

Mapping Africa's Infrastructure Potential with Geospatial Big Data and Causal ML

Sebastian Krantz*

November 5, 2024

Abstract

Using rich geospatial data and causal machine learning (ML), this paper maps potential economic benefits from incremental investments in all major types of public and economic infrastructure across Africa. These 'infrastructure potential maps' cover all African populated areas at a spatial resolution of 9.7km . They show that the local returns to infrastructure are highly variable and context-specific. For example 'hard infrastructure' such as paved roads and communications is more beneficial in cities, whereas 'social infrastructure' such as education, health, public services and utilities is more critical in rural areas. Market access and agglomeration forces largely govern these returns. The open *Africa Infrastructure Database* built for this project provides granular data classified into 54 economic categories. Its exploration further reveals that Africa's infrastructure is concentrated and often inefficiently allocated. All notable findings are consistent with economic literature, highlighting causal ML's ability to extract insights from geospatial data and assist spatial planning.

Keywords: Africa, infrastructure, investment potential, geospatial big data, causal ML

JEL Classification: O18; R11; R40; C14

1 Introduction

It is well recognized in economic literature and policy discourse that Africa has great public infrastructure deficits. The African Development Bank (ADB) estimates that Africa's infrastructure needs amount to \$130-170 billion (2018 USD) per year, with a financing gap in the range of \$68-108 billion (ADB, 2018). Recent estimates are even higher at \$181-\$221 billion (2024 USD) annual needs (ADB, 2024). An influential World Bank report from 2010 states that Sub-Saharan Africa (SSA) has 31 paved roads km per 100km^2 of land, compared to 134km in other low-income countries, estimates the annual infrastructure gap at \$93 billion, and urges SSA countries to spend one percent of GDP on roads (Foster & Briceño-Garmendia, 2010). Large gaps also remain in other areas; for example, according to World Bank statistics, by 2022, 51% of Sub-Saharan Africans had access to electricity, 36% were using the internet (although 84% had mobile phones), and 34% were using basic sanitation services. These deficits suggest that public infrastructure investments in Africa may have high returns for economic activity and wealth generation.

Much academic research also finds sizeable economic returns to infrastructure—such as historical railway access increasing real agricultural incomes by 16% (Donaldson, 2018) or land values by 60% (Donaldson & Hornbeck, 2016), power cuts reducing firm revenues and producer surplus by 5-10% (Allcott et al., 2016), internet availability inducing 2% higher growth and structural change (Goldbeck & Lindlacher, 2021), joint roads/power investments yielding an 11% increase in welfare (Moneke, 2020), road network inefficiencies implying a 1.3% welfare loss (Graff, 2024), and changes in transport costs invoking large reshufflings of population, wealth, and economic activity (Storeygard, 2016; Jedwab & Storeygard, 2022; Faber, 2014; Baum-Snow et al., 2020). Foster et al. (2023) and Gorgulu et al. (2023) provide qualitative and quantitative reviews of the effects of infrastructure on development outcomes and find sizeable and overwhelmingly positive effects.

*Kiel Institute for the World Economy
Address: Haus Welt-Club, Kiellinie 66, D-24105 Kiel
E-mail: sebastian.krantz@ifw-kiel.de

This paper contributes to the debate on the returns to infrastructure investments by utilizing the wealth of granular geospatial data that has become available in recent years to observationally estimate local returns to different types of public and economic infrastructures across Africa. It first introduces the *Africa Infrastructure Database*, the largest open compilation of categorized infrastructure assets in Africa to date, comprising 15.1 million points/buildings of interest (POIs) and 4.4 million km of network infrastructure classified into 54 economic categories. It then applies recently developed causal machine learning (ML) methods (Chernozhukov, Chetverikov, et al., 2018; Athey et al., 2019; Nie & Wager, 2021), enabling more credible and targeted inferences from observational datasets. These methods, also known as 'double' or 'debiased' ML (DML), employ ML models to remove factors confounding the relationship between a treatment (infrastructure) and an outcome (economic activity or wealth). Following this first-stage 'debiasing' step, a second-stage 'causal model' estimates a heterogeneous treatment effect. With only a large spatial cross-section, identification cannot be established as in dynamic settings. However, I demonstrate causal ML's ability to remove confounding influences, such as political favouritism (Dreher et al., 2019).

The paper is among the first to apply causal ML to large-scale geospatial data. Many papers use ML methods to map poverty (Jean et al., 2016; Yeh et al., 2020; Chi et al., 2022; Lee & Braithwaite, 2022), or improve aid targeting (Aiken et al., 2022). Others use ML to track infrastructure, e.g., road expansion and changes in buildup (Peng & Chen, 2021), or infrastructure quality (Oshri et al., 2018). These advances led to various 'hybrid' impact evaluations. For example Ratledge et al. (2022) use ML to track changes in wealth following electricity grid expansion in Uganda. The impact assessment is done with a difference-in-difference estimator. Conversely, Peng & Chen (2021) track roads and buildup and evaluate the outcomes using nightlights and DHS data among other indicators with a market-access IV approach. Pollmann (2020) uses ML methods to find counterfactual treatment locations to evaluate spatial treatments. As one of the few spatial applications of causal ML, Gilbert et al. (2021) develop a spatial causal inference framework and examine the heterogeneous effects of pollution via DML. To my knowledge, causal ML has not yet been used to estimate economic returns to infrastructure. Most applications of causal ML have been in non-spatial settings, e.g. health (Chernozhukov, Demirer, et al., 2018) and education (Athey & Wager, 2019), business analytics (Huenermund et al., 2021) and marketing (Huber, 2024), sociology (Brand et al., 2023), and increasingly in broader social science (Imbens, 2024).

Why should causal ML be used to evaluate returns to infrastructure—alongside classical impact evaluations and hybrid approaches? One reason is that causal identification is difficult to establish, particularly with infrastructure. Infrastructure interventions are costly, rarely (quasi-)random, and seldomly extend beyond a single type of infrastructure, country or region. Few studies are able to provide local heterogeneity in infrastructure effects (Foster et al., 2023). Limited external validity and differences in data and methodology complicate drawing broad inferences from such studies (Gorgulu et al., 2023). Conflicting studies of the same intervention in different settings also produce conundrums in academic knowledge, exemplified by the RCT literature on household electrification (Lee et al., 2020; Bayer et al., 2020) and some work on rural roads (Asher & Novosad, 2020). In contrast, granular geospatial data has recently become very rich and uniform and is globally available. By analyzing this data with causal ML, I can make a contribution vastly exceeding the scope of econometric research. In particular, I provide localized (1) marginal effects and counterfactual predictions for (2) 14 different types of public and economic infrastructure as well as local and overseas market access, at (3) $9.7km$ ($96km^2$) spatial resolution,¹ for (4) $>100,000$ populated locations across Africa. My results thus exhibit comparability between different types of infrastructure and across space, and suggest that heterogeneity in returns to infrastructure is indeed eminent. These high-resolution 'infrastructure potential maps' thus present an advance toward policymakers asking about the returns to an additional road, generator, school, or hospital in a specific city, village, or suburb. They are limited by the quality of geospatial data for Africa and of spatial measures of wealth/activity, by the identifying assumptions of causal ML,² and by their static partial equilibrium nature. Nevertheless, they represent a flexible empirical advance with potential to inform policymakers' spatial and sectoral infrastructure allocation problems.

¹Which is roughly the size of a medium-sized city, or of the center of a smaller capital city such as Kigali.

²Well known in the literature as the *unconfoundedness assumption*, stipulating the conditional independence of treatment and outcome based on observables. I will present evidence that, in the face of rich geospatial data, this assumption is not easily dismissed via classical arguments such as political favouritism (Dreher et al., 2019).

Apart from estimating marginal returns to infrastructure, granular geospatial data is also useful for gaining deeper insights into Africa's spatial economy. Section 3 of this paper demonstrates that explainable AI (XAI) and nonparametric methods such as PCA and clustering, can help understand the concentration and the local and global economic significance of different types of infrastructure, as well as the spatial economic efficiency implied by current infrastructure allocations. Hall et al. (2022) call for the use of XAI to better understand wealth/poverty prediction models, and in Section 3, I explain the model of Lee & Braithwaite (2022), the current state-of-the-art in Africa.³

The remainder of this paper is structured as follows: Section 2 introduces the detailed geospatial data and describes how it is processed into analyzable form. Section 3 studies the processed data and expounds several stylized facts about Africa's spatial economy. Section 4 introduces the causal ML framework and presents average and heterogeneous partial infrastructure effects. Section 5 does the same for counterfactual predictions. Section 6 summarizes the findings and concludes.

2 Data and Preprocessing

The *Africa Infrastructure Database*⁴ constructed for this project is large and built up from granular data sources. It combines detailed geospatial point and vector data with raster data on population, accessibility, and various economic outcome measures. In the following I introduce these different types of geospatial data, harmonize, and aggregate them into a precise spatial grid.

2.1 Geospatial Data on Infrastructure

The main source of open geospatial data on infrastructure is Open Street Map (OSM). The Africa OSM has expanded rapidly over the past years. The growing reliability of OSM is also reflected in the increased research use of the map. For example, Peng & Chen (2021) use the Zambia OSM in 2019 to train their image segmentation model, and Graff (2024) uses the OSM routing service to generate network connectivity data for his spatial model. As a statistical point of reference, Microsoft Research released a [global dataset of segmented roads](#) in 2022 indicating that globally only 2.3% of roads are missing from OSM, and in Africa around 2.6%. To optimally utilize the map, I develop a basic functional classification of OSM features and apply it to the entire Africa OSM, yielding a database of \sim 12.6 million points/buildings and \sim 3.8 million km of lines of economic interest. An R package *osmclass* (Krantz, 2023) to classify OSM features was built for this task.

Apart from OSM, the Overture Maps Foundation aims to create comprehensive open maps data for developers, with steering members Amazon, Meta, Microsoft, and TomTom. In July 2023, the first production-grade maps dataset was released. With all transport-related data taken from OSM and buildings untagged, the biggest addition is a global dataset of 57 million places of interest (POIs) combined from OSM, Meta, and Microsoft. Within Africa, this adds an additional 1.3 million places mostly from Meta, of which 824k have a minimum confidence score above 0.4. In addition to OSM and Overture, the [Alltheplaces](#) open-source project provides web-scraped POI data, adding 114k POIs mostly from established brands like Starbucks or KFC in Africa.

These sources do not reliably map all infrastructures of economic significance, thus I complement them with several curated datasets covering specific infrastructures. In particular, I assemble data on cell towers ([OpenCellid](#)), health facilities managed by the public sector (Maina et al., 2019), power plants ([Global Integrated Power Tracker](#) and [WRI database](#) (Byers et al., 2018)), steel plants ([Global Steel Plant Tracker](#)), special economic zones (SEZs) ([Open Zone Map](#)), and ports (2015 World Port Index (MSI, 2019) and [World Bank](#)). Table 1 summarizes the places dataset by source.

While the *osmclass* R package provides a basic classification of features closely aligned with the OSM tagging system, and curated datasets have a limited number of functionally distinct features, the Overture places (OVP) data has $>$ 1000 primary place categories. I thus perform a second, more rigorous, classification step to categorize features across sources into 47 specific

³This explanation, based on an XGBoost model and SHAP values (Lundberg et al., 2020), is credible as Lee & Braithwaite (2022) also use XGBoost on many of the same features (alongside a CNN) to predict the wealth index.

⁴Available at <https://drive.google.com/drive/folders/1hpROhpjQ3UHzOTYvzPwnJdEs5dpZP584?usp=sharing>

Table 1: Africa Infrastructure Database: Places Dataset by Source

Source	Count	of which Polygons	Categories
Open Street Map (OSM)	12,221,198	9,038,206	45
OpenCellid (Cell Towers)	1,894,356	0	1
Overture Maps Places (confidence > 0.4)	823,786	0	44
All The Places (Open Web-Scraped POIs)	114,382	0	12
Health Facilities in SSA (Nature Scientific Data)	96,290	0	1
Global Integrated Power Tracker	871	0	1
WRI Global Power Plants Database	363	0	1
Global Steel Plant Tracker	41	0	1
Open Zone Map (Special Economic Zones)	387	0	1
World Port Index	235	0	1
WorldBank Global International Ports	9	0	1
SUM	15,151,918	9,038,206	47

Notes: Table shows places of interest (POIs) data collected from different sources. In OSM POIs may be tagged buildings/have geometries.

economic categories and 26 simplified ones. The full classification process is detailed in Appendix A. Appendix Table A3 shows the final harmonized classification.

Following harmonization, I also deduplicate the data across sources by allowing only features of the same category from one source within a 10m radius.⁵ Curated datasets thereby take precedence over OSM, which in turn supersedes OVP. Tables 1 and A3 summarize already deduplicated data.

In addition to POIs, I extract linestring features such as (non-residential) roads, larger waterways, power lines, railways, aeroways, pipelines, and telecommunication lines from OSM. I complement OSM power lines with electricity grid maps from the European Commission ([Kakoulaki & Moner-Girona, 2020](#)) and the [World Bank](#). Table 2 provides a breakdown. In Total, I collect ~ 4.4 million km of network infrastructure, of which ~ 1.6 million km are roads, ~ 1.5 million km are waterways, and 967 thousand km are power lines. The other categories sum to 272 thousand km.

Table 2: Africa Infrastructure Database: Lines Data by Category

Category	Count	Length (Km)
road	763,912	1,621,144
waterway	359,756	1,507,112
power ¹	1,013,150	967,317
railway	84,707	128,408
aeroway	27,360	11,019
pipeline	9,453	55,394
storage	8,551	389
ferry	2,412	48,259
aerialway	171	175
telecom	87	28,682
SUM	2,269,903	4,379,392

¹ OSM power lines were combined with datasets from the EC's Joint Research Centre (JRC) and the World Bank.

Lastly, I also obtain 2022 fixed and mobile download and upload speeds from [OOKLA](#) via the [EU Africa Knowledge Platform](#) - within map tiles at a resolution of around 610m at the equator.

2.2 Spatial Measures of Wealth and Economic Activity

To study the returns to infrastructure, accurate spatial measures of quantities of interest such as economic activity/value-added or household wealth are needed. A popular spatial proxy for economic activity, following the seminal work of [V. Henderson et al. \(2011, 2012\)](#), is remotely sensed nightlight luminosity ([Donaldson & Storeygard, 2016](#)). Since 2011, nightlights data is available at high resolution (15 arc seconds or ~ 500 m at the equator) from the Visible Infrared Imaging

⁵This is done by shifting a 10m grid over the POI features in steps of 1m and deduplicating features within each 10m \times 10m square, thus there is some path dependence in terms of which POIs are compared first.

Radiometer Suite (VIIRS) onboard the Suomi satellite (Gibson et al., 2020). Recently, NASA's 'Black Marble' product offers a more processed version of the VIIRS imagery for monitoring human activities (Román et al., 2018). Early adopters Peng & Chen (2021) show that this data is substantially more accurate than the conventional VIIRS data in tracking Zambian GDP over time.

Several contributions also spatially distribute GDP (value-added) from national and/or regional accounts via high-resolution data on population, geophysical features, and nightlights. The widely used G-Econ database (Nordhaus et al., 2006) provides global GDP estimates for 1-degree grid cells from 1990 to 2005. More recently, Kummu et al. (2018) distribute national GDP in constant 2015 PPP dollars at 5 arc-min (0.0833 degrees or 9.3km at the equator) resolution, using subnational value-added estimates from Gennaioli et al. (2013) and population from the HYDE 3.2 database.

Recent efforts, starting with Jean et al. (2016), also use richer data sources to predict wealth at high spatial resolutions. Notably, Chi et al. (2022) combine vast and heterogeneous data from satellites, mobile phone networks, topographic maps, as well as connectivity data from Meta to estimate nationally comparable estimates of wealth - a Relative Wealth Index (RWI) - for all low and middle-income countries at 2.4km resolution. Focussing on SSA, Lee & Braithwaite (2022) develop a cross-country prediction methodology combining day- and nighttime satellite imagery, high-resolution population estimates, and OSM to predict the International Wealth Index (IWI) - a comparable asset-based wealth index calculated from DHS Surveys for 25 SSA countries since 2017 - for 929,295 populated places in 44 SSA countries at 1-square mile resolution.⁶ They obtain a cross-country R^2 of 91.7% for the IWI, outperforming all previous research results.

Appendix Figure A1 shows the wealth/activity estimates by Román et al. (2018), Kummu et al. (2018), Chi et al. (2022), and Lee & Braithwaite (2022). None of these measures is ideal to study the returns to infrastructure. Nightlights are, by definition, correlated with power infrastructure and also relatively sparse since very low-light areas are set to 0 in the Black Marble product. Gridded GDP is, by definition, highly correlated with population and may thus be biased towards residential areas. The RWI is not constructed to be comparable across countries and is not available for (South-)Sudan, whereas the IWI is not available for North Africa and uses parts of OSM and population in its construction. In the following, I use all 4 estimates shown in Figure A1 to determine weights applied during the aggregation of granular data, but focus on the IWI for final estimation since it is an accurate high-resolution and cross-country comparable estimate. I also conduct robustness exercises with nightlights and ground truth IWI estimates from DHS surveys conducted since 2010 to ensure that key results are not driven by the ML model of Lee & Braithwaite (2022).

2.3 Covariate Raster Layers

Many further sources of high-resolution data layers about geophysical features, agriculture, climate, and conflict could be included as covariates in an analysis of infrastructure and wealth/economic activity. But economic research such as Storeygard (2016), Jedwab & Storeygard (2022), Donaldson (2018) and Peng & Chen (2021) has focussed on two particularly important dimensions of spatial variation affecting economic outcomes: population (urbanization) and market access.

I obtain population estimates for 2020 from the Gridded Population of the World Version 4 (GPW4) project (CIESIN, 2016), which is based on administrative data. I broadly distinguish between the infant (0-14 years) and working-age (15-49 years) population to allow for local variation in demographic characteristics. To approximate market access, I consider global accessibility indicators from Weiss et al. (2018), who develop a global map of travel time (in minutes) to cities with more than 50,000 people in the year 2015 at 1 km^2 resolution. The map is based on a global friction surface constructed from detailed spatial data on transport networks and geophysical features. Nelson et al. (2019) expands this work to settlements of 9 different sizes, from towns of 5000 inhabitants to megacities with more than 5 million inhabitants. Nelson (2022) further computes travel time to ports of 4 different sizes (very small, small, medium, and large) using data from the 2015 (26th) edition of the WPI. From these 12 accessibility maps, I compute 4 which appear most

⁶From OSM, Lee & Braithwaite (2022) employ the total length of roads, distance to the closest road, number of junctions, distance to the closest junction, total building area, and the number of buildings for each 1 square-mile populated area, and the number of and distance to 24 locations of interest such as schools, hospitals, and markets.

relevant in Africa: (1) travel time to cities >50,000 as in Weiss et al. (2018); (2) travel time to cities >1 million; (3) travel time to the nearest port, and (4) travel time to one of 43 medium or large African ports.⁷ Figure A3 shows 3 of these accessibility maps and total GPW4 population.

2.4 The Ideal Spatial Grid

Jointly analyzing infrastructure and wealth/activity requires spatial binning and data aggregation. For accurate spatial analysis, an equal area grid is desirable. Discrete Global Grid Systems (DGGS) enable this via hierarchical tessellation of cells partitioning the globe. Sahr et al. (2003) propose the Icosahedral Snyder Equal Area Aperture 3 Hexagon (ISEA3H) as a good general-purpose geodesic DGGS.⁸ ISEA3H is available at 31 different resolutions, from 12 global cells spaced 7054km apart to 2059 trillion cells spaced 0.5m apart. High spatial resolutions are desirable but increase the computational burden and reduce the number of features in each cell, limiting statistical models' ability to learn about the spatial economy. I thus empirically gauge the highest resolution grid that still yields acceptable predictions for wealth/economic activity by counting POIs in each cell and category and computing the average correlation of these counts with the 4 indicators in Figure A1. I also compute the average R^2 of linear models predicting the wealth/activity indicator from all category counts. Appendix Table A8 reports the results for ISEA3H grids of 7 different resolutions, ranging from 3,901 cells 87km apart down to 557,766 cells 3.2km apart.

Both individual and joint predictions become less accurate with increasing grid resolution. The largest drop occurs when moving from a resolution 11 grid (16.8km) to a resolution 12 grid (9.7km). Resolution 12 cells have a size comparable to the city center of Kigali (a larger city like Kampala being covered by 3-4 cells) and contain 94 POIs on average (10 in the median cell). To enable high-resolution estimation for city centers and suburban regions, I opt for the resolution 12 grid and mitigate the drop in predictive performance and the effects of hard cell borders by allowing spatial spillovers from up to 2nd-order neighbours. The implementation of these spillovers is described below. Figure 1 visualizes the grid with GPW4 2020 total population estimates.

2.5 Data Aggregation

I first aggregate the raster data over the 96 km^2 ISEA3H grid by taking the mean of wealth indices, travel times, and internet speed, and the sum of GDP, nightlights, and population within each cell. I also compute the total length in m of network features (Table 2) per cell, distinguishing paved from unpaved roads and combining 3 types of waterways using average harmonized coefficients from a Ridge Regression against the outcomes in Figure A1.⁹ I count POIs in each cell and category (Table A3), but also consider a weighted approach where I compute quartiles of the building-areas of all OSM features tagged to buildings, and use counts of 2/3/4 if the area is within the 2nd/3rd/4th quartile. In this way, large features such as large school buildings receive up to 4 times the weight of small school buildings or schools that are just points. Similarly, I also apply these quartiles counts to the areas of SEZ's, the capacity of power and steel plants, and the outflow of ports. The quartile method thus takes into account the intensive margin of features.

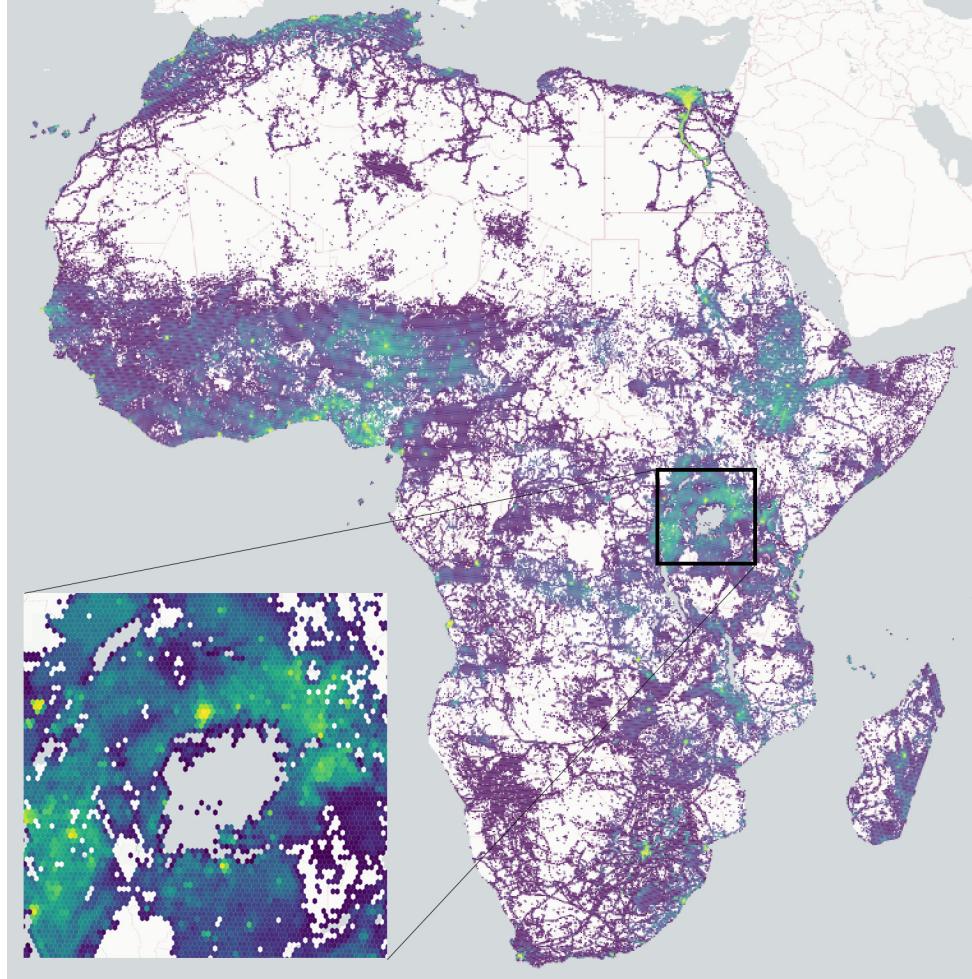
To aggregated the 47 detailed categories in Table A3 into the 26 simplified ones, which are more independent and thus more useful for analysis, I employ weights derived from penalized regressions predicting the 4 outcomes from Figure A1. Formally, let there be P_j detailed categories for simplified category j indexed by p . For example the 'communications' category combines $P_j =$

⁷This choice of accessibility maps is sensible as diversified economic activity in SSA tends to take place in the largest urban centers, of which most countries have one or two. Towns of 50,000 people often function as important hubs to gather agricultural produce from the region for sale on local markets or transport to larger cities and ports (Jedwab & Storeygard, 2022). The vast majority of exporting or importing in Africa also happens through medium and large-sized ports, with modern container terminals and often intersecting international shipping routes.

⁸An aperture of 3 implies that hexagon areas decrease by a factor of 3 as grid resolution increases. The grid is implemented in the DGGRID C++ library (Sahr, 2022), and accessed via the 'dggridR' package (Barnes, 2020).

⁹Namely rivers, man-made canals, and man-made waterways relating to utilities or agricultural activities such as drains and ditches, which are initially summed into a variable called 'waterway_other'. I then use Ridge Regression to estimate an equation of the form $y = \beta_0 + \beta_1 + \text{river} + \beta_2 + \text{canal} + \beta_3 + \text{waterway_other} + \epsilon$, where y is the log of GDP, the IWI or the RWI (nightlights is too sparse). The coefficients are restricted to be greater than zero, and the optimal Ridge penalty is chosen by 10-fold cross-validation. I then obtain relative coefficients by dividing through by β_1 , and averaging them across the 3 outcomes. The result is $\beta_1 = 1$, $\beta_2 = 9.035$, and $\beta_3 = 6.13$, indicating that man-made features are much more important for spatial activity. Their length is thus increased by a factor β_j .

Figure 1: GPW4 2020 Population in 160,719 96km² ISEA3H Cells Covering Areas with any POI



Notes: Figure shows GPW4 population summed over an ISEA3H 96km² discrete global grid covering areas with any POI.

3 detailed categories: 'communications_network' (cell towers, antennas), 'communications_other' (TV or radio station, newspaper, publisher), and 'telecom_len', which is the length (in m) of OSM telecommunications lines in each cell. I combine these detailed categories into a simplified one (\mathbf{x}_j) using appropriate linear weights β_{jp} maximizing the correlation with the outcomes (\mathbf{y})

$$\max_{\beta_{jp}} \text{cor}(\mathbf{x}_j, \mathbf{y}) \quad \forall j \quad \text{s.t.} \quad \beta_{jp} \geq 0 \quad \text{where} \quad \mathbf{x}_j = \sum_{p=1}^{P_j} \beta_{jp} \mathbf{x}_{jp}. \quad (1)$$

To prevent overfitting and negative coefficients β_{jp} , this problem is solved using a Ridge Regression, restricting β_{jp} to be positive, and choosing the optimal penalty parameter (λ^*) via 10-fold cross-validation. The resulting coefficients β_{jp} are normalized by the coefficient of the most populous category ('communications_network'), and surprisingly consistent across outcomes. I thus average them to compute final weights. Appendix Table A6 provides three examples. For 'communications', the coefficient on 'communications_other' is 11.7 and the coefficient on 'telecom_len' is 0.1, which are sensible in relation to a cell tower ('communications_network') having a weight of 1.

Finally, I account for economic geography and mitigate cell-border effects by creating additional 'spillover' variables (i.e., spatial lags) as inverse-distance-weighted average of neighbouring cells

$$\mathbf{x}_j^{\text{neigh}} = \sum_{i \neq j} \frac{\mathbf{x}_i}{\delta_{ij}} / \sum_{i \neq j} \frac{1}{\delta_{ij}} \quad \forall i \quad \text{where} \quad \delta_{ij} < \tau. \quad (2)$$

I choose $\tau = 24.2\text{km}$, which includes all second-order neighbours. The results are robust to the absence of spillover variables and also to using simple counts instead of quartile counts, but richer processing results in increased predictive power and spatial lags help limit confounding influences.

Appendix Table A4 shows summary statistics for the final gridded dataset (simple counts). Appendix Figure A2 additionally shows histograms of the aggregated wealth/activity measures, and Appendix Table A5 shows pairwise Pearson's correlations of these measures aggregated over the grid. All measures are moderately correlated but follow slightly different distributions.

3 Africa's Spatial Economy

With very rich data on infrastructure, population, and wealth/activity in Africa at hand, I start off by examining Africa's spatial economy and the current allocation of infrastructure. I first analyze the overall allocation and concentration of infrastructure(s), and zoom in on 5 African capital cities to uncover urban heterogeneity. I then examine the spatial clustering of different infrastructures and create an index of spatial efficiency measuring the proximity of residential areas, core infrastructures, and clusters of economic activity. I find that this index is correlated with development indicators, logistic performance, and Graff (2024)'s road network inefficiency measure. Finally, I predict the IWI using ML models and interpret them with XAI methods, yielding a global and local characterization of important infrastructure predictors of wealth.

3.1 Spatial Concentration

To study spatial concentrations, I take 88,960 grid cells with more than 10 inhabitants/ km^2 ¹⁰ according to both GPW4 and WorldPop 2020¹⁰ estimates, count POIs in simplified categories and compute empirical CDFs. Appendix Figure A4 shows the results, both for individual feature categories, some of which are highlighted in colour, and for averages across categories, computed before or after the CDF calculation. GDP and GPW4 population are also included as a reference.

The top panel reveals that POIs in Africa are highly concentrated - more than population and, for most categories, GDP. The top 1000 cells (1.12% of 88,960) account for 62% of POIs in the average feature category, but only 49% of total GDP and 27% of total population. Only education and power infrastructure are less concentrated than GDP. The average CDF suggests that close to 100% of any given infrastructure is allocated in less than 10,000 cells. The only widely mapped feature present in nearly all cells is residential buildings. When infrastructure is pooled across categories, it is less concentrated, and the top 1000 cells only account for 30% of infrastructure. If residential buildings, farmland, power, and construction are excluded, this share rises to 66%. Excluding these four categories also yields 21,891 populated cells (24.6%) that have no other POI.

