1)	(2)	(3)	(4)	(5)
Intercept	75.36**	75.55**	65.68**	58.01**	65.68**
	(0.02)	(0.22)	(2.33)	(0.02)	(2.74)
Soil quality index	27.8**	25.63**	25.83**		25.83**
	(2.58)	(4.04)	(4.69)		(4.06)
Road cost	13.68**	13.31**	7.02**	2.62*	7.02**
	(2.2)	(2.45)	(1.53)	(1.05)	(1.65)
Soil quality index by road cost		4.8	17.21*		17.21*
		(7.74)	(8)		(7.48)
Rural pop. density			5.75**		5.75**
			(0.7)		(1.28)
Soil production index				3.22**	
				(1.23)	
Soil production index:road cost				8.29+	
				(4.37)	
Urbanization					-0.98**
					(0.07)
Cultivation					0.08
					(0.07)

Instrumental variable models, with district-level poverty rate as dependent variable, and geological dummies as instruments for soil quality. Standard errors in parentheses. +: statistically significant at the 10% level. ∗: statistically significant at the 5% level. ∗∗: statistically significant at the 1% level.

instrument for current infrastructure provision.

5.1 First-stage relationship

In Table 7 we report the first-stage relationship between contemporary roads and colonial roads.[8] If we regress the (standardized) measure of road cost on dummies for countries and a dummy for whether a district had a primary road in colonial times, the estimate is negative (implying that transportation costs are lower today in districts that had a primary road in the colonial era) and highly statistically significant (even after correcting for clustering by country of the errors). From the substantive point of view, having a primary road in colonial times leads to a reduction of one-tenth of a standard deviation in contemporary road costs. The results are analogous if instead

[8]These are not the actual first stages, because the ones actually used in the IV models also include the control variables included in the second stage. These pseudo-first-stages are reported here to show how colonial infrastructure induces variation in our contemporary infrastructure provision measures.

of the dummies for presence of colonial roads, we use (log) distance from a primary colonial road: higher distance is statistically significantly associated with higher road costs. The third model includes both the dummy for primary colonial road in the district and (log) distance from a primary colonial road. In this case, only the coefficient on the distance measure is statistically significant. The result is robust to the inclusion of a dummy for presence of secondary colonial roads and (log) distance from a secondary colonial road. The estimates for this model are reported in the fourth column of Table 7. Again the (log) distance from a primary colonial road has a highly statistically significant positive effect on current road cost. In addition, this relationship is not driven by observable third variables that affect both colonial and contemporary infrastructure. The model in columns 5 and 6 of Table 7 show how the relationship survives the inclusion of many geographic characteristics of the district. The coefficient on the dummy for colonial primary road is -0.04 (clustered-robust standard error 0.02), and the coefficient on (log) distance from a colonial road is .06 (with clustered standard error .01) after controlling for the following: quadratic polynomials for elevation and slope, ruggedness (standard deviation of slope within district), distance from the coast, presence of a river, and area of the district. Similarly, the coefficient on the (log) distance of the district from a colonial primary road is still positive, and highly statistically significant.

The estimates reported in column 6 show that the results do not depend on the specific measure of contemporary infrastructure we use: in this case, the response variable is not the road cost index, but the (log) distance of the center of the district from a contemporary major road, based on the FAO Major Roads map. Finally, the model in the last column of the table reports the estimates of a regression of (standardized) road cost on the two colonial infrastructure measures, without country fixed effects. Again, current transportation costs are predicted by the distance from a primary colonial road.

The instruments based on colonial roads are far from weak. In the model with no country fixed effects, but *only* the colonial roads measures, the R^2 is 0.13, and the F-statistic for the whole regression is 356.8. In the model of column 5, with geographic controls and country fixed effects, the R^2 is 0.37, and the F-statistic is 39.7. A considerable portion (13 percent) of the variation in current road costs is explained by colonial patterns alone, and more than a third of it can be explained by country effects, colonial patterns, and stable geographic characteristics of the district.

Table 7: First stage estimates for road cost instrumented by colonial roads

DV: Road cost	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	0.41**	0.52**	0.51**	0.51**	0.41**	-0.83*	0.22**
	(0.01)	(0.02)	(0.02)	(0.02)	(0.08)	(0.37)	(0.05)
Colonial road dummy	-0.21**		0.02	0.02	-0.04*	-0.01	0.09
	(0.02)		(0.03)	(0.03)	(0.02)	(0.06)	(0.06)
Distance primary colonial		0.12**	0.12**	0.12**	0.06**	0.2**	0.14**
		(0.01)	(0.02)	(0.02)	(0.01)	(0.03)	(0.02)
Distance secondary colonial				0.01			
				(0.01)			
Secondary colonial road dummy				0			
				(0.04)			
Distance coast					0.01**	-0.03**	
					(0)	(0.01)	
River district					0.04**	0.09	
					(0.01)	(0.07)	
Elevation					0.01	0.16**	
					(0.01)	(0.03)	
Elevation^2					0$^+$	-0.01**	
					(0)	(0)	
Median slope					-0.03	-0.22	
					(0.07)	(0.31)	
Median slope^2					0.01	0.06	
					(0.01)	(0.07)	
Standard deviation slope					0.03	0.03	
					(0.02)	(0.09)	
Area					0.02**	0	
					(0)	(0.01)	

First stage estimates, with road cost instrumented by the colonial roads measures. Standard errors in parentheses. $+$: statistically significant at the 10% level. $*$: statistically significant at the 5% level. $**$: statistically significant at the 1% level.

5.2 Plausibility of the "as good as randomly assigned" assumption: further evidence

In the previous subsection, we showed how contemporary infrastructure is affected by patterns in transportation infrastructure provision in the colonial era. We now want to demonstrate that one main driver of the decisions made in the colonial era is the location of natural resources that can be exploited by the extractive/mining industry. For this purpose, we rely on the data assembled by the U.S. Geological Survey. This reports the location of several mineral-related activities (like mines, quarries, refineries, smelters). The data refers to the year 2003. We only focus on mines, quarries, and wells. The data refer to the current knowledge about location of minerals, and the current location of extractive sites, while ideally we'd like to use information about a) knowledge *in the colonial period* regarding the location of minerals, and b) actual extractive activities in the

colonial era. Unfortunately, information of that kind is not available. Historical data are not readily available but the presence of mines and quarries in the colonial era should be highly correlated with their presence in 2003, the year to which our map refers. The use of this imperfect source adds some measurement error but, if anything, this should lead to attenuation bias (as long as the measurement error does not depend itself on colonial roads). We create some variables based on the geographic information about the location of extractive activities. In particular, we compute two summaries. The first is the distance between the geometric center of the district, and the closest mining location, regardless of whether the mine lies within or outside the district. The second is just a dummy that takes the value of one if there is a mine in the district. [9]

We then regress the various summaries of colonial infrastructure availability on the summaries that reflect the presence of valuable minerals. The results of these estimations are reported in Table 8. Consistently with our expectation, there is a very high, and statistically significant, correlation between the location of mines, quarries and wells, and the presence of infrastructure in the colonial era. A logit regression of the dummy for whether a district has a primary road in the colonial era on a dummy for whether there is an extractive site in the district (and accounting for separate intercepts by country) yields a coefficient of 0.89 (standard error 0.17). A rough approximation of the effect on probability scale is that the probability of having a colonial road is around 22 percentage points higher in mining districts than in non-mining districts.[10]

In addition, we want to test whether colonial decisions were affected by the quality of farmland. Indeed, it would pose problems to our IV strategy if one of the main objectives colonial powers had when building transportation infrastructure was connecting high-quality farmland to cities or seaports. In order to assess what are the factors associated with better infrastructure in colonial times, we estimate regressions analogous to those in the previous paragraph, but we also include measures of soil quality. We also estimate simpler regressions with just the soil quality measures, excluding the mining activity variables, to show how the lack of association between colonial infrastructure and soil quality is negligible regardless of the specification. The results are

[9] We cannot differentiate between types of minerals extracted, because there are too many categories, and it is not obvious how to allocate them to a smaller number of categories that would be manageable for econometric estimation.

[10] Similarly, in a regression of (log) distance of the district center from a "primary" colonial road on (log) distance of the district from a mine, with fixed effects by country, the coefficient is approximately .2 (with cluster-robust standard error .06), implying that a one-percent increase in distance from a mine leads on average to a .2 percent increase in distance from primary colonial roads.

Table 8: Location of colonial roads

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Distance mine	0.2**		0.15*		0.17*			
	(0.06)		(0.07)		(0.07)			
Mining district dummy		0.89**		0.47*		0.52**		
		(0.17)		(0.18)		(0.18)		
Distance coast			0.05**	-0.06*				
			(0.02)	(0.02)				
Elevation			-0.03*	0.12**				
			(0.02)	(0.03)				
Median slope			-0.42*	1.93**				
			(0.19)	(0.34)				
Elevation^2			0	0**				
			(0)	(0)				
Median slope^2			0.11**	-0.43**				
			(0.04)	(0.07)				
Soil production index					0.04	-0.05	0.04	-0.05
					(0.1)	(0.07)	(0.09)	(0.07)

Models for the location of the colonial roads. The models in odd-numbered columns are linear models with log distance from a first-class colonial road as dependent variable; the models in even-numbered columns are logit models with the presence of a first-class colonial road in the district as dependent variable. Standard errors in parentheses. $+$: statistically significant at the 10% level. $*$: statistically significant at the 5% level. $**$: statistically significant at the 1% level.

reported in Table 8.

