Achten, Sandra; Leßmann, Christian

Working Paper

Spatial inequality, geography and economic activity

CESifo Working Paper, No. 7547

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Achten, Sandra; Leßmann, Christian (2019) : Spatial inequality, geography and economic activity, CESifo Working Paper, No. 7547, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/198907

CESifo WORKING PAPERS

7547
2019

March 2019

Spatial inequality, geography and economic activity

Sandra Achten, Christian Lessmann

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7547*
Category 6: Fiscal Policy, Macroeconomics and Growth

Spatial inequality, geography and economic activity

Abstract

We study the effect of spatial inequality on economic activity. Given that the relationship is highly simultaneous in nature, we use exogenous variation in geographic features to construct an instrument for spatial inequality, which is independent from any man-made factors. Inequality measures and instruments are calculated based on grid-level data for existing countries as well as for artificial countries. In the construction of the instrumental variable, we use both a parametric regression analysis as well as a random forest classification algorithm. Our IV regressions show a significant negative relationship between spatial inequality and economic activity. This result holds if we control for country-level averages of different geographic variables. Therefore, we conclude that geographic heterogeneity is an important determinant of economic activity.

JEL-Codes: R120, O150.

Keywords: regional inequality, spatial inequality, economic activity, development, geography, machine learning.

Sandra Achten
Technische Universität Braunschweig
Carl-Friedrich-Gauss-Faculty
Institute of Economics
Germany – 38106 Braunschweig
s.achten@tu-braunschweig.de

*Christian Lessmann**
Technische Universität Braunschweig
Carl-Friedrich-Gauss-Faculty
Institute of Economics
Germany – 38106 Braunschweig
c.lessmann@tu-braunschweig.de

*corresponding author

This version: March 2019
This research was supported by the German Research Foundation (grant BL1502/1-1). We would like to thank the participants of the 8th ifo Dresden Workshop ‘Regional Economics’ for helpful comments.

1 Introduction

This paper studies the relationship between spatial inequality and economic activity. Inequality is one of the main topics of research in economics and other social sciences, and researchers study different types of inequality. Vertical inequality describes inequality between persons or households within a particular social group, e.g. within a country. In contrast, horizontal inequality is concerned with differences among individuals of different groups that may be based on race, gender, ethnicity or place of habitation (Stewart (2000), Stewart (2005)). Our analysis is concerned with the latter issue, i.e. the regional or spatial inequality between groups of individuals living in different regions. This particular type of inequality has been attracting increasing interest from academics, given that inequality between groups may cause serious problems in a society – from inefficient redistribution politics to separatist movements and internal conflict (Østby (2008), Lessmann (2016)). These threats are, of course, more pronounced in ethnically heterogeneous countries due to the coincidence of ethnic inequality and spatial inequality (Kanbur and Venables (2005), Alesina et al. (2016)).

From a theoretical viewpoint, there are various channels through which different types of inequality may effect development and economic activity. In the literature on personal inequality, many authors stress the socio-political or the political instability channel (see Alesina and Rodrik (1994), Alesina and Perotti (1996), Perotti (1996), Benabou (2000) and Galor and Moav (2004) and Galor (2009) for an overview). Some of the arguments, particularly the redistribution channel or increased political instability, can easily be transferred into the context of spatial inequality. Redistribution policies are often accompanied by the higher taxes that are needed to finance equalization programs (Alesina and Rodrik (1994)). This may cause a decrease in the efficiency of factor allocation, which might ultimately harm overall economic development. Political stability often has a regional dimension not only in ethnically divided countries (one might think of the separatist movements in Catalonia) but also in relatively homogeneous countries like Great Britain, where lagging regions tend to vote for nationalist policies. This division harms economic integration and long-run development. In Great Britain, these movements fuelled the decision for Brexit, which might be very costly, with up to 2.6% of income losses as measured by the national GDP (Dhingra et al. (2016)). Andrés Rodríguez-Pose calls these effects the "revenge of the places that don't matter" (Rodríguez-Pose (2018) p. 189). Therefore, regional inequality is of interest not only for reasons of equity but also for the development of an economy as a whole.