The bottom panel shows analogous results for line (network) features, indicating that roads, waterways, power, and railways are less concentrated than GDP, and, in the case of unpaved roads and waterways, also than population. The top 1000 cells only account for 10% of waterways and unpaved roads, 17% of power lines, 23% of paved roads, and 40% of railways. Concentration also proceeds gradually: the top 3000 cells account for 20% of unpaved roads and waterways, 39% of paved roads and 70% of railways, and the top 10,000 cells account for 40% of unpaved roads and 70% of paved roads. Excluding waterways yields 27.3% of populated cells with no other line feature.

To also provide a spatial overview, Appendix Figure A5 plots the spatial allocation of 9 critical infrastructures on a log10 scale. The top panel shows paved roads, power, and communications, indicating large gaps in central Africa, the Sahel, and the Horn of Africa regions. The great lakes region zoomed in in these plots is well connected in the central populated areas, but still lacks connectivity in sparsely populated peripheries such as eastern Congo and Northern Kenya. The middle and bottom panels of Appendix Figure A5 show education and health facilities, public services and utilities (excl. power), automotive facilities, public transport, and financial services. Exempting education and health facilities, these features are largely concentrated in urban areas.¹¹

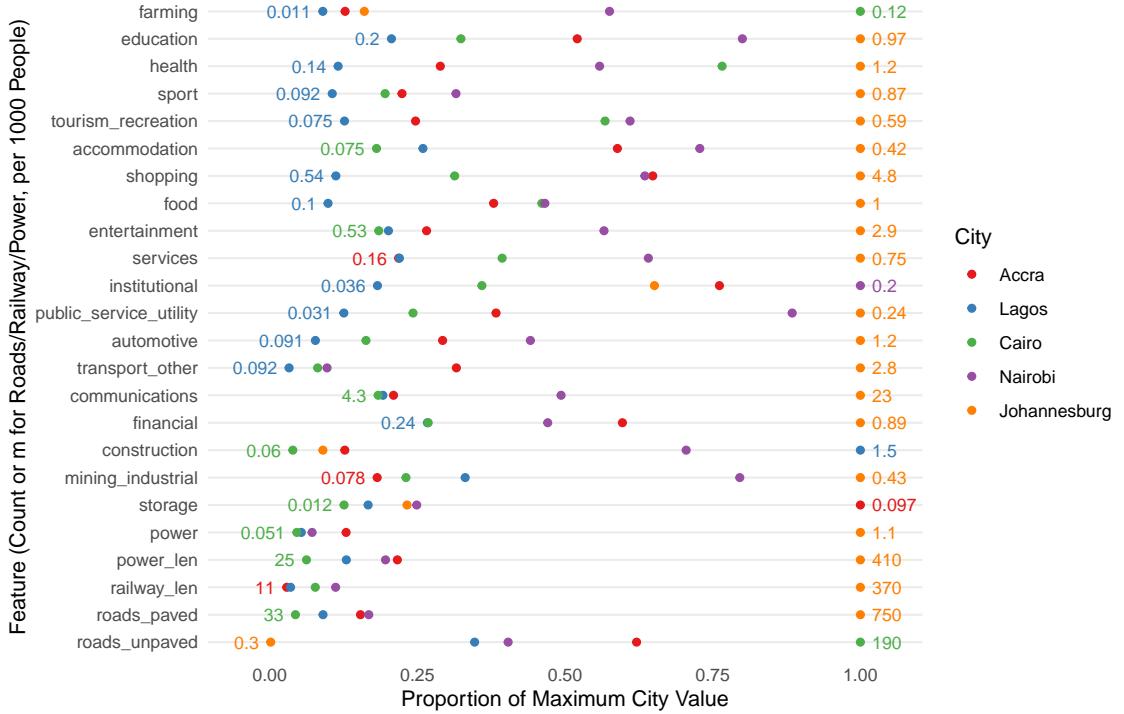
¹⁰WorldPop (<https://www.worldpop.org/>) provides high-resolution ($1km^2$) global population estimates using geospatial big data (including infrastructure) and ML models to distribute administrative estimates. Thus, it is locally more accurate than GPW4 data based solely on administrative sources, but ultimately a function of infrastructure and thus only used for descriptive analysis in this paper.

¹¹Some of these plots evidence differences in data coverage; for example, Tanzania and Uganda both have a very large amount of educational facilities, averaging around 1 facility per 1000 people, suggesting that OSM mappers have been more active in these countries. Health seems more balanced across countries partly due to the data contribution by Maina et al. (2019). Public services (e.g., post-office), automotive (e.g., gas station) and public

3.2 Urban Heterogeneity

To uncover heterogeneity in urban areas, I compare 5 large African cities: Accra, Cairo, Lagos, Nairobi, and Johannesburg. I take 7 hexagons covering the central parts of each city, 692km^2 in total, compute the counts of POIs and the length of lines within these areas and divide them by the WorldPop 2020 population estimate. Figure 2 shows the feature intensities per 1000 people.

Figure 2: Feature Density in 5 Major African Capital Cities per 1000 People



Notes: Figure shows feature counts per 1000 inhabitants (WorldPop 2020 estimates) for 5 significant African cities. For each city, seven 96km^2 hexagons covering essential parts of the city are considered. Features are counted in detailed categories and then combined into simplified categories using the weighted aggregation procedure described in Section 2.5.

Ostensibly, the cities are very heterogeneous. Johannesburg leads in most feature categories, providing 750m of paved transport roads, 370m of railway, and 410m of power lines per 1000 inhabitants, as well as significantly higher automotive, other (public) transport, communications, education, and health infrastructure per capita than the other cities. Conversely, Lagos lags in many categories, providing less than a quarter of the services per capita than Johannesburg. Between these two, Nairobi and Accra are performing well, with Nairobi having the most institutions par capita alongside high densities of education, public services, industrial facilities, construction, accommodation (hotels), services, tourism and recreation. Accra also has a high level of institutions, shopping, financial services, accommodation, and education facilities per capita. It is noteworthy that, apart from Johannesburg, the level of infrastructure per capita in these cities does not align with their IWI (Accra: 73.5, Lagos: 67.4, Nairobi: 64.2, Cairo: 70.7 [RF prediction], Johannesburg: 81.6) or GDP per Capita¹² estimates, with Lagos and Cairo relatively underperforming. Infrastructure concentration in Africa is thus a complex function of agglomeration economies, economic development and institutions, history, and urban planning.

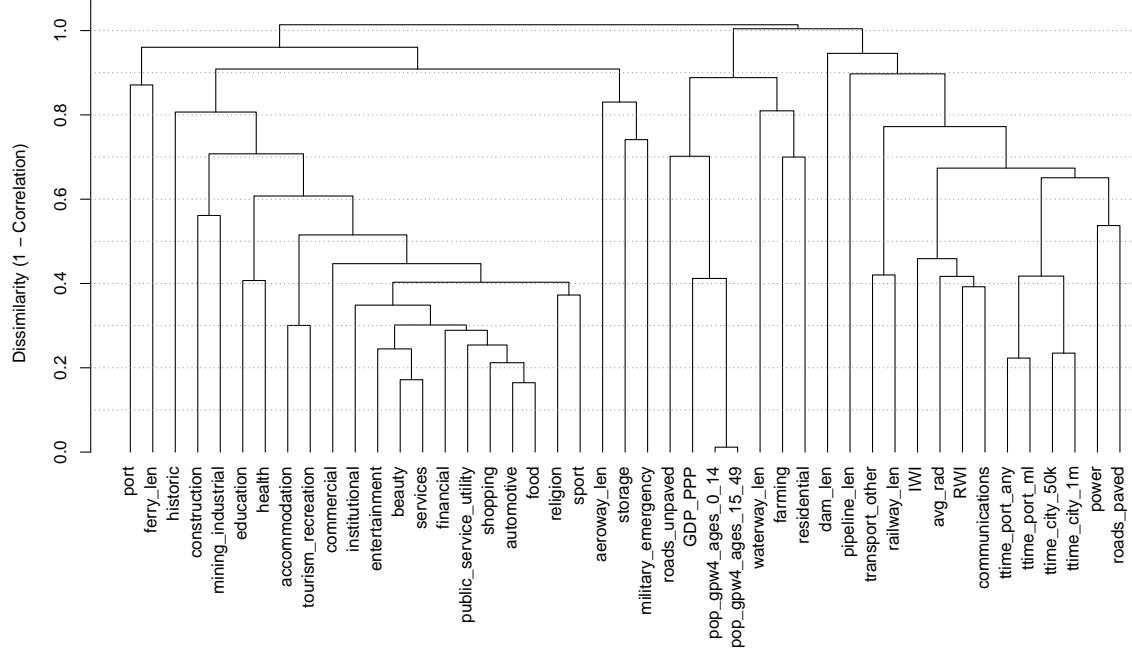
transport (e.g., train/bus station) facilities as well as financial services (banks/ATMs) are more consistently mapped across countries and largely concentrated in cities/towns. Public transport facilities are concentrated along railway lines in some countries. In general, OSM is far from perfect, and national differences in data coverage may entail empirical problems which econometrically can be alleviated via fixed effects. In a ML setting, this amounts to including country dummies (One Hot Encoding). Practically, I find that such dummies do not alter the results but significantly increase training times, and hence omit them. The flexible ML approach with high-dimensional data (including spatial lags) appears to deals gracefully with spatial differences in mapping intensity. Yet, comprehensive geospatial information would increase the reliability and robustness of the results. Commercial data providers like Google or Dataplor may offer significantly improved coverage, but are too expensive to purchase at scale.

¹² Accra: 3953, Lagos: 7872, Cairo: 9672, Nairobi: 10439, Johannesburg: 14317 in 2015 USD PPP. These estimates are based on coarse administrative data by Gennaioli et al. (2013), scaled by Kummu et al. (2018).

3.3 Spatial Clustering

Still more can be learned about infrastructure concentrations in Africa by clustering features using spatial correlations—to determine which types of infrastructure cluster together and which are often found in populated and/or high-income locations. I take the natural log of feature counts in the simplified classification, compute Pearson’s correlations among all variables, and use one minus the correlations as a distance matrix for hierarchical clustering with complete linkage. Figure 3 shows a dendrogram, and Appendix Figure A6 the corresponding clustered correlation matrix.

Figure 3: Hierarchical Clustering of Variables using Correlation and Complete Linkage



Notes: Figure shows dendrogram from hierarchical clustering with complete linkage using a correlation-based distance metric, see also Appendix Figure A6. Variables in simple counts correspond to Appendix Table A4 and are cast in logs before computing correlations, except for the IWI and RWI which are included in levels. Also, travel time estimates (in minutes) are first log-transformed and then negated to yield positive correlations with other variables.

Different dissimilarity cutoffs reveal distinct groups of correlated variables. Notably, setting the cutoff around $h = 0.98$ reveals three prominent groups in the dendrogram, corroborated by the correlation matrix (Figure A6). The group on the RHS includes travel times, household wealth, nightlights, power infrastructure, paved roads, and communications. These variables seem to be a proxy for physical infrastructure, which in turn correlates highly with nightlights and wealth. The second group, comprising most variables on the LHS of the dendrogram, includes economic activities in the broadest sense. Finally, the middle cluster includes population, residential areas, farming, and GDP, which is interpolated across space using population data. The dendrogram and correlation matrix (Figure A6) thus suggest spatial disparities between where people live, where they work, and where most physical infrastructure is located. Presumably, more efficient spatial organizations imply stronger spatial correlations among these three groups of features.

3.4 Spatial (In)Efficiency

Informed by these observations, I compute an index of spatial efficiency (ISE) along 3 dimensions: the GPW4 working age (15-49) population, the first principal component (PC1) of paved roads, power, and communications as a compound measure of (hard) infrastructure, and the PC1 of education, institutional, health, religion, public_service_utility, food, shopping, beauty, services, commercial, mining_industrial, tourism_recreation, sport, construction, farming, entertainment, financial, and accommodation as a compound measure of (broadly conceived) economic activity. The PC1 of roads, power, and communications accounts for 63% of their joint variance, and the PC1 of the activity variables captures 56% of their joint variance. Table 3 shows Pearson’s correlations among these components. The ISE is then computed as the geometric mean of these correlations

Table 3: Correlations of ISE Dimensions

$N = 160,499$	P	I	A
GPW4 POP 15-49 (P)	1.000		
Roads & Power PC1 (I)	0.589	1.000	
Economic Activity PC1 (A)	0.634	0.708	1.000

Notes: Table reports Pearson's correlations among dimension indices (PC1) and population. Their geometric mean is the ISE (Eq. 3).

$$\text{ISE} = \text{cor}(P, I)^{\frac{1}{3}} \text{cor}(P, A)^{\frac{1}{3}} \text{cor}(I, A)^{\frac{1}{3}} = 0.642. \quad (3)$$

It is thus an index on the range [0, 1], a value of 1 indicating perfect spatial efficiency with all productive resources concentrated in the same locations. This is practically unachievable, but an all-Africa ISE of 0.64 suggests much room for improvement. Table 4 shows country-level ISE estimates, computed from cells in each country, and Appendix Figure A7 a corresponding map.

 Table 4: Country-Level ISE Estimates ($\mu = 0.646$, $\sigma = 0.193$)

#	ISO3	ISE								
1	SYC	0.950	TUN	0.788	ZAF	0.719	COD	0.624	BWA	0.508
2	MLI	0.908	MAR	0.783	BFA	0.717	MDG	0.621	LBY	0.486
3	MUS	0.889	TGO	0.779	SWZ	0.715	DJI	0.619	LBR	0.440
4	UGA	0.868	EGY	0.779	GMB	0.712	BDI	0.617	NAM	0.429
5	RWA	0.860	BEN	0.774	GAB	0.694	CAF	0.612	ESH	0.410
6	KEN	0.849	MWI	0.763	NGA	0.693	MOZ	0.593	ERI	0.360
7	GHA	0.830	LSO	0.750	ZWE	0.688	AGO	0.573	SOM	0.284
8	CIV	0.822	TZA	0.745	GIN	0.686	SDN	0.568	CPV	0.252
9	SLE	0.813	ETH	0.744	MRT	0.675	CMR	0.564	COM	0.214
10	STP	0.812	DZA	0.741	COG	0.672	TCD	0.543	GNQ	0.153
11	SEN	0.809	NER	0.729	ZMB	0.651	GNB	0.542	SSD	0.105

Notes: Table reports sorted country-level ISE estimates (Eq. 3) computed from cells within each country.

The country-level ISE estimates are mildly correlated with key development indicators in 2020, such as GDP per Capita PPP ($r = 0.159$), Life-Expectancy at Birth ($r = 0.215$), and the Human Development Index ($r = 0.185$). Interestingly, they show stronger correlations with the 2018 Logistics Performance Index ($r = 0.396$) and the 2020 Doing Business Index ($r = 0.592$). They are also negatively correlated to the hypothetical welfare gains (in percent) from an optimal reallocation of the road network in each country as calculated by Graff (2024) ($r = -0.360$). The stronger association of the ISE with these indicators vis-a-vis development outcomes suggests that it indeed measures spatial (in)efficiency. The index is uncorrelated with total land area ($r = -0.033$), although several small states like Seychelles, Mauritius, and Rwanda score particularly high.

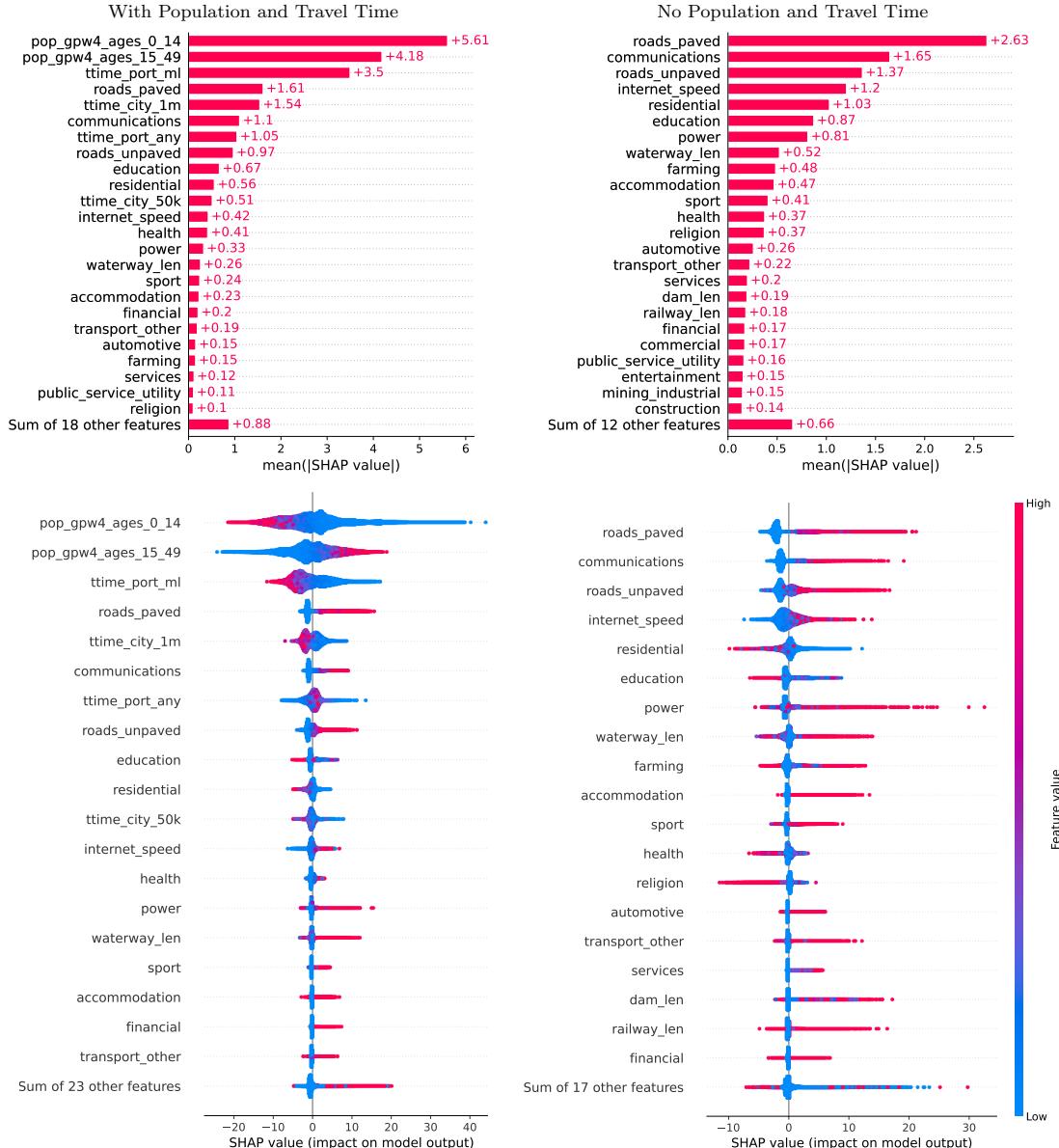
3.5 Wealth Prediction and Interpretation

Given the rich nature of Africa's spatial economy, a ML approach to predict wealth from infrastructure that is able to capture non-linear associations in the data can yield further insights. Borisov et al. (2021) show that gradient-boosting machines (GBMs) (J. H. Friedman, 2001) still outperform most deep learning methods on tabular data. Thus, I employ the competition-winning XGBoost algorithm (Chen & Guestrin, 2016) and tune its hyperparameters with Optuna (Akiba et al., 2019) on a test set containing 25% of the data.¹³ The IWI model trained with early stopping achieves a test set R^2 of 78.8%. For comparison, I also train a Random Forest (RF) (Breiman, 2001) model with default parameters and 1000 trees, which achieves a test set R^2 of 73.3%. Appendix Figure A8 shows empirical CDFs of the absolute values of the residuals, indicating that the XGB model predicts 76% of test-set observations with an error of less than 5 IWI points.

¹³The optimal hyperparameters generally feature a low learning rate ($\eta = 0.01 - 0.02$), deep trees ($\text{max_depth} = 9 - 15$), significant randomization over samples ($\text{subsample} = 0.5 - 0.8$) and significant regularization, especially through constraints on the minimal size of final nodes ($\text{min_child_weight} = 6 - 20$).

To accurately attribute predictions across the different variables, I compute Shapely Values, a game theoretic approach to fairly attribute the contribution of variables to a single prediction. In particular, SHAP (SHapley Additive exPlanations) following Lundberg & Lee (2017) give additive variable contributions for each instance that sum to the difference of the prediction from the average model prediction. I use the TreeSHAP algorithm of Lundberg et al. (2020) to compute (interventional) SHAP values for two different XGBoost models: the model trained on the full dataset evaluated above, and a model excluding population and travel time estimates. The latter removes strong correlations of these variables with wealth. The top panel of Figure 4 summarizes overall variable importance for the IWI based on the average absolute SHAP value across instances.

Figure 4: (Average) SHAP Values for XGBoost Models Predicting the IWI



Notes: Figure shows (average absolute) interventional SHAP values for XGBoost models following Lundberg et al. (2020) (i.e., using the TreeSHAP algorithm to compute SHAP values on tree-based ensemble models like XGBoost efficiently). Shapely Values are a game-theoretic approach to fairly attribute ML model predictions to the predictors (variables) at the instance level. SHAP values quantify such attributions as the contribution of each predictor to the difference between the prediction and the average model prediction across all instances. The average absolute SHAP value across instances (top panel) thus gives a precise summary of a predictor's global significance, whereas plotting individual SHAP values using a Beeswarm plot and colouring them by the feature level (bottom panel) compactly summarizes the direction of the effect (which may vary non-linearly with the predictor level, more clearly visible in scatterplots as in Appendix Figure A14).

Figure 4 suggests that, apart from population and travel time to major cities and ports, roads, communications, education, power infrastructure, and residential buildings are the most significant

predictors of wealth. Many further features such as accommodation (hotels), sport, health, automotive, and public transport facilities are also important. With nightlights as outcome (Appendix Figure A10), population, travel time, communications, roads, and power remain important, but sports, automotive facilities, and industrial areas rise in the ranking, being large emitters of light.

To gauge the direction of the effects, which may be heterogeneous as SHAP values are computed at the instance level and tree ensembles may be highly non-linear, the bottom panel of Figure 4 provides a beeswarm plot of the SHAP values coloured by feature intensity. The plot suggests that most variables have the intended effect, with higher intensity translating into larger SHAP values (predictions). Defying the intuition, residential buildings, education, health, and religion have largely negative effects. This spatial pattern needs to be cautiously interpreted *ceteris paribus*, i.e., in the presence of other correlated features often found in wealthy urban locations (e.g., paved roads, power, communications), more education and health facilities may decrease model predictions. Appendix Figure A11 shows the same plot for nightlights, with similar results.

Appendix Figures A14 and A15 additionally provide detailed scatterplots between feature intensity and SHAP values for key predictors. They are particularly useful towards detecting thresholds where a feature's effect on the prediction begins to change. For example, the SHAP value of paved roads strongly increases above $10^4 = 10\text{km}$ in a cell, suggesting that roads are more important for prediction in urban areas. The same applies to power beyond a threshold of $10^{2.5} \approx 300$ facilities (e.g., transformers, generators) per cell. With education, the opposite is the case: up to 10 facilities per cell, the SHAP value is high, but beyond that, it reduces, indicating that in central urban spaces, the presence of schools decreases wealth predictions. Health and religious facilities exhibit similar but weaker dynamics. In contrast, communications (mainly cell towers) have an almost log-linear positive effect on model predictions. In models with population, the infant (0-14) population has a strong negative effect on predictions, whereas the adult population (15-49) has a strong positive effect, implying significant Malthusian dynamics in the data.

4 Estimating Marginal Infrastructure Benefits

Having explored the data in some depth, this section advances by asking about the marginal effects of infrastructure on wealth/activity. In the absence of a viable identification strategy (such as RCT, IV, RDD, DID) to causally identify infrastructure effects at spatial scale and across different categories, I use observational causal inference-causal ML—to approximate such effects.

The traditional econometric view, still held to some extent in other disciplines, is that if one can control for the most important confounding factors in an observational setting, a careful *ceteris paribus* interpretation of the partial effect is possible. So far, this premise has been applied almost exclusively in the context of linear regression, implying that if all confounders are observed and the population model is linear-additive, a careful *ceteris paribus* interpretation is possible.

However, the *ceteris paribus* statement need not be that strong, as one can relax the assumptions of linearity and additivity. The premise of causal ML (also known as 'double' or 'debiased' ML) is that if one observes all factors that confound or proxy for confounding influences, the relationship between treatment and outcome can be specified conditional on an optimal ML prediction of both from observables. In the setting at hand, this means that if Africa's spatial economy is sufficiently observed through the available granular data on infrastructure, POIs, population, and market access, and if appropriate ML models are deployed, it may be possible to identify marginal partial-equilibrium effects of individual infrastructures on wealth/economic activity. Before examining these identification assumptions in more detail, I introduce this estimation strategy more formally.

I follow Nie & Wager (2021) and Chernozhukov et al. (2017); Chernozhukov, Chetverikov, et al. (2018), and adopt some notation from Hirano & Imbens (2004). Let Y be an outcome of interest, W a continuous treatment of interest (a specific infrastructure) with possible values $\omega \in \Omega$, and $\mathbf{X} = [\mathbf{X}'_H, \mathbf{X}'_C]'$ be a vector of observed confounders, where \mathbf{X}_H includes covariates that also affect treatment effect heterogeneity. The *unconfoundedness assumption* is then formally stated as

$$Y(\omega) \perp W \mid \mathbf{X} \quad \forall \omega \in \Omega, \tag{4}$$

such that the potential outcome $Y(\omega)$ is independent of the treatment realization $W = \omega$ conditional on characteristics \mathbf{X} . The quantity of interest is the Conditional Average Partial Effect (CAPE) $\tau(\mathbf{X}_H) = E[\partial Y / \partial W | \mathbf{X}_H]$. I further make the *strong* assumption that the CAPE can be additively separated from the effect of \mathbf{X} on Y , such that the following partial linear specification holds¹⁴

$$Y = \tau(\mathbf{X}_H) \cdot W + g(\mathbf{X}) + \epsilon \quad (5)$$

$$W = f(\mathbf{X}) + \eta \quad (6)$$

$$E[\epsilon | \mathbf{X}] = E[\eta | \mathbf{X}] = E[\eta \cdot \epsilon | \mathbf{X}] = 0. \quad (7)$$

The *nuisance functions* $g()$ and $f()$, and also the treatment effect function $\tau()$, are assumed to be general functions that can be approximated by appropriate ML estimators. Taking the expectations of Eq. 5 conditional on \mathbf{X} and subtracting it yields

$$Y - E[Y | \mathbf{X}] = \tau(\mathbf{X}_H) \cdot (W - E[W | \mathbf{X}]) + \epsilon. \quad (8)$$

This is the specification of [Robinson \(1988\)](#). Assuming one can estimate $m(\mathbf{X}) = E[Y | \mathbf{X}]$ and $f(\mathbf{X}) = E[W | \mathbf{X}]$, and denoting the 'debiased' variables by $\tilde{Y} = Y - m(\mathbf{X})$ and $\tilde{W} = W - f(\mathbf{X})$ one can then, following [Nie & Wager \(2021\)](#), estimate the CAPE $\hat{\tau}(\mathbf{X}_H)$ by minimizing the R-Loss

$$\tau^*(\cdot) = \operatorname{argmin}_{\tau} \left\{ E[(\tilde{Y} - \tau(\mathbf{X}_H)\tilde{W})^2] \right\} = \operatorname{argmin}_{\tau} \left\{ E[\tilde{W}^2(\tilde{Y}/\tilde{W} - \tau(\mathbf{X}_H))^2] \right\}. \quad (9)$$

The RHS of Eq. 9 demonstrates the so-called 'weight trick' that minimizing the R-Loss with a nonparametric (ML) estimator $\tau(\mathbf{X}_H)$ amounts to predicting the target \tilde{Y}/\tilde{W} using sampling weights \tilde{W}^2 . For the practical implementation of Eq. 9, [Chernozhukov, Chetverikov, et al. \(2018\)](#) show that it is important to estimate the first-stage *nuisance functions* \hat{m}, \hat{f} via cross-fitting. In that case, a Neyman Orthogonality condition exists to ensure that the estimate $\hat{\tau}(\mathbf{X}_H)$ from minimizing the R-Loss is insensitive to biases (e.g., regularization bias) in \hat{m}, \hat{f} .¹⁵ The final stage estimate $\hat{\tau}(\mathbf{X}_H)$ should also be cross-validated to ensure credible out-of-sample CAPE predictions.

Once $\hat{\tau}(\mathbf{X}_H)$ is obtained, further quantities such as the Average Partial Effect (APE) or Group Average Partial Effects (GAPE) can be obtained by doubly-robust methods analogous to Augmented Inverse Probability Weighting (AIPW) ([Robins et al., 1994](#)) in the binary treatment case. Following [Chernozhukov et al. \(2022\)](#), [Athey & Wager \(2021\)](#), and using notation from the *grf* R package ([Tibshirani et al., 2023](#)), the general form of such an AIPW estimator is

$$\bar{\tau} = \frac{1}{n} \sum_{i=1}^n \Gamma_i, \quad \Gamma_i = \tau_i(\mathbf{X}_{Hi}) + h_i(\mathbf{X}_i, W_i)(\tilde{Y}_i - \tau_i(\mathbf{X}_{Hi})\tilde{W}_i). \quad (10)$$

In the binary W case, the debiasing weight $h(\mathbf{X}, W)$ amounts to the so-called Horvitz-Thompson transformation.¹⁶ In the continuous W case, [Tibshirani et al. \(2023\)](#) propose $h(\mathbf{X}, W) = \tilde{W}/\hat{W}^2$, where \hat{W}^2 denotes a cross-fitted estimate of \tilde{W}^2 from \mathbf{X} , which [Tibshirani et al. \(2023\)](#) obtain via a honest RF ([Athey & Imbens, 2016](#); [Breiman, 2001](#)) Out-of-Bag (OOB) predictions. To assess treatment heterogeneity, [Athey & Wager \(2019\)](#) average the doubly robust scores $\hat{\Gamma}_i$ in high and low regions of $\hat{\tau}(\mathbf{X}_H)$ and test the difference in means. They also compute the Best Linear Prediction (BLP) from $\hat{\tau}(\mathbf{X}_H)$ following [Chernozhukov, Demirer, et al. \(2018\)](#) via Eq. 9 as a calibration test.

To satisfy the unconfoundedness assumption, it is critical to choose appropriate estimators for $m()$, $f()$ and $\tau()$, and a suitable cross-fitting strategy. Because I have no strong prior beliefs on how infrastructure (\mathbf{X}) affects wealth/economic activity (Y) and other infrastructure (W), I remain as model-agnostic as possible about these functional forms by employing a *Super Learner* approach ([Van der Laan et al., 2007](#)) to create an optimal weighted combination of several candidate learners

¹⁴This assumption is relaxed in Section 5 estimating counterfactual predictions (causal dose-response functions).