When we include the Soil production index alongside the measures based on extractive activities, the estimate of the effect of soil quality is nowhere near statistical significance. In the model with (log) distance as the dependent variable, the point estimate is positive. The magnitude is quite small, implying that a move from one standard deviation below to one standard deviation above the mean of the Soil production index would lead to an expected increase by less than 0.05 of a percentage point in distance from a colonial road. In the logistic model, the point estimate of the coefficient on soil productivity is negative, implying an approximate decrease in the probability that the district has a colonial road by one percentage point following a move from one standard deviation below to one above the mean of soil quality. The coefficient on the dummy for an extractive site is still statistically significant, even though the magnitude of the point estimate is slightly smaller, 0.52, with standard error 0.17, implying a rough estimate of a 13 percentage point increase in the probability of a district having a primary colonial road in mining districts. These results are not driven by the inclusion of both the mine measure and the soil measure: when included in isolation, soil productivity has no predictive power for distance from a primary colonial

road or for the probability of having a primary colonial road. The coefficients are unchanged from the models that include both extractive activities and soil quality.

The evidence points to the fact that colonial decisions regarding where to build roads were not driven by the desire to connect high-quality farmland but, rather, to reach mining and quarrying areas. In addition, regressions of soil productivity on distance from mines (not reported) show that there is no relationship between the presence of mines and soil quality. This evidence is far from surprising, but reassuring for our identification strategy, and it justifies our decision to use colonial roads as an instrument. Roads today depend on roads in the colonial period, but the factors that influenced the colonial powers' decisions regarding where to build roads are not those we consider (and show to be) important to understand contemporary patterns of rural poverty.

In any case, we should not lose sight of the fact that what we need is a source of variation in infrastructure that is unaffected by current levels of rural poverty and by unobserved current factors that also affect rural poverty. For this purpose, it is sufficient that colonial infrastructure is "randomized" by the presence of valuable minerals. There might be alternative paths through which the presence of colonial roads or the presence of valuable minerals might affect contemporary levels of rural poverty. Subsection A.1 deals with possible violations of the exclusion restriction.

5.3 IV REGRESSION RESULTS

The estimates of the 2SLS regressions instrumenting for current infrastructure with colonial-era infrastructure are reported in Tables 9 (with default 2SLS standard errors) and 10 (with standard errors clustered by country). In all the models, the excluded instruments are the (log) distance of the center of the district from the closest primary colonial road, and the dummy for districts that had a primary colonial road.

The model in the first column of Table 9 is the IV counterpart of Model 1 in Table 2: it includes only country fixed effects, a measure of soil quality (the one based on the first principal component of seven soil constraint measures), and the measure of road cost. The interaction is not included in this model. Confirming the results of the non-IV model, both road cost and soil quality have a *positive* effect on poverty. The models in the second and third columns are the IV counterparts of Models 4 and 5 in Table 2. Again, better soil (whether measured with the Soil production index

or the first principal component) is associated with higher poverty. In addition, the coefficients on transportation costs and their interaction with soil quality are again positive.

One can infer that high transportation costs – and poor transportation infrastructure – contribute in a substantial way to creating or perpetuating rural poverty; the absence of infrastructure has a much stronger effect in districts with better soil; finally, lack of infrastructure potentially makes a soil rich district pooer. The models in the fourth and fifth columns restrict the analysis to districts that are not located in the proximity of a major urban area: the coefficients are identified only from the comparison of districts that are approximately equally remote. In this model, the main effect of soil quality (as measured by the first principal component of the seven constraints) and its interaction with road cost are positive and statistically significant (regardless of whether one uses default or clustered standard errors). The main effect of road cost is negative (and statistically significant), though.

Model 5 includes a host of control variables related to physical and political geography. The coefficients on these variables are broadly analogous to those estimated in the "vanilla" regressions of Table 2. Also, in this model both soil quality and the interaction between soil quality and road cost are positive (and statistically significant regardless of the estimator used for the standard errors) while the main effect of road cost is not statistically significant.

Model 5 estimates the IV model instrumenting both for soil quality (using geological measures as instruments, as in the previous section) and for infrastructure using the colonial roads instruments. In this model (that includes no additional controls other than the country dummies) the three coefficients are estimated to be positive, and both the interaction and the main effect of road cost are highly statistically significant (when using the default 2SLS standard errors; only the main effect is statistically significant with the clustered standard errors).[11]

[11] In this model, all the excluded instruments – and their interactions – are used to instrument each of the three endogenous variables: soil quality, road cost, and their interaction. It is worth reporting that the R^2 of the first stage for soil quality is unchanged whether one uses only the geological measures or also the colonial infrastructure instruments. This is far from surprising, but confirms that we are using only geology-induced variation to estimate the coefficient on soil quality. Similarly, the inclusion of the geology measures do not increase the fit of the first stage for infrastructure, which confirms that the variation we are using to estimate the effect of road cost comes from variation in the presence of colonial roads only.

Table 9: IV models, with road cost instrumented by location of colonial roads

DV: Poverty rate	(1)	(2)	(3)	(4)	(5)
Intercept	77.59**	77.81**	79.36**	82.96**	79.93**
	(1.59)	(1.59)	(1.71)	(1.56)	(3.62)
Soil quality index	4.3**	3.77**		2.9**	2.9**
	(0.13)	(0.22)		(0.36)	(0.43)
Road cost	5.57**	5.18*	1.52	-11.54**	-1.82
	(2.04)	(2.04)	(2.17)	(2.57)	(3.4)
Soil quality index by Road cost		2.23**		7.93**	4.64**
		(0.75)		(1.33)	(1.5)
Soil production index			1.51*		
			(0.67)		
Soil production index by Road cost			12.87**		
			(2.66)		
Distance coast					0.95**
					(0.15)
River district					1.28*
					(0.59)
Elevation					1.89**
					(0.22)
Elevation^2					-0.08**
					(0.01)
Median slope					4.52
					(2.83)
Median slope^2					-1.42*
					(0.62)
Distance town					-1.13**
					(0.14)
Distance border					-1.01*
					(0.43)
Distance capital					-0.23*
					(0.1)

Instrumental variable estimates, with poverty rate as dependent variable and road cost instrumented by measures of colonial infrastructure. Standard errors in parentheses. $+$: statistically significant at the 10% level. $*$: statistically significant at the 5% level. $**$: statistically significant at the 1% level.

5.4 FURTHER TESTS: CONDITIONING ON MINING SITES; SENSITIVITY TO VIOLATIONS

Above, we show that the location of colonial infrastructure is driven, to a significant extent, by the location of mining and other extractive sites. Now, one could wonder whether this constitutes a challenge to the exclusion restriction (which requires that the location of colonial infrastructure does not affect contemporary rural poverty other than through the persistence of colonial transportation infrastructure on contemporary post-colonial infrastructure). We address this issue in several ways.

First of all, we condition on the location of mines in the 2SLS models for rural poverty. Indeed, if the challenge has to do with the fact that mines affect colonial roads, colonial roads affect

Table 10: IV models, with road cost instrumented by location of colonial roads

DV: Poverty rate	(1)	(2)	(3)	(4)	(5)
Intercept	77.59**	77.81**	79.36**	82.96**	79.93**
	(0.09)	(0.25)	(0.06)	(0.39)	(3.7)
Soil quality index	4.3**	3.77**		2.9**	2.9**
	(0.32)	(0.6)		(0.73)	(0.5)
Road cost	5.57*	5.18*	1.52	-11.54**	-1.82
	(2.44)	(2.31)	(2.28)	(2.44)	(2.01)
Soil quality index by Road cost		2.23		7.93**	4.64*
		(1.85)		(2.5)	(1.83)
Soil production index			1.51		
			(1.3)		
Soil production index by Road cost			12.87		
			(8.21)		
Distance coast					0.95**
					(0.12)
River district					1.28+
					(0.75)
Elevation					1.89**
					(0.36)
Elevation^2					-0.08**
					(0.02)
Median slope					4.52
					(3.4)
Median slope^2					-1.42+
					(0.72)
Distance town					-1.13**
					(0.25)
Distance border					-1.01
					(1.48)
Distance capital					-0.23
					(0.17)

Instrumental variable estimates, with poverty rate as dependent variable and road cost instrumented by measures of colonial infrastructure. Standard errors clustered by country in parentheses. +: statistically significant at the 10% level. ∗: statistically significant at the 5% level. ∗∗: statistically significant at the 1% level.

contemporary roads, but mines also affect rural poverty independently, then conditioning on the presence of mines "blocks" the direct path that runs from the existence of a mining industry to contemporary rural poverty. Yet, we note that this also reduces the amount of variation in colonial roads that we are exploiting for the IV analysis. The model is reported in the first column of Table 11. Again we estimate both the main effects of soil quality and road cost, and their interaction, to be positive and statistically significant. This provides evidence that our IV strategy is not invalid due to the direct effect of mining on rural poverty. At the same time, the coefficient on the mining district dummy is negative and statistically significant: mining districts experience, on average and all else equal, lower levels of rural poverty. The effect is of some (albeit modest) economic

significance: rural poverty rates in mining districts are approximately 3.7 percentage points lower than in non-mining districts *located in the same country*.[12]

In addition, we control for some observable characteristics of the district in recent times while including also the dummy for mining districts. In particular, if colonial infrastructure, or the location of mines, affects contemporary rural poverty through the presence of urbanized areas, this would challenge our assumption. It should also be kept in mind that we control for urbanization levels, and in some models we also exclude observations that are very close to urban areas.[13] The second column of the table reports one of these models, that controls for distance from towns, distance from the border, distance from the coast, presence of a river, distance from the capital, the polynomials for elevation and slope, and the dummy for mining districts. As usual, contemporary infrastructure is instrumented with the two measures based on colonial roads. The main effect of soil quality, and the interaction of soil quality and road cost, are both positive and statistically significant, while the main effect of road cost does not reach conventional levels of statistical significance (with a standard error larger in absolute value than the point estimate). Again this corroborates the robustness of the interaction effect on which we focus. In this model, too, the coefficient on the dummy for mining districts is negative and statistically significant, implying that, all else equal, rural poverty in mining districts is around 4.3 percentage points lower than in non-mining districts.