Empirical research on this topic can be divided into two branches. The first branch examines the shape of the relationship between inequality and growth. Its origin is the hypothesis of Kuznets (1955), stating that the relationship has the shape of an inverted U. Williamson (1965) transfers Kuznets' considerations to the case of regional inequality. The main argument is the following. At early stages of economic development, inequality increases with growth, since growth usually takes place in a particular region of a country, e.g. due to differences in factor endowments. In the progress of economic development, people migrate, wages adjust and redistribution takes place; therefore, inequality tends to decrease again. It is obvious that in this theory, there is a mutual

dependence between growth and inequality. Inequality affects economic outcomes and vice versa. Empirical studies find some evidence of an inverted U (Barrios and Strobl (2009), Lessmann (2014)) or even an N-shaped relationship with increasing inequality at very high levels of development (Amos (1988) and Lessmann and Seidel (2017)). However, the Kuznets hypothesis is ultimately a statement on a correlation between inequality and development. The existing empirical studies do not take the simultaneity of both variables into account. Therefore, they cannot make any statement on causality. Addressing this gap is the main point of the current analysis.

The causality issue brings us to the second strand of the literature, which focuses on the question of whether there is a causal relationship between inequality and growth. Concerning vertical inequality, the results are quite ambiguous. Some empirical studies support an insignificant or even a positive impact of inequality on growth (Li and Zou (1998), Barro (2000), Forbes (2000)), whilst others find a negative relationship (Alesina and Rodrik (1994), Persson and Tabellini (1994)). Neves et al. (2016) conduct a meta-analysis on this relationship and find a negative effect of inequality on growth that is stronger in less developed countries and that weakens if regional dummies are included. Importantly, concerning our topic of spatial inequality, no serious attempt has been made to study a causal relationship. Alesina et al. (2016) investigate correlations between regional as well as ethnic inequalities and national development. Their regressions show that both types of inequality are significantly negatively related to national income. However, they do not go beyond presenting partial correlations in a cross-section of countries. Given that unobserved heterogeneity, particularly with respect to spatial differences in initial geographic fundamentals, might affect both economic activity and spatial inequality simultaneously, we aim to go one step further. The main contribution of our study is to analyse the causal impact of spatial inequality on economic activity.

The major challenge for our study is to find a suitable instrumental variable for spatial inequality. Our approach is to some extent inspired by Easterly (2007) who argues that while market-based inequality has ambiguous effects on economic development, inequality of agricultural endowments is structural and leads unambiguously to lower growth. Based on this line of reasoning, he uses a geographical variable – the wheat-sugar suitability ratio – as an instrument for personal inequality. Following his arguments, cultivating wheat is associated with family farms, the middle class and an equal society, whereas the cultivation of sugarcane is associated with slave labour and inequality. Easterly's instrumental variable regressions point to a significant negative effect of personal inequality on economic development.

It is obvious that there are many geographical features that make some places better than others for human habitation and output production. This is mainly due to so-called first nature determinants of development, which describe the physical setting of a location independently of man-made – second nature – factors (Krugman (1993)). First nature aspects include, for example, geography, climate, and resource endowments (Williamson (1965), Diamond (1997)). These are purely exogenous natural factors of development. Second nature determinants, such as market size effects, factor mobility and infrastructure, ultimately depend on man-made endogenous factors.

According to the New Economic Geography (NEG) approach, a small advantage in these first nature conditions may lead to agglomeration of economic activity in these areas (Krugman (1991, 1997) and Fujita et al. (1999)). Thus, differences in economic activity within countries initially depend on natural factors, whose role may be amplified or mitigated by man-made factors. In our empirical analysis, we are going to use these regional differences in geographic pre-conditions in order to create an instrument for spatial inequality, which is purely based on exogenous first nature factors.

We analyse the impact of spatial inequality on economic activity using an instrumental variable approach, where the instrument for spatial income dispersion is constructed from the exogenous 'first nature' determinants of regional development. The general idea for the instrument is inspired by Easterly (2007) and Henderson et al. (2017). We use a large set of geographic variables that are geocoded and provide information on the first nature determinants. The data are gridded; therefore, we can analyse different cell sizes as well as synthetic countries, ignoring any administrative boundaries. We combine these geodata with satellite night light data as a proxy for economic activity. Elvidge et al. (1997), Chen and Nordhaus (2011) and Henderson et al. (2012) show that night lights are a fairly good proxy for economic growth at the national level, while Henderson et al. (2017) provide a detailed analysis of the relationship between geography and lights at the regional level. The latter study is the starting point of our IV approach.