¹⁵The Neyman Orthogonality Condition states that $\partial_{m,f}(m_0, f_0)E[\epsilon \tilde{W}] = \partial_{m,f}(m_0, f_0)E[(\tilde{Y} - \tau(\mathbf{X}_H) \cdot \tilde{W})\tilde{W}] = 0$, where $\partial_{m,f}(m_0, f_0)$ is the (Gateaux) derivative of the moment condition w.r.t. to the nuisance parameters m, f evaluated at their true values denoted by m_0, f_0 . This is zero, such that the moment conditions are not sensitive to small perturbations (biases) in the nuisance parameter estimates \hat{m}, \hat{f} . Such moment conditions and the final-stage (GMM) estimators they produce are called Neyman Orthogonal.

¹⁶The Horvitz-Thompson transformation is given by $h(\mathbf{X}, W) = (W - e(W))/(e(W)(1 - e(W)))$ where $e(W)$ denotes the propensity score.

via cross-validation. The most commonly used functional forms in the heterogeneous treatment effect literature are the LASSO, Random Forests, and Gradient Boosting. Recently, Friedberg et al. (2020) have proposed Local Linear Forests (LLF) as a variant of Generalized Random Forests (Athey et al., 2019) more accurate with smooth targets. A causal version of LLF is also available in the *grf* R package. I thus create a *Super Learner* from the predictions of these 4 algorithms

$$\hat{SL} = \beta_0 + \beta_1 \text{LA}\hat{\text{SSO}} + \beta_2 \hat{\text{RF}} + \beta_3 \hat{\text{LLF}} + \beta_4 \hat{\text{GBM}}, \quad (11)$$

where the weights β_i are determined by a relaxed Elastic Net tuned with 10-fold cross-validation. Each algorithm is also cross-validated/fitted: LASSO is tuned with 10-fold cross-validation using the built-in functionality of the *glmnet* R package (J. Friedman et al., 2010). RF and LLF are fit using the *grf* package with honest trees and produce OOB predictions. For GBM, I use XGBoost and employ a 3-fold cross-fitting approach, where an XGBoost model is trained on two folds, using the excluded fold as a validation set for early stopping and then predicting the outcome in the excluded fold. The ensemble estimator of the CAPE $[\tau()]$ similarly combines estimates from Causal Forests (Athey et al., 2019) and Local-Linear Causal Forests (Friedberg et al., 2020) obtained via *grf*, with cross-validated/fitted LASSO and XGBoost estimates from minimizing the R-Loss (Eq. 9). A cross-validated relaxed Elastic Net is again used as the final stage estimator and corresponds to the BLP of the CAPE from the four model's CAPE estimates. All infrastructure, population, and travel time estimates enter the model in natural logs. This allows for the interpretation of $\hat{\tau}(\mathbf{X}_H)$ as a semi-elasticity and improves the fit of the LASSO and LLF models.

The main threat to this DML strategy is the existence of unobserved factors, such as political, historical, or ethnic contingencies influencing the local placement of certain infrastructures without affecting the spatial distribution of infrastructure as a whole. Recent contributions such as Dreher et al. (2019) and Graff (2024) provide evidence of favouritism and colonial legacy having an impact on activity and infrastructure in Africa, but they do not investigate whether this occurs on a local selective basis. To provide an example in the spirit of Dreher et al. (2019): if an African leader supports his or her birth region by only building additional schools, then this may be problematic because the ML prediction of schools from all other observable characteristics would not be able to capture the favouritism in education. If, however, the support is more broad-based and the region has not only more schools but also more hospitals, roads, and access to power, then these other features increase the ML prediction of schools (and of wealth) in that region and thus favouritism is absorbed by the ML control functions. Even if the first case holds, my nuisance models include a spatial lag (neighbouring cell average) of the IWI to eliminate spatial autocorrelation, thus favouritism in education alone would also have to increase wealth only in the current cell.

To provide empirical evidence, I take data on political and ethnic favouritism and colonial railroad construction from Graff (2024). I aggregate the infrastructure data, debiased at high resolution using the ensemble ML estimator, into his 0.5° grid. I then run regressions similar to Tables 1 and 2 in Graff (2024) with both raw and debiased infrastructure quantities as outcome variables. Appendix Table A9/A10 reports results for simple/quantile counts. The top half considers raw infrastructure data and provides strong evidence for colonial history and favouritism extended by leaders towards their birth regions affecting the spatial distribution of infrastructure. With the debiased data in the bottom half, most coefficients become zero and insignificant. A multiple testing adjustment following Clarke et al. (2020) renders all debiased coefficients insignificant. Thus, I argue that favouritism is an aggregate phenomenon and not a threat to high-resolution causal ML estimates controlling for other infrastructure and spatial autocorrelation.

Another caveat is that infrastructure also enters high-resolution outcome measures such as the IWI (through complex ML models learning from OSM and satellite imagery) or nightlights. However, the *Super Learner* can likely recover these 'unobserved nuisance functions', such that the effect of the treatment infrastructure can still be measured, even though the signal-to-noise ratio in the debiased data may be higher than with direct ground-truth measurements.

To test this, I also estimate (C)APEs using IWI estimates from DHS surveys conducted in SSA since 2010. I obtain 17,396 cells with more than 5 households and $10 \text{ persons}/\text{km}^2$, for which I compute the average household IWI, vs. 89,044 cells with IWI predictions by Lee & Braithwaite (2022). Both estimates are correlated but not perfectly ($r = 0.814$), reflecting slight

methodological differences and Lee & Braithwaite (2022)'s substantially higher resolution and use of DHS surveys only from 2017.¹⁷ Appendix Table A11 shows the Median CAPE prediction (for all 103,922 populated cells) obtained from both IWI measures, and Appendix Table A12 shows doubly robust APE estimates computed using only cells where the respective measure is available. Both tables signify quite similar estimates, especially for important predictors such as roads, communications, travel times, accommodation, health, and automotive facilities. Curiously, education has no effect on the DHS-based IWI. The DHS-derived median absolute coefficient size is also slightly larger. This is likely due to the use of standard mean squared error loss functions for the predicted IWI, which Ratledge et al. (2022) show generate a narrower distribution of model-predicted values than the true distribution. It may also be due to the reduced cell sample for the DHS-based IWI, or differences in methodology or timing, as elucidated above. Notwithstanding, the estimates are highly correlated across feature categories ($r \geq 0.86$ for the APE). Appendix Tables A11 and A12 thus suggest that using the predicted IWI doesn't significantly alter the results.

A final obstacle is reverse causality between infrastructure and wealth/activity. This is impossible to rule out in a pure cross-section. However, controlling for the spatial lags of the IWI and of other infrastructure is likely to mitigate its effects, as similar wealth levels may be present in adjacent cells and wealthy inhabitants of cell i may also build infrastructure in neighbouring cells.

In summary, I argue that the causal ML strategy absorbs the most significant spatial planning decisions, and the first stage residuals from the *Super Learners* are likely to represent some noisy local idiosyncrasies that can be used to estimate at least a partial equilibrium APE for different types of infrastructure. Notwithstanding, as I cannot formally establish causality or identification, the estimates should be interpreted with caution. I now proceed to report the estimates.

4.1 Average Partial Effects

Table 5 reports the APE of the natural log of infrastructure on the IWI. The LHS shows results when features are simply counted in each category and cell, whereas the RHS shows results with quantile counts applied during aggregation, as detailed in Section 2.5. The estimates can be interpreted as the change in IWI points [0, 100] induced by a 100% increase in the corresponding feature intensity. Following Athey & Wager (2019), I also calculate APEs for cells above and below the median CAPE estimate and test for the difference between them as evidence for heterogeneity.¹⁸

Table 5: Average Partial Effects on IWI

Feature	ATE	Simple Counts			Quantile Counts			
		High	Low	Diff.	ATE	High	Low	Diff.
roads_paved	0.218***	0.283***	0.153***	0.131***	0.216***	0.28***	0.151***	0.128***
power	0.0683***	0.0849***	0.0517***	0.0332	0.0675***	0.0758***	0.0592***	0.0166
education	0.563***	0.958***	0.168***	0.79***	0.436***	0.774***	0.0973***	0.677***
health	0.803***	1.25***	0.359***	0.887***	0.772***	1.15***	0.389***	0.765***
communications	0.689***	0.951***	0.427***	0.525***	0.689***	1.01***	0.371***	0.635***
public_service_utility	0.327***	0.452***	0.203	0.248	0.331***	0.482***	0.179***	0.303*
automotive	0.873***	0.721***	1.02*	-0.299	0.444**	0.668**	0.22	0.447
transport_other	0.332***	0.395***	0.27***	0.125	0.305***	0.402***	0.207***	0.194**
financial	0.626***	0.884***	0.369***	0.515**	0.543***	0.791***	0.295	0.496**
services	1.38	—	1.38	—	0.748*	1.27*	0.222	1.05
ttime_city_1m	-0.513***	-0.335***	-0.691***	0.356***	-0.502***	-0.192***	-0.812***	0.62***
ttime_port_any	-0.293***	—	-0.293***	—	-0.295***	-0.141**	-0.448***	0.307***
residential	-0.0679***	0.0271	-0.163***	0.19***	-0.0519***	0.0391**	-0.143***	0.182***
accommodation	0.807***	0.778**	0.835***	-0.0568	0.771***	0.725***	0.817***	-0.0925
tourism_recreation	0.184**	0.472***	0.147*	0.325*	0.309***	0.509***	0.108	0.401**
mining_industrial	0.541***	0.878***	0.204***	0.674***	0.46***	0.698***	0.221***	0.477***

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1. A ‘-’ indicates missing estimates (signifying underidentification/collinearity).

Notes: Table shows doubly-robust APE estimates of the log feature intensity (simple counts or quantile counts in each cell, see Section 2.5) on the International Wealth Index [0, 100] by Lee & Braithwaite (2022) (covering 42 SSA countries). The “High” and “Low” estimates report the APE above and below the median CAPE estimate. The “Diff.” column indicates their difference to test for heterogeneity. All terms are tested using a two-sided t-test with standard errors derived from the doubly robust scores following Athey & Wager (2019).

¹⁷I use DHS surveys from 2010 for this exercise to have more data.

¹⁸Since the reported CATE is already a BLP from multiple CATE estimators, the BLP cannot be computed as an additional test for effect heterogeneity.

Table 5 suggests that, controlling for all other features, including automotive and public transport facilities and travel time to major cities and ports, paved roads have a moderate partial effect on wealth amounting to a 0.22-point IWI increase to a doubling of paved road length. The effect of power is even smaller at only 0.07. On the other hand, education, health, communications, automotive facilities, financial services and accommodation (hotels) have rather large effects around 0.56-0.87. Public services and (other) transport have smaller effects around 0.33. Travel time to major cities and ports, proxying for exogeneous/extracellular changes in market access, have the expected negative effect of around -0.5 for major cities and -0.3 for ports. Residential areas also have a slight negative effect around -0.07. This must be understood *ceteris paribus*: in the first-stage models I control for GPW4 population, but this is only available at the smallest administrative units. Thus, residential buildings may proxy for the remaining uncontrolled variation in population. The negative coefficient thus suggests that when all other infrastructure is held fixed, adding people does not increase individual wealth, e.g., 'urbanization without growth' à la Fay & Opal (2000).¹⁹

4.2 Effect Heterogeneity

Cell-level CAPE estimates are arguably more policy-relevant than an all-Africa infrastructure APE. This subsection investigates their distribution and correlates before the next subsection analyzes their spatial patterns. Figure 5 plots kernel density estimates of the IWI CAPEs, indicating that for most infrastructure categories, there is considerable spatial heterogeneity in the conditional partial effect. Only paved roads and power appear to yield no negative CAPEs and have very narrow distributions. Most other features, except for residential buildings and travel times, have largely positive but very heterogeneous effects. Education and financial services show some negative CAPEs. In theory these shouldn't exist, but ML models are unrestricted. Negative values thus indicate that the partial effects can be zero or close to zero in some locations.

To compactly summarize important determinants of effect heterogeneity, I compute Pearson's correlations between the CAPEs and all predictors in logs. I report the top heterogeneity variables across infrastructure categories based on their average squared correlation with the CAPEs. Figure 6 visualizes the results, where the columns are the CAPEs, and the rows are covariates in decreasing order of average squared correlations. Rows and columns are clustered with complete linkage.²⁰

At first sight, Figure 6 exhibits rather simple heterogeneity patterns: most CAPEs are either increasing or decreasing in the level of infrastructure/agglomeration. In particular, the CAPEs of paved roads, communications, and, to a lesser extent, power and financial services are increasing in the level of agglomeration. For other types of infrastructure the opposite appears to be the case. In particular education, public services and utilities, and mining/industrial facilities have stronger marginal effects on household welfare if found in structurally weak/rural areas. Interestingly, travel time to cities (market access) also has a stronger negative effect in rural areas²¹, whereas travel time to ports has a larger negative effect inside urban agglomerations. Overall, the results suggest infrastructure substitution effects governed by agglomeration forces, e.g., in rural areas, the proximity to cities as well as the presence of a school or mine/industrial plant as employer has larger wealth effects than in cities with diverse economic/employment opportunities.²²

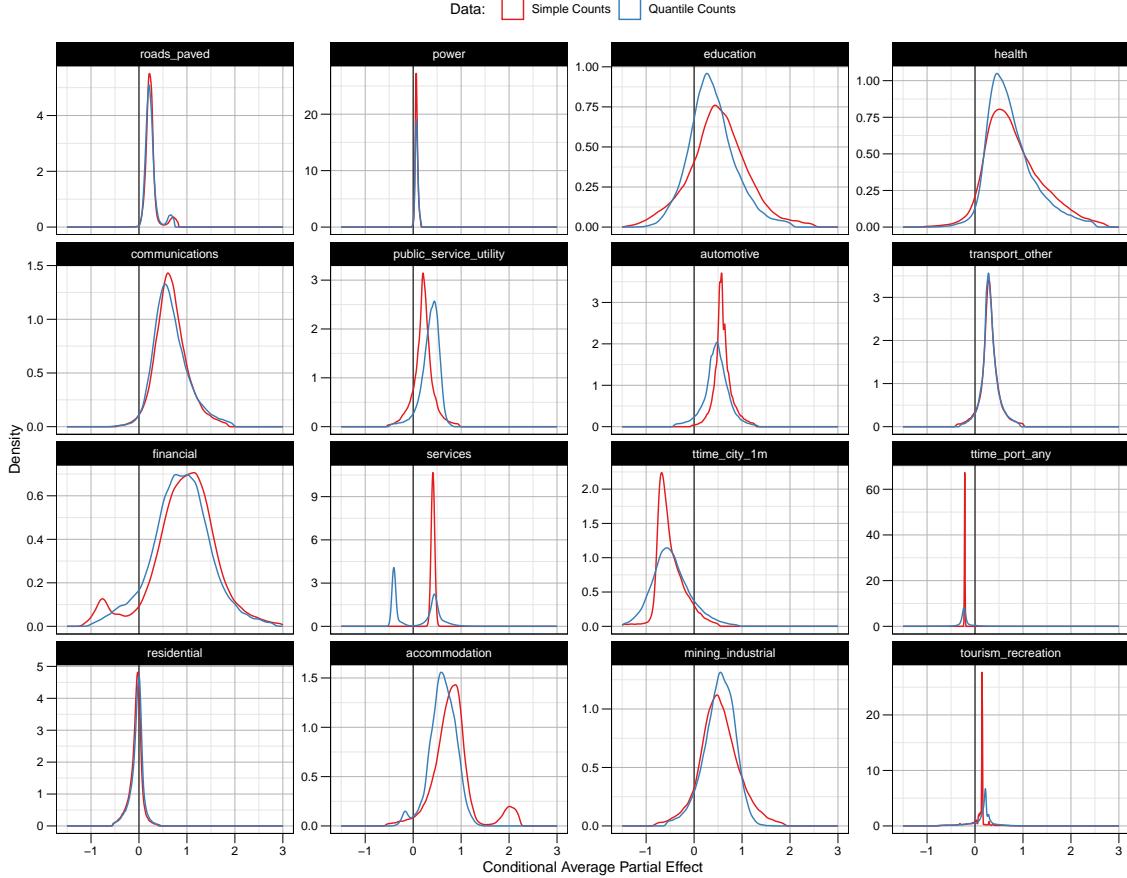
¹⁹For reference, I also estimated APEs using the log of nightlights as outcome measures. The results, reported in Appendix Table A15, are broadly consistent with Table 5, but emphasize light-generating activities such as automotive facilities (traffic), industrial facilities, and financial services (CBDs), and yield smaller effects for roads, communications, health, and education. Surprisingly, power infrastructure does not have large effects either.

²⁰The number reported in Figure 6 is the average correlation across the simple and quantile counts datasets (both are very similar permitting such averaging).

²¹Since the CAPE of travel time to major cities is negative, the positive correlation with spatial features implies that it is smaller inside agglomerations. Intuitively, travel time to the city center (on a log-scale) is less important for wealth generation inside cities than it is in rural areas which may be near or far from major cities.

²²As a further robustness exercise, Appendix Figures A17 and A18 show analogous CAPE estimates from direct DHS wealth estimates, and Tables A13-A14 show correlations among CAPEs from both IWI measures. On average, the CAPEs are similar, particularly for important features with a robust effect, such as paved roads, communications, and mining/industrial facilities. The median CAPE correlation is only around 0.13, and 0.25-0.3 if spillover variables are omitted. Appendix Figures A19 and A20 further show the densities and main correlates of nightlights CAPEs. While the densities are broadly similar, the correlation are almost all negative. This may be due to the unavailability of data in low-lit areas (zero observations are not considered for training and NASA's Black Marble product truncates noisy nightlights estimates in low-lit areas to zero).

Figure 5: DML CAPE Kernel Density Estimates for IWI



Notes: Figure shows Gaussian kernel density estimates of the Conditional Average Partial Effect (CAPE) of log features' on the International Wealth Index (IWI) [0, 100] by Lee & Braithwaite (2022) (covering 42 SSA countries). Data are aggregated in each cell using either simple counts or quantile counts (see Section 2.5) before taking the natural log.

There is also interesting spatial variation in the CAPEs not fully conveyed by correlations with infrastructure variables. Therefore, I now proceed with a visual examination of the estimates.

4.3 Spatial Examination of Effect Heterogeneity

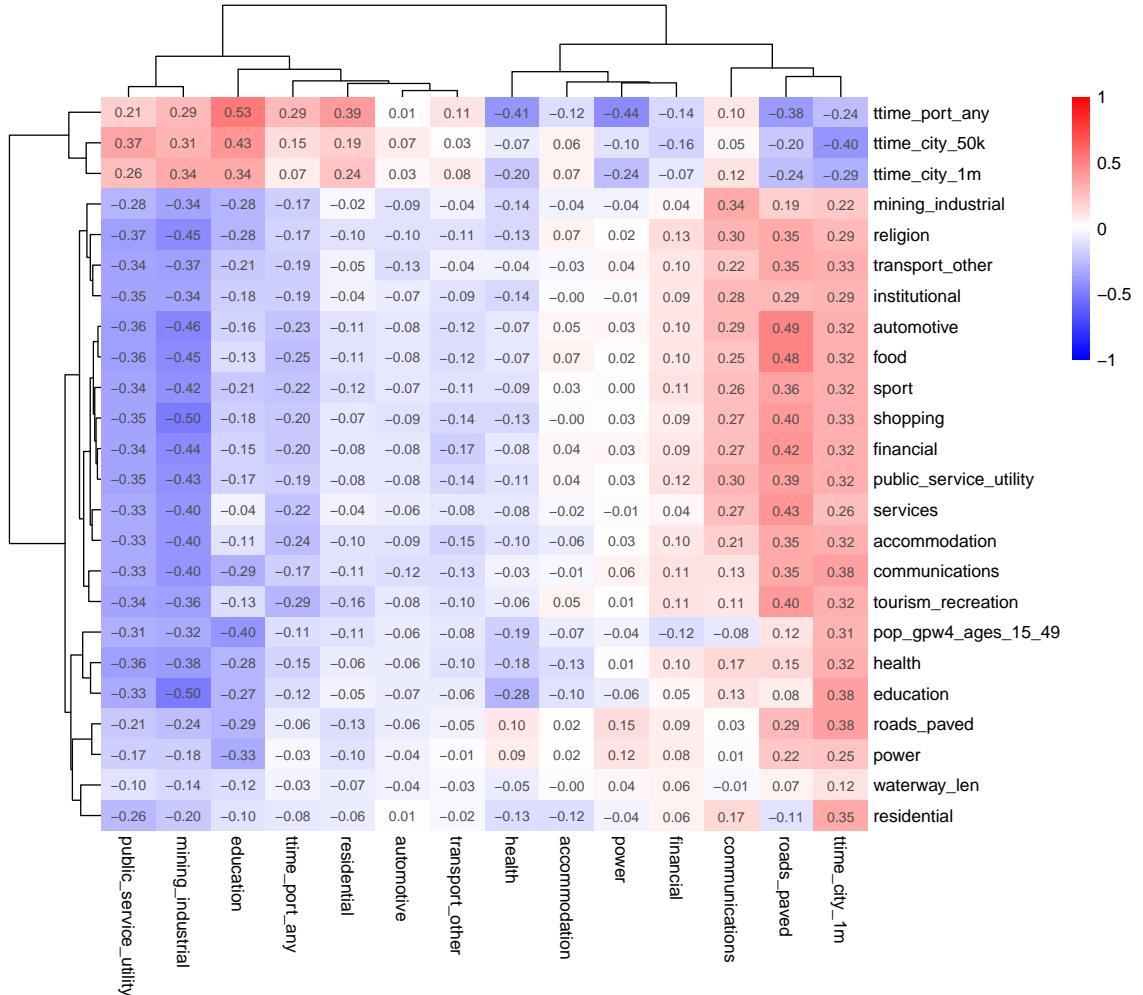
Figure 7 visualizes the geometric mean CAPE across simple and quantile count-based estimates for important features—to smooth outliers, both estimates are highly correlated at $r > 0.8$. Overall, there is considerable and complex spatial heterogeneity in the CAPEs. The top panel shows the CAPEs for roads, power, and communications. Exempting power,²³ the effects are higher in urban areas, particularly for paved roads. The higher effects of communications in urban areas is consistent with Masaki et al. (2020)'s study of broad-band expansion in Senegal and explicable by higher shares of young skilled workers in cities. Since the IWI estimate by Lee & Braithwaite (2022) is only available for SSA, CAPEs for North Africa are pure predictions from the SSA-trained model. This seems problematic for paved roads given their much greater density in North Africa.

The middle panel of Figure 7 shows the CAPEs for education, health, and public services and utilities (excl. power). Exempting health,²⁴ these broadly follow the opposite spatial pattern,

²³In previous versions of this paper, the CAPEs of power was concentrated in cities as well, see Appendix Figure A16. That version only used power infrastructure as reported in OSM. The present version incorporates some higher-resolution (sometimes household-level) grid data from the EU's Joint Research Centre, and yields quite different results. Thus, the coverage and definition of infrastructures can significantly affect the outcomes.

²⁴In a previous version of this paper (available on the [authors website](#) as 'Version 2'), the CAPE of health also very much resembled the one of public services and utilities, see Appendix Figure A16, but that version did not include significant data on public health facilities in SSA by Maina et al. (2019). This indicates that comprehensive

Figure 6: Correlates of IWI CAPE Estimates (\mathbf{X}_H): Average Across Datasets



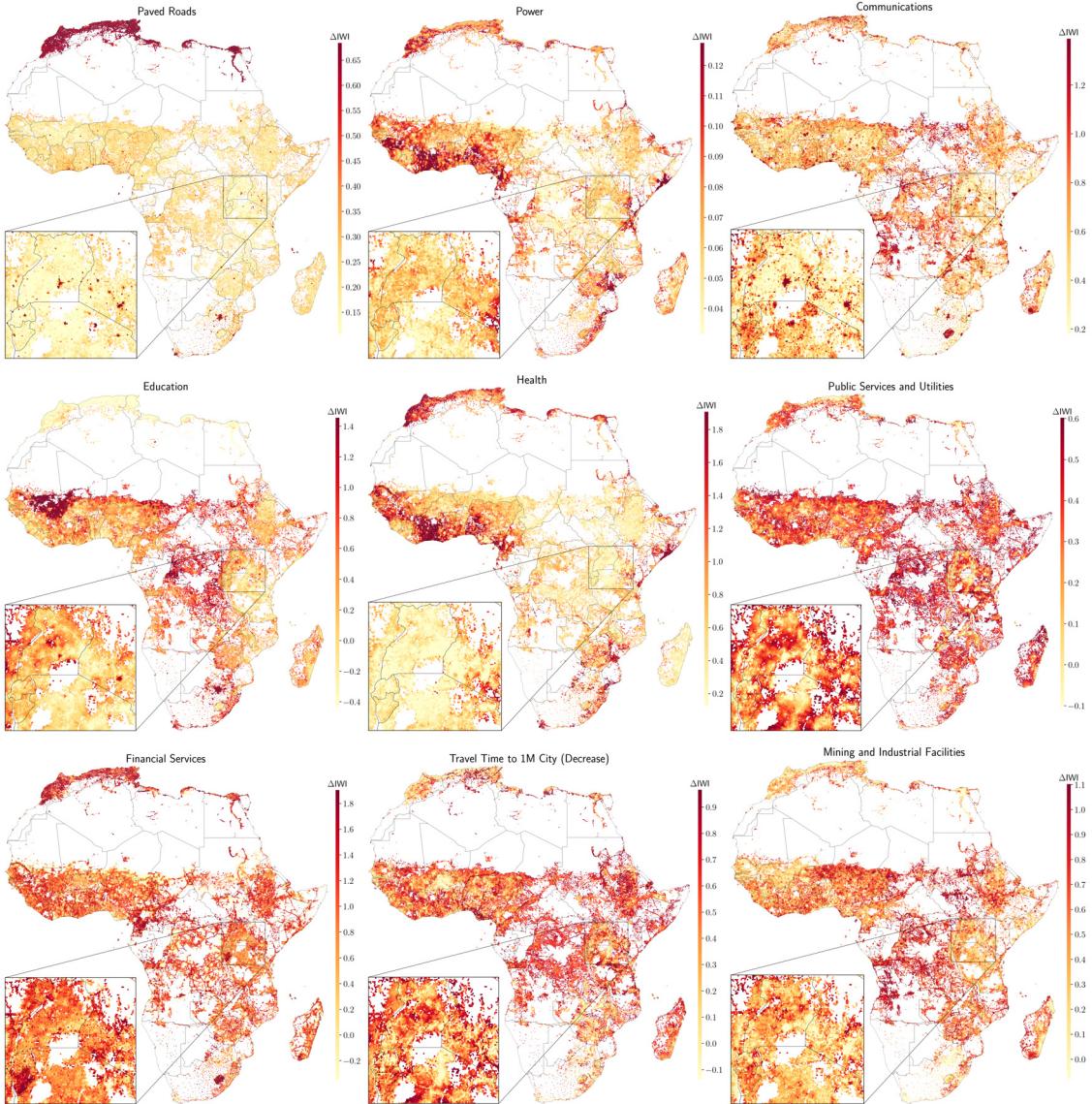
Notes: Figure shows average Pearson’s correlations across the simple counts and quantile counts datasets of CAPEs (columns) with the features in logs (rows). A positive correlation implies that the CAPE is increasing in feature intensity.

being highest in underdeveloped regions such as the Sahel region, Congo, and rural areas in Ethiopia, (South-)Sudan and Somalia. The bottom panel shows CAPEs for financial services, travel time to large cities, and mining and industrial facilities. The financial services CAPEs suggest that some countries like Burundi or Lesotho are financially underdeveloped. Travel time to cities CAPEs are probably the better estimates for what regional connectivity improvements through new roads could yield in terms of welfare. They are large in many areas outside of urban centers, consistent, e.g., with Nakamura et al. (2019)’s study of rural road expansion in Ethiopia finding that remote areas benefitted the most. Dorosh et al. (2012) also find sizeable increases in local crop production from reduced travel times to nearby towns/cities in SSA, and Kebede (2024) finds sizeable agricultural income increases attributable to increased production of comparative advantage crops, and suggests a tradeoff between rural roads benefiting agriculture and highways and railroads benefitting urban populations and manufacturing. Regarding the latter, Duranton & Turner (2012) find that a 10% increase in a US city’s stock of highways causes a 1.5% increase in its employment. This tradeoff also seems evident in these CAPE estimates for the stock of paved roads and travel time to major cities. Finally, the mining and industrial facilities CAPEs are also high in many structurally weak areas, presumably due to their role as major employers.

In summary, CAPE estimates indicate complex associations between distinct infrastructure classes and the spatial distribution of welfare. The estimates for roads and communications generally suggest stronger marginal effects inside urban spaces, whereas education, public services,

and uniform geospatial data is important for robust and policy-amenable causal ML estimates.

Figure 7: Spatial CAPE Estimates: Geometric Mean



Notes: Figure shows the Conditional Average Partial Effect (CAPE) of log features' on the International Wealth Index (IWI) [0, 100] by Lee & Braithwaite (2022) (covering 42 SSA countries). A geometric mean across estimates derived from simple and quantile counts data (Section 2.5) is reported to limit outliers. Both are correlated ($r > 0.8$ for all features). For travel time to major (1M) cities, the CAPE was negated, i.e., dark spots indicate large negative CAPEs.

and market access show stronger effects in rural areas. These findings are broadly consistent with the literature on urbanization summarized by J. V. Henderson & Turner (2020) identifying poor access to services in rural areas as a key driver of rural-to-urban migration. While data limitations preclude their policy use, the results allow for a hopeful outlook that comprehensive spatial datasets,²⁵ and a refined spatial ML methodology,²⁶ could generate policy-relevant CAPEs.

Apart from their highly partial nature,²⁷ another limitation of CAPEs from a policy point of view is that marginal effects, and semi-elasticities in particular, are difficult to interpret. Policymakers are often more interested in counterfactual predictions (CPs), such as the simulated effects of different levels of infrastructure in the same locations. It is possible to generate CPs with causal ML, as demonstrated below, but the properties of such estimates still need to be established.

²⁵Quite comprehensive location intelligence data is already available in the commercial realm, e.g., by Google Maps or Dataplor. The Overture Maps Foundation aims to create comprehensive open maps data for developers.

²⁶Such as Graph Neural Networks or Convolutional Neural Networks which are better suited to spatial datasets.

²⁷As pointed out before CAPEs don't consider general equilibrium effects, such as the relocation of populations and economic activities following infrastructure investments. They also hold fixed potentially useful spatial variation, such as the partial effect of roads holding fixed market access and automotive facilities.