Third, we treat the location of mines as an instrument, and we include it in the set of excluded instruments alongside colonial-era infrastructure. As discussed above, the presence of natural resources can affect current levels of rural poverty. In order to test the sensitivity of our results to violations of the exclusion restriction, we perform sensitivity analysis using the "local to zero" approach proposed by Conley et al. (2012). The model in the third column of the table uses the two measures based on colonial infrastructure, and the dummy for mining districts, as instruments for contemporary infrastructure. Again the three coefficients are positive, and statistically significant. (The coefficient on the interaction, but not those on the main effects, drops out of statistical significance when clustered standard errors are used.) The model in the fourth column shows that

[12]The estimate of the effect of mining on rural poverty in a district is net of the country-wide (spillover) benefits that several countries, e.g., Botswana and South Africa, might derive from the presence of an extractive industry, or national-level resource-curse effects that might obtain elsewhere.

[13]In addition, when we include rural population density as a control, completely urban districts drop out of the analysis because rural population density is missing by construction (in the original data) for urban areas.

the result is robust even if using a broad set of controls (all those included in the model in the second column) and the other measure of soil quality (the soil production index). Again, there's a positive, and statistically significant, interaction between soil quality and infrastructure.

Table 11: Robustness checks for the instrumental variable models

DV: Poverty rate	(1)	(2)	(3)	(4)
Intercept	78.18**	80.14**	77.84**	60.21**
	(1.6)	(3.62)	(1.59)	(3.8)
Soil quality index	3.78**	2.89**	3.78**	
	(0.22)	(0.43)	(0.22)	
Road cost	5.2*	-1.93	5.06*	1.29
	(2.04)	(3.4)	(2.04)	(3.37)
Mining district dummy	-3.74**	-4.29**		
	(1.36)	(1.18)		
Soil quality index by road cost	2.25**	4.7**	2.2**	
	(0.75)	(1.5)	(0.75)	
Distance coast		0.94**		0.99**
		(0.15)		(0.16)
River district		1.29*		1.44*
		(0.59)		(0.62)
Elevation		1.93**		1.73**
		(0.22)		(0.23)
Elevation^2		-0.08**		-0.06**
		(0.01)		(0.01)
Median slope		4.41		23.15**
		(2.83)		(2.89)
Median slope^2		-1.38*		-5.96**
		(0.62)		(0.62)
Distance town		-1.12**		-1.57**
		(0.14)		(0.14)
Distance border		-0.99*		-0.7
		(0.43)		(0.45)
Distance capital		-0.23*		-0.1
		(0.1)		(0.1)
Soil production index				-0.03
				(0.69)
Urbanization				-19.34**
				(1.54)
Soil production index by road cost				8.66**
				(3.23)

Further robustness checks: controlling for the direct effect of mining, and using mining as an instrument. Default 2SLS standard errors in parentheses. $+$: statistically significant at the 10% level. $*$: statistically significant at the 5% level. $**$: statistically significant at the 1% level.

Table 12: Robustness checks for the instrumental variable models

DV: Poverty rate	(1)	(2)	(3)	(4)
Intercept	78.18**	80.14**	77.84**	60.21**
	(0.34)	(3.66)	(0.25)	(5.5)
Soil quality index)	3.78**	2.89**	3.78**	
	(0.67)	(0.5)	(0.6)	
Road cost	5.2+	-1.93	5.06*	1.29
	(2.71)	(2.01)	(2.32)	(1.83)
Mining district dummy	-3.74*	-4.29**		
	(1.73)	(1.08)		
Soil quality index by Road cost	2.25	4.7*	2.2	
	(2.13)	(1.81)	(1.85)	
Distance coast		0.94**		0.99**
		(0.12)		(0.12)
River district		1.29+		1.44+
		(0.75)		(0.85)
Elevation		1.93**		1.73**
		(0.36)		(0.44)
Elevation^2		-0.08**		-0.06**
		(0.02)		(0.02)
Median slope		4.41		23.15**
		(3.33)		(5.14)
Median slope^2		-1.38+		-5.96**
		(0.7)		(1.19)
Distance town		-1.12**		-1.57**
		(0.26)		(0.29)
Distance border		-0.99		-0.7
		(1.49)		(1.52)
Distance capital		-0.23		-0.1
		(0.17)		(0.18)
Soil production index				-0.03
				(0.98)
Urbanization				-19.34**
				(2.89)
Soil production index by Road cost				8.66
				(5.91)

Further robustness checks: controlling for the direct effect of mining, and using mining as an instrument. Standard errors clustered by country in parentheses. +: statistically significant at the 10% level. $*$: statistically significant at the 5% level. $**$: statistically significant at the 1% level.

6 Exploring the mechanism: the role of education in a case study of Kenya

6.1 The data

Human capital might provide a possible mechanism mediating the relationship between soil quality, isolation, and rural poverty, and potentially leading to the "curse" of good soil in isolated areas that we detect in the continent-wide data. We use a case study of Kenya to explore the role played by human capital investment. Kenya is a good choice for a case study for various reasons. First of all, it is a large country, divided in a large number of districts (311), guaranteeing that there is sufficient variation in the explanatory variables to arrive at precise estimates. In addition, Kenya Open Data (https://opendata.go.ke) publishes education information that can be combined with the georeferenced information on soil quality and roads that we use in the main analysis. This provides a unique opportunity to study the mechanism in detail. In particular, Kenya Open Data provides two school-level datasets, one for primary and one for secondary schools, with geographic coordinates of each school. For primary schools, information about total enrollment, the pupil teacher ratio, and the pupil classroom ratio is available. For secondary schools, total enrollment in 2007, total teaching staff, and the pupil teacher ratio are reported. Both datasets also report other variables that are not directly relevant for our analysis.

We overlay the map of the level 3 administrative districts we use in the rest of the analysis on the school locations. We then compute summaries by district: the mean of pupil-teacher ratio and pupil-classroom ratio, the total enrollment in the district, and the total number of teachers in the district. We also compute the distance of the geometric centroid of the district to the closest school, and the number of primary and secondary schools in the district. We merge these new variables to the data for Kenya from the main dataset used in the rest of the paper. We also add the population of the district, computed based on the spatial data on the total population in Sub-Saharan Africa, distributed by FAO.

We also create a second dataset for Kenya, based on different information published by Kenya Open Data. In this dataset, the country is subdivided in 71 districts, whose boundaries do not coincide with the level-3 administrative divisions used in the rest of the paper. The names of these

districts can be matched with those used in the data on the percentage distribution of school age population who never attended school (for 2005/6). In addition, the School Attendance Data by District dataset, that is georeferenced, can be combined with the spatial data on soil and infrastructure. We can therefore compute, for each of these 71 districts, the percentages and the total counts for school attendance and non-attendance at various levels of education. We also recalculate, based on the sources discussed in the main data section of the paper, the summaries of soil quality (based on the FAO Soil production index), transportation infrastructure (the Road cost index), elevation, urbanization, and rural population density. Furthermore, we also recreate the soil type dummies, based on the Digital Soil Map of the World (DSMW) geological data we used in the main IV estimation for soil quality, and the summaries for the presence of transportation infrastructure in the colonial era used in the IV for infrastructure.

6.2 THE RESULTS

Two main results emerge from the analysis of the education data for Kenya. First of all, areas with better soil tend to experience lower school enrollment rates, after accounting for a basic set of district-level features; second, these same areas seem to have worse and fewer schools.

Table 13 reports the estimates of a first set of models, based on the data created with the more fine-grained map. The first two columns report the coefficient estimates for linear regression models with log total enrollment respectively in primary and secondary schools as response variable; the outcome in the models reported in the next two columns is distance from the closest school (respectively primary and secondary); the outcome in the models in the last two columns is the (log) number of schools in the district (respectively primary and secondary). All the models control for (log) population in the district, and for urbanization.

Districts with better soil quality seem to experience worse educational outcomes. Secondary enrollment is significantly lower in districts with better soil, which also have significantly fewer schools. In these models, no clear evidence emerges for primary schools, with the exception that the (geometric) center of the district is significantly farther from primary schools (implying that reaching a school might be in general more difficult in districts with better soil). Table 14 reports the two-stage least squares estimates for the models reported in Table 13: the soil production

measure is instrumented by the set of dummy variables for dominant soil class discussed above. The estimates are remarkably close to those of the OLS estimates, and the substantive implications are unchanged. The only exception is the coefficient on soil quality in the model for the number of primary schools, which is now statistically significantly negative, implying that districts with better soil have fewer primary schools.

Table 13: Educational outcomes in Kenya

DV:	Prim enrol	Sec enrol	Dist prim	Dist sec	Num prim schools	Num sec schools
Intercept	-0.484	-5.588**	0.247**	0.373**	-2.512**	-3.688**
	(0.498)	(0.75)	(0.023)	(0.044)	(0.36)	(0.36)
Soil production index	0.027	-0.231*	0.011**	0.018**	-0.062	-0.211**
	(0.066)	(0.099)	(0.003)	(0.006)	(0.048)	(0.048)
Road cost	-0.061	-0.179	0.005	0.014	-0.022	-0.081
	(0.083)	(0.126)	(0.004)	(0.007)	(0.06)	(0.06)
Urbanization	0.005	0.111	-0.005	-0.011	0.057	0.124*
	(0.074)	(0.112)	(0.003)	(0.007)	(0.054)	(0.054)
Population	0.935**	1.172**	-0.019**	-0.027**	0.613**	0.573**
	(0.045)	(0.067)	(0.002)	(0.004)	(0.032)	(0.032)

Models for educational outcomes in Kenya. Dependent variables are respectively primary enrollment, secondary enrollment, log distance from a primary school, log distance from a secondary school, log number of primary schools, and log number of secondary schools. $+$: statistically significant at the 10% level. $*$: statistically significant at the 5% level. $**$: statistically significant at the 1% level.