Our empirical strategy follows a three-step procedure that we illustrate in figure 1 as an example for the Iberian peninsula. First, we investigate the relationship between grid-cell light emissions and first nature geography. We use a 1×1 degree grid instead of subnational administrative units to avoid potential political influences. For every grid cell, we calculate the average values for the different geographic determinants such as precipitation, temperature, soil constraints, and biomes. These geographic determinants serve as the basis of predicting night-time lights on the grid-cell level. The prediction gives us an indicator for the natural potential of economic activity. We use two different methods for the prediction. We begin our analysis using a standard regression model, which closely follows Henderson et al. (2017). This model can explain about 49% of the variation in grid-cell lights. However, regressions are a parametric estimation procedure that assumes strong linear relationships between dependent and independent variables. Still, it is very likely that first-nature variables form a complex non-linear system that makes human habitation and production more or less effective. This interdependence of factors cannot be captured in OLS regressions. Therefore, we apply a machine learning algorithm – random forest – that combines first-nature variables in a flexible way in order to make a prediction of grid-cell light that performs much better than a linear prediction.[1] These estimations explain about 82% of the variation in grid-cell lights, which is a significant improvement compared with the linear regressions. We use the regression coefficients and the machine learning classification results, respectively, in order to predict night lights that reflect only the component of grid-cell light caused by exogenous physical geography.

Second, based on these predictions on the grid-cell level, we calculate measures of spatial inequality

[1] Random forest is shown to produce improved predictions in a non-linear setting (Varian, 2014).

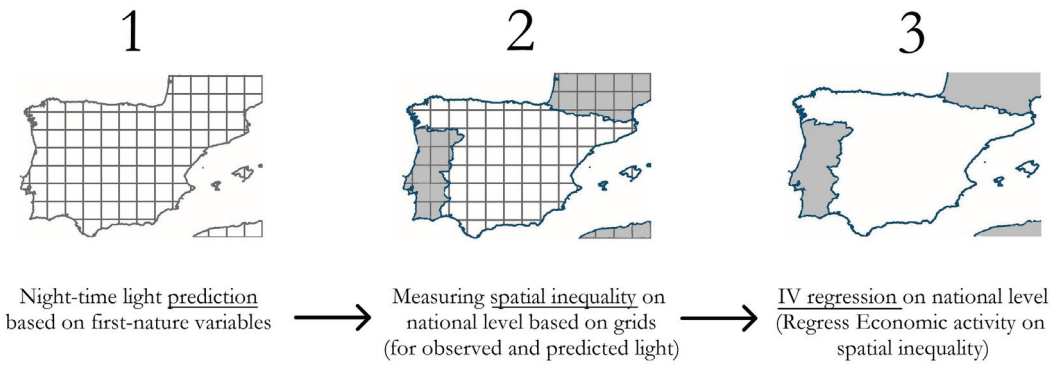


Figure 1: Process-diagram (Grid cell example: Spain and Portugal)

(i.e. the Gini coefficient) for every country (Lessmann (2014), Lessmann and Seidel (2017)). In our example, the inequality measure is based on 12 cells in the case of Portugal and 67 in Spain, respectively. We calculate inequality for observed as well as for predicted light according to both prediction methods (OLS and Random Forest). We account for cells that are smaller due to coast locations by area weighting. The first two steps of our analysis are presented in section 2.

Third, we run instrumental variable regressions from economic activity (measured by night-time lights) on measures of observed spatial inequality, again on the national level. Thereby, we instrument observed spatial inequality based on measured night lights, with inequality based on predicted lights. We control for mean values of the geographic variables in order to distinguish between average effects of geography on national development and the effect stemming from geographic heterogeneity. We have complete information about all geographical variables for 15,664 landscape grid cells (1×1 degree) located in 212 countries. We calculate spatial inequality only for countries that consist of more than 3 grid cells, which reduces the sample size to 184 countries. Note that geographic variables have only a very small variation over time; therefore, our analysis is restricted to a cross-sectional perspective.[2] In order to rule out extreme weather events or other local shocks, we concentrate on a 5-year period average for the years 2008 to 2012. The third 'main' step of our analysis is presented in section 3.

The results of our study are the following. We find a highly significant negative relationship between spatial inequality and economic activity, i.e. the higher the spatial inequalities are, the lower the national economic activity. This result is confirmed for synthetic countries and when we include additional standard growth determinants following Barro (2000). The size of the effect suggests that a 0.01 unit increase in spatial inequality measured by the Gini coefficient results in a 1.7% to 3.8% lower income, depending on the specification and prediction method. Our study implies that regional inequality is an obstacle for economic development. However, we cannot causally identify those factors moderating the relationship.