5 Counterfactual Predictions

Counterfactual predictions (CPs) can be obtained by increasing infrastructure quantities in the data and comparing causal ML model predictions from this altered dataset to the baseline prediction. This approach is similar to the single-model or 'S-learner' in the context of CATE estimation with binary treatments - applied by [Hill \(2011\)](#) with Bayesian Additive Regression Trees (BART) and discussed in [Jacob \(2021\)](#). I follow [Faccure & Germano \(2021\)](#) in using debiased data to avoid confounding influences in the model. Concretely, I estimate

$$\tilde{Y} = \theta(\tilde{W}, \mathbf{X}_H), \quad (12)$$

where $\theta()$ is again an ensemble ML estimator (Eq. 11) and \tilde{Y} , \tilde{W} are debiased. In contrast to CAPE estimation, this formulation allows for the effect to be non-linear, i.e., to change with the level of \tilde{W} . After estimating $\theta()$ in a cross-fitting manner, I evaluate it at different levels of \tilde{W} . I compute the 10%, 25%, 50%, 75%, and 90% quantiles of the positive distribution of W across cells, alternately add them to \tilde{W} , and obtain a prediction. Table 6 reports these quantile increases.

Table 6: Counterfactual Increase in Feature Quantity/Cell for Selected Features

Feature	10%	25%	50%	75%	90%
roads_paved (m)	2111	6295	10371	14493	24119
power	59	168	314	570	980
education	1	1	3	9	24
health	1	1	2	3	8
communications	1	2	6	23	85
public_service_utility	1	1	2	4	10
automotive	1	1	2	7	23
transport_other	1	1	2	7	20
financial	1	1	2	6	21
services	1	1	2	6	24
ttime_city_1m (min)	104	203	363	602	930
ttime_port_any (min)	130	289	564	951	1520
residential	1	3	8	24	84
accommodation	1	1	2	7	24
mining_industrial	1	1	3	8	19
tourism_recreation	1	1	3	8	30

Notes: Table shows sample quantiles on simple-counts data, computed across cells with positive infrastructure for each feature category, respectively. The quantiles are rounded to the full number. They are used for counterfactual predictions, i.e., increasing the infrastructure quantity in all cells by this amount and making a prediction using the causal single-learner (Eq. 12).

As Table 6 indicates, a 10% increase amounts to one additional facility in most categories, 59 additional power-related items (transformers, generators), or a 2111m increase in the length of paved roads in each cell. Higher quantiles imply more substantive increases per cell.²⁸

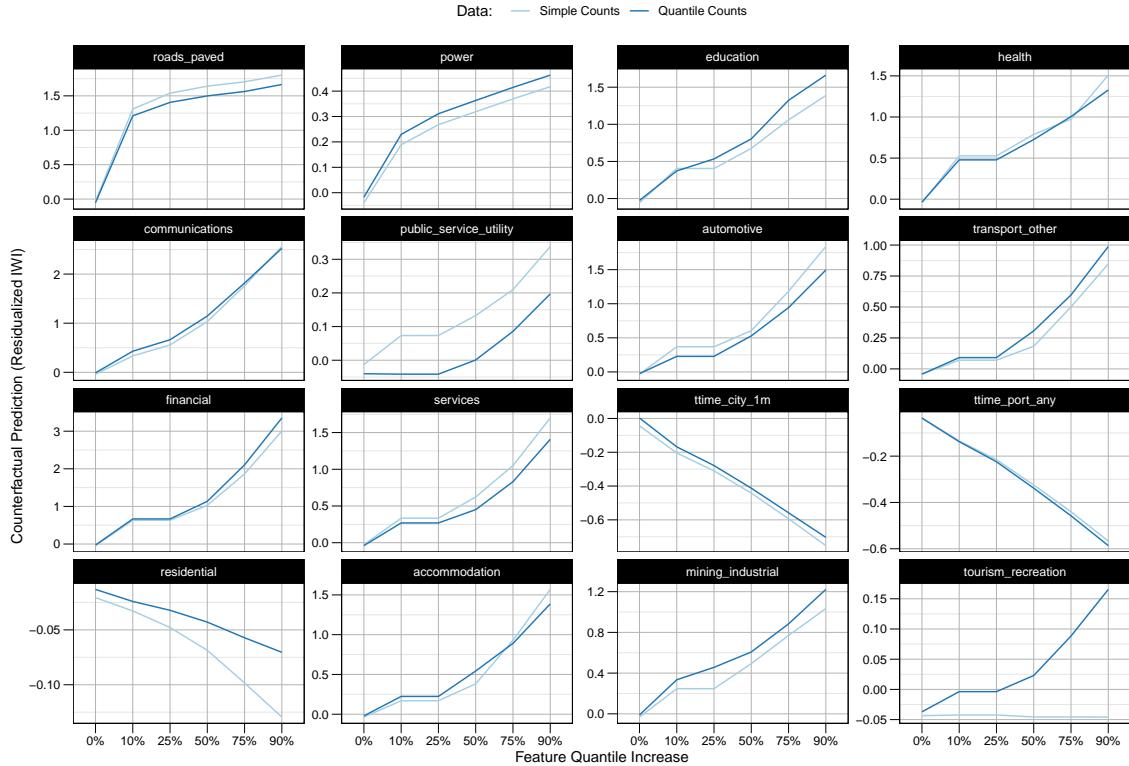
Figure 8 shows average CP (ACP) curves with the IWI as outcome measure, and Appendix Figure A21 with the log of nightlights. Exempting residential buildings, which have a positive effect with nightlights, the results from both outcomes are very similar. The relative magnitudes across features are also broadly consistent with the APEs. Appendix Tables A16 and A17-A18 show ACP curves derived from the DHS-based IWI, which are also similar, instilling some confidence in the methodology. Since for many sparse features, both the 10% and 25% treatment levels imply one additional feature, some ACP curves in Figure 8 have flat segments.

The ACP curves generally suggest decreasing returns. This is particularly pronounced for paved roads and power, where the 10% increase has a large initial effect, and further increases add much

²⁸Only the respective treatment infrastructure, \tilde{W} , is increased; all other infrastructures \mathbf{X}_H are unaltered. Due to debiasing, the OOB R^2 for the IWI residuals is only 2-5%, but in-sample fits are considerably higher and model-based simulation results appear meaningful and are very consistent with (C)APE estimates.

lower wealth gains. The main reason is that most cells are scarcely populated, without paved roads or power supply, and do not have many types of features at all. Thus, the existence of any paved roads or power suggests to the model that more diverse activity takes place in the cell, implying an increase in wealth. However, further increases in roads/power don't yield proportional wealth responses, and also make these rural cells very different from any cells the model has learned.²⁹

Figure 8: Average Counterfactual Predictions for IWI



Notes: Figure shows average counterfactual predictions (across all cells) derived from simple and quantile counts data.

5.1 Spatial Examination of 50% Counterfactual Predictions

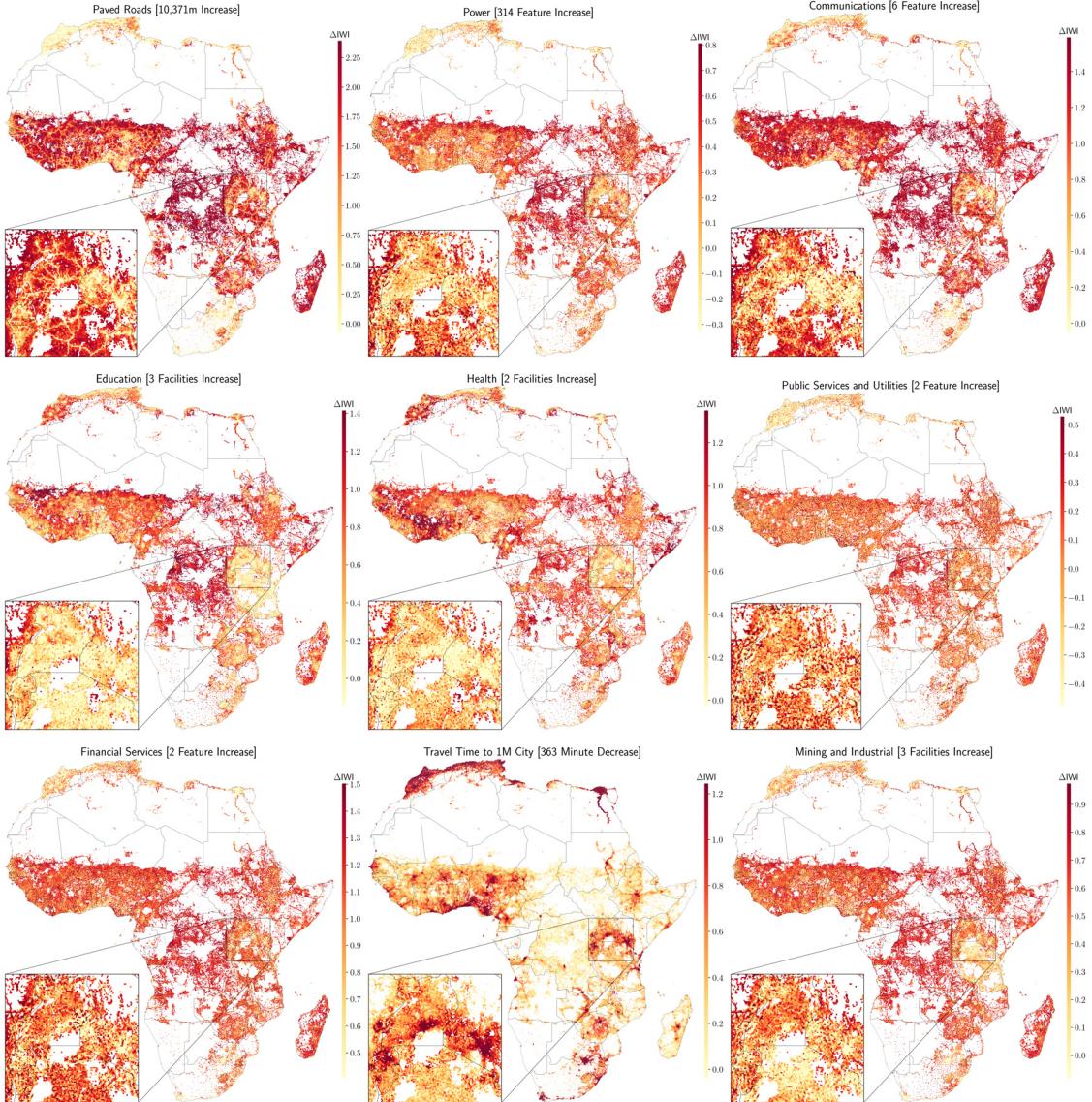
It remains to examine the spatial distribution of these counterfactual predictions. Since predictions from simple and quantile count data are very similar, I again report a geometric mean across the two which slightly downweights outliers. Since an appraisal of all different treatment levels in the spatial dimension would be overwhelming, I only plot the 50% level in Figure 9.

Evidently, CPs are more similar across different infrastructure types than CAPEs, indicating that generally the household wealth returns to building a fixed quantity of infrastructure is higher in rural areas than in cities. A notable exception is travel time to major cities, which the CPs suggest should rather be reduced in the vicinity of such cities. Building infrastructure in rural areas as suggested by most of these maps is, however, not aggregate welfare maximizing since most of these areas are scarcely populated. It likely is more sensible to build infrastructure in populated places even though per-household welfare gains are considerably lower.

To provide at least a heuristic, partial equilibrium appraisal of how a social planner seeking to maximize aggregate welfare might allocate investments, I multiply the 50% CFPRs of Figure 9 by the WorldPop 2020 population measure in each cell. Appendix Figure A24 shows the outcome. As expected, investments in populated areas generally yield higher aggregate welfare returns, but there remains considerable heterogeneity across different infrastructures; for example power and

²⁹Because cells have heterogeneous populations, Appendix Figure A22 also provides average total wealth effects obtained by multiplying the cell-level predictions with the WorldPop 2020 population estimate and computing the average across cells. The relative magnitudes are broadly similar.

Figure 9: Spatial 50% Counterfactual Predictions: Geometric Mean



Notes: Figure shows 50% counterfactual predictions of the International Wealth Index (IWI) [0, 100] by Lee & Braithwaite (2022), i.e. the predicted wealth increase (Eq. 12) from an increase in each cell amounting to the median of the non-negative feature density, summarized in Table 6. A geometric mean across estimates derived from simple and quantile counts data (Section 2.5) is reported to limit outliers. Both are correlated ($r > 0.8$ for all features).

education investments are primarily directed to populated rural areas and less to city centers.

Overall CPs emphasize the benefits of investments in rural areas for poverty alleviation. But policymakers also consider population, inequality, and other social and political objectives in determining infrastructure allocations. With more careful efforts at monetizing them, considering not only the beneficiaries and policy objectives but also heterogeneous construction costs, CPs appear able to provide useful guidance for spatial planning. They are easier to interpret than CAPEs, allow simulating different treatment levels, and do not require a restrictive additive separability assumption. On the other hand, the single-learner approach needs to be better studied and has significant shortcomings. In particular, the relevance of the treatment variable is not guaranteed in an unrestricted ML model. Estimating a CAPE using multiple estimators and combining them via a BLP model has the advantage of at least providing well-calibrated estimates.

6 Conclusion

Compiling rich geospatial data on infrastructure and high-resolution data on household wealth, economic activity, population, and accessibility in Africa and analyzing it with nonparametric and causal machine learning (ML) methods yields detailed insights into Africa’s spatial economy and the local and global returns to different public and economic infrastructures. Notably, infrastructure in Africa is concentrated in cities and often inefficiently allocated across space. Paved roads, communications, education, power, and residential buildings are the best infrastructure predictors of wealth. Other features such as accommodation (hotels), automotive facilities, sports facilities, commercial buildings, public services and utilities, and health facilities are also important. Travel time to major cities and ports (market access) is similarly essential to physical infrastructure for wealth prediction. The relative importance of these variables varies across locations. In cities, the quantity of paved roads, communications, and, to a lesser extent, power are key for wealth prediction and yield higher marginal returns. In contrast, education, health, and public service facilities yield higher marginal returns in rural areas. Access to cities is more critical in rural areas, whereas access to ports is more beneficial in cities. There is considerable spatial heterogeneity in the marginal returns to different infrastructures governed by agglomeration economies and market access, as well as by specific location characteristics and the availability of other infrastructures.

Spatial causal ML methods taking into consideration detailed economic geography to estimate heterogeneous local returns, while at the same time using the available information to net out confounding influences, appear capable of uncovering this heterogeneity and provide useful partial equilibrium perspectives. Infrastructure potential maps visualizing such Conditional Average Partial Effects can assist spatial economic planning by facilitating comparisons of returns to small public infrastructure investments across different locations and sectors. However, infrastructure spending should not be allocated directly based on the empirical results of this paper without improving the data and methodology. Causal ML findings should also be supported by econometric evidence, which is generally unavailable at scale but uncovers key mechanisms. The findings of this paper are broadly consistent with econometric findings across several countries.

An improved methodology could utilize ML methods explicitly capable of learning spatial dependencies—such as Graph Neural Networks or Convolutional Neural Networks. More comprehensive spatial datasets would also strengthen the *unconfoundedness* assumption stipulating sufficient observance of the spatial economy. In the commercial realm (e.g., Google), comprehensive data is already available, but also open data (e.g., OpenStreetMap, Overture Maps) is growing rapidly. The open *Africa Infrastructure Database* built for this paper compiles and economically classifies open data from several sources. It can serve as a basis for further methodological experimentation or other spatial/sectoral analyses of infrastructure and economic activity in Africa.

With open geospatial data growing, spatial measurement of wealth/activity improving, and rapid methodological innovation at the intersection of observational causal inference and machine learning, the pioneering results presented in this paper allow for a hopeful outlook on causal machine learning’s potential to enhance spatial economic and development planning.

References

- ADB. (2018). Africa's infrastructure: great potential but little impact on inclusive growth. *African Economic Outlook*.
- ADB. (2024). Scaling up financing is key to accelerating africa's structural transformation. *African Development Bank News and Events*. Retrieved from <https://www.afdb.org/en/news-and-events/scaling-financing-key-accelerating-africas-structural-transformation-73244> (Accessed: October 14, 2024)
- Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903), 864–870.
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*.
- Allcott, H., Collard-Wexler, A., & O'Connell, S. D. (2016, March). How do electricity shortages affect industry? evidence from india. *American Economic Review*, 106(3), 587-624. Retrieved from <https://www.aeaweb.org/articles?id=10.1257/aer.20140389> doi: 10.1257/aer.20140389
- Asher, S., & Novosad, P. (2020, March). Rural roads and local economic development. *American Economic Review*, 110(3), 797-823. Retrieved from <https://www.aeaweb.org/articles?id=10.1257/aer.20180268> doi: 10.1257/aer.20180268
- Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353–7360.
- Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2), 1148.
- Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2), 37–51.
- Athey, S., & Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1), 133–161.
- Barnes, R. (2020). dggridr: Discrete global grids [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=dggridR> (R package version 2.0.4)
- Baum-Snow, N., Henderson, J. V., Turner, M. A., Zhang, Q., & Brandt, L. (2020). Does investment in national highways help or hurt hinterland city growth? *Journal of Urban Economics*, 115, 103124.
- Bayer, P., Kennedy, R., Yang, J., & Urpelainen, J. (2020). The need for impact evaluation in electricity access research. *Energy Policy*, 137, 111099.
- Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2021). Deep neural networks and tabular data: A survey. *arXiv preprint arXiv:2110.01889*.
- Brand, J. E., Zhou, X., & Xie, Y. (2023). Recent developments in causal inference and machine learning. *Annual Review of Sociology*, 49(1), 81–110.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Byers, L., Friedrich, J., Hennig, R., Kressig, A., Li, X., McCormick, C., & Valeri, L. M. (2018). A global database of power plants. *World Resources Institute*, 18.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5), 261–265.

- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1-C68. Retrieved from <https://doi.org/10.1111/ectj.12097> doi: 10.1111/ectj.12097
- Chernozhukov, V., Demirer, M., Duflo, E., & Fernandez-Val, I. (2018). *Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india* (Tech. Rep.). National Bureau of Economic Research.
- Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., & Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, 90(4), 1501–1535.
- Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3), e2113658119.
- CIESIN. (2016). *Gridded population of the world, version 4 (gpwv4): Population count* (Tech. Rep.). NASA Socioeconomic Data and Applications Center (SEDAC). Retrieved from <http://dx.doi.org/10.7927/H4X63JVC>
- Clarke, D., Romano, J. P., & Wolf, M. (2020). The romano–wolf multiple-hypothesis correction in stata. *The Stata Journal*, 20(4), 812–843.
- Donaldson, D. (2018). Railroads of the raj: Estimating the impact of transportation infrastructure. *American Economic Review*, 108(4-5), 899–934.
- Donaldson, D., & Hornbeck, R. (2016). Railroads and american economic growth: A “market access” approach. *The Quarterly Journal of Economics*, 131(2), 799–858.
- Donaldson, D., & Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4), 171–198.
- Dorosh, P., Wang, H. G., You, L., & Schmidt, E. (2012). Road connectivity, population, and crop production in sub-saharan africa. *Agricultural Economics*, 43(1), 89-103. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1574-0862.2011.00567.x> doi: <https://doi.org/10.1111/j.1574-0862.2011.00567.x>
- Dreher, A., Fuchs, A., Hodler, R., Parks, B. C., Raschky, P. A., & Tierney, M. J. (2019). African leaders and the geography of china’s foreign assistance. *Journal of Development Economics*, 140, 44–71.
- Duranton, G., & Turner, M. A. (2012). Urban growth and transportation. *Review of Economic Studies*, 79(4), 1407–1440.
- Faber, B. (2014, 03). Trade Integration, Market Size, and Industrialization: Evidence from China’s National Trunk Highway System. *The Review of Economic Studies*, 81(3), 1046-1070. Retrieved from <https://doi.org/10.1093/restud/rdu010> doi: 10.1093/restud/rdu010
- Facure, M., & Germano, M. (2021). *Python Causality Handbook: First Edition*. Zenodo. Retrieved from <https://doi.org/10.5281/zenodo.4445778> doi: 10.5281/zenodo.4445778
- Fay, M., & Opal, C. (2000). *Urbanization without growth: A not so uncommon phenomenon* (Vol. 2412). World Bank Publications.
- Foster, V., & Briceño-Garmendia, C. (2010). *Africa’s infrastructure: a time for transformation*. World Bank.
- Foster, V., Gorgulu, N., Straub, S., & Vagliansindi, M. (2023). *The impact of infrastructure on development outcomes: A qualitative review of four decades of literature* (Policy Research Working Paper No. WPS 10343). Washington, D.C.: World Bank Group. Retrieved from <http://documents.worldbank.org/curated/en/099529203062342252/IDU0e42ae32f0048304f74086d102b6d7a900223>
- Friedberg, R., Tibshirani, J., Athey, S., & Wager, S. (2020). Local linear forests. *Journal of Computational and Graphical Statistics*, 30(2), 503–517.

- Friedman, J., Tibshirani, R., & Hastie, T. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22. doi: 10.18637/jss.v033.i01
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.
- Gennaioli, N., La Porta, R., Lopez-de Silanes, F., & Shleifer, A. (2013). Human capital and regional development. *The Quarterly journal of economics*, 128(1), 105–164.
- Gibson, J., Olivia, S., & Boe-Gibson, G. (2020). Night lights in economics: Sources and uses 1. *Journal of Economic Surveys*, 34(5), 955–980.
- Gilbert, B., Datta, A., Casey, J. A., & Ogburn, E. L. (2021). A causal inference framework for spatial confounding. *arXiv preprint arXiv:2112.14946*.
- Goldbeck, M., & Lindlacher, V. (2021). *Digital infrastructure and local economic growth: Early internet in sub-saharan africa*. ifo Institute for Economic Research, University of Munich.
- Gorgulu, N., Foster, V., Jain, D., Straub, S., & Vagliasindi, M. (2023). *The impact of infrastructure on development outcomes: A meta-analysis* (Policy Research Working Papers No. 10350). Washington, DC: World Bank. Retrieved from <http://hdl.handle.net/10986/39534> (License: CC BY-NC 3.0 IGO)
- Graff, T. (2024). Spatial inefficiencies in africa's trade network. *Journal of Development Economics*, 103319.
- Hall, O., Ohlsson, M., & Rögnvaldsson, T. (2022). A review of explainable AI in the satellite data, deep machine learning, and human poverty domain. *Patterns*, 3(10).
- Henderson, J. V., & Turner, M. A. (2020). Urbanization in the developing world: too early or too slow? *Journal of Economic Perspectives*, 34(3), 150–173.
- Henderson, V., Storeygard, A., & Weil, D. N. (2011). A bright idea for measuring economic growth. *American Economic Review*, 101(3), 194–199.
- Henderson, V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *American economic review*, 102(2), 994–1028.
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1), 217–240.
- Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164, 73–84.
- Huber, M. (2024, July). Causal machine learning in marketing. *International Journal of Business & Management Studies*, 5(7), 1–6. doi: 10.56734/ijbms.v5n7a1
- Huenermund, P., Kaminski, J. C., & Schmitt, C. (2021). Causal machine learning and business decision making. *Academy of Management Proceedings*, 2021(1), 12517. doi: 10.5465/AMBPP.2021.12517abstract
- Imbens, G. W. (2024). Causal inference in the social sciences. *Annual Review of Statistics and Its Application*, 11.
- Jacob, D. (2021). Cate meets ml: Conditional average treatment effect and machine learning. *Digital Finance*, 3(2), 99–148.
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790–794.
- Jedwab, R., & Storeygard, A. (2022). The average and heterogeneous effects of transportation investments: Evidence from sub-saharan africa 1960–2010. *Journal of the European Economic Association*, 20(1), 1–38.

- Kakoulaki, G., & Moner-Girona, M. (2020). *Electricity grid africa* [Dataset]. <http://data.europa.eu/89h/624c6e71-3b9c-4f48-8c67-645911798d41>. European Commission, Joint Research Centre (JRC).
- Kebede, H. A. (2024). Gains from market integration: Welfare effects of new rural roads in ethiopia. *Journal of Development Economics*, 168, 103252.
- Krantz, S. (2023). osmclass: Classify open street map features [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=osmclass> (R package version 0.1.3)
- Kummu, M., Taka, M., & Guillaume, J. H. (2018). Gridded global datasets for gross domestic product and human development index over 1990–2015. *Scientific data*, 5(1), 1–15.
- Lee, K., & Braithwaite, J. (2022). High-resolution poverty maps in sub-saharan africa. *World Development*, 159, 106028.
- Lee, K., Miguel, E., & Wolfram, C. (2020, February). Does household electrification supercharge economic development? *Journal of Economic Perspectives*, 34(1), 122-44. Retrieved from <https://www.aeaweb.org/articles?id=10.1257/jep.34.1.122> doi: 10.1257/jep.34.1.122
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1), 56–67.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Maina, J., Ouma, P. O., Macharia, P. M., Alegana, V. A., Mitto, B., Fall, I. S., ... Okiro, E. A. (2019). A spatial database of health facilities managed by the public health sector in sub saharan africa. *Scientific data*, 6(1), 134.
- Masaki, T., Granguillhome Ochoa, R., & Rodriguez-Castelan, C. (2020). *Broadband internet and household welfare in senegal* (Policy Research Working Paper No. 9386). Washington, D.C.: World Bank. Retrieved from <http://hdl.handle.net/10986/34472> (License: CC BY 3.0 IGO)
- Moneke, N. (2020). *Infrastructure and structural transformation: evidence from ethiopia* (Unpublished doctoral dissertation). London School of Economics and Political Science.
- MSI. (2019). *World port index 2015* (Tech. Rep.). Retrieved from <https://msi.nga.mil/Publications/WPI>
- Nakamura, S., Bundervoet, T., & Nuru, M. (2019). *Rural roads, poverty, and resilience: Evidence from ethiopia* (Policy Research Working Paper No. 8800). Washington, D.C.: World Bank. Retrieved from <http://hdl.handle.net/10986/31495> (License: CC BY 3.0 IGO)
- Nelson, A. (2022, 10). *Travel time to cities and ports in the year 2015* (Tech. Rep.). Retrieved from https://figshare.com/articles/dataset/Travel_time_to_cities_and_ports_in_the_year_2015/7638134 doi: 10.6084/m9.figshare.7638134.v4
- Nelson, A., Weiss, D. J., van Etten, J., Cattaneo, A., McMenomy, T. S., & Koo, J. (2019). A suite of global accessibility indicators. *Scientific data*, 6(1), 1–9.
- Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2), 299–319.
- Nordhaus, W., Azam, Q., Corderi, D., Hood, K., Victor, N. M., Mohammed, M., ... Weiss, J. (2006). The g-econ database on gridded output: methods and data. *Yale University, New Haven*, 6, 11.
- Oshri, B., Hu, A., Adelson, P., Chen, X., Dupas, P., Weinstein, J., ... Ermon, S. (2018). Infrastructure quality assessment in africa using satellite imagery and deep learning. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 616–625).

- Peng, C., & Chen, W. (2021). Roads to development? Examining the Zambian context using AI-Sat.
- Pollmann, M. (2020). Causal inference for spatial treatments. *arXiv preprint arXiv:2011.00373*.
- Ratledge, N., Cadamuro, G., de la Cuesta, B., Stigler, M., & Burke, M. (2022). Using machine learning to assess the livelihood impact of electricity access. *Nature*, 611(7936), 491–495.
- Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427), 846–866.
- Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 931–954.
- Román, M. O., Wang, Z., Sun, Q., Kalb, V., Miller, S. D., Molthan, A., ... others (2018). Nasa's black marble nighttime lights product suite. *Remote Sensing of Environment*, 210, 113–143.
- Sahr, K. (2022). User documentation for discrete global grid generation software. *Southern Oregon Univ., Ashland, OR, USA, Tech. Rep. Dggrid version, 7.5*.
- Sahr, K., White, D., & Kimerling, A. J. (2003). Geodesic discrete global grid systems. *Cartography and Geographic Information Science*, 30(2), 121–134.
- Stekhoven, D. J., & Bühlmann, P. (2012). Missforest: non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112–118.
- Storeygard, A. (2016). Farther on down the road: transport costs, trade and urban growth in sub-saharan africa. *The Review of economic studies*, 83(3), 1263–1295.
- Tibshirani, J., Athey, S., Sverdrup, E., & Wager, S. (2023). grf: Generalized random forests [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=grf> (R package version 2.3.0)
- Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1).
- Weiss, D. J., Nelson, A., Gibson, H., Temperley, W., Peedell, S., Lieber, A., ... others (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, 553(7688), 333–336.
- Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., ... Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1), 2583.

Appendix

Constructing the Africa Infrastructure Database

The basis for the database is the Open Street Map (OSM) of Africa from April 2024, downloaded from Geofabrik.de. OSM has three basic data structures (nodes, ways, and relations) labeled through tags. A tag consists of a key and a value. Under the free tagging system, an object can have unlimited tags. However, the community agrees on certain key-value combinations for the most commonly used tags, which act as informal standards. In particular, the [OSM Feature Documentation](#) lists 29 primary tags, such as amenity, building, highway, water, landuse, shop, craft, etc. These primary tags are used with specified values to classify certain features, for example, amenity = school, shop = bakery, or building = hotel, and often accompanied by supplementary tags providing more precise information about a feature, e.g., name, description, denomination, etc.

There are three main obstacles to reconciling this tagging system with the economic significance of map features: (1) a feature can be classified according to multiple primary tags, e.g., amenity = school, building = education, landuse = education or amenity = hospital, healthcare = laboratory, emergency = yes; (2) sometimes classifications based on primary tags can conflict, e.g., for a religious school amenity = school and religion = christian, or for a hotel with restaurant building = hotel and amenity = restaurant; (3) the proximity of the object to the economic activity concerned may not be very clear, e.g., amenity = school and landuse = education both signify the primary use of an object for educational purposes, but the 'school' tag is a lot more specific. The tagging system of OSM is also subject to changes, and due to the crowdsourced nature of the map, some features are not classified according to current standards.

Constructing a functional classification of OSM is thus more an art than a science. One needs to devise a system of economic categories and lay out all tags according to which an economic category is to be assigned and the order in which these tags and categories are to be matched. This needs to be informed by both the OSM tagging standards and empirical accounts of current mapping practice. Once a classification scheme has been specified and applied to the map, features assigned to multiple categories need to be investigated, and the classification needs to be iteratively refined to minimize overlap and misclassification.

Following this process, I have developed a classification scheme to classify point and polygon (closed ways) map features into 33 economic categories based on 33 (mostly primary) tags and 341 values to be matched, including matching on any value for specific tags like 'sport' or 'power.' Within each category, tags providing precise information about the nature of the feature are matched first, and more general tags like 'building' or 'landuse' are matched last. The classification excludes natural features like mountains or lakes, natural or administrative boundaries, and minor features with little economic significance, such as traffic signs or flag poles. Minor infrastructure related to power, telecommunications, and military purposes is, however, included.