Table 14: Instrumental variable models for educational outcomes in Kenya

DV:	Prim enrol	Sec enrol	Dist prim	Dist sec	num prim schools	num sec schools
Intercept	-0.641	-5.82**	0.261**	0.405**	-2.629**	-3.861**
	(0.509)	(0.768)	(0.024)	(0.047)	(0.369)	(0.373)
Soil production index	-0.124	-0.454**	0.024**	0.049**	-0.175*	-0.377**
	(0.102)	(0.154)	(0.005)	(0.009)	(0.074)	(0.075)
Road cost	-0.046	-0.157	0.003	0.011	-0.011	-0.065
	(0.085)	(0.127)	(0.004)	(0.008)	(0.061)	(0.062)
Urbanization	-0.017	0.079	-0.003	-0.007	0.041	0.1+
	(0.076)	(0.114)	(0.004)	(0.007)	(0.055)	(0.056)
Population	0.95**	1.193**	-0.02**	-0.03**	0.623**	0.589**
	(0.046)	(0.069)	(0.002)	(0.004)	(0.033)	(0.033)

Models for educational outcomes in Kenya. Dependent variables are respectively primary enrollment, secondary enrollment, log distance from a primary school, log distance from a secondary school, log number of primary schools, and log number of secondary schools. IV estimates: soil production instrumented by dominant soil class dummies. $+$: statistically significant at the 10% level. $*$: statistically significant at the 5% level. $**$: statistically significant at the 1% level.

We also repeat the estimation using the saturated approach described above: we discretize the soil quality and road cost variables, and then we create dummies for each combination of soil and infrastructure categories. The results for models like those in Table 13 using this approach

are reported in the plots in Figure 4 and in Table 15. These report the estimates from the random effects models (see Section 3.3 above for an explanation). The message that emerges from these is (unsurprisingly) consistent with that from the models in which soil production enters linearly: education outcomes are worse in districts with better soil. In addition, it seems that the effect is more pronounced in particular in districts with relatively bad quality transportation infrastructure.

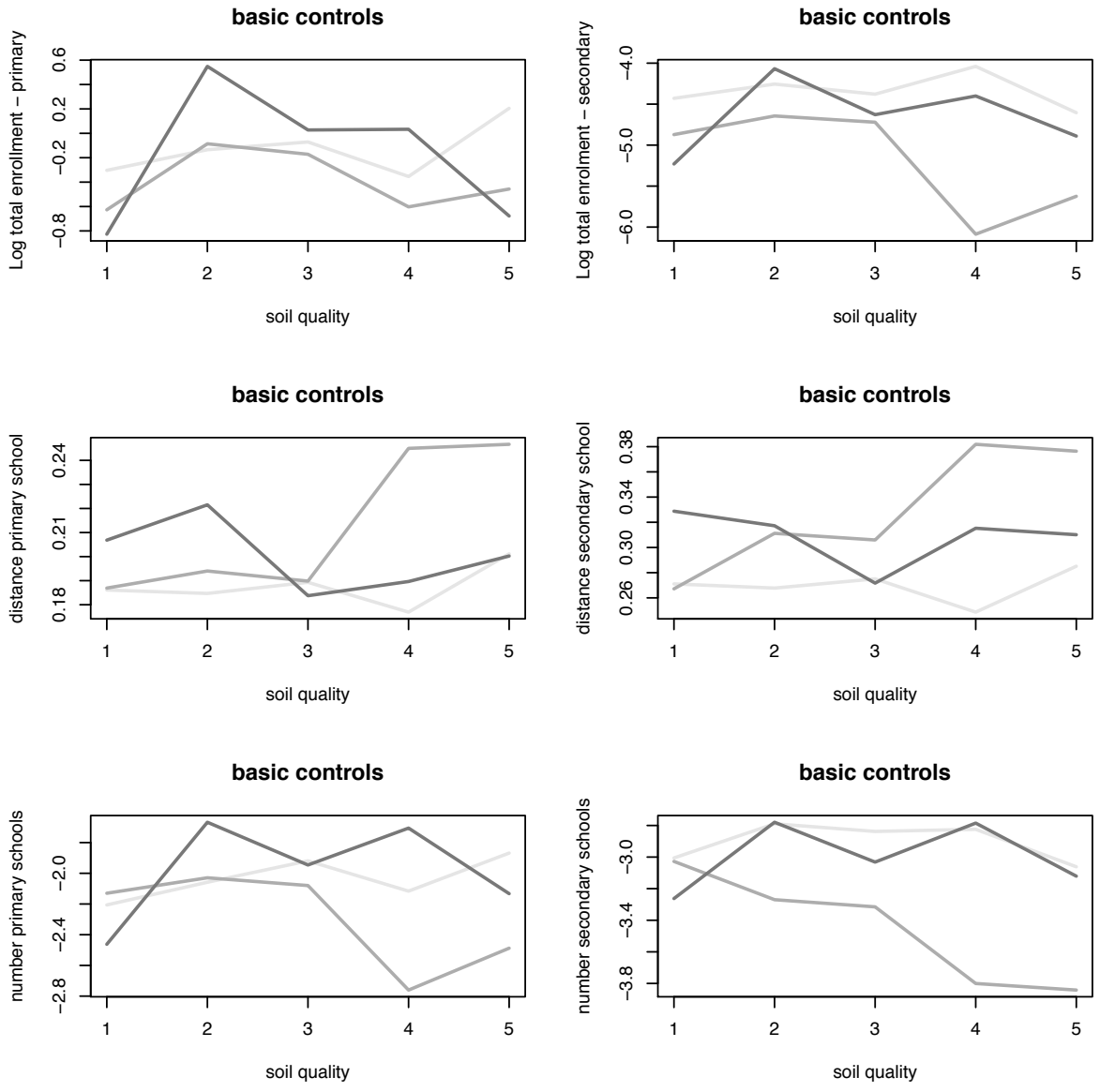

Figure 4: Saturated models for the measures of school enrollment and school provision. Darker lines represent districts with worse transportation infrastructure.

We also have data that capture the quality of the educational provision in a given district. In

Table 15: Non-parametric interactive models for educational outcomes in Kenya

road cost	soil	fe	se	fe	se	fe	se	fe	se	fe	se	fe	se
1.00	1.00	-0.30	0.58	-4.43	0.90	0.19	0.03	0.27	0.05	-2.21	0.42	-3.01	0.42
1.00	2.00	-0.13	0.58	-4.25	0.89	0.18	0.03	0.27	0.05	-2.06	0.42	-2.79	0.42
1.00	3.00	-0.07	0.55	-4.38	0.84	0.19	0.02	0.28	0.05	-1.92	0.40	-2.84	0.40
1.00	4.00	-0.35	1.11	-4.04	1.72	0.18	0.05	0.25	0.10	-2.12	0.80	-2.82	0.81
1.00	5.00	0.20	0.65	-4.60	1.01	0.20	0.03	0.29	0.06	-1.87	0.47	-3.06	0.47
2.00	1.00	-0.63	0.58	-4.87	0.89	0.19	0.03	0.27	0.05	-2.13	0.42	-3.03	0.42
2.00	2.00	-0.09	0.58	-4.64	0.90	0.19	0.03	0.31	0.05	-2.03	0.42	-3.27	0.42
2.00	3.00	-0.17	0.55	-4.72	0.85	0.19	0.02	0.31	0.05	-2.08	0.40	-3.32	0.40
2.00	4.00	-0.60	0.57	-6.09	0.88	0.24	0.03	0.38	0.05	-2.76	0.41	-3.80	0.41
2.00	5.00	-0.46	0.51	-5.62	0.79	0.25	0.02	0.38	0.05	-2.49	0.37	-3.84	0.37
3.00	1.00	-0.83	0.50	-5.23	0.78	0.21	0.02	0.33	0.04	-2.46	0.36	-3.26	0.37
3.00	2.00	0.55	0.52	-4.07	0.81	0.22	0.02	0.32	0.05	-1.67	0.38	-2.78	0.38
3.00	3.00	0.03	0.54	-4.63	0.84	0.18	0.02	0.27	0.05	-1.95	0.39	-3.03	0.39
3.00	4.00	0.03	0.79	-4.40	1.22	0.19	0.04	0.32	0.07	-1.71	0.57	-2.78	0.57
3.00	5.00	-0.68	0.58	-4.89	0.90	0.20	0.03	0.31	0.05	-2.13	0.42	-3.12	0.42

Estimates from the saturated models, with district-level enrollment and school provision as dependent variables. In the columns labeled "fe" the estimates come from the fixed-effects estimates. The columns labeled "se" report the standard errors from the fixed-effects estimation.

particular, we estimate regression models with response variable respectively the pupil-teacher ratio (at the primary and secondary level), the pupil-classroom ratio, and the (log) number of teachers in the district. All the models control for (log) population in the district. In the case of the model for number of teachers, this accounts for the fact that larger districts need more teachers. In the other cases, it accounts for possible scale economies effects. The results of the estimates are reported in Table 16.

Higher soil quality is statistically significantly associated with a higher pupil-teacher ratio (PTR) and with a lower number of secondary school teachers. No clear evidence emerges for the pupil teacher ratio at the secondary level or the pupil-classroom ratio at the primary level. Table 17 reports the two-stage least squares estimates of these models. The only noticeable difference is that the coefficient on soil production in the model for primary PTR, while unchanged in magnitude, is estimated somewhat less precisely and it is statistically significant only at the 10% level.

It is important to note that in the models for enrollment presented in Table 13, soil quality is not statistically significantly associated with primary school enrollment. In practice, good soil districts do not experience significantly lower enrollment in primary schools than districts with worse soil. Yet, the model for PTR shows that primary schools in better-soil districts tend to be understaffed. On the other hand, the models in Table 13 show that secondary enrollment is lower

in good soil districts.