[2] The only time-variant variables are temperature and precipitation, which have insufficient power to explain changes in night lights over time.

The paper is organized as follows. In section 2, we analyse the relationship between grid-cell light and first-nature determinants using standard regressions and the machine learning algorithm. The results are used to create an instrument for spatial inequality, which solely depends on exogenous geographical factors. In section 3, we show the results of IV regressions, where we regress night light densities on spatial inequality at the country level. Inequality is instrumented with predicted inequality from the previous section. We also provide a number of robustness tests on artificial countries, subgroups of countries, etc. Section 4 concludes the paper and offers an outlook on future research.

2 Creating the Instrument for Regional Inequality

2.1 Prediction of Economic Activity

The first step of our analysis is to investigate the impact of geography on regional night-time lights as proxy for economic activity. We use two different methods: (1) standard regression analysis and (2) random forest, a machine learning algorithm. By means of the selection of variables, we closely follow Henderson et al. (2017).

2.1.1 Methodology

Importantly, the aim of this part of our study is very different to Henderson et al. (2017). In economics, we are usually interested in marginal effects, i.e. how does one independent variable affect the dependent variable holding all other variables constant. This is what a standard regression analysis does. It imposes the (strong) assumption of linearity between variables on the model and data, which might be justifiable in many circumstances. If we have a question in mind of the form *how does one additional year of education affect lifetime earnings?*, this seems to be a valid approach. Here, regressions are the method of choice, since we would expect from theory that there is a linear or quadratic relationship (decreasing returns from education), which could be easily investigated with linear regressions, potentially supplemented by polynomial functions of variables or variable interactions. The results are easy to interpret and could be used for policy advice.

In contrast to the general thoughts above, in our study, we are not interested in the marginal effect of any single independent variable on the outcome. Instead, we are interested in making an improved prediction of grid-cell light density from exogenous first-nature geographic variables. Keep in mind that – at this stage – we only aim to construct an instrumental variable for spatial inequality. Importantly, the relationships between first-nature characteristics and economic activity must not be linear. Take precipitation, for example. Would we expect a linear relationship between precipitation and economic activity? The answer is no, since zero rainfall means droughts, high rainfall is associated with floods, and on top of this, whether rainfall fosters or harms economic activity depends on terrain issues and soil, amongst other factors. These complex relationships

are hard to capture in linear regressions, even if we work with polynomial functions of variables or multiple-variable interactions.

We investigate the relationship between lights and first-nature geography using a machine learning classifier algorithm, i.e. random forest. This method is commonly used in computer sciences, photogrammetry, and GIS applications, and it is receiving increasing interest among economists working with large and complex data sets (for an overview see Varian (2014) and Einav and Levin (2014)). The main advantage compared to predictions based on linear regressions is the non-parametric nature of the final model. We do not need to specify polynomials of variables or interaction terms in our model, since by construction, the algorithm addresses non-linearities and interdependencies. Moreover, we have a large set of variables in our data that are not necessarily continuous, which applies to the dependent variable (night light density) as well as the independent variables (first-nature geography). For example, it is questionable whether a light value of 60 implies twice as much economic activity compared to a value of 30. While a regression assumes exactly this, our machine learning procedure does not. We will discuss this issue in the data section below.

Random forest relies on trees, which describe a sequence of decisions. The general idea of a regression or classification tree is the following. In a regression tree, the data are split according to values of the target variable (grid-cell lights) in order to create homogeneous subgroups of observations according to the different input variables (first-nature geography). The regression or classification tree consists of different *branches*, *internal nodes*, and *terminal nodes* (for details see James et al. (2013)).

Let us use an arbitrary example suited to our context. If light were a simple binary variable (a grid cell is just dark or illuminated), then the algorithm would be a simple classification tree that splits the data into two sub-groups (branches) which have – according to a particular input variable (e.g. a biome type) – the highest possible group homogeneity in the target variable (light). It might be that the decisions of having light or no light are efficiently split by the fact of whether the grid-cell is classified as biome-type *Desert* or as something else. This creates an internal node that is split again according to the variable that divides the subsample into the most homogeneous subgroups (e.g. by temperature or distance to equator). The algorithm stops partitioning the data if, e.g. a predetermined number of internal nodes is reached or if a predetermined number of observations within a terminal node is reached. In the final prediction, an observation is then assigned the mean value of the terminal node, which fits to its different input criteria (biome type, average temperature above a certain value, etc.). Figure 5 in the Appendix shows an example of how a decision tree might look in our context.[3] One disadvantage of such a regression tree is a high variance. To reduce the variance, a method called bagging takes the average of several regression trees that use different training samples. Note that we refer to the random forest algorithm, which is a special form of bagging where each tree only uses a subgroup of predictors. Unfortunately,