The [*osmclass*](#) R package developed for this purpose helps apply such classifications - defined as nested lists of categories, tags, and values - to OSM PBF files imported as spatial data frames. It includes the classifications used in this paper. Table A1 summarizes the classification and features extracted from the Africa OSM of April 2024, sorted by the number of features on the map.³⁰

³⁰This is not the order in which categories were matched, which was chosen to minimize misclassification in Africa. A detailed view of the classification is provided in the R package at <https://github.com/SebKrantz/osmclass/blob/main/R/classifications.R>.

Table A1: Classification of Africa OSM Point and Polygon Features: April 2024

Category	NTags	NVals	Tags and Number of Matched Values	N
residential	3	11	building (8), building:use (2), landuse (1)	6,382,886
power	4	4	power (all), utility (1), building (1), tower:type (1)	1,672,204
farming	4	19	place (1), man_made (1), building (9), landuse (8)	1,252,658
construction	2	2	building (1), landuse (1)	566,912
transport	12	49	amenity (21), highway (9), railway (all), aerialway (all), waterway (6), aeroway (all), public_transport (all), bridge (all), junction (all), office (2), man_made (2), building (3)	453,510
shopping	4	9	amenity (2), shop (2), building (4), landuse (1)	358,407
education	3	7	amenity (3), building (3), landuse (1)	328,016
sports	3	13	leisure (7), sport (all), building (5)	177,134
facilities	2	19	amenity (18), building (1)	161,083
religion	6	21	amenity (6), building (11), office (1), landuse (1), religion (all), denomination (all)	118,861
food	1	7	amenity (7)	115,087
health	3	11	amenity (9), healthcare (all), building (1)	106,353
utilities_other	5	18	man_made (12), water (1), office (1), building (3), landuse (1)	95,801
commerical	2	2	building (1), landuse (1)	90,549
industrial	4	7	industrial (all), man_made (2), building (1), landuse (3)	84,928
recreation	4	20	amenity (5), leisure (13), landuse (1), building (1)	79,447
historic	1	1	historic (1)	75,332
accommodation	2	8	tourism (7), building (1)	67,054
craft	1	1	craft (all)	59,246
tourism	3	4	tourism (1), shop (1), office (2)	47,617
financial	2	11	amenity (6), office (5)	46,040
institutional	3	7	office (5), building (1), landuse (1)	42,775
public_service	2	12	amenity (10), building (2)	36,658
mining	2	5	man_made (4), landuse (1)	29,711
communications	7	17	amenity (2), telecom (all), communication (all), utility (1), man_made (6), office (1), tower:type (5)	28,740
storage	3	6	man_made (2), building (3), landuse (1)	28,308
office_other	2	2	office (all), building (1)	26,158
waste	4	8	amenity (5), water (1), man_made (1), landuse (1)	21,313
education_alt	3	9	amenity (6), office (2), building (1)	13,852
military	3	5	military (all), building (3), landuse (1)	12,696
entertainment	2	17	amenity (13), leisure (4)	6,028
emergency	1	1	emergency (1)	4,622
creativity	3	8	amenity (2), leisure (1), office (5)	2,142
SUM	33	341		12,592,128

Table A2 shows a similar classification of line-based features. Residential roads are excluded as they are irrelevant to trade and strongly overlap with residential buildings recorded on the map. Smaller natural water features such as streams, wadis, and ponds are also excluded.

Table A2: Classification of Africa OSM Line Features: April 2024

Category	NTags	NVals	Tags and Number of Matched Values	N	Length (Km)
road	1	10	highway (10)	763,912	1,621,144
waterway	3	12	waterway (8), water (1), man_made (3)	359,756	1,507,112
power	1	1	power (all)	120,052	422,504
railway	1	1	railway (all)	84,707	128,408
aeroway	1	1	aeroway (all)	27,360	11,019
pipeline	1	1	man_made (1)	9,453	55,394
storage	1	2	man_made (2)	8,551	389
ferry	1	1	route (1)	2,412	48,259
aerialway	1	1	aerialway (all)	171	175
telecom	2	2	telecom (all), communication (all)	87	28,682
SUM	11	32		1,376,805	3,834,579

Harmonized Classification

I then combine the point/polygon data from OSM with the 824 thousand POIs from the Overture Places layer to create a uniform classification. This involves creating 47 detailed categories and matching features to them based on OSM tags or primary POI categories in Overture maps, of which there are more than 1000. To simplify the dataset again for most analytical use cases and ensure that each category has sufficient features, I combined some of the 47 categories, yielding a simplified classification of 26 categories. The other POI data, e.g., from All The Places, is more limited and easier to classify (e.g., mostly shops, restaurants, and hotels clearly tagged) and can seamlessly be assigned to the 47 categories.

Deduplication

Having classified POIs from 11 different sources (Table 1) into 47 categories, it remains to sort out duplicates across sources. The order of precedence is to favour curated data (such as Global Integrated Power Tracker or health facilities by [Maina et al. \(2019\)](#)) above OSM, which in turn takes precedence over Overture places (which various [online appraisals](#) found less accurate). POIs are then deduplicated within each category and 10m square. This resolution is motivated by considering a dense shopping mall where stores may be only 10m apart. POIs are resolved to grid cells by dividing their coordinates by the degree-equivalent of 10m at the equator $[10/(40075017/360)] \approx 9e-5$, subtracting the modulus of this division from the coordinates and using them to group and deduplicate features. To ensure equal distance representation across Africa, longitudes are multiplied beforehand by $\cos(lat \times \pi/180)$, where lat is the latitude. After a deduplication round, the coordinates are incremented by 1m degree-equivalent ($\approx 9e-6$), and the process is repeated. A sequence of such 1m nudges to the lon and lat coordinates is used to 'shift the grid' across space in a structured way until no more duplicates can be found. Since which POIs are first compared depends on the initial position of the 'grid,' there is some path dependence in this process. However, the clear hierarchy across sources ensures that curated datasets are generally fully retained.

Table A3 summarizes the finally classified and deduplicated POI data, including the corresponding number of POIs and the share of OSM and OSM polygons (tagged buildings).

Linestring (network) data is mainly taken from OSM, which is a reliable data source for roads, waterways, and railways. However, its coverage of power lines is limited to major lines in most regions. To increase granularity, I add Africa electricity grid maps from the EU's Joint Research Center ([Kakoulaki & Moner-Girona, 2020](#)) and the [World Bank](#), which more than doubles the total length of power lines observed from 423 thousand km in OSM to 967 thousand km (Table 2) following harmonization. To combine/harmonize the linestrings, I compute a geometric union between all three data sources. This obscures definitions of individual power lines from different datasets and breaks up the linestrings into smaller segments, but for the purposes of this research, only the total length of lines per cell matters. Table 2 summarizes the final lines dataset.

Table A3: Africa Infrastructure Database: Places Dataset by Category

detailed (47)	simplified (26)	count	perc	polygons	poly_perc	osm	osm_perc	src_cat
residential	residential	6,276,302	41.42	6,256,643	99.69	6,267,969	99.87	12
communications_network	communications	1,914,448	12.64	502	0.03	20,092	1.05	31
communications_other	communications	14,245	0.09	826	5.80	8,239	57.84	31
power	power	1,654,373	10.92	1,711	0.10	1,651,175	99.81	92
farming	farming	1,228,969	8.11	1,219,018	99.19	1,223,781	99.58	28
construction	construction	585,168	3.86	565,752	96.68	574,167	98.12	90
transport_other	transport	303,186	2.00	21,166	6.98	293,233	96.72	188
automotive	transport	164,845	1.09	61,571	37.35	118,311	71.77	84
education_essential	education	392,361	2.59	216,007	55.05	326,848	83.30	27
education_other	education	27,943	0.18	2,520	9.02	10,811	38.69	36
shopping_essential	shopping	227,820	1.50	43,393	19.05	173,233	76.04	42
shopping_other	shopping	176,718	1.17	5,459	3.09	106,535	60.29	1,239
wholesale	shopping	9,994	0.07	5,094	50.97	6,185	61.89	18
facilities	public_service_utility	173,669	1.15	50,344	28.99	170,252	98.03	37
utilities_other	public_service_utility	109,465	0.72	18,261	16.68	108,419	99.04	48
public_service	public_service_utility	58,369	0.39	18,571	31.82	40,192	68.86	32
health_essential	health	247,386	1.63	33,318	13.47	95,917	38.77	233
health_specialized	health	8,295	0.05	38	0.46	538	6.49	51
health_other	health	5,722	0.04	125	2.18	1,037	18.12	68
sport	sport	181,758	1.20	123,156	67.76	158,453	87.18	242
industrial	mining_industrial	138,407	0.91	62,032	44.82	121,246	87.60	156
mining	mining_industrial	30,209	0.20	22,378	74.08	29,824	98.73	10
SEZ	mining_industrial	387	0.00	0	0.00	0	0.00	6
religion	religion	161,716	1.07	59,682	36.91	116,015	71.74	84
food	food	160,645	1.06	7,602	4.73	74,966	46.67	141
financial	financial	151,493	1.00	3,289	2.17	42,018	27.74	37
accommodation	accommodation	123,248	0.81	19,853	16.11	61,228	49.68	34
parks_and_nature	tourism_recreation	66,925	0.44	43,918	65.62	57,798	86.36	32
tours_and_sightseeing	tourism_recreation	40,796	0.27	1,698	4.16	26,200	64.22	42
museums	tourism_recreation	5,096	0.03	812	15.93	1,805	35.42	28
beaches_and_resorts	tourism_recreation	4,856	0.03	171	3.52	312	6.43	8
outdoor_activities	tourism_recreation	3,110	0.02	157	5.05	1,033	33.22	53
commercial	commercial	109,580	0.72	97,251	88.75	108,351	98.88	11
historic	historic	100,738	0.66	27,458	27.26	72,683	72.15	94
beauty	beauty	65,711	0.43	769	1.17	27,305	41.55	29
professional_services	services	30,948	0.20	413	1.33	2,427	7.84	65
home_services	services	27,575	0.18	391	1.42	5,466	19.82	78
business_services	services	6,168	0.04	56	0.91	382	6.19	31
institutional	institutional	62,313	0.41	9,974	16.01	44,450	71.33	65
drinking	entertainment	35,314	0.23	2,256	6.39	23,858	67.56	18
performing_arts	entertainment	11,584	0.08	1,415	12.22	2,785	24.04	19
nightlife	entertainment	7,865	0.05	302	3.84	2,011	25.57	14
gaming	entertainment	1,425	0.01	65	4.56	250	17.54	13
storage	storage	28,077	0.19	22,761	81.07	27,757	98.86	17
military	military_emergency	13,104	0.09	9,953	75.95	12,441	94.94	41
emergency	military_emergency	3,348	0.02	75	2.24	3,200	95.58	39
port	port	244	0.00	0	0.00	0	0.00	9
total	total	15,151,918	100.00	9,038,206	59.65	12,221,198	80.66	35

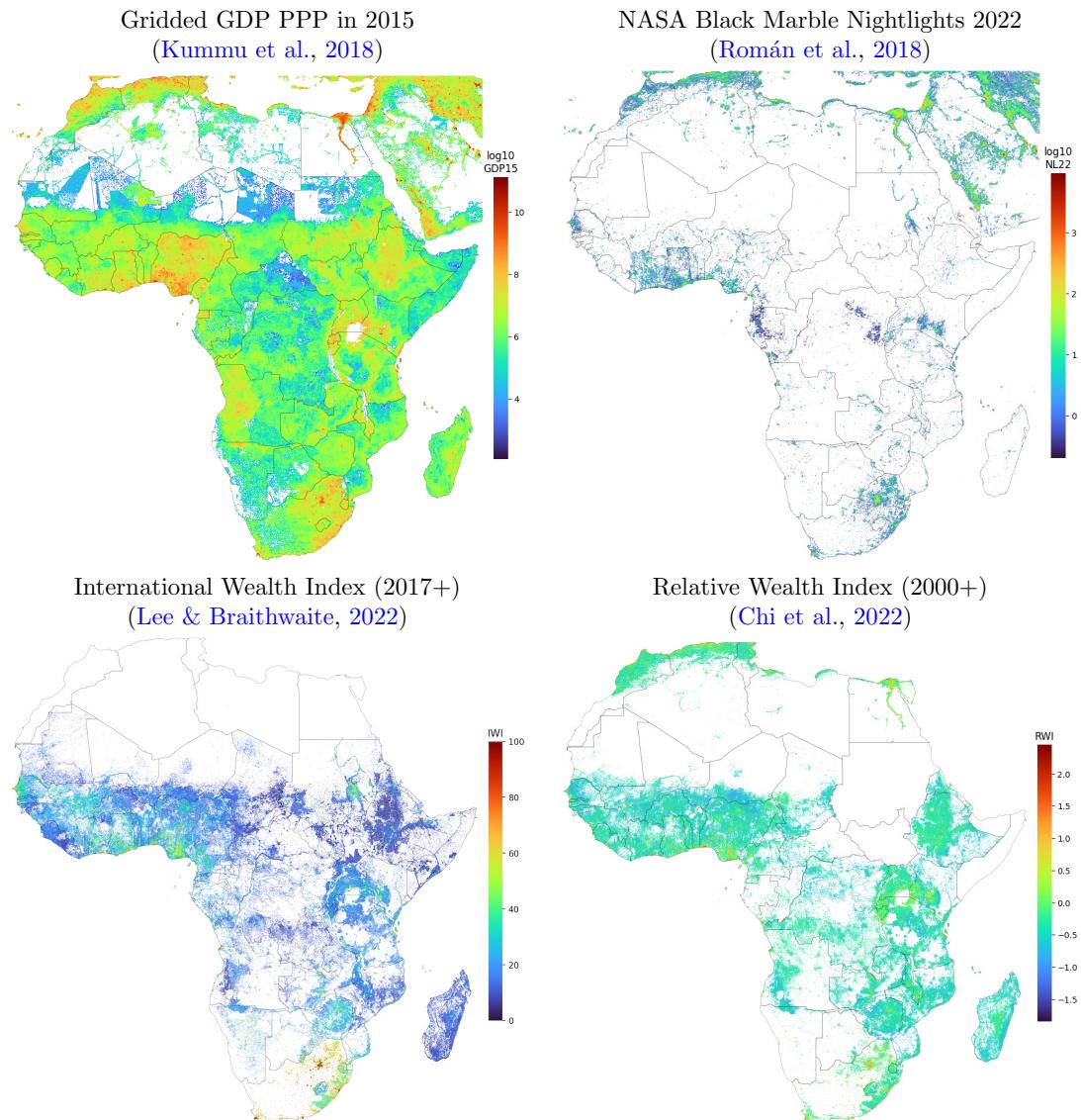
Notes: Table summarizes places of interest (POIs) data collected from different sources (Table 1), classified into 47 economic categories and deduplicated by category and location (as described above). Column ‘count’ records the number of POIs, ‘perc’ the percentage in total POIs, ‘polygons’ the number of tagged OSM polygons (buildings) interpreted as POIs, ‘poly_perc’ the percentage of POIs that are polygons, ‘osm’ the number of POIs taken from OSM, ‘osm_perc’ the fraction of POIs from OSM, and ‘src_cat’ the number of primary categories (across sources) mapping to a specific detailed category (e.g., there are 1,239 different kinds of small/specialized shops mapping to the ‘shopping_other’ category).

Descriptive Statistics

Table A4: Dataset of Spatial Predictors (Simplified Categories) over 96km² Grid, $N = 160,499$

#	Variable	Ndist	Mean	SD	Min	25%	50%	75%	Max
<i>POIs: simple counts combined across detailed categories</i>									
1	accommodation	224	0.8	9.8	0.0	0.0	0.0	0.0	1399.0
2	automotive	307	1.0	17.7	0.0	0.0	0.0	0.0	2114.0
3	beauty	207	0.4	10.3	0.0	0.0	0.0	0.0	1382.0
4	commercial	205	0.7	34.7	0.0	0.0	0.0	0.0	8264.0
5	construction	536	3.6	86.6	0.0	0.0	0.0	0.0	13209.0
6	farming	770	7.7	49.8	0.0	0.0	0.0	1.0	2581.0
7	financial	276	0.9	18.1	0.0	0.0	0.0	0.0	2208.0
8	food	316	1.0	19.0	0.0	0.0	0.0	0.0	1696.0
9	historic	183	0.6	15.6	0.0	0.0	0.0	0.0	5580.0
10	institutional	169	0.4	6.6	0.0	0.0	0.0	0.0	794.0
11	port	4	0.0	0.0	0.0	0.0	0.0	0.0	3.0
12	power	2435	115.9	319.9	0.0	0.0	0.0	20.0	14536.0
13	religion	248	1.0	11.0	0.0	0.0	0.0	0.0	1079.0
14	residential	1739	39.1	488.8	0.0	0.0	2.0	10.0	65378.0
15	sport	280	1.1	31.9	0.0	0.0	0.0	0.0	5112.0
16	storage	133	0.2	5.9	0.0	0.0	0.0	0.0	765.0
17	transport_other	383	1.9	32.3	0.0	0.0	0.0	0.0	2999.0
18	communications	1194	14.9	198.7	0.0	0.0	0.0	0.0	22867.0
19	education	363	2.6	20.9	0.0	0.0	0.0	0.0	2414.0
20	health	332	1.9	22.9	0.0	0.0	0.0	1.0	3521.0
21	entertainment	420	1.7	32.4	0.0	0.0	0.0	0.0	3944.0
22	services	197	0.4	9.4	0.0	0.0	0.0	0.0	1573.0
23	shopping	698	5.5	126.8	0.0	0.0	0.0	0.0	21543.0
24	public_service_utility	140	0.4	5.7	0.0	0.0	0.0	0.0	952.0
25	mining_industrial	272	1.2	11.5	0.0	0.0	0.0	0.0	1027.0
26	military_emergency	75	0.1	1.7	0.0	0.0	0.0	0.0	249.0
27	tourism_recreation	288	1.0	15.8	0.0	0.0	0.0	0.0	2441.0
<i>Lines: total length in km in each cell</i>									
28	dam_len	1868	29.5	431.6	0.0	0.0	0.0	0.0	79889.0
29	aeroway_len	2607	66.6	805.7	0.0	0.0	0.0	0.0	68671.0
30	ferry_len	1362	57.4	958.9	0.0	0.0	0.0	0.0	95065.0
31	pipeline_len	4031	281.0	2675.5	0.0	0.0	0.0	0.0	331527.0
32	railway_len	7707	776.7	4975.3	0.0	0.0	0.0	0.0	706405.0
33	roads_paved	18393	2825.1	9028.5	0.0	0.0	0.0	0.0	465351.0
34	roads_unpaved	27757	6361.4	10741.4	0.0	0.0	0.0	10768.0	625255.0
35	waterway_len	36356	23069.8	150937.0	0.0	0.0	0.0	13141.5	11753756.0
<i>Raster Layers: sum of population and mean of travel time and internet speed (bytes/s) in each cell</i>									
36	internet_speed	53324	20800.5	22664.3	151.0	6209.0	12031.0	27115.0	278876.0
37	pop_gpw4_ages_0_14	14147	2323.1	9063.3	0.0	192.0	716.0	2013.0	729958.0
38	pop_gpw4_ages_15_49	15038	2786.9	14553.7	0.0	207.0	726.0	2071.0	1503562.0
39	ttime_city_50k	3380	281.8	457.2	0.0	76.0	156.0	312.0	9303.0
40	ttime_city_1m	4333	610.5	631.5	0.0	245.0	451.0	779.0	9868.0
41	ttime_port_any	4723	811.5	758.8	0.0	316.0	602.0	1054.0	10065.0
42	ttime_port_ml	5229	1093.5	909.3	1.0	453.0	849.0	1472.0	10216.0

Figure A1: Spatial Measures of Wealth and Economic Activity



Notes: Figure shows different measures of wealth and economic activity at their original resolution, with blank missing or zero values. The IWI is the main outcome measure, but all measures assist data aggregation (Section 2.5, Table A6).

Figure A2: Histogram of Wealth/Activity Measures in 96km^2 Grid, Avg. Obs.: 101,417

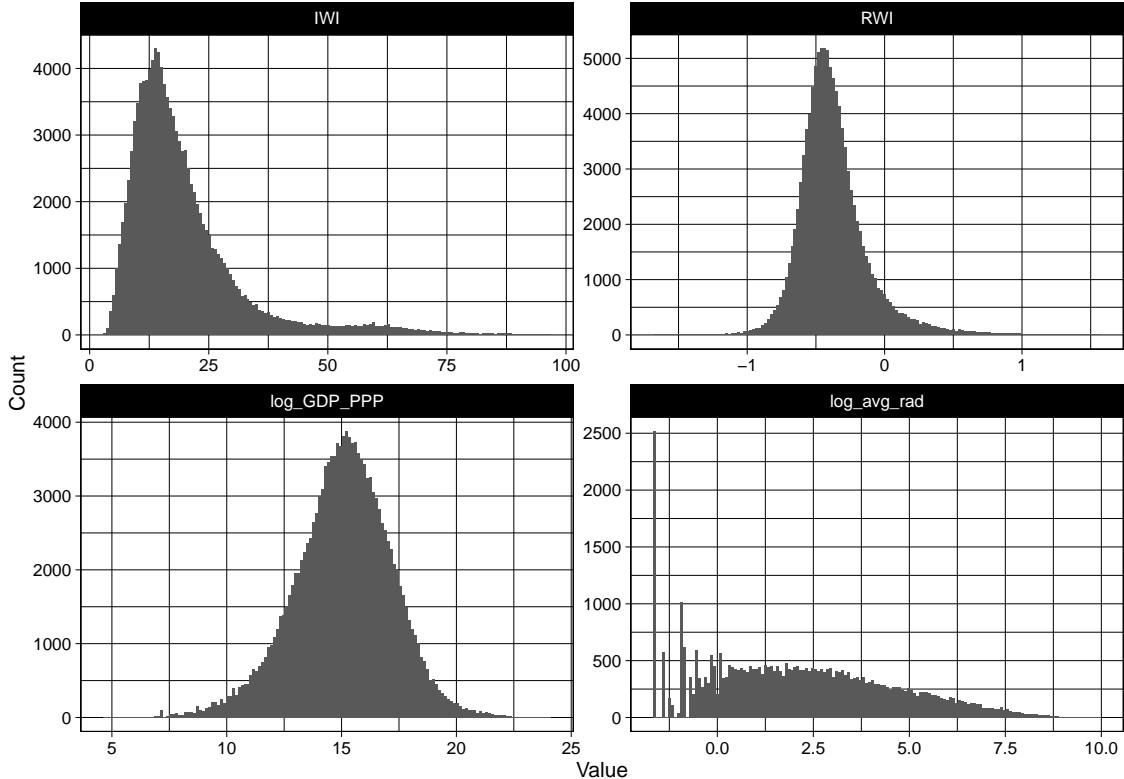
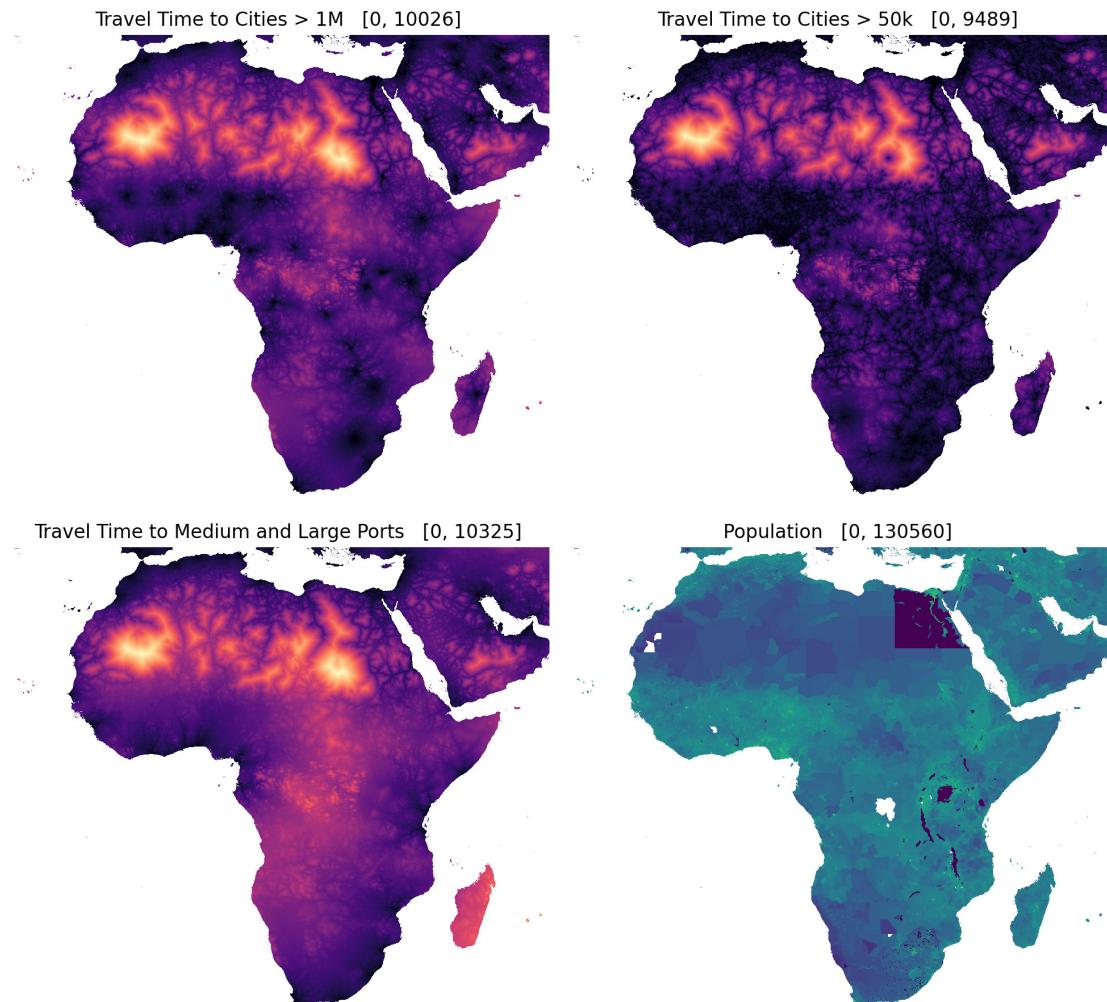


Table A5: Pearson's Correlations of Wealth/Activity Measures in 96km^2 Grid

		log_GDP_PPP	log_avg_rad	IWI	RWI
Pooled	log_GDP_PPP	1 (126123)			
Across	log_avg_rad	.453* (33293)	1 (36656)		
Countries	IWI	.397* (98948)	.576* (25608)	1 (105924)	
	RWI	.449* (95295)	.577* (31858)	.550* (88101)	1 (101456)
Scaled	log_GDP_PPP	1 (126123)			
and	log_avg_rad	.436* (33293)	1 (36656)		
Centered	IWI	.363* (98948)	.578* (25608)	1 (105924)	
by Country	RWI	.349* (95295)	.560* (31858)	.533* (88101)	1 (101456)

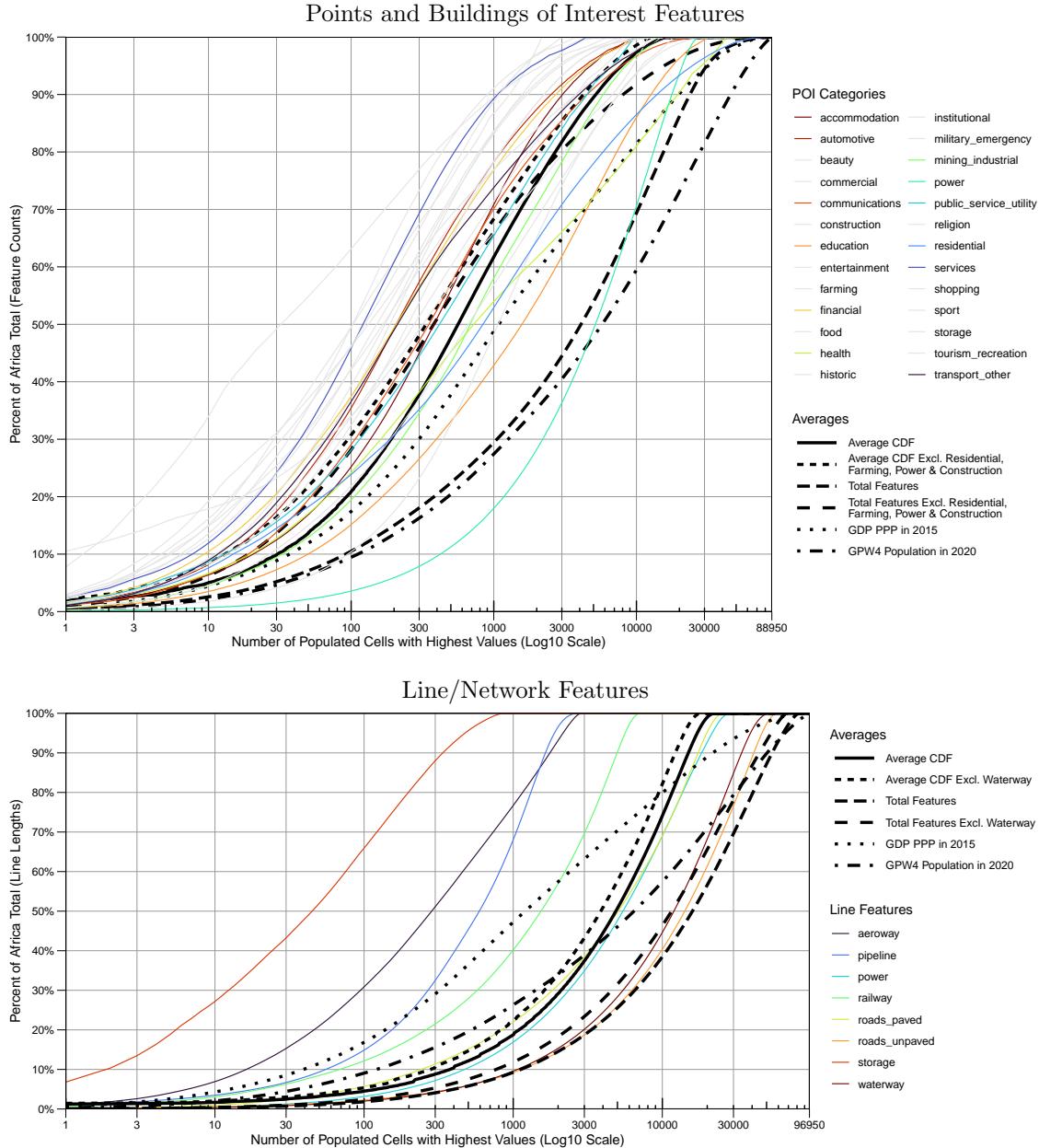
Notes: N. obs in parentheses, a '*' denotes significance at the 5% level.