Taken together, these models point to poorer education outcomes in districts with higher-quality soil. In good-soil districts, as many children enroll in primary schools as those in otherwise similar districts with worse soil quality, but schools tend to be significantly understaffed. At the secondary school level, on the other hand, fewer children in the relevant demographic enroll in school: this leads to a constant pupil-teacher ratio in spite of the significantly lower number of teachers detected in the model in the fourth column of Table 16.

Table 16: Provision of education in Kenya

DV:	PTR (prim)	PTR (sec)	PCR (prim)	Tot teachers (sec)
Intercept	39.44**	12.91**	23.09**	-4.75**
	(6.27)	(2.94)	(5.38)	(0.52)
Soil production index	2.2**	-0.08	0.66	-0.25**
	(0.73)	(0.29)	(0.62)	(0.07)
Road cost	0.2	0.09	-0.83	-0.15+
	(0.87)	(0.33)	(0.74)	(0.09)
Urbanization	-2.61**	-0.37	-1.95**	0.15+
	(0.77)	(0.29)	(0.66)	(0.08)
Population	-0.23	0.24	0.97*	0.86**
	(0.56)	(0.26)	(0.48)	(0.05)

Models for educational outcomes in Kenya. Dependent variables are respectively primary pupil/teacher ratio, secondary pupil/teacher ratio, primary pupil/classroom ratio, and log number of teachers. +: statistically significant at the 10% level. ∗: statistically significant at the 5% level. ∗∗: statistically significant at the 1% level.

Table 17: IV models for provision of education in Kenya

DV:	PTR (prim)	PTR (sec)	PCR (prim)	Tot teachers (sec)
Intercept	39.5**	13.67**	23.93**	-4.96**
	(6.29)	(3)	(5.48)	(0.54)
Soil production index	2.06+	-0.69	-1.11	-0.46**
	(1.13)	(0.46)	(0.98)	(0.11)
Road cost	0.22	0.14	-0.7	-0.13
	(0.87)	(0.33)	(0.76)	(0.09)
Urbanization	-2.62**	-0.43	-2.14**	0.12
	(0.78)	(0.3)	(0.68)	(0.08)
Population	-0.23	0.17	0.9+	0.88**
	(0.56)	(0.26)	(0.49)	(0.05)

Models for educational outcomes in Kenya. Dependent variables are respectively primary pupil/teacher ratio, secondary pupil/teacher ratio, primary pupil/classroom ratio, and log number of teachers. IV estimates: soil production instrumented by dominant soil class dummies. +: statistically significant at the 10% level. ∗: statistically significant at the 5% level. ∗∗: statistically significant at the 1% level.

Figure 5 and Table 18 report the results of the saturated interactive models for the variables

that measure quality of education. Darker lines represent districts with higher road cost (hence worse roads) and higher values of soil quality represent better soil. The message that emerges is the same, unsurprisingly: better soil districts have higher pupil-teacher ratios than worse soil districts, and this effect is more pronounced in districts with worse roads; similarly, the number of secondary school teachers is at its lowest in districts with good soil and with roads in the intermediate category.

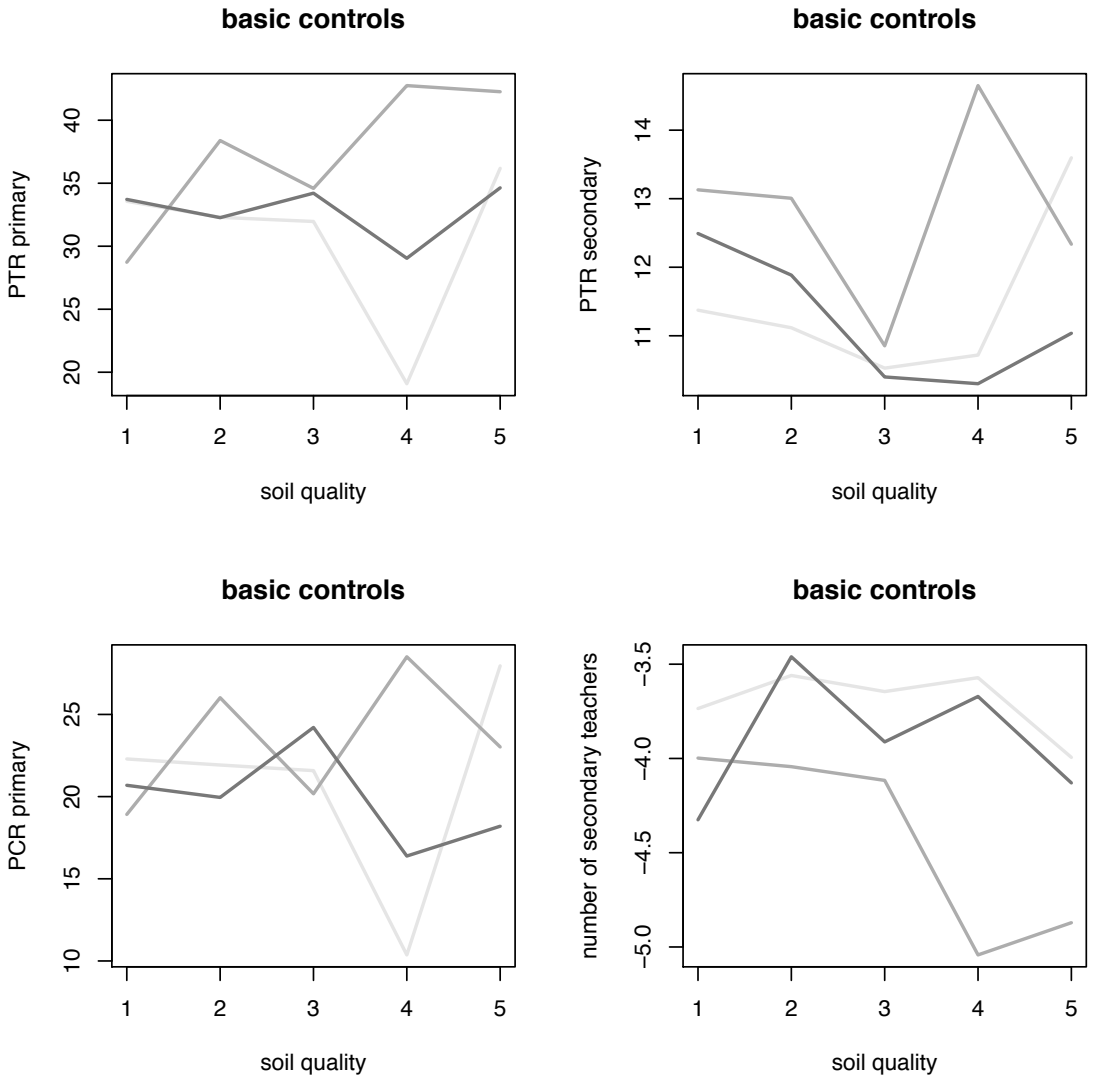

Figure 5: Saturated models for the measures of education quality. Darker lines represent districts with worse transportation infrastructure.

The response variable in the models reported in Table 19 is school attendance respectively in

Table 18: Non-parametric interactive models for educational provision in Kenya

road cost	soil	fe	se	fe	se	fe	se	fe	se
1.00	1.00	33.57	7.33	11.37	3.31	22.29	6.33	-3.74	0.62
1.00	2.00	32.29	7.30	11.12	3.31	21.92	6.31	-3.56	0.61
1.00	3.00	31.96	7.02	10.53	3.21	21.58	6.06	-3.65	0.58
1.00	4.00	19.09	12.33	10.72	4.98	10.37	10.65	-3.57	1.18
1.00	5.00	36.18	7.99	13.60	3.53	27.96	6.91	-3.99	0.70
2.00	1.00	28.73	7.49	13.13	3.48	18.92	6.47	-4.00	0.61
2.00	2.00	38.39	7.29	13.01	3.27	26.02	6.29	-4.04	0.62
2.00	3.00	34.59	6.96	10.85	3.14	20.17	6.01	-4.12	0.59
2.00	4.00	42.75	6.92	14.65	3.39	28.51	5.98	-5.04	0.61
2.00	5.00	42.26	6.50	12.34	2.97	23.02	5.61	-4.87	0.55
3.00	1.00	33.72	6.74	12.49	3.11	20.69	5.82	-4.33	0.54
3.00	2.00	32.27	6.52	11.88	3.07	19.95	5.63	-3.46	0.56
3.00	3.00	34.21	6.89	10.40	3.18	24.21	5.95	-3.91	0.58
3.00	4.00	29.04	9.25	10.30	3.95	16.38	7.99	-3.67	0.84
3.00	5.00	34.64	7.40	11.04	3.30	18.20	6.39	-4.13	0.62

Estimates from the saturated models, with district-level pupil/teacher ratio as dependent variables. In the columns labeled "fe" the estimates come from the fixed-effects estimates. The columns labeled "se" report the standard errors from the fixed-effects estimation.

primary, secondary, and tertiary education, as a percentage of the relevant school-age population, and (log) count of school-age children who do not attend school either because parents did not allow them or they had to work for money or help at home. The coefficient on soil quality is negative (implying lower school attendance) in the three models for attendance. It is clearly statistically significant only for primary school attendance, and marginally significant for tertiary attendance. No clear evidence emerges for the number of children classified as "not allowed" to attend school. The results are not very different in the two-stage least squares estimation, reported in Table 19. In particular, the coefficient on soil quality is still statistically significantly negative only in the model for primary attendance.

The plots in Figure 6 and the estimates in Table 21 come from the saturated models with the dummies for combinations of discretized soil quality and road cost and the specifications of Table 19. These broadly confirm the results of the regression models. Interestingly, when looking at disaggregated combinations of soil and roads, it seems that the number of children who are not allowed to attend school is quite higher in good-soil districts with bad roads. The pattern is the opposite in the case of good-soil districts with good roads.