Figure A3: Raster Covariate Layers: GPW4 Population (2020) and Market Access (2015)



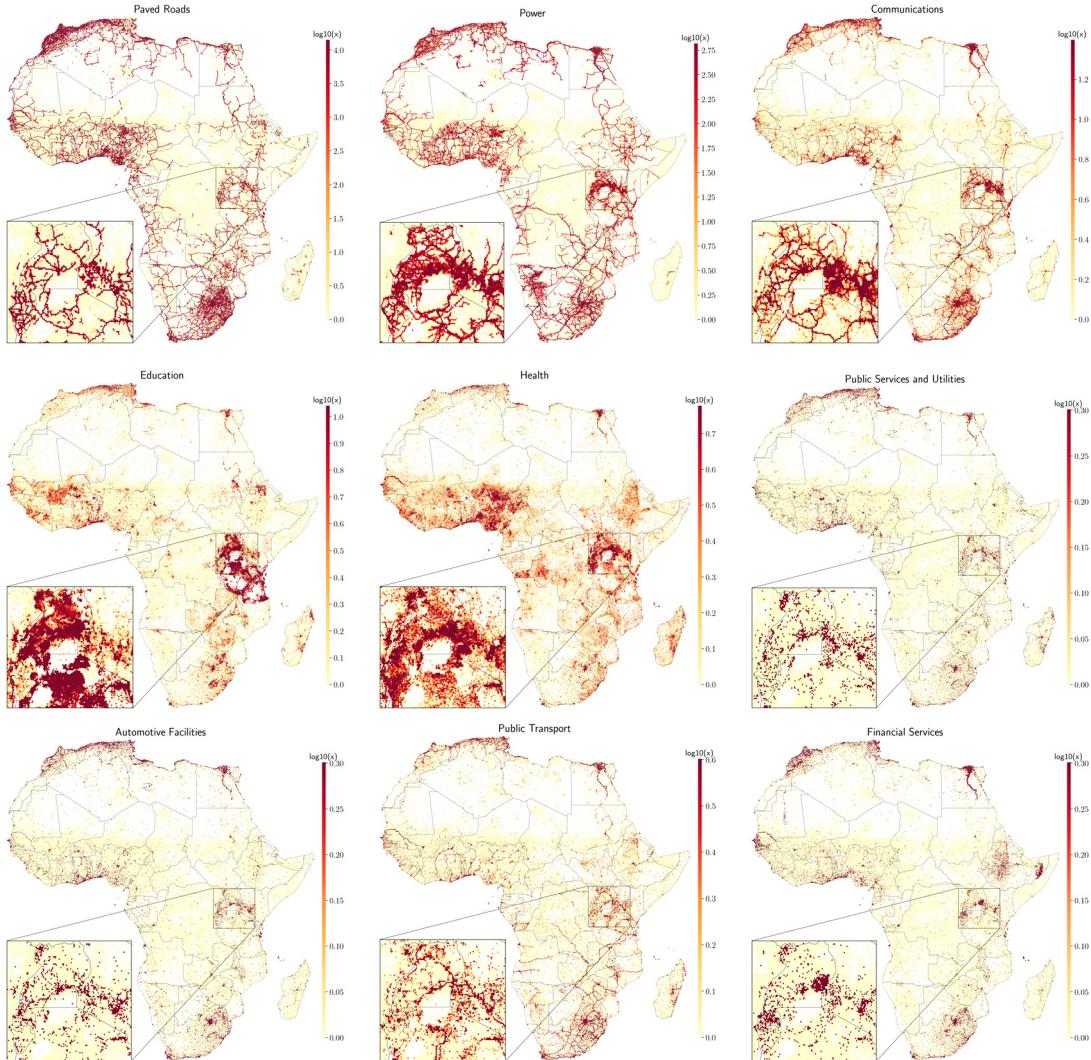
Notes: Figure shows accessibility maps from [Nelson et al. \(2019\)](#); [Nelson \(2022\)](#), and GPW4 population ([CIESIN, 2016](#)). The population layer has some missing data. These gaps are imputed in the final dataset following aggregation over a 96kkm^2 grid (in Section 2.4) using the 'missForest' algorithm by [Stekhoven & Bühlmann \(2012\)](#), yielding a 97% OOB- R^2 .

Figure A4: Feature Concentrations: Empirical CDF's



Notes: Figure shows empirical CDF's (with cell-counts on the x-axis) of POIs (top) and lines (bottom) in simplified economic categories. The CDFs are obtained by considering 88,960 grid cells with more than 10 persons/km² according to both GPW4 and WorldPop 2020 estimates, counting/summing the lengths of features in simplified categories, computing proportions across cells, sorting these proportions in descending order, and cumulatively summing them to create a CDF, which is plotted against the cell index indicated on a Log10 scale.

Figure A5: Spatial Distribution of Selected Infrastructures: Length/Count (Log10 Scale)



Notes: Figure shows total length in m (of paved roads and power lines) or count in 96km² ISEA3H cells for 9 key infrastructures on a log10 scale.

Table A7: Correlations of Total Infrastructure with Wealth/Activity and Population

	IWI	RWI	GDP_PPP	avg_rad	pop_gpw4	pop_wpop
<i>Cell-level, N = 88,961 populated cells (>10 persons/km²)</i>						
Total PP Feature Count	0.302	0.401	0.481	0.484	0.546	0.562
Total PP Features ERFPC	0.249	0.351	0.586	0.590	0.603	0.617
Total Line Length	0.519	0.578	0.477	0.623	0.491	0.515
Total Line Length EW	0.558	0.588	0.482	0.640	0.490	0.515
<i>Country-level, N = 55</i>						
Total PP Feature Count	-0.009	-0.263	0.538	-0.055	0.800	0.799
Total PP Features ERFPC	-0.010	-0.265	0.535	-0.057	0.799	0.799
Total Line Length	0.412	-0.040	0.730	0.265	0.719	0.721
Total Line Length EW	0.530	0.049	0.769	0.332	0.671	0.667

ERFPC = excluding residential buildings, farmland, power and construction

EW = excluding waterways

Table A6: Detailed Category Combination Weights (β_{jp}) for 3 Categories

communications	communications_network	communications_other	telecom_len
IWI	1.00	5.70	0.24
RWI	1.00	13.65	0.10
GDP_PPP	1.00	8.86	0.02
avg_rad	1.00	18.65	0.03
Average	1.00	11.72	0.098
mining_industrial	industrial	mining	SEZ
IWI	1.00	3.43	89.13
RWI	1.00	0.93	66.93
GDP_PPP	1.00	0.00	52.49
avg_rad	1.00	0.86	23.97
Average	1.00	1.31	58.13
shopping	shopping_essential	shopping_other	wholesale
IWI	1.00	1.46	15.41
RWI	1.00	1.46	14.73
GDP_PPP	1.00	3.82	52.37
avg_rad	1.00	2.57	18.71
Average	1.00	2.33	25.31

Notes: The weights are normalized, then averaged across outcomes.

Table A8: POIs in ISEA3H Grids of Different Resolutions

Res.	Area (km ²)	Spacing (km)	N. Cells (N)	N. Features	Mean[Feat./Cat.]	Corr. (r)	LM <i>R</i> ²
				Mean	Median	Mean	Median
8	7774.21	87.08	4,091	3,703.720	509	116.978	33.235
9	2591.40	50.28	11,243	1,347.676	144	58.674	16.200
10	863.80	29.03	29,247	518.067	48	33.677	9.250
11	287.93	16.76	71,192	212.832	20	21.489	6.000
12	95.98	9.68	160,719	94.276	10	14.657	4.000
13	31.99	5.59	336,587	45.016	6	10.154	3.000
14	10.66	3.23	646,959	23.420	4	7.121	2.000
						0.195	0.278

Notes: Res. is the grid resolution, also given in terms of cell area and spacing between the centroids of adjacent cells. N. Cells is the number of grid cells containing any POI in Africa. N. Features is average/median number of POIs in a cell. Mean[Feat./Cat.] is the average number of POIs per detailed category (47) in Table A3, computed for each cell and aggregated across cells using the mean or median. Corr. gives the average Pearson's correlation of the feature count within each category with each of the 4 spatial wealth/activity measures in Figure A1. Thus, it is an average of $47 \times 4 = 188$ correlation coefficients. LM gives the average R^2 of a linear model of the same 4 outcome measures against the features counted in 47 categories, thus it is the average of 4 R^2 estimates. The chosen resolution is highlighted.

Figure A6: Clustered Correlations of Variables

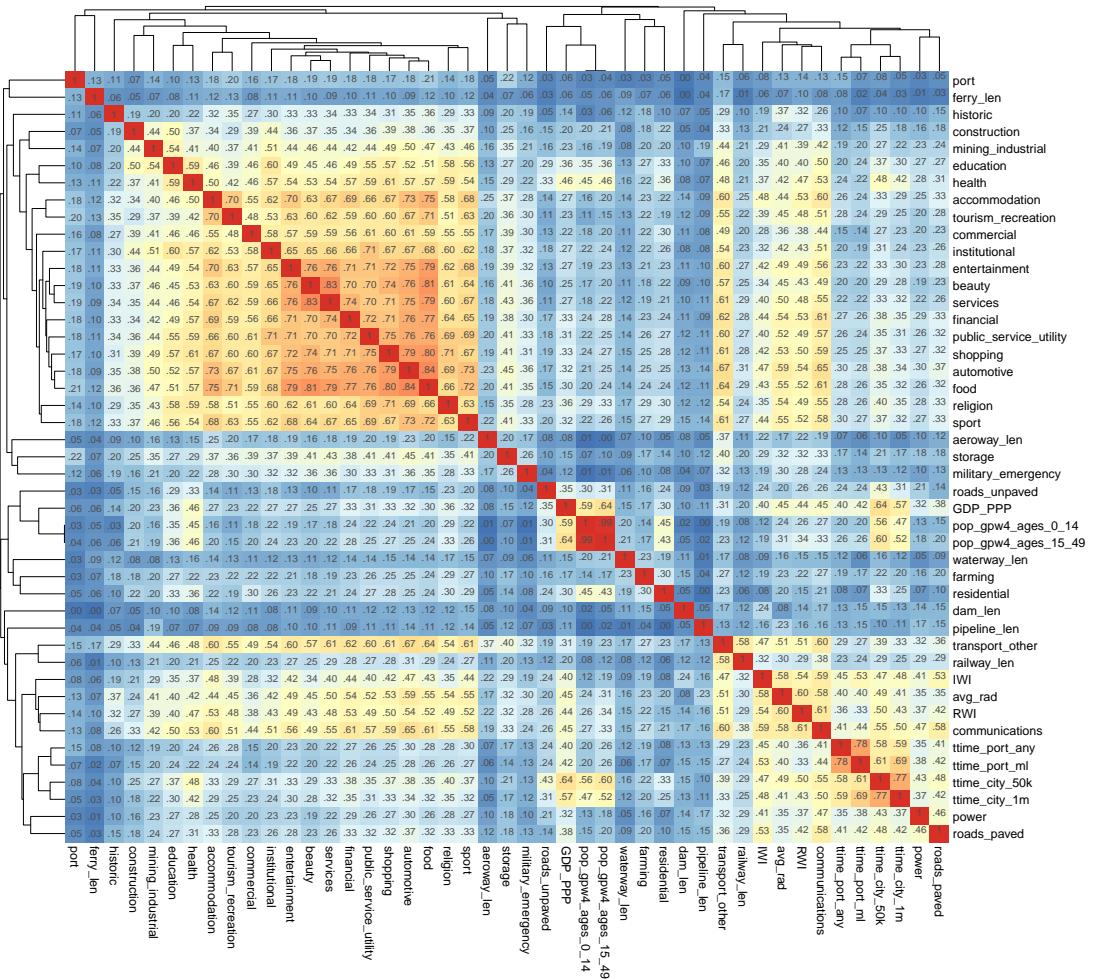


Figure A7: Index of Spatial Efficiency

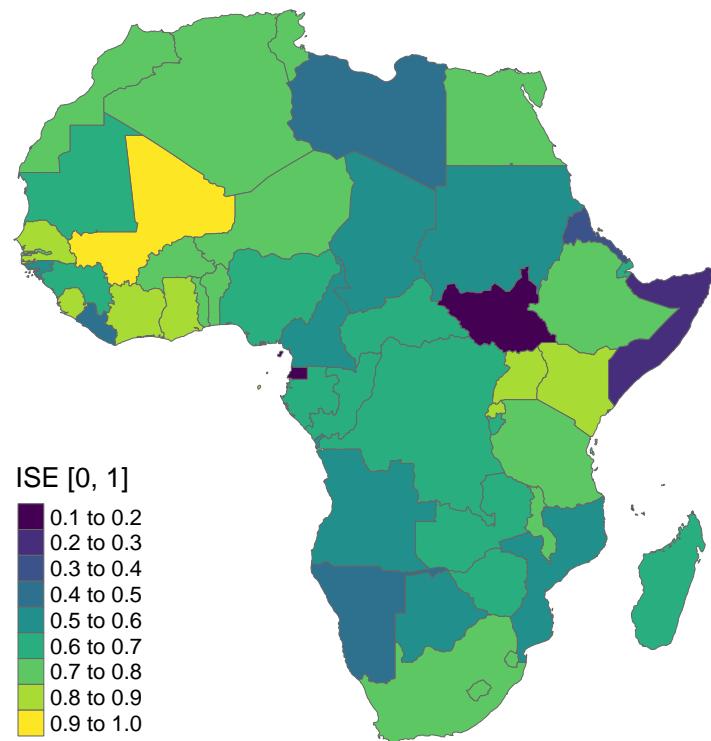
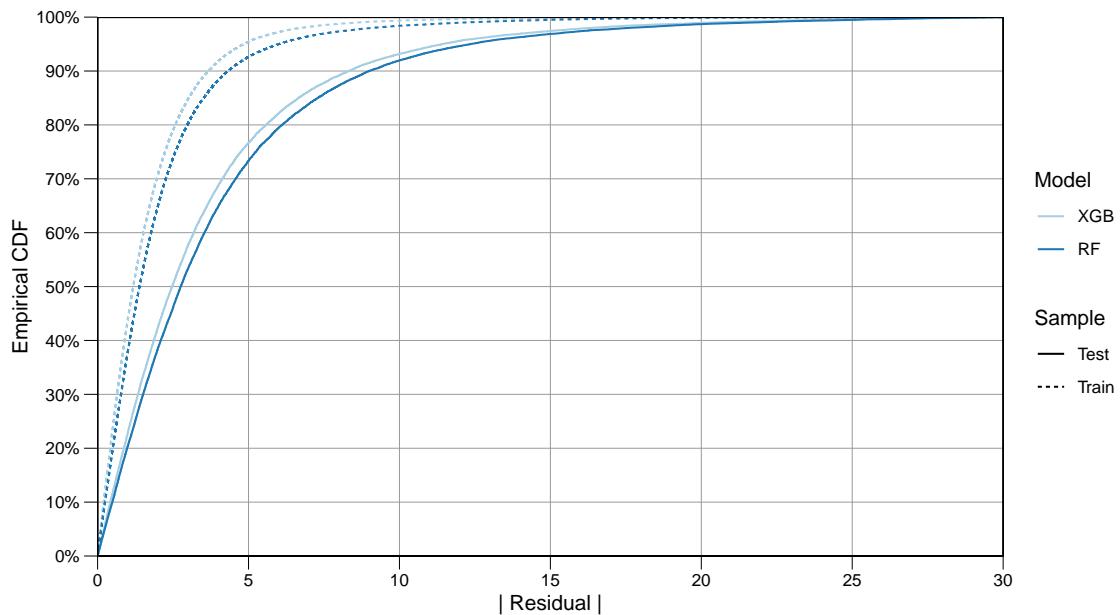
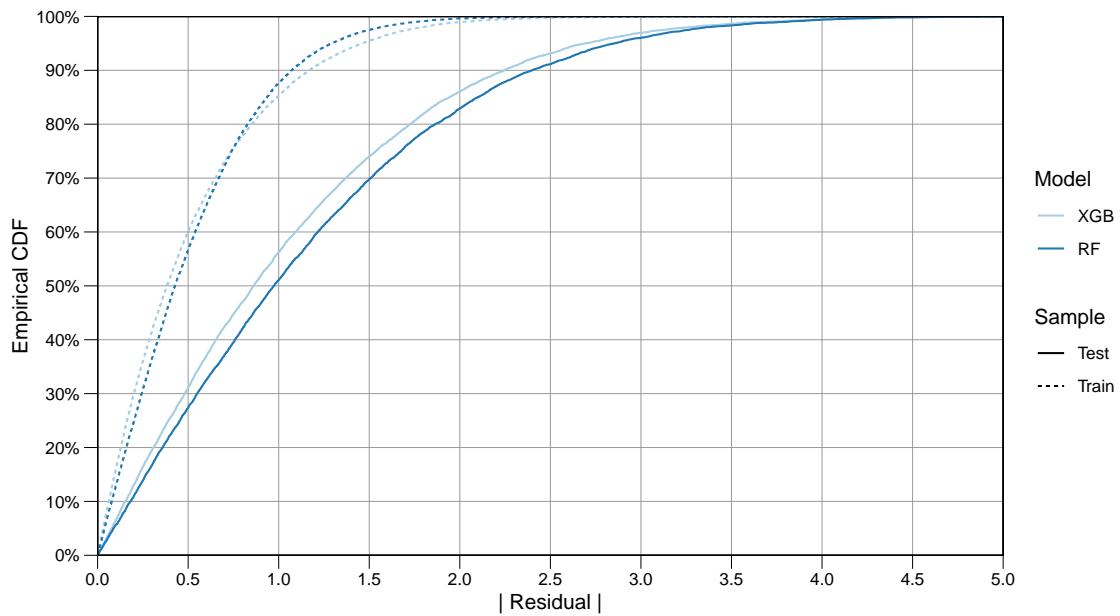


Figure A8: Empirical CDF's of Residuals from ML Models Predicting the IWI [0, 100]



Notes: Dataset has 42 predictors, not including spatial spillover variables. The training set has 86,354 observations, the test set 28,782. Training set R^2 is 96.0% for XGB and 93.9% for RF. Test set R^2 is 78.8% for XGB and 73.3% for RF.

Figure A9: Empirical CDF's of Residuals from ML Models Predicting the Log of Nightlights 2022



Notes: Dataset includes 42 predictors, excluding spillover variables. The training set has 28,976 observations, the test set 9,655. The training set R^2 is 91.8% for XGB and 92.7% for RF. The test set R^2 is 68.4% for XGB and 63.6% for RF.

SHAP Values for Other Outcomes

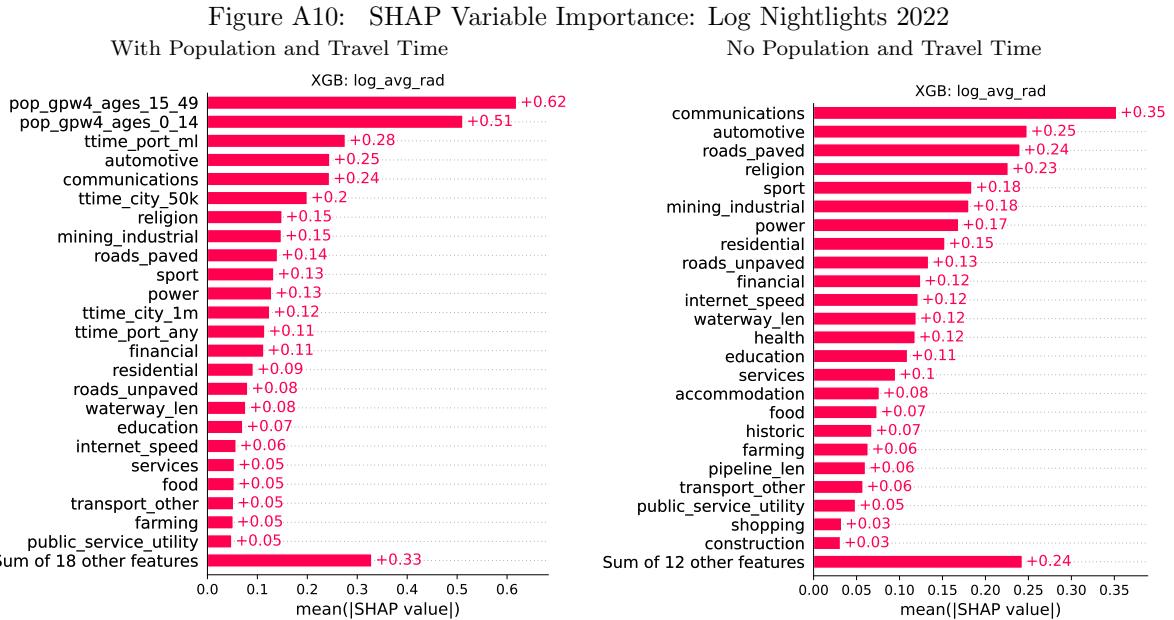
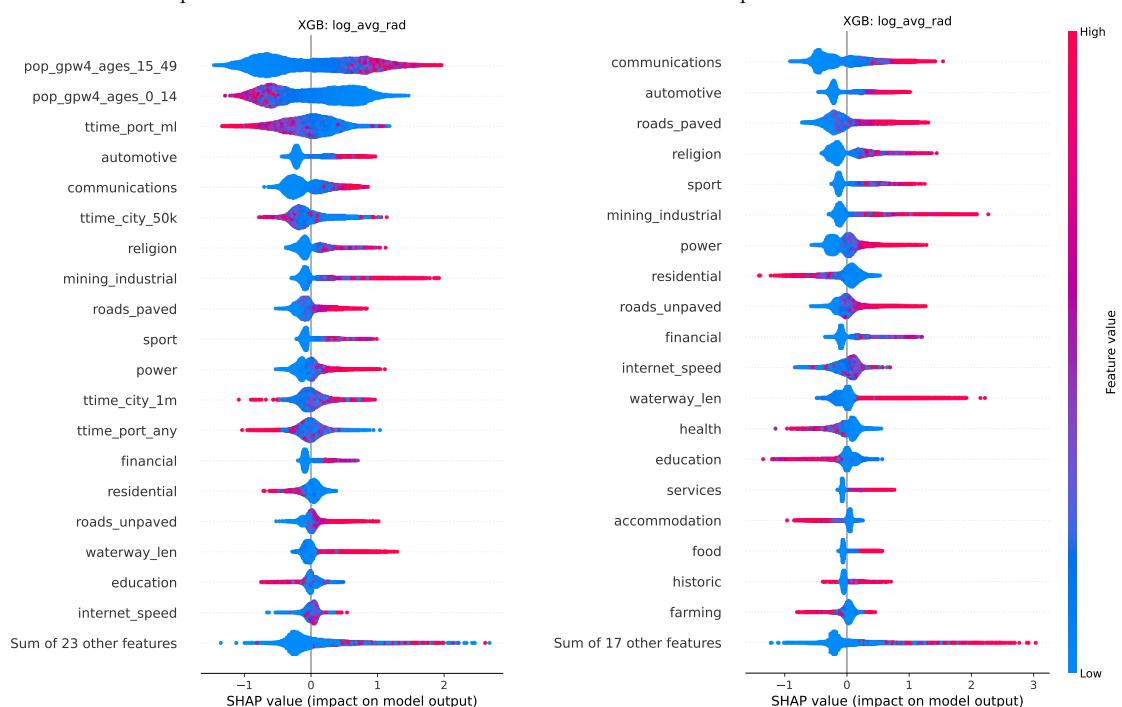


Figure A11: SHAP Effect Distribution: Log Nightlights 2022



SHAP Values from Feature Counts

Figure A12: SHAP Variable Importance: Simple Feature Counts
IWI, no POPTT Log Nightlights, no POPTT

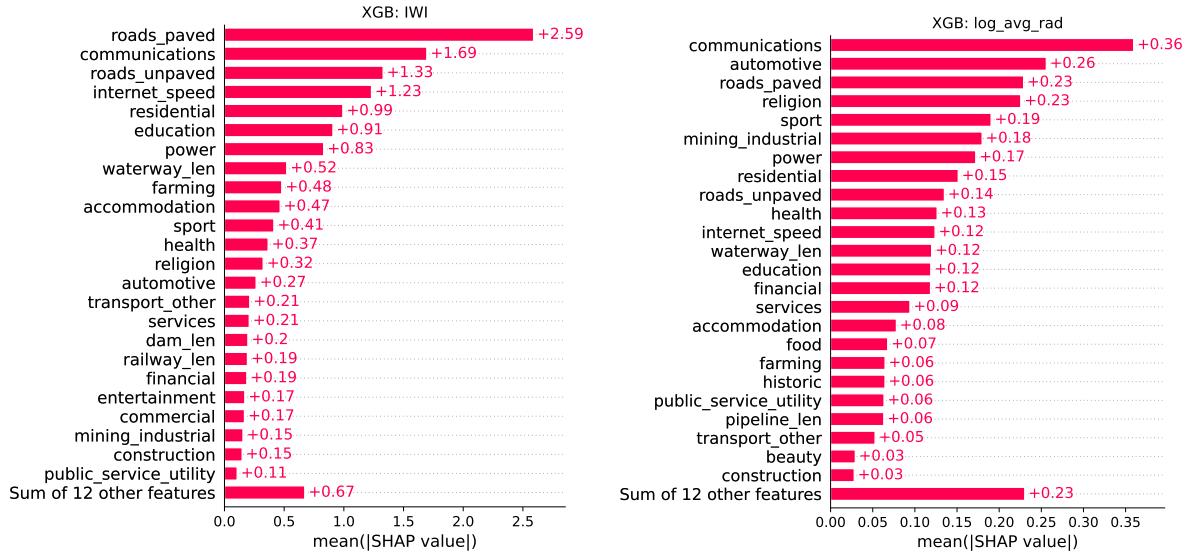
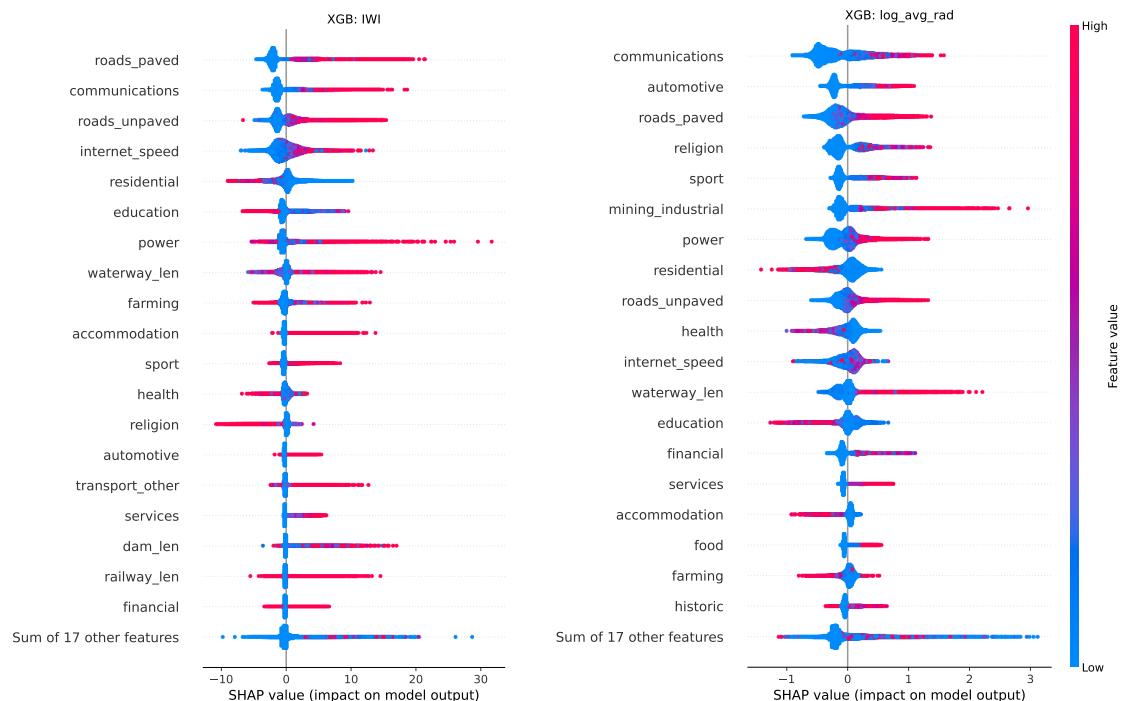


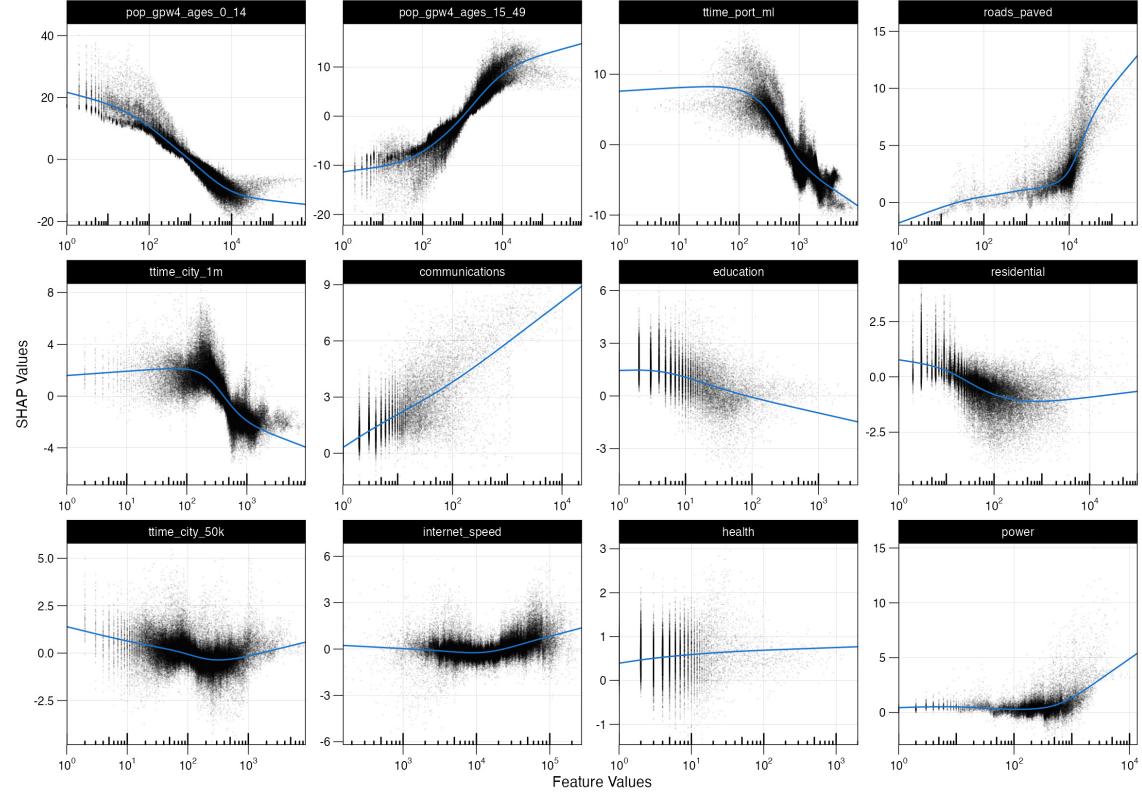
Figure A13: SHAP Effect Distribution: Simple Feature Counts
IWI, no POPTT Log Nightlights, no POPTT