Table 19: School attendance in Kenya

DV: Attendance	Primary	Secondary	Tertiary	Not allowed
Intercept	111.13**	44.06**	5.64**	5.9
	(2.65)	(2.43)	(0.78)	(10.62)
Road cost	-1.27	0.69	-0.33	-0.68
	(2.2)	(2.01)	(0.65)	(0.63)
Soil production index	-4.31*	-2.17	-0.94+	0.21
	(1.85)	(1.69)	(0.54)	(0.53)
Urbanization	0.48	1.67**	0.25*	-0.02
	(0.37)	(0.33)	(0.11)	(0.11)
Number school age children				-0.13
				(0.9)

Models for school attendance. +: statistically significant at the 10% level. ∗: statistically significant at the 5% level. ∗∗: statistically significant at the 1% level.

Table 20: IV models for school attendance in Kenya

Attendance	Primary	Secondary	Tertiary	Not allowed
Intercept	111.82**	44.05**	5.54**	6.67
	(2.82)	(2.44)	(0.79)	(10.96)
Road cost	-0.67	0.68	-0.41	-0.71
	(2.34)	(2.02)	(0.66)	(0.64)
Soil production index	-9.57**	-2.09	-0.21	0.37
	(2.87)	(2.48)	(0.81)	(0.77)
Urbanization	0.35	1.67**	0.27*	-0.02
	(0.39)	(0.34)	(0.11)	(0.11)
Number school age children				-0.2
				(0.94)

Models for school attendance. IV estimates. +: statistically significant at the 10% level. ∗: statistically significant at the 5% level. ∗∗: statistically significant at the 1% level.

Table 21: Non-parametric interactive models for school attendance in Kenya

road cost	soil	fe	se	fe	se	fe	se	fe	se
1.00	1.00	117.46	6.28	39.09	6.95	9.84	2.33	9.83	11.68
1.00	2.00	121.27	5.20	47.22	5.75	5.53	1.93	6.50	11.84
1.00	3.00	117.52	4.79	47.23	5.30	6.30	1.78	5.84	11.51
1.00	5.00	116.62	12.40	37.11	13.73	3.52	4.61	2.26	12.82
2.00	1.00	119.71	7.17	50.93	7.94	8.32	2.66	7.22	11.56
2.00	2.00	120.28	6.25	51.84	6.92	6.21	2.32	3.54	11.86
2.00	3.00	114.58	4.14	50.03	4.58	7.04	1.54	5.60	11.56
2.00	4.00	125.09	12.40	58.55	13.73	5.21	4.61	2.32	13.08
2.00	5.00	95.60	5.05	38.23	5.59	4.03	1.88	8.19	11.91
3.00	1.00	111.89	4.87	45.91	5.39	4.44	1.81	5.92	11.11
3.00	2.00	116.14	7.14	40.34	7.90	4.94	2.65	4.65	11.48
3.00	3.00	119.26	6.34	41.77	7.02	3.58	2.35	3.51	11.32
3.00	4.00	63.46	8.73	21.46	9.67	1.91	3.24	8.78	12.66
3.00	5.00	104.40	4.72	42.56	5.23	4.80	1.75	8.08	11.80

Estimates from the saturated models, with district-level school attendance as dependent variable. In the columns labeled "fe" the estimates come from the fixed-effects estimates. The columns labeled "se" report the standard errors from the fixed-effects estimation.

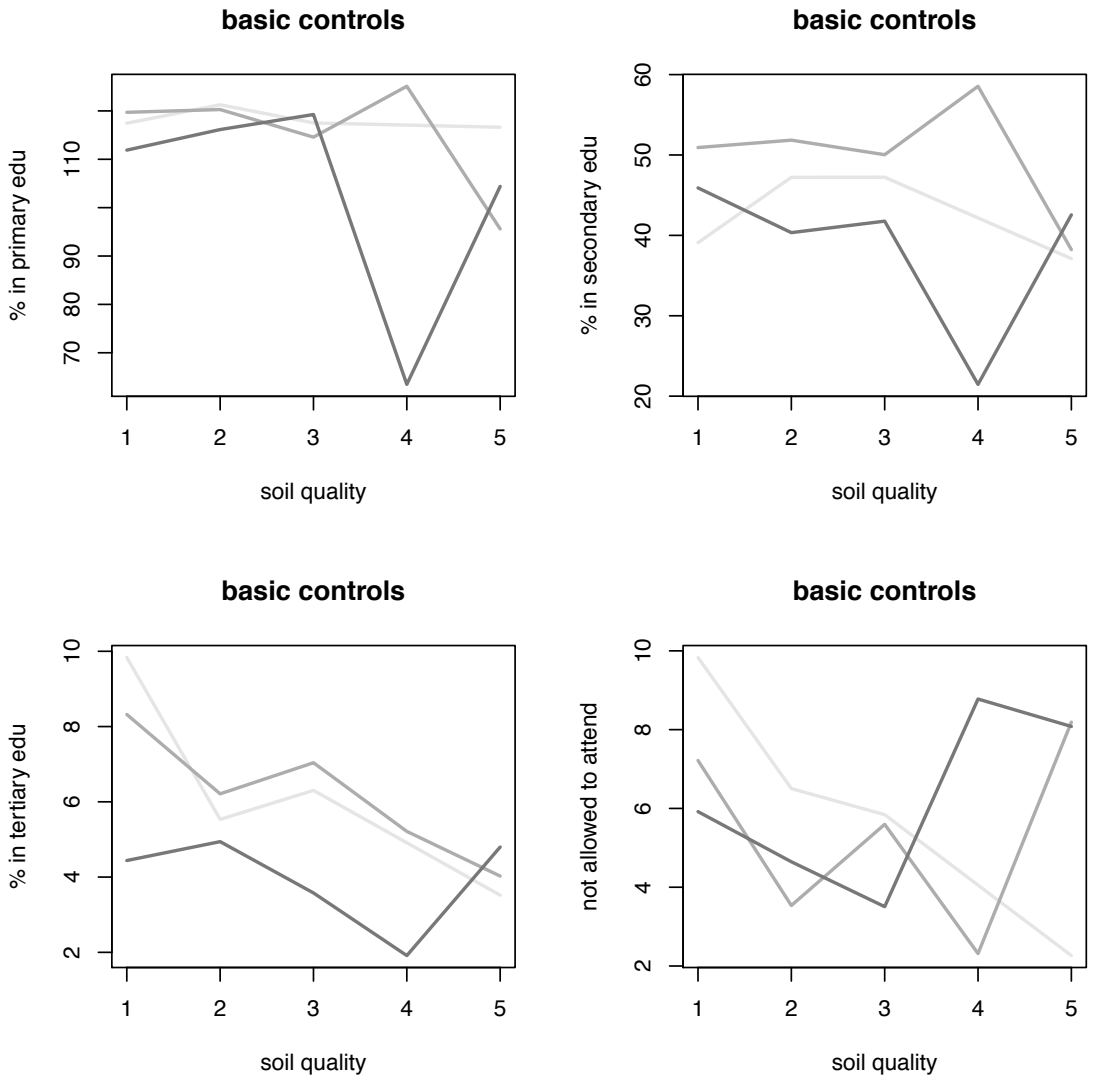

Figure 6: Saturated models for the measures of education attendance from the smaller dataset. Darker lines represent districts with worse transportation infrastructure.

6.3 Summary of the results

We explore the mechanism for the counter-intuitive relationship between soil quality and poverty. In order to do so, we exploit the detailed data on education that, unfortunately, is only available for Kenya. We find that districts with good soil tend to experience worse educational outcomes, both in terms of school attendance and in terms of education supply. Fewer children attend school, and few schools are available, in these districts that are "blessed" with higher-quality soil. There is also evidence that the effect is stronger, to an extent, in more isolated districts: in other words, in districts with lower quality transportation infrastructure soil quality is more tightly associated with worse education. There is some evidence, in particular, that the number of children who are not allowed to attend school is systematically higher in districts with bad roads and good soil than both in districts with good soil and good roads, and in districts with bad roads and also worse soil. In other words, there is lower human capital accumulation in districts with high soil quality but poor transportation infrastructure.

6.4 Political Economy of Rural Infrastructure: Avenues for Further Research

Thus far we have illustrated the role of education as a potential mechanism to explain poverty in isolated places with good soil. Given poor infrastructure, households in soil-rich districts may choose to invest less in human capital than they would otherwise because the returns to this investment are low, while the opportunity costs foregone from agricultural activities are high. This vicious cycle may further aggravate the high level of poverty in high soil quality areas. By contrast, households in soil-poor districts who also have poor infrastructure face lower opportunity costs from investing in alternative income generating activities such as education. These households may therefore be more likely to enroll their children in school and further contribute to human capital accumulation.

Yet the question remains as to why, if there are such potentially high returns to investing in infrastructure in soil-rich areas, these goods continue to be underprovided. Answering this question represents a promising avenue for further research and requires us to turn to more proximate political factors to explain the persistent under-provision of roads to areas that need it the most. Sev-

eral extant studies have considered such politically-driven factors to explain the (mis)allocation of infrastructure goods.

A number of articles have explored the role of ethnic favoritism in determining road allocation. For example, Burgess et al (2014) find that during non-democratic periods in Kenya, districts that share the ethnicity of the president receive significantly more investment in and construction of paved roads. Ethnic favoritism is not relegated solely to roads and infrastructure, but to the provision of other goods and services such as education, health, electricity and water (Kramon and Posner 2013). In a large-N panel study, Franck and Rainer (2012) find evidence of ethnic favoritism with respect to primary education and infant mortality in 18 African countries. Together these studies support the claim that social group membership can explain why some areas consistently do not receive public infrastructure goods: many groups simply do not have a representative in power to redirect targeted goods toward them.

Blimpo, Harding and Wantchekon (2013) further highlight the importance of political representation by examining the relationship between political marginalization, road infrastructure, and food security. They argue that marginalization is negatively correlated with investment in roads, thereby decreasing access to food supplies and contributing to food insecurity in the politically marginalized districts of four African countries. Politicians do not need to be responsive to the needs of these marginalized constituencies because they have no weight in determining political outcomes. Instead, politicians wish to direct infrastructure projects to voters who can improve their electoral chances. We have previously shown in this paper that soil-rich districts tend to be more isolated , which could contribute to their marginalization. If this is the case, voters in soil-rich but isolated areas may continually have their demands for quality roads unmet. Even if politicians do respond to demands for infrastructure, they may simply turn these projects into an opportunity for local rent-seeking and elite capture that keeps them in office (Khemani 2010). Thus marginalized voters may find it difficult to hold their political representatives accountable or punish them for their lack of provision.

Another potential reason for this lack of accountability is that politicians can change voter incentive structures by offering substitutable private goods in lieu of investments in infrastructure. For example, seeking to explain the lack of infrastructure in India despite high voter demand among the poor, Khemani (2010) argues that politicians replace large public investment projects

with targeted social welfare transfers in the face of rising electoral competition in weakly governed states. That is, politicians eschew the provision of transport infrastructure in favor of less expensive alternatives that are electorally popular. These alternatives could entail the distribution of targeted private benefits or clientelistic goods funded by rents appropriated from public projects. Boone (2009) posits that politicians also rely on land property rights as a source of patronage in rural constituencies. This tactic may be especially useful in soil-rich areas where land may be more valuable to poor individuals who rely on the land for subsistence, and also much cheaper for politicians to provide. In the long run, the use of these alternative strategies can explain both the continual under-provision of roads and the survival of politicians unresponsive to constituent needs.

7 Conclusion

In this paper, we detect a surprising empirical regularity, the positive correlation between soil quality and poverty in Africa, and then provide rigorous evidence about how this pattern emerges by linking it to resource under-utilization due to lack of infrastructure and limited access to markets.

In light of these results, the current emphasis on fertilizer use among a number of policymakers and agricultural economists could be misguided; more attention should be paid to rural infrastructure and human capital investment. Roads facilitate access to markets, and the presence of schools locally provides incentives for investment in human capital. The low rate of adoption of new technologies and the reluctance to use fertilizer in many rural communities might be driven, paradoxically, by the combination of poor rural infrastructure and the availability of relatively high-quality land. Therefore the growing interest among development agencies on rural infrastructure and local public goods in general is a promising avenue for rural development and poverty alleviation in Africa.

REFERENCES

Ali, Rubaba. 2010. "Impact of rural road improvement on high yield variety technology adoption: Evidence from Bangladesh." Unpublished manuscript, University of Maryland Department of Agricultural and Resource Economics.

Amaral, Pedro S., and Quintin Erwan. 2006. "A competitive model of the informal sector." *Journal of Monetary Economics* 53(7):1541-1553.

Angrist, Joshua. D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton: Princeton University Press.

Arellano, Manuel. 1987. "Computing Robust Standard Errors for Within-groups Estimators." *Oxford Bulletin of Economics and Statistics* 49(4)431-434.

Ayogu, Melvin. 2007. "Infrastructure and Economic Development in Africa: A Review." *Journal of African Economies* 16: 75–126.

Banerjee, Abhijit, Esther Duflo, and Nancy Qian, 2012. *On the Road: Access to Transportation Infrastructure and Economic Growth in China.* National Bureau of Economic Research Working Paper No. 17897.

Barbier, E.B. 2010. "Poverty, development and environment." *Environment and Development Economics* 15 (Special issue 06): 635-660.

Bell, Clive, and Susanne van Dillen. 2012. "How does India's rural roads program affect the grassroots? Findings from a survey in Orissa." Policy Research Working Paper Series 6167, The World Bank.

Blimpo, M. P., Harding, R., and Wantchekon, L. 2013. Public Investment in Rural Infrastructure: Some Political Economy Considerations. *Journal of African Economies*, 22(2), 57-83.

Boone, C. 2009. Electoral populism where property rights are weak: land politics in contemporary sub-Saharan Africa. *Comparative Politics*, 183-201.

Burgess, R., Jedwab, R., Miguel, E., and Morjaria, A. 2014. The Value of Democracy: Evidence from Road-Building in Kenya, Working Paper, LSE

Casaburi, Lorenzo, Rachel Glennerster, and Tavneet Suri. 2010. "Rural roads and intermediated trade: Regression discontinuity evidence from Sierra Leone." Working Paper.

Christiaensen L. and L. Demery. 2007. "Down to earth: Agriculture and Poverty Reduction in Africa, Directions in development," World bank: Washington D.C.

Conley, Timothy G., Christian B. Hansen, and Peter E. Rossi. 2012. "Plausibly exogenous." *The Review of Economics and Statistics* 94 (1):260-272.

Dercon S. and L. Christiansen. 2011. "Consumption risk, technology adoption and poverty traps: Evidence from Ethiopia." *Journal of development Economics* 96(2): 159-173.

Dreschel, Pay Lucy Gyiele, Dagmar Kunze, and Olufunke Cofie. 2001. "Population density, soil nutrient depletion, and economic growth in sub-Saharan Africa." *Ecological Economics* 38(2):251-258.

Ehui, S. and Pender, J. 2005. "Resource degradation, low agricultural productivity, and poverty in sub-Saharan Africa: pathways out of the spiral." *Agricultural Economics*, 32: 225-242.

Ephraim Nkonya, Nicolas Gerber, Philipp Baumgartner, Joachim von Braun, Alex De Pinto, Valerie Graw, Edward Kato, Julia Kloos, and Teresa Walter. 2011. "The Economics of Land Degradation." Peter Lang Internationaler Verlag der Wissenschaften, 2011.

Elvidge, Christopher D., Paul C. Sutton, Tilottama Ghosh, Benjamin T. Tuttle, Kimberly Baugh, Buhendra Bhaduri, and Edward Bright. 2009. "A global poverty map derived from satellite data." *Computers & Geosciences* 35: 1652-60.

Environmental Systems Research Institute, Inc. 1992. *Esri ArcWorld Database* [computer file]. Redlands, CA: Environmental Systems Research Institute.

Faber, Benjamin. 2012. "Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System." Unpublished manuscript, London School of Economics.

FAO. 2007. *Soil Production Index* [computer file]. Rome, Italy: FAO.

`http://www.fao.org/geonetwork/srv/en/resources.get?id=30577`

FAO-UNESCO Soil Map of the World. 1974. 1:5.000.000. UNESCO, Paris.

Gelman, Andrew, Aleks Jakulin, Maria G. Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics* 2(4):1360-1383.

Gelman, Andrew, David Park, Boris Shor, Joseph Bafumi, and Jeronimo Cortina. 2008. *Rich State, Poor State, Red State, Blue State: Why Americans Vote the Way They Do.* Princeton: Princeton University Press.

Ghitza, Yair, and Gelman, Andrew. 2013. "Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups". American Journal of Political Science 57(3):762-776.

Gibson J. and S. Rozelle. 2003. "Poverty and Access to roads in Papua New Guinea." *Economic Development and Cultural Change* 52(1): 159-185.

Gollin Douglas, and Richard Rogerson, 2014. "Productivity, Transport Costs and Subsistence Agriculture". Journal of Development Economics 107: 38-48.

Goyal, Aparajita. 2010. "Information, direct access to farmers, andrural market performance in central India." Policy Research Working Paper Series 5315, The World Bank.

Jacoby, H. 2000. "Access to markets and the benefits of rural roads." *Economic Journal* 465:713-737.

Jacoby, H. and Minten, B. 2009. "On Measuring the benefits of lower transport costs." *Journal of Development Economics*, 89:28-38.

Khandker, S., Bakht, Z., and Boolwal, G. 2009. "The Poverty Impact of Rural Roads: Evidence from Bangladesh." *Economic Development and Cultural Change*, 57(4):685-722.

Khemani, S. (2010). "Political economy of infrastructure spending in India." World Bank. `http://openknowledg`

Kramon, E., and Posner, D. N. (2013). Who benefits from distributive politics? How the outcome one studies affects the answer one gets. *Perspectives on Politics*, 11(02):461-474.

Lewis, W. A. (1955). *The Theory of Economic Growth.* Homewood, IL: Richard Irwin.

McMillan, M.S., and Rodrik, D. 2011. "Globalization, Structural Change and Productivity Growth." NBER Working Paper Series 17143.

Minten, B., Koru, B. and D. Stifel. 2013. "The last mile(s) in modern input distribution: Pricing, profitability, and adoption." *Agricultural Economics* 44: 1-18.

Morris, M., Kelly, V.A., Kopicki, R.J. and R.J. Byerlee. 2007. "Fertilizer Use in African Ariculture - Lessons Learned and Good Practice Guidelines, Directions in Development, Agriculture and Rural development." World Bank: Washington D.C.

Mu, R. and D. van de Walle. 2011. "Rural Roads and Local Market development in Vietnam." *Journal of Development Studies* 47(5): 709-734.

Nunn, N., and D. Puga. 2012. "Ruggedness: The Blessing of Bad Geography in Africa." *Review of Economics and Statistics* 94(1): 20-36.

Okwi, Paul O., Godfrey Ndeng'e, Patti Kristjanson, Mike Arunga, An Notenbaert, Abisalom Omolo, Norbert Henninger, Todd Benson, Patrick Kariuki, and John Owuor. 2007. "Spatial determinants of poverty in rural Kenya." *Proceedings of the National Academy of Sciences of the USA (PNAS)* 104 (43):16769-16774.

Radeny, Maren, and Bulte, Erwin. 2012. "Determinants of Rural Income: The Role of Geography and Institutions in Kenya." *Journal of African Economies* 21(2):307-341.

Sanchez, Pedro A. 2002. "Soil fertility and hunger in Africa." *Science* 295(5562):2019-2020.