SHAP Values Scatter (ALE) Plots

Figure A14: IWI: SHAP Values and Feature Levels - Top 12 Predictors

With Population and Travel Time



No Population and Travel Time

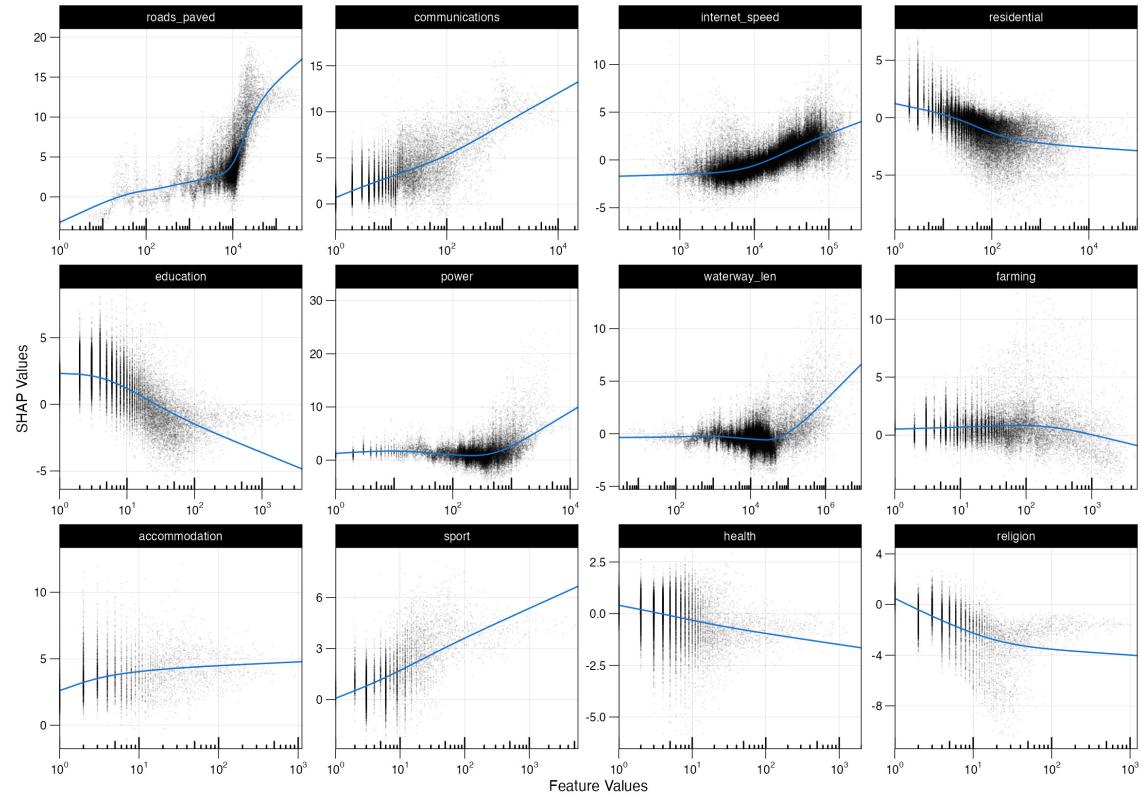
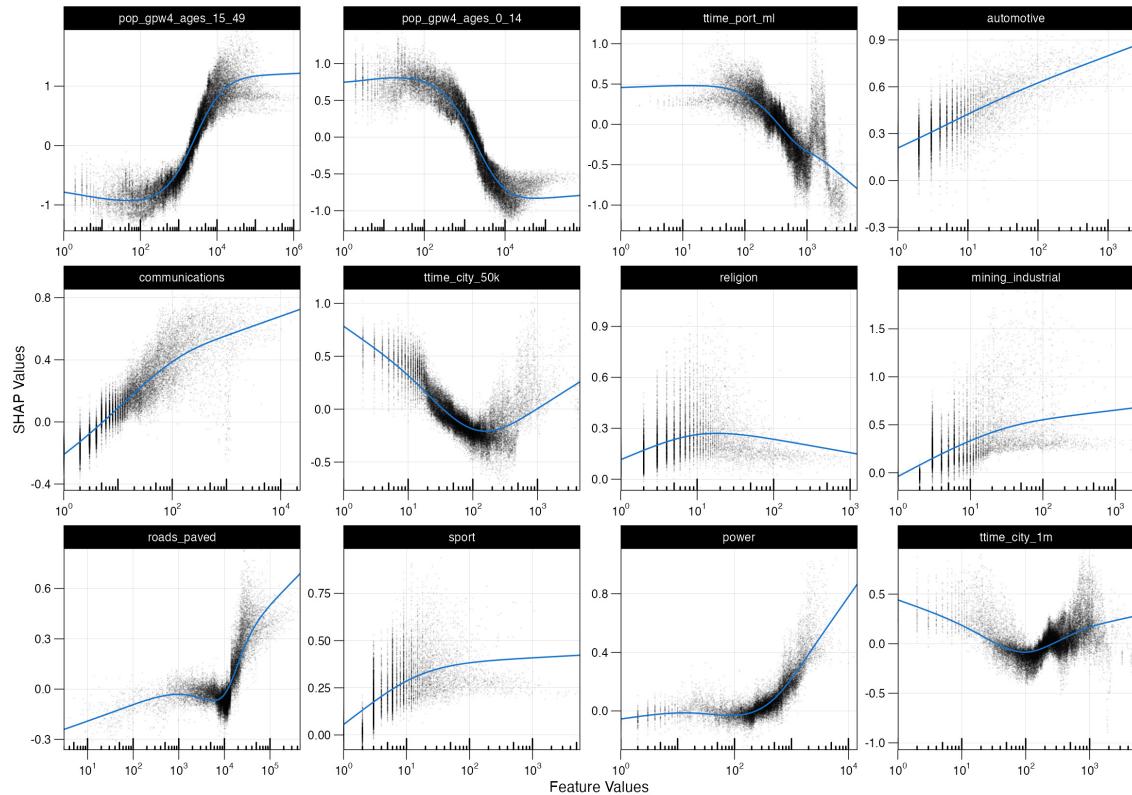
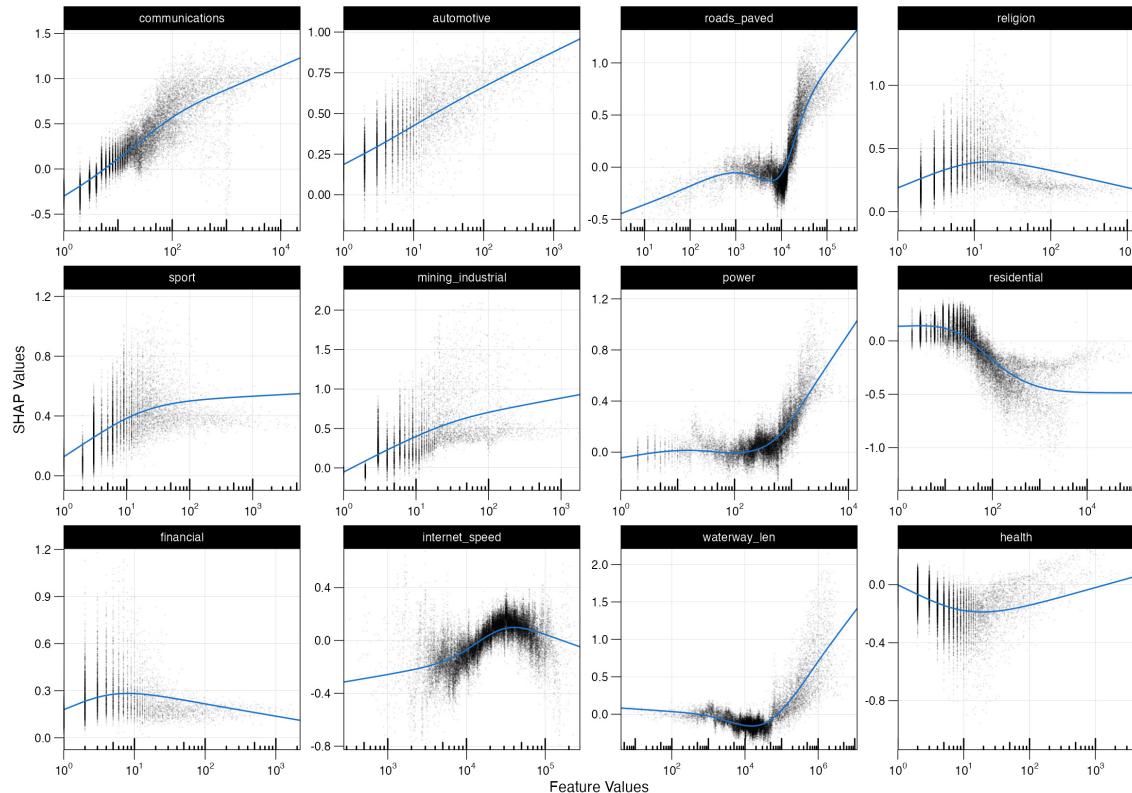


Figure A15: Log Nightlights 2022: SHAP Values and Feature Levels - Top 12 Predictors

With Population and Travel Time



No Population and Travel Time



Robustness Checks and Additional Results

Figure A16: CAPE Estimates for Power and Health using Only OSM Data (February 2024 Version)

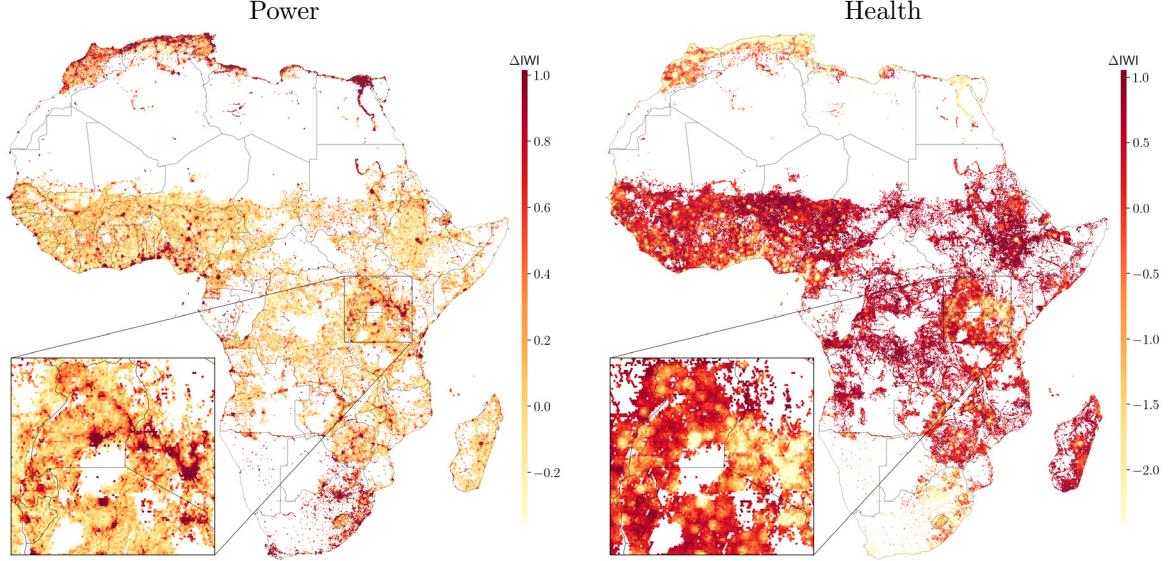


Table A9: ML Debiased Infrastructure and Favouritism: Simple Feature Counts

feature	Colonial Railroads		Political Favouritism				Ethnic Favouritism			
	Ln Kilometers Raw	Ctrl	Ln Years in Power Raw	Ctrl	Ever in Power Raw	Ctrl	Discriminated Raw	Ctrl	Excl. from Gov. Raw	Ctrl
<i>Raw Infrastructure Data, in Natural Logs</i>										
roads.paved	.517***	.29***	.521***	.02	1.506***	.355	-.164	.225	-.674***	-.061
power	.374***	.287***	.417***	.066	1.024***	.224	.571**	.304	-.186	.338*
education	.097***	.041***	.149***	-.02	.369***	-.005	.052	.038	-.29***	.001
health	.076***	.029***	.122***	-.048**	.324***	-.076	.048	-.05	-.128***	.037
communications	.212***	.115***	.236***	-.013	.622***	.078	.106	.118	-.229***	.133
public_service.utility	.037***	.02***	.063***	-.016	.161***	-.018	.01	.006	-.018	.037*
automotive	.062***	.036***	.08***	-.023	.203***	-.031	.045	.054	-.021	.065**
transport_other	.142***	.102***	.114***	-.021	.292***	-.013	.06	.071	-.031	.027
financial	.057***	.025***	.067***	-.049***	.169***	-.091**	.066**	.031	-.002	.009
services	.039***	.015***	.045***	-.034***	.104***	-.074***	.017	.03*	-.003	.028*
ttime.city.1m	-.162***	-.088***	-.189***	.051*	-.477***	.076	.098	.108	.316***	.066
ttime_port.any	-.121***	-.069***	-.21***	-.017	-.585***	-.104*	-.003	-.004	.366***	-.128***
residential	.032**	.052***	.167***	-.075	.444***	-.129	-.085	-.007	-.257**	.019
accommodation	.074***	.036***	.081***	-.013	.24***	.019	.07*	.07*	-.025	.008
<i>Debiased Data via Ensemble ML Models</i>										
roads.paved	.015	.012	-.031	-.067	.025	-.063	-.034	.191	.006	-.186
power	0	.003	.023	.028	.021	.024	.205**	.186	.048	.051
education	0	-.003	-.014	-.011	-.048	-.031	-.011	.002	0	-.007
health	-.005*	-.007**	-.005	-.007	-.005	-.014	-.014	-.046*	-.005	-.021
communications	-.007*	-.007	-.001	.016	-.005	.043	.018	-.011	-.032	-.039
public_service.utility	-.003*	-.001	.005	.006	.012	.015	-.003	-.009	.006	.022*
automotive	-.002	0	-.002	0	-.021	-.013	.018*	.019	-.005	.002
transport_other	.003	.003	-.004	-.001	-.02	-.012	.042***	.017	.021*	-.005
financial	-.002	-.001	-.012*	-.016**	-.024	-.034*	.018*	.02	.002	-.011
services	.001	-.002*	0	0	.001	.003	.002	.008	.001	.008
ttime.city.1m	-.005	-.002	.002	.002	-.025	-.024	.011	.015	.012	.046
ttime_port.any	-.014***	-.006*	-.033**	-.015	-.125***	-.081**	.008	.046	.034	-.003
residential	-.002	-.002	-.011	-.044	-.053	-.108	-.039	-.035	-.027	-.008
accommodation	.005***	.004*	0	0	.008	.007	.007	.027	-.003	-.01
Observations	5346	5346	5346	5346	5346	5346	385	385	385	385
Country FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Geographic Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Simulation Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1; based on cluster-robust standard-errors

Notes: See Graff (2024) Tables 1 and 2 for further details about the variables and controls.

Table A10: ML Debiased Infrastructure and Favouritism: Quantile Feature Counts

feature	Colonial Railroads		Political Favouritism				Ethnic Favouritism			
	Raw	Ctrl	Ln Kilometers	Ln Years in Power	Ever in Power	Discriminated	Excl. from Gov.	Raw	Ctrl	
<i>Raw Infrastructure Data, in Natural Logs</i>										
roads_paved	.517***	.29***	.521***	.02	1.506***	.355	-.164	.185	-.674***	-.067
power	.373***	.286***	.416***	.065	1.022***	.222	.571**	.273	-.186	.31*
education	.111***	.046***	.186***	-.012	.465***	.025	.047	.014	-.341***	.034
health	.079***	.031***	.137***	-.047*	.375***	-.063	.04	-.083	-.147***	.023
communications	.212***	.115***	.237***	-.012	.624***	.078	.106	.072	-.229***	.086
public_service_utility	.045***	.026***	.079***	-.011	.213***	.01	.037	.005	-.026	.029
automotive	.069***	.04***	.093***	-.022	.238***	-.022	.062*	.069*	-.02	.07**
transport_other	.146***	.105***	.119***	-.02	.306***	-.009	.075	.074	-.027	.048
financial	.058***	.026***	.067***	-.049***	.171***	-.091**	.064**	.028	-.005	.007
services	.039***	.015***	.045***	-.035***	.103***	-.075***	.017	.027	-.004	.024
ttime_city_1m	-.162***	-.088***	-.189***	.051*	-.477***	.076	.098	.116	.316***	.089
ttime_port_any	-.121***	-.069***	-.21***	-.017	-.585***	-.104*	-.003	-.002	.366***	-.124***
residential	.01	.049***	.153**	-.102	.413**	-.195	-.094	-.058	-.176	.021
accommodation	.076***	.037***	.088***	-.01	.261***	.031	.07*	.06	-.03	.012
mining_industrial	.073***	.045***	.096***	-.022	.219***	-.054	-.077	.009	-.151***	.17***
tourism_recreation	.044***	.027***	.086***	-.012	.241***	.009	0	.022	-.031	.014
construction	.056***	.027***	.114***	-.012	.286***	-.004	-.056	.085	-.359***	.107
<i>Debiased Data via Ensemble ML Models</i>										
roads_paved	.018	.015	-.025	-.06	.027	-.059	-.057	.161	.002	-.159
power	-.005	-.002	.041	.043	.076	.07	.246***	.202	.073	.087
education	.001	-.003	-.019	-.018	-.068	-.057	-.026	-.029	-.009	-.024
health	-.006*	-.006*	-.012	-.015	-.02	-.033	-.019	-.068**	.001	-.019
communications	-.01***	-.009**	-.004	.013	-.016	.03	.035	.012	-.033	-.053*
public_service_utility	-.003	-.001	.011	.012	.035*	.04*	.007	-.001	-.004	.002
automotive	-.003	-.001	.001	.004	-.016	-.008	.026**	.028*	-.001	.003
transport_other	.003	.003	-.004	0	-.023	-.015	.052***	.021	.029**	.006
financial	-.002	-.001	-.009	-.014**	-.02	-.03*	.02**	.02	.003	-.009
services	0	-.002*	-.001	0	-.002	.001	-.002	.004	0	.007
ttime_city_1m	-.006**	-.004	-.002	-.003	-.033	-.034	.007	.002	.009	.053*
ttime_port_any	-.015***	-.006*	-.033**	-.017	-.127***	-.082**	-.002	.048	.032	.001
residential	-.003	-.001	-.021	-.058	-.068	-.134	-.034	-.091	-.016	-.062
accommodation	.005**	.004	.002	.004	.014	.017	.004	.022	-.004	-.004
mining_industrial	-.003	-.002	.003	.004	-.023	-.023	-.045*	-.04	.01	.031
tourism_recreation	-.004*	-.001	0	-.002	.005	-.005	-.018	-.009	-.011	-.022
construction	-.006	-.003	-.005	.006	-.023	.001	.044	.117*	-.007	.126**
Observations	5346	5346	5346	5346	5346	5346	385	385	385	385
Country FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Geographic Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Simulation Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1; based on cluster-robust standard-errors

Notes: See Graff (2024) Tables 1 and 2 for further details about the variables and controls.

Table A11: Median CAPE: DHS IWI vs. IWI Prediction by [Lee & Braithwaite \(2022\)](#)

Feature	Simple Counts		Quantile Counts	
	DHS	L&B	DHS	L&B
roads_paved	0.248	0.231	0.179	0.225
power	0.197	0.068	0.221	0.071
education	0.009	0.509	-0.032	0.347
health	0.932	0.724	1.033	0.661
communications	0.798	0.655	0.571	0.624
public_service_utility	0.770	0.214	0.149	0.381
automotive	0.615	0.580	0.437	0.468
transport_other	0.593	0.300	0.686	0.296
financial	3.054	0.992	2.302	0.891
services	1.766	0.412	0.651	0.262
ttime_city_1m	-1.285	-0.566	-1.323	-0.521
ttime_port_any	-0.965	-0.215	-1.003	-0.229
residential	-0.068	-0.047	0.007	-0.025
accommodation	1.507	0.802	1.356	0.654
mining_industrial	0.841	0.509	0.497	0.544
tourism_recreation	-0.061	0.149	-0.185	0.198
construction	0.246	0.153	0.093	0.180
Median Abs. Coef.	0.770	0.412	0.497	0.347
Corr. Spearman		0.865		0.809
Corr. Pearson		0.857		0.902

Abbreviations: DHS = IWI estimate from Demographic and Health Survey; L&B = Predicted IWI estimate by [Lee & Braithwaite \(2022\)](#).

Notes: the DHS-based IWI estimate is computed from all DHS surveys conducted in SSA since 2010. It is averaged across 96km^2 hexagonal grid cells just like the IWI estimate of [Lee & Braithwaite \(2022\)](#). Grid cells with less than 5 households or less than 10 people per km^2 population density are excluded. This yields 17,396 cells with DHS-based IWI estimates used for training, versus 89,048 cells available with the predicted IWI by [Lee & Braithwaite \(2022\)](#). After the ensemble CAPE model is trained, a CAPE prediction is made for 103,922 cells which have a population density above 10 persons/ km^2 and any POI feature. The median of these CAPE estimates is reported in this table.

Table A12: Average Partial Effect: DHS IWI vs. IWI Prediction by [Lee & Braithwaite \(2022\)](#)

Feature	Simple Counts		Quantile Counts	
	DHS	L&B	DHS	L&B
roads_paved	0.202	0.218	0.173	0.216
power	0.087	0.068	0.114	0.068
education	0.031	0.563	-0.073	0.436
health	0.585	0.803	0.645	0.772
communications	1.037	0.689	1.031	0.689
public_service_utility	0.842	0.327	0.470	0.331
automotive	1.482	0.873	0.952	0.444
transport_other	0.382	0.332	0.324	0.305
financial	1.857	0.626	1.641	0.543
services	1.911	1.380	4.506	0.748
ttime_city_1m	-1.158	-0.513	-1.174	-0.502
ttime_port_any	-1.014	-0.293	-0.937	-0.295
residential	-0.097	-0.068	-0.101	-0.052
accommodation	1.518	0.807	1.073	0.771
mining_industrial	0.740	0.541	0.508	0.460
tourism_recreation	0.137	0.184	0.094	0.309
construction	0.211	0.161	0.154	0.186
Median Abs. Coef.	0.740	0.513	0.508	0.436
Corr. Spearman		0.863		0.870
Corr. Pearson		0.895		0.874

Abbreviations: DHS = IWI estimate from Demographic and Health Survey; L&B = Predicted IWI estimate by [Lee & Braithwaite \(2022\)](#).

Notes: the DHS-based IWI estimate is computed from all DHS surveys conducted in SSA since 2010. It is averaged across 96km^2 hexagonal grid cells just like the IWI estimate of [Lee & Braithwaite \(2022\)](#). Grid cells with less than 5 households or less than 10 people per km^2 population density are excluded. This yields 17,396 cells with DHS-based IWI estimates used for training, versus 89,048 cells available with the predicted IWI by [Lee & Braithwaite \(2022\)](#). After the ensemble CAPE model is trained, a CAPE prediction is made for 103,922 cells which have a population density above 10 persons/ km^2 and any POI feature. The APE is obtained through augmented inverse probability weighting following Eq. 10, and thus only considers cells where outcome data is available. Differences in the estimates are thus expected because with L&B much more cells are available for APE calculation.

Table A13: CAPE Correlations: DHS IWI and IWI Prediction by [Lee & Braithwaite \(2022\)](#)

Feature	Simple Counts	Quantile Counts
roads_paved	0.593	0.119
power	0.157	0.040
education	0.317	0.024
health	0.013	-0.038
communications	0.271	-0.171
public_service_utility	0.100	0.386
automotive	0.108	0.114
transport_other	0.126	0.001
financial	0.287	0.362
services	0.000	-0.006
ttime_city_1m	0.038	-0.022
ttime_port_any	-0.118	-0.025
residential	-0.000	0.383
accommodation	0.246	0.013
mining_industrial	0.422	0.271
tourism_recreation	0.076	0.000
construction	0.488	-0.213
Median Corr.	0.126	0.0128
Median Abs. Corr.	0.126	0.040

Notes: the DHS-based IWI estimate is computed from all DHS surveys conducted in SSA since 2010. It is averaged across 96km^2 hexagonal grid cells just like the IWI estimate of [Lee & Braithwaite \(2022\)](#). Grid cells with less than 5 households or less than 10 people per km^2 population density are excluded. This yields 17,396 cells with DHS-based IWI estimates used for training, versus 89,048 cells available with the predicted IWI by [Lee & Braithwaite \(2022\)](#). After the ensemble CAPE model is trained, a CAPE prediction is made for 103,922 cells which have a population density above 10 persons/ km^2 and any POI feature. Pearson's correlation of these CAPE estimates is reported in this table.

Table A14: CAPE Correlations: DHS IWI and IWI Prediction by [Lee & Braithwaite \(2022\)](#): Previous Estimates without Separate Spillover Variables (One Variable Per Infrastructure)

	Count	Count + SS	Tag Weights	Weights + SS
roads_paved	0.213	0.305	0.245	0.246
power	-0.337	0.249	-0.001	0.078
education	0.550	0.379	0.632	-0.338
health	0.319	0.280	0.251	0.708
communications	0.559	0.271	0.399	0.305
public_service_utility	0.292	0.542	0.802	0.267
automotive	0.301	-0.070	0.575	0.161
transport_other	-0.083	0.329	-0.031	0.181
financial	0.253	0.128	0.402	0.303
services	0.231	0.003	0.025	-0.313
ttime_city_1m	0.325	0.353	0.288	0.419
ttime_port_any	0.500	0.257	0.508	0.375
residential	0.556	0.135	0.454	-0.009
accommodation	0.305	0.481	0.325	0.260
industrial	0.132	0.196	0.727	0.510
tourism_recreation	0.158	0.069	0.374	0.373
construction	0.296	-0.153	0.037	-0.000
Median Corr.	0.294	0.244	0.284	0.195
Median Abs. Corr.	0.298	0.244	0.299	0.232

Abbreviations: SS = Spatial Spillovers.

Figure A17: DML CAPE Kernel Density Estimates for DHS-Based IWI

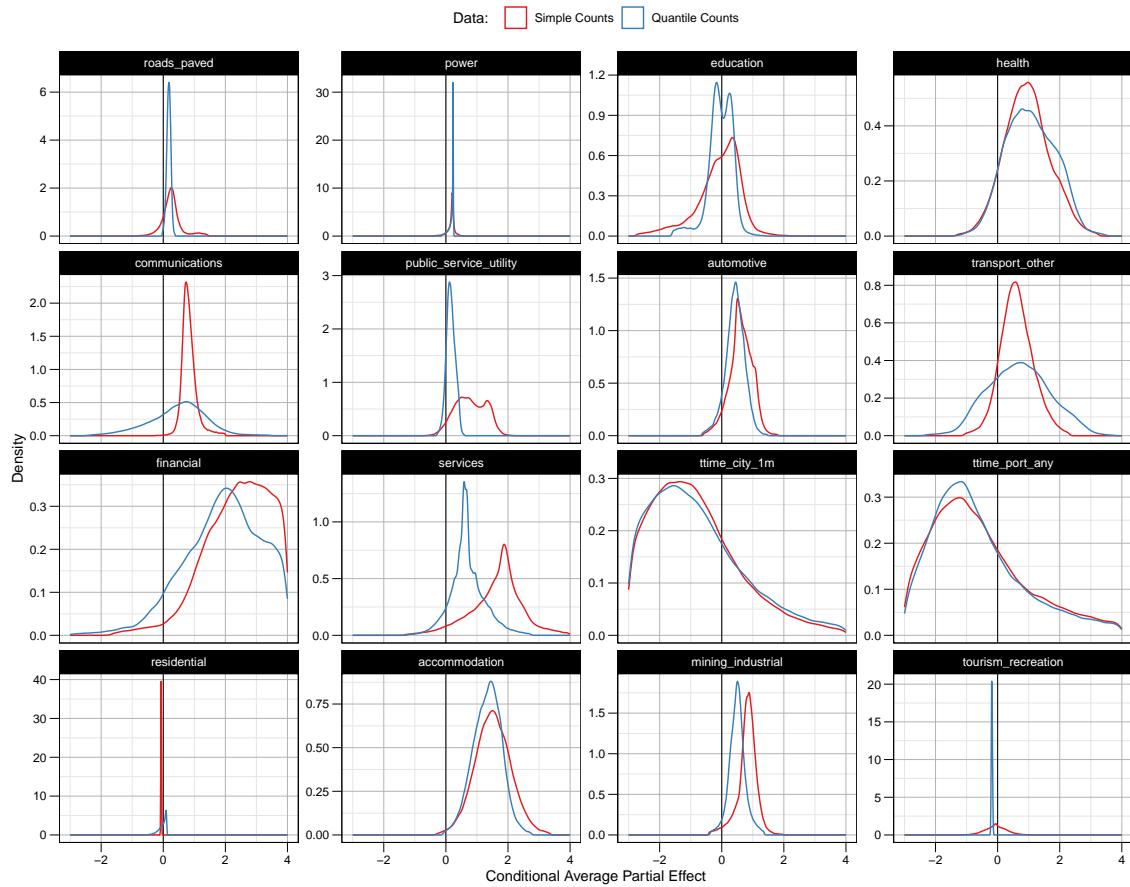


Figure A18: Top 25 Correlates of DHS-Based IWI CAPE Estimates: Average Across Datasets

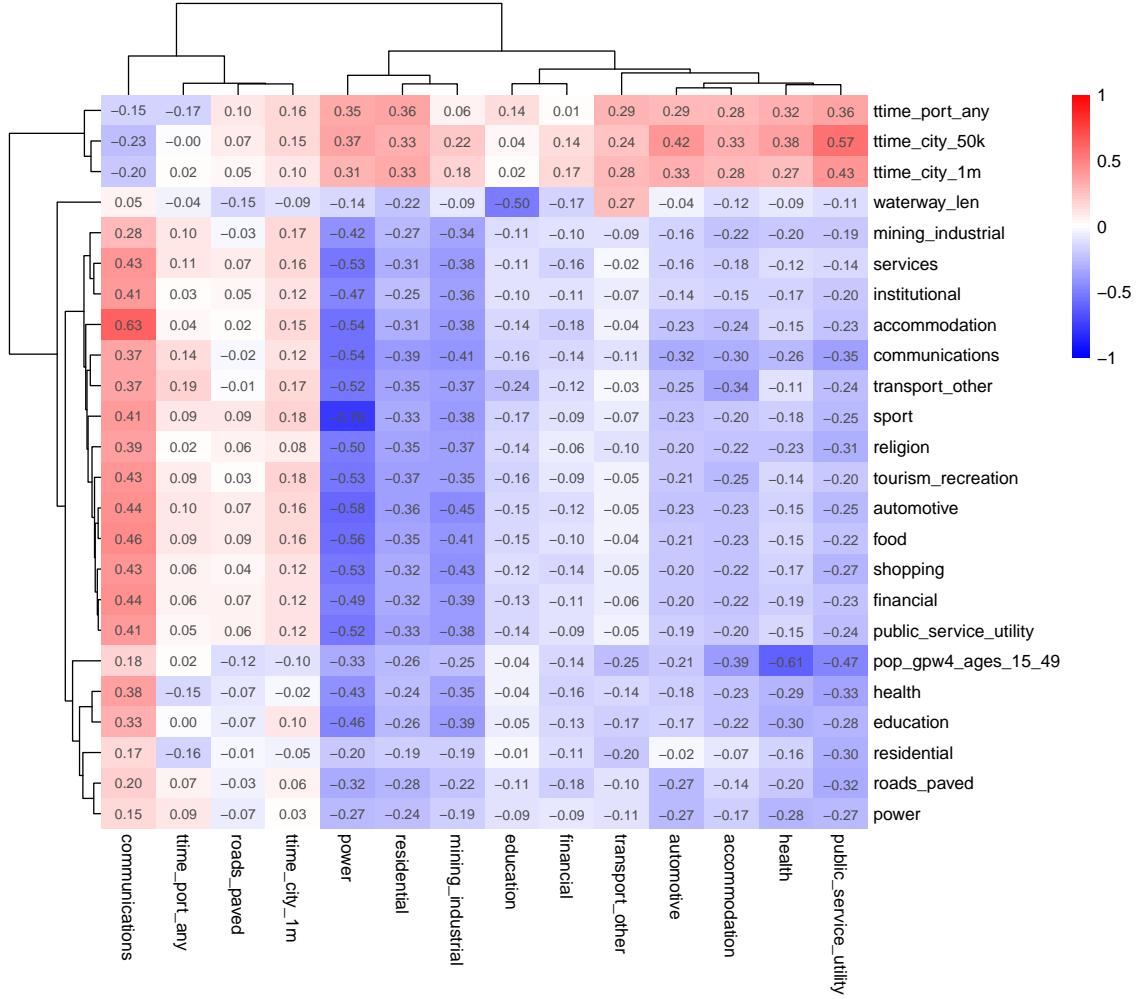


Table A15: Average Partial Effects on Log Nightlights 2022

Feature	ATE	Simple Counts			ATE	Quantile Counts			Diff.
		High	Low	Diff.		High	Low	Diff.	
roads_paved	0.0198***	0.0328***	0.00682**	0.026***	0.0193***	0.0351***	0.00358	0.0315***	
power	0.0235***	0.0436***	0.00338	0.0402***	0.0247***	0.0445***	0.00497	0.0395***	
education	0.0825***	0.144***	0.0209	0.123***	0.0532***	0.104***	0.00198	0.102***	
health	0.107***	0.252***	-0.0382**	0.29***	0.101***	0.243***	-0.0405**	0.284***	
communications	0.17***	0.212***	0.128***	0.0847***	0.168***	0.226***	0.111***	0.115***	
public_service_utility	0.112***	0.214***	0.00991	0.204***	0.106***	0.21***	0.00209	0.208***	
automotive	0.247***	0.313***	0.181***	0.132**	0.201***	0.272***	0.131***	0.141***	
transport_other	0.107***	0.144***	0.0696***	0.0744**	0.111***	0.157***	0.0646***	0.0928***	
financial	0.246***	0.414***	0.0787***	0.336***	0.231***	0.451***	0.0109	0.441***	
services	0.184**	0.277**	0.0908	0.186	0.133**	0.214**	0.0527	0.161	
ttime_city_1m	-0.0285	0.0928***	-0.15***	0.243***	-0.0119	0.129***	-0.152***	0.281***	
ttime_port_any	-0.0795***	0.0781***	-0.237***	0.315***	-0.0764***	0.0507**	-0.204***	0.254***	
residential	0.0105*	0.0346***	-0.0137	0.0483***	0.016***	0.0444***	-0.0124	0.0568***	
accommodation	0.1***	0.197***	0.0035	0.194***	0.0877***	0.206***	-0.0302	0.236***	
mining_industrial	0.235***	0.372***	0.097***	0.275***	0.185***	0.282***	0.0878***	0.195***	
tourism_recreation	0.0139	0.0769**	-0.0491*	0.126***	-0.000934	0.0523	-0.0542*	0.107**	
construction	0.0505***	0.0921***	0.00898	0.0831***	0.0515***	0.093***	0.0101	0.0829***	

Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Notes: Table shows doubly-robust APE estimates of the log feature intensity (simple counts or quantile counts in each cell, see Section 2.5) on the log of NASA nightlights 2022 (Román et al., 2018). The "High" and "Low" estimates report the APE above and below the median CAPE estimate. The "Diff." column indicates their difference to test for heterogeneity. All terms are tested using a two-sided t-test with standard errors derived from the doubly robust scores following Athey & Wager (2019).

Figure A19: DML CAPE Kernel Density Estimates for Log Nightlights 2022

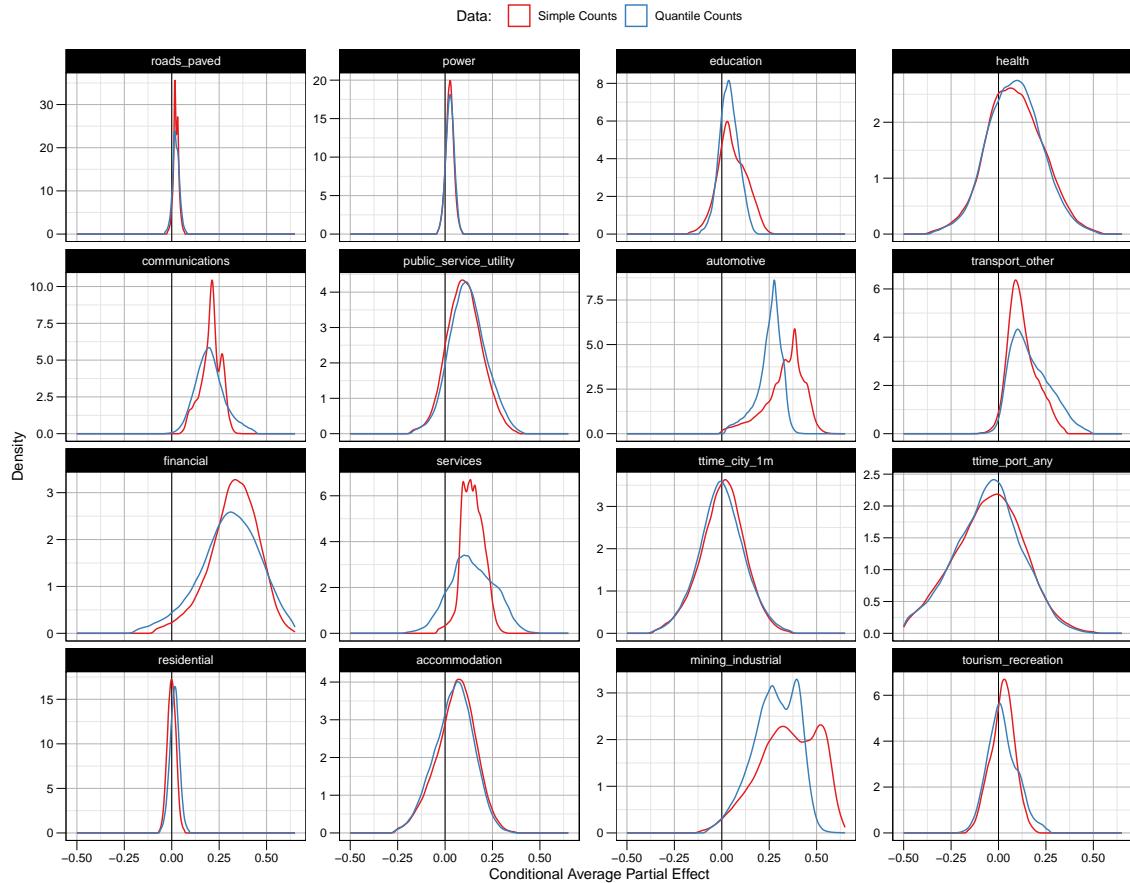


Figure A20: Top Correlates of Nightlights CAPE Estimates: Average Across Datasets

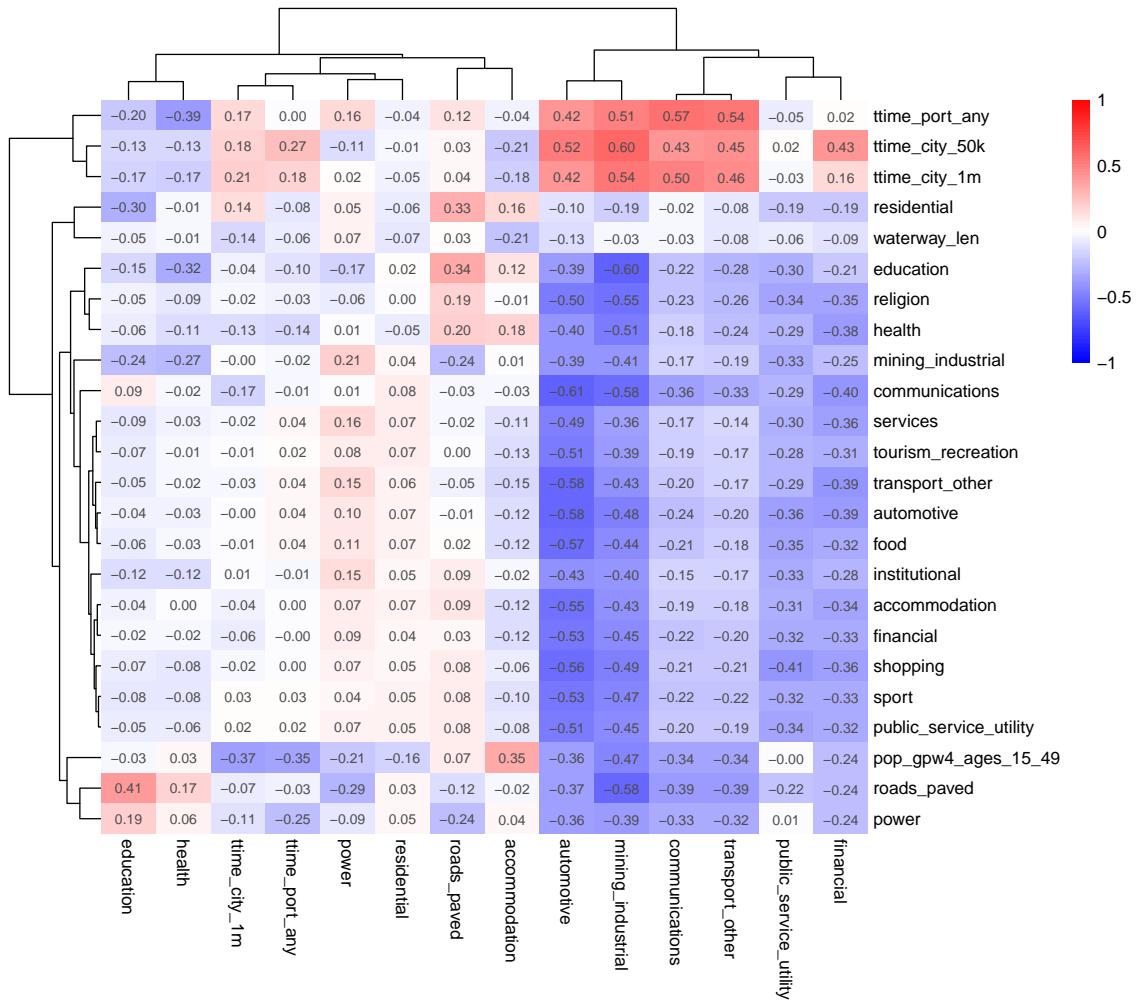


Figure A21: Counterfactual Predictions for Log of Nightlights

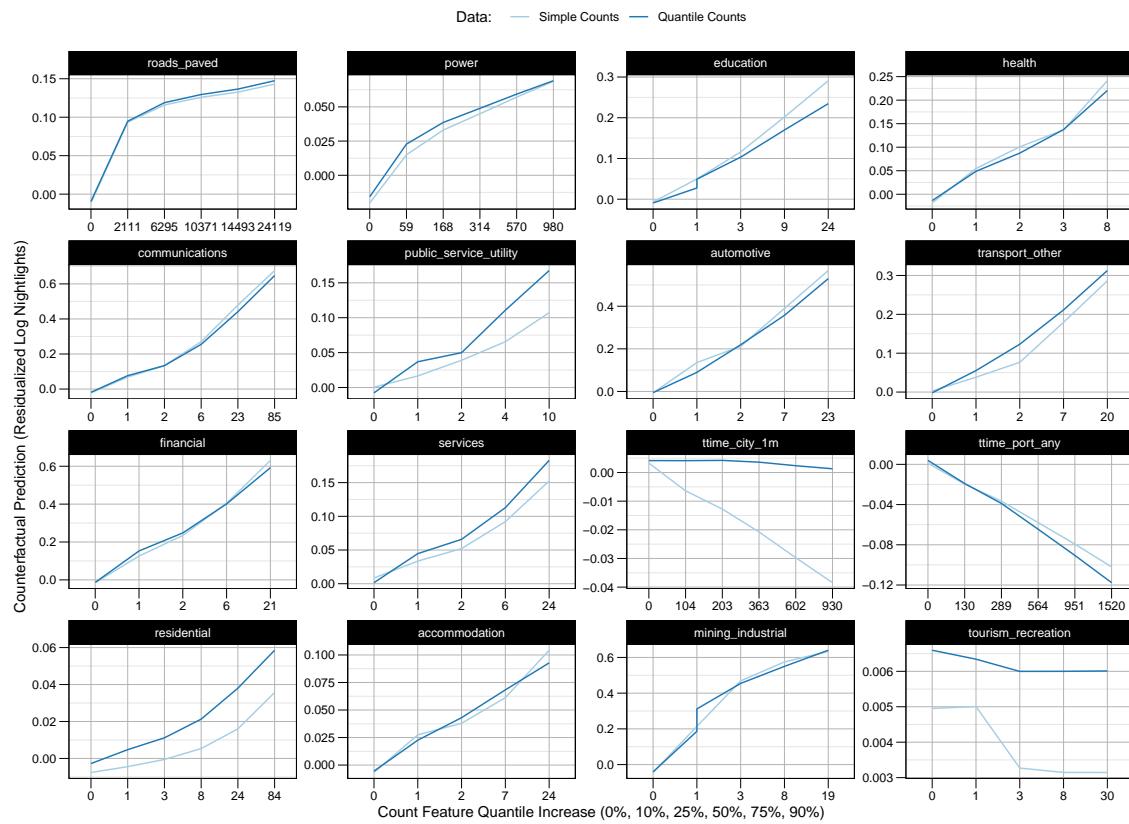


Figure A22: Counterfactual Predictions for IWI: Average Total Wealth Effects per Cell

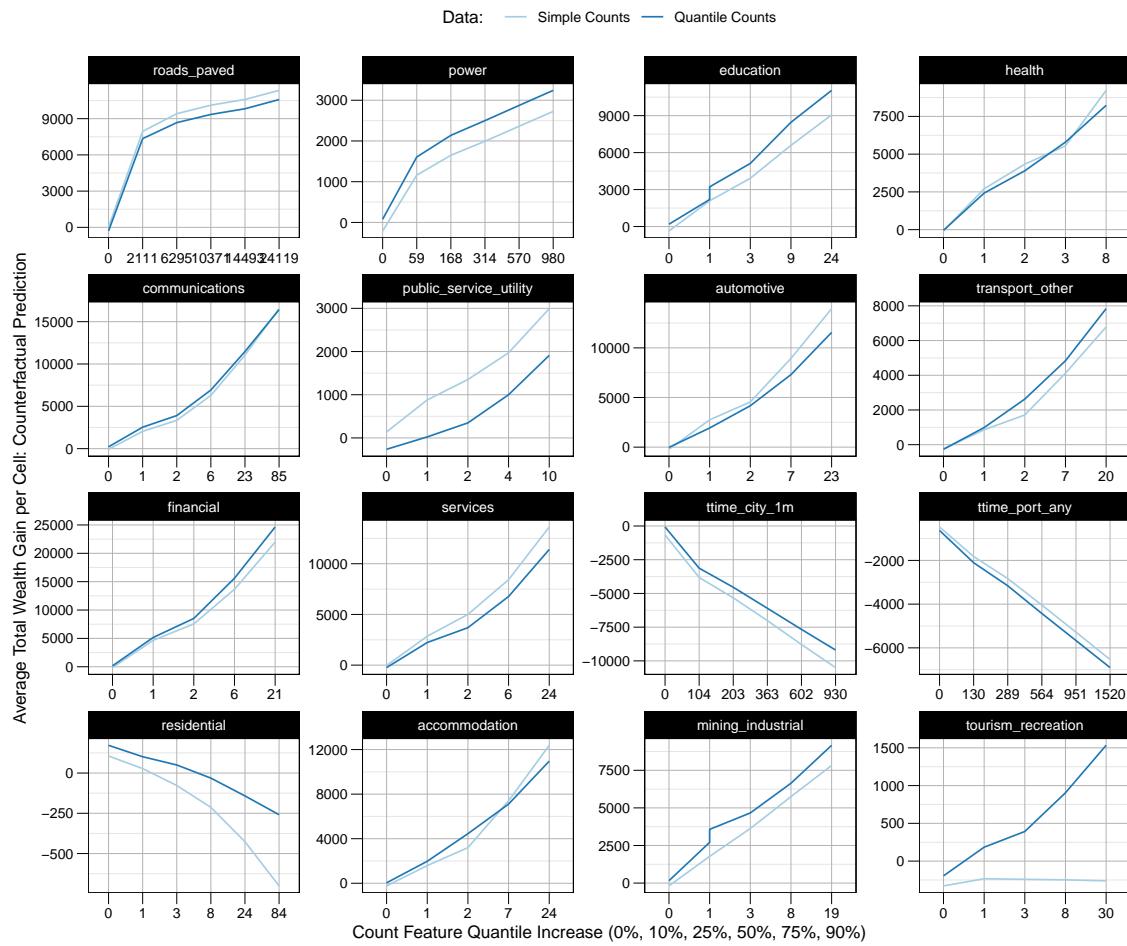


Figure A23: Counterfactual Predictions for DHS-Based IWI

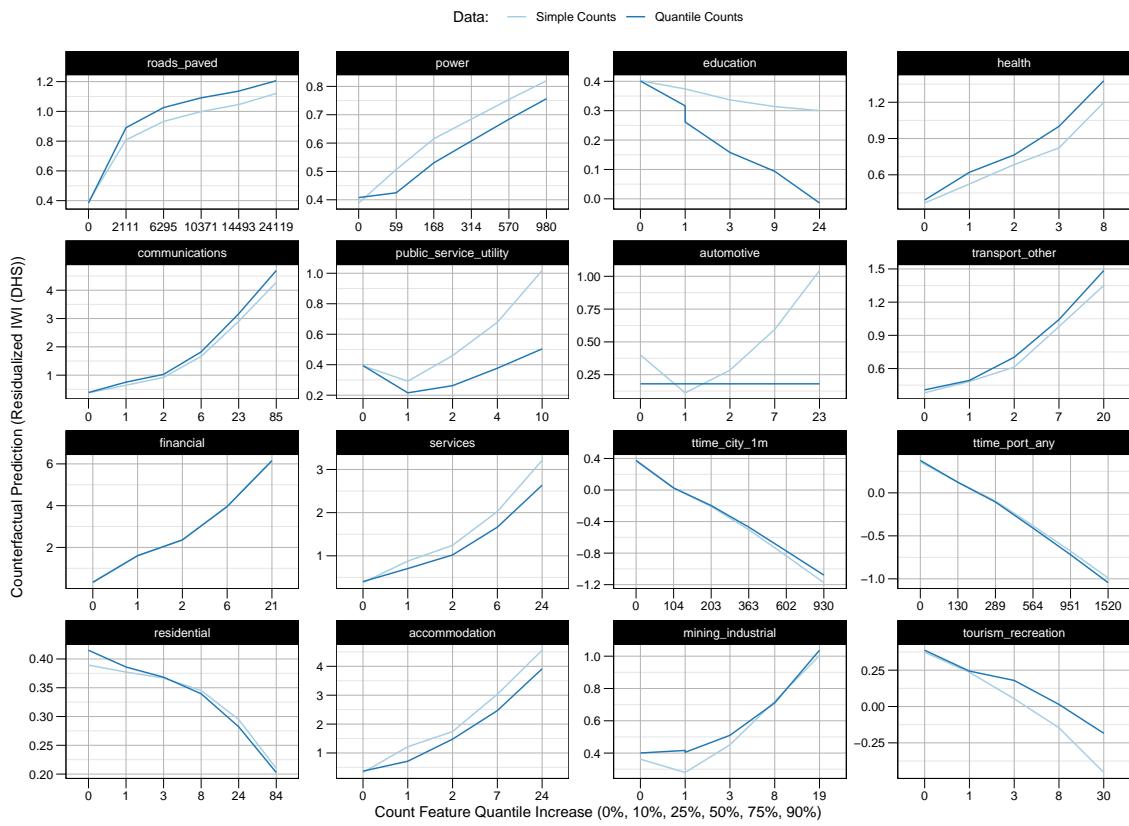


Table A16: Median 50% Counterfactual Prediction: DHS IWI vs. IWI by [Lee & Braithwaite \(2022\)](#)

Feature	Simple Counts		Quantile Counts	
	DHS	L&B	DHS	L&B
roads_paved	1.087	1.622	1.165	1.470
power	0.762	0.302	0.694	0.354
education	0.405	0.662	0.242	0.802
health	0.739	0.776	0.821	0.713
communications	1.710	0.996	1.871	1.112
public_service_utility	0.534	0.131	0.330	-0.007
automotive	0.363	0.583	0.176	0.512
transport_other	0.674	0.171	0.773	0.291
financial	2.368	1.000	2.377	1.112
services	1.285	0.609	1.068	0.435
ttime_city_1m	-0.505	-0.481	-0.469	-0.442
ttime_port_any	-0.365	-0.351	-0.386	-0.371
residential	0.398	-0.068	0.406	-0.044
accommodation	1.762	0.371	1.517	0.533
mining_industrial	0.506	0.477	0.575	0.595
tourism_recreation	0.108	-0.058	0.232	0.012
Median Abs. Coef.	0.604	0.479	0.635	0.477
Corr. Spearman		0.715		0.724
Corr. Pearson		0.700		0.715

Abbreviations: DHS = IWI estimate from Demographic and Health Survey; L&B = Predicted IWI estimate by [Lee & Braithwaite \(2022\)](#).

Notes: the DHS-based IWI estimate is computed from all DHS surveys conducted in SSA since 2010. It is averaged across 96km^2 hexagonal grid cells just like the IWI estimate of [Lee & Braithwaite \(2022\)](#). Grid cells with less than 5 households or less than 10 people per km^2 population density are excluded. This yields 17,396 cells with DHS-based IWI estimates used for training, versus 89,048 cells available with the predicted IWI by [Lee & Braithwaite \(2022\)](#). Counterfactual predictions are made for 103,922 cells which have a population density above 10 persons/ km^2 and any POI feature. The median of the 50% increase counterfactual prediction is reported in this table.

Table A17: Correlations of Counterfactual Predictions: DHS IWI and IWI by [Lee & Braithwaite \(2022\)](#)

Feature	Simple Counts	Quantile Counts
roads_paved	-0.045	-0.115
power	-0.019	0.136
education	0.069	0.349
health	0.272	0.219
communications	0.281	0.410
public_service_utility	0.299	0.029
automotive	-0.016	0.323
transport_other	0.109	0.006
financial	0.385	0.505
services	0.177	0.058
ttime_city_1m	0.345	0.540
ttime_port_any	0.127	0.118
residential	0.304	0.271
accommodation	0.298	0.364
mining_industrial	0.145	0.208
tourism_recreation	0.062	0.091
Median Abs. Corr.	0.161	0.214
Median Abs. Corr.	0.161	0.214

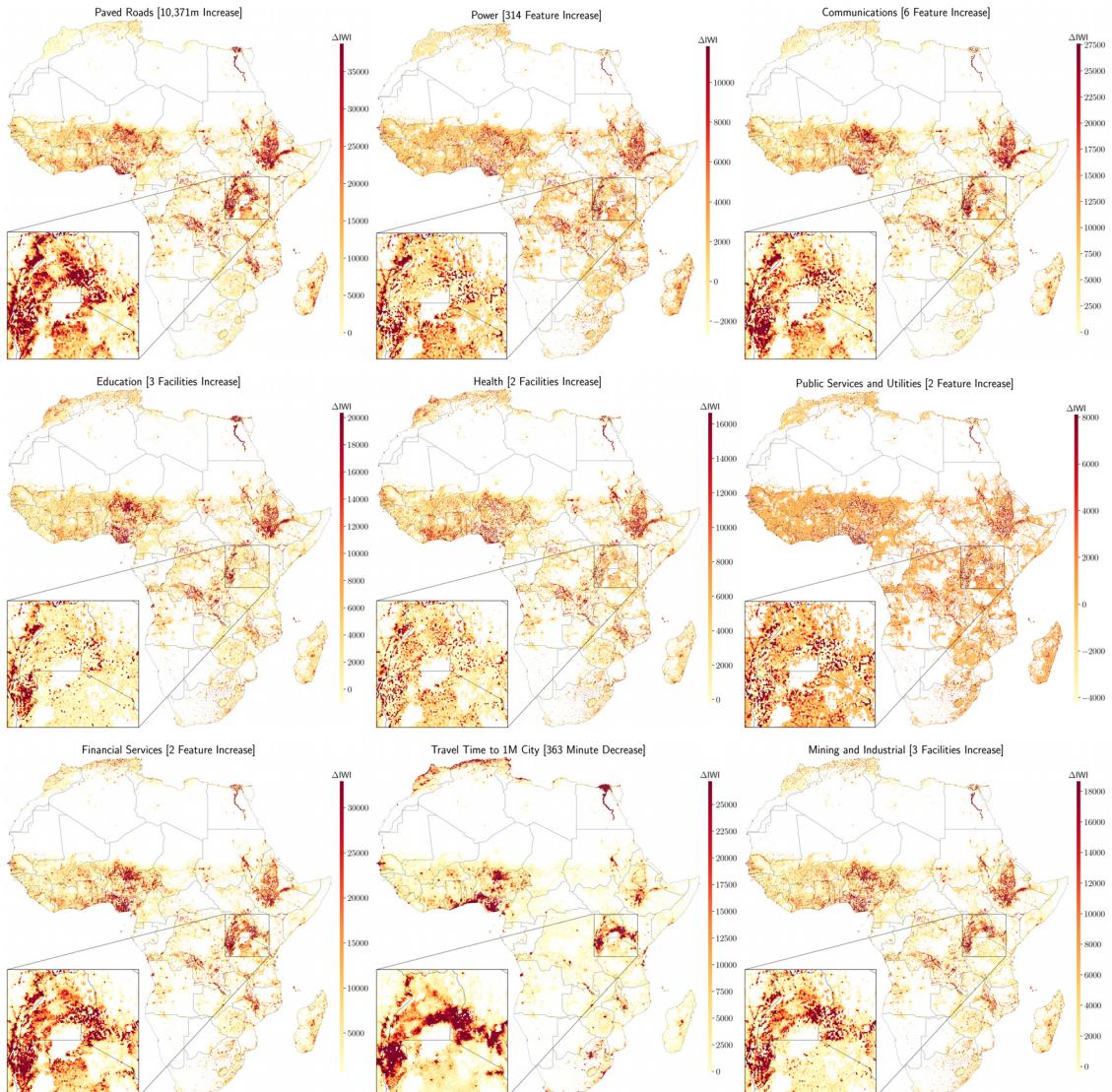
Notes: the DHS-based IWI estimate is computed from all DHS surveys conducted in SSA since 2010. It is averaged across 96km^2 hexagonal grid cells just like the IWI estimate of [Lee & Braithwaite \(2022\)](#). Grid cells with less than 5 households or less than 10 people per km^2 population density are excluded. This yields 17,396 cells with DHS-based IWI estimates used for training, versus 89,048 cells available with the predicted IWI by [Lee & Braithwaite \(2022\)](#). Counterfactual predictions are made for 103,922 cells which have a population density above 10 persons/ km^2 and any POI feature. The average Pearson's correlation of these counterfactual predictions across the two IWI estimates is reported in this table.

Table A18: Correlations of Counterfactual Predictions: DHS IWI and IWI by [Lee & Braithwaite \(2022\)](#): Previous Estimates without Separate Spillover Variables (One Variable Per Infrastructure)

	Count	Count + SS	Tag Weights	Weights + SS
roads_paved	0.396	0.743	0.471	0.610
power	0.313	0.141	0.166	0.195
education	0.054	0.297	0.238	0.378
health	0.147	0.091	0.152	0.336
communications	0.184	0.282	0.264	0.345
public_service_utility	0.237	0.166	0.193	0.321
automotive	0.129	0.502	0.215	0.497
transport_other	0.267	0.165	0.218	0.240
financial	0.269	0.057	0.376	0.244
services	0.206	0.302	0.325	0.368
ttime_city_1m	0.216	0.131	0.257	0.125
ttime_port_any	0.226	0.036	0.240	0.108
residential	-0.060	0.057	-0.032	0.081
accommodation	0.233	0.574	0.363	0.651
industrial	0.147	0.218	0.146	0.170
tourism_recreation	0.239	0.161	0.308	0.302
Median Corr.	0.216	0.163	0.248	0.298
Median Abs. Corr.	0.216	0.163	0.248	0.298

Abbreviations: SS = Spatial Spillovers.

Figure A24: Spatial 50% CFPRs (Geometric Mean) \times WorldPop 2020 Population



Notes: Figure shows 50% counterfactual predictions of the International Wealth Index (IWI) [0, 100] by Lee & Braithwaite (2022), i.e., the predicted wealth increase (Eq. 12) from an increase in each cell amounting to the median of the non-negative feature density, summarized in Table 6, multiplied with the WorldPop 2020 population measure. It is thus an estimate of the partial effect of the investment on total cell welfare.