Scherr, Sara J. 1999. "Soil degradation: A threat to developing country food security in 2020?" Food, Agriculture and the Environment Discussion Paper 27. Washington DC: IFPRI.

Stifel, David, and Minten, Bart. 2008. "Isolation and agricultural productivity." *Agricultural Economics* 39(1):2019-2020.

Woomer, Paul L., and Michael J. Swift. 1994. "The Biological Management of Tropical Soil Fertility." *Tropical Soil Biology and Fertility Programme.* John Wiley, New York.

World bank. 2008. World Development Report: Agriculture for development. World Bank, Washington D.C.

Yamano, Takashi, and Yoko Kijima. 2010a. "The associations of soil fertility and market access with household income: Evidence from rural Uganda." *Food Policy* 35(1):51-59.

Yamano, Takashi, and Yoko Kijima. 2010b. "Market Access, Soil Fertility, and Income in East Africa." GRIPS Discussion Papers 10-22, National Graduate Institute for Policy Studies.

A FURTHER EVIDENCE FROM THE IV

A.1 CHECKING THE ROBUSTNESS OF THE IV ESTIMATES

We further probe the robustness of the results we report. In Table 22 and 23 we report the estimates of additional IV models (again estimated via 2SLS), with (log) distance from a colonial primary road and the dummy for presence of a primary road in the district as excluded instruments.

The model in the first column of Table 22 includes controls for a large set of district-level observable characteristics. In addition to the polynomials for elevation and terrain slope, we include distance from the coast, the dummy for districts with a river, (log) rural population density, distance from towns, distance from the capital, average and maximum level of urbanization in the district, (standardized) length of the growing period, number of towns over 10,000 inhabitants (normalized by area of the district), (standardized) presence of livestock, and (standardized) rainfall. In this model, too, soil quality has a positive association with rural poverty, and the interaction between lack of infrastructure and soil quality is positive. The main effect of infrastructure (the effect of poor infrastructure for a district with average soil quality) is negative (which is hard to rationalize, but one has to keep in mind we are using many other measures of "remoteness" and in particular, the presence of towns, that, on their own, have a negative association with poverty –more towns in the district, less poverty).

The model in the second column of Table 22 uses an alternative measure of contemporary infrastructure provision, the (log) distance of the center of the district from the closest road (based on the roads featured in the VMAP0 Major Road Library published by FAO). Again, soil quality and the interaction between soil quality and lack of infrastructure (measured by the distance of the district from a "major road") are positively related to rural poverty. In this model, too, remoteness in a district with average soil quality is negatively associated with rural poverty. It is unclear to

us what drives this result. In this model, we do not include any controls other than the country fixed effects. The third model uses yet a third measure of contemporary infrastructure: the (log) distance of the center of the district from roads of category 1 and 2 in the RWDB roads shapefile. The results here are that (as usual) the infrastructure-soil quality interaction is positive (statistically significant with regular standard errors only) but the main effects are not statistically significant (and negative in both cases). This model does not include any controls other than the country fixed effects. The fourth model uses the poverty measure based on percentage of population under two dollars a day as dependent variable. The main effects of soil quality and road cost are not statistically significant, but again the coefficient on the interaction term is positive (and statistically significant with the default 2SLS standard errors).

Sensitivity checks for the exclusion restriction We perform some sensitivity analysis to assess to what extent the results of the IV estimation depend on the exclusion restriction holding exactly rather than approximately. For this purpose, we follow the approach of Conley et al. (2012), which makes it possible to identify IV coefficients using a relaxed version of the exclusion restriction. This restriction amounts to the assumption that the direct effect of the instrument on the outcome is exactly zero. Conley et al. (2012) suggest to replace the assumption of zero effect with a distribution of plausible effects. One can estimate the model under different distributions, to assess how serious the violation of the exclusion restriction needs to be for the results of the 2SLS estimation to be misleading.

We use the "local-to-zero" method of Conley et al., specifying a full distribution for the direct effects of the instrument on the outcome. The distribution captures the prior knowledge regarding the direct effect of the instruments on the outcome; in other words, it reflects the fact that there is uncertainty regarding whether the exclusion restriction holds.

As a result of the sensitivity analysis, one obtains, for every distribution of the direct effect of the instruments on the outcome, a distribution of the IV effect. This distribution reflects the uncertainty regarding the effect of the instrumented endogenous variable on the outcome variable deriving both by sampling variation and by the uncertainty regarding the strength of the departure from the exclusion restriction.

In our analysis, the most important instance in which the exclusion restriction might be vio-

lated is that of the model that uses mining activity in the district as an instrument for the presence of infrastructure.

We perform the Conley et al. (2012) analysis for this model under different scenarios. The first one assumes that the direct effect of being a mining district on poverty has mean -5 (meaning that being a mining district reduces poverty by 5 percentage points, all else equal) with a standard deviation 2.5 (so that with some –small– probability the effect is zero or positive). We also allow for the interaction between mining and soil productivity to be negative, and for the other variables to have smaller direct effects (respectively 1 and -1 for the terms involving log distance from a colonial road and for the dummy for presence of a colonial road in the district).

If this were the case, the inference about a positive effect of transportation costs and of the interaction between soil quality and transportation costs would still be valid. If anything, the 2SLS estimate we get (and for which the exclusion restriction is assumed to hold exactly) would be an underestimation of the effect of these variables, if mining had such a strong negative effect on rural poverty.

The second scenario assigns to all of the direct effects of instruments on outcome a normal distribution with mean 0 and standard deviation .5 for the main effects, .25 for the interactions between soil productivity and the excluded instruments. This allows for mining and for colonial infrastructure to have, potentially, a positive or a negative direct effect on rural poverty. In this case, we are probing sensitivity to small departures – in either direction – from the exclusion restriction.

In this scenario, the inference regarding the positive interactive effect between road cost and soil quality would still be valid, with a 95 percent interval for this effect between 0.91 and 16.15, while the interval for the main effect of transportation infrastructure would span zero.

The third scenario assigns to the direct effects a uniform distribution (on the (-.25, .25) support). Also in this case, the inference for the interaction effect would be the same, and the 95 percent interval for the effect would be [2.07, 15.5]. The inference of a positive main effect of road cost would not be justified, as again the 95 percent interval spans zero.

What these sensitivity checks tell us is that even if the exclusion restriction (and in particular that for the direct effect of the presence of mines and quarries in the district) did not hold exactly, but were subject to (relatively small) violations, the IV inference regarding the interac-

tion between soil quality and road costs would still be valid. In addition, if mining activities had a direct poverty-reducing effect, then our IV estimates of the effect of infrastructure would be *under*-estimates of the true effect.

Table 22: Further robustness checks

robustness1	(1)	(2)	(3)	(4)	(5)
Intercept	55.85**	76.17**	94.62**	58.12**	77.78**
	(3.4)	(1.71)	(11.75)	(0.87)	(1.7)
Soil quality index	2.48**	6.59**	-2.76+		
	(0.4)	(0.41)	(1.41)		
Road cost	-8.58*			0.49	7.29**
	(3.53)			(1.43)	(1.96)
Distance coast	0.95**				
	(0.14)				
River district	1.29*				
	(0.51)				
Elevation	1.52**				
	(0.2)				
$Elevation^2$	-0.07**				
	(0.01)				
Rural pop. density	11.12**				
	(0.4)				
Median slope	4.97*				
	(2.48)				
$Median\ slope^2$	-1.54**				
	(0.54)				
Distance town	-0.77**				
	(0.12)				
Distance.capital	0.06				
	(0.08)				
Urbanization	-3.94**				
	(0.23)				
Max urbanization	-0.4**				
	(0.05)				
Length growing period	-1.45				
	(1.23)				
towns_10K_area	-0.12**				
	(0.03)				
Livestock	0.76				
	(0.68)				
Rainfall	2.37*				
	(1.05)				
Soil quality index by Road cost	4.07**				
	(1.34)				
Distance road		-5.92**			
		(1.19)			
Soil quality index by Distance Road		4.03**			
		(0.71)			
Roads			-5.31		
			(4.15)		
Soil quality index by Roads			3.31**		
			(0.61)		
Soil production index				0.29	1.13
				(0.41)	(2.14)
Soil production index by Road cost				12.96**	22.61**
				(2.14)	(6.36)

Table 23: Further robustness checks

robustness1	(1)	(2)	(3)	(4)	(5)
Intercept	55.85**	76.17**	94.62**	58.12**	77.78**
	(5.18)	(1.17)	(2.22)	(0.1)	(0.2)
Soil quality index)	2.48**	6.59**	-2.76		
	(0.44)	(0.77)	(6.01)		
Road cost	-8.58**			0.49	7.29*
	(1.92)			(1.22)	(3.64)
Distance coast	0.95**				
	(0.12)				
River district	1.29				
	(0.78)				
Elevation	1.52**				
	(0.32)				
Elevation^2	-0.07**				
	(0.02)				
Rural pop. density	11.12**				
	(0.93)				
Median slope	4.97				
	(3.24)				
Median slope^2	-1.54*				
	(0.68)				
Distance town	-0.77**				
	(0.21)				
Distance capital	0.06				
	(0.09)				
Urbanization	-3.94**				
	(0.5)				
Max urbanization	-0.4**				
	(0.09)				
Length growing period	-1.45				
	(2.4)				
Towns over 10K inhabitants	-0.12*				
	(0.06)				
Livestock	0.76				
	(0.59)				
Rainfall	2.37				
	(2.09)				
Soil quality index by Road cost	4.07**				
	(1.27)				
Distance road		-5.92**			
		(1.45)			
Soil quality index by Distance road		4.03**			
		(1.28)			
Roads			-5.31**		
			(0.79)		
Soil quality index by Roads			3.31		
			(2.14)		
Soil production index				0.29	1.13
				(1.06)	(2.68)
Soil production index by Road cost				12.96	22.61
				(9.47)	(17.45)