[3] The example shown in Figure 5 uses the biome type temperate broadleaf as the first criterion to split the data; one subgroup is split based on the average temperature above a certain value; the other based on the biome type Mediterranean forest, etc.

there is no graphical representation for the results of random forests, and we therefore revert to the simple regression tree results.

The random forest algorithm does not provide results that can be directly compared to the regression coefficients of single variables. However, we can use the R^2 as criterion for the overall explanatory power of the model. In addition, we could ask the question: how important is a particular input variable for the quality of the prediction of the target variable? In order to create such an importance measure, the algorithm completely excludes a particular variable from the model, makes the prediction and calculates the mean squared prediction error. This result is then compared with a model that includes this variable and evaluates how the mean squared prediction error changes. This gives us an indication of whether a variable is a very important factor in predicting the target variable or whether it is less important. This importance measure is abbreviated by *IncMSE*. A second measure of importance is the so called *IncNodePurity*, which measures the total increase in the homogeneity of nodes (or the decrease in the so-called node impurities) which would arise if the sample were split according to a certain variable. This measure is based on the residual sum of squares and averaged over all trees.

The estimation procedure using regression analysis is straightforward and closely follows Henderson et al. (2017). We regress nighttime light on the different first-nature geographical variables using the following model:

$$\log(Light_{ij}) = \beta_0 + \beta_k \cdot FirstNature_{kij} + \theta_j + \varepsilon_{ij} \tag{1}$$

where $Light_{ij}$ is the sum of light divided by the number of pixels in cell i of country j[4], $First\ Nature_{kij}$ is a matrix which contains k different explanatory variables, θ_j are country fixed effects, and ε_{ij} is the error term.

Concerning the random forest algorithm, there is no single-equation representation of the model. We use light as the target variable and predict it with the same first nature variables as input variables, including country dummies. The calculations are employed using the R-package *randomForest*. We use the standard settings, i.e. the number of trees (*ntree*) is 500, the random set of variables for each additional tree (*mtry*) is the square root of the number of features, and the number of terminal nodes is not restricted.

2.1.2 Data

The main variable of interest is nighttime light, which we use as a proxy for overall economic activity at the regional level. Here, we follow a growing literature building on Chen and Nordhaus (2011), Henderson et al. (2012), Alesina et al. (2016), Henderson et al. (2017), and Lessmann and Seidel (2017), amongst others. Given that nighttime lights could be said to be a widely accepted indicator for (regional) GDP, we do not comment on all features of the data. Please see Donaldson

[4] Cell sizes are fixed by definition. Dividing by the pixels means that we consider only the landmass within each cell, which is important at the coast.

and Storeygard (2016) and Krause and Bluhm (2016) for an overview of the properties of the data and its pros and cons.

Nevertheless, one feature of the data is important in our context: the scale. The NOAA scientists assign a digital number to each pixel reflecting luminosity between $DN = 0$ (no light) and $DN = 63$ (maximum light intensity). This raises two questions: is the scale continuous and linear, and are the DN values of an ordinal nature with 64 categories (including zero)? Given that there is no clear physical representation of the DN values (like watt/pixel), one might treat the variable as ordinal. A regression analysis, however, does the contrary. First, it assumes that a DN value of 60 is twice as high as a value of 30, which could be challenged, and, second, linear regressions do not take the lower- and upper-bound truncation of the data into account. The second issue is very important in our case, since the prediction of lights from the geographic determinants might yield predicted values smaller than zero or greater than 63, both of which are impossible. By contrast, the random forest algorithm only makes predictions out of observed mean outcomes of terminal nodes; therefore, the predicted values always lie in the possible range of data. Moreover, the algorithm works with both continuous and ordinal data. These characteristics are likely to improve the quality of our predictions.

We refer to the so-called "Stable Lights" product distributed by NOAA. The original data have a resolution of 30×30 arc seconds, which is about 1 km^2 at the equator. We focus on a $1 \times 1^\circ$ grid, which covers an area of about 12,000 km^2 and corresponds to the size of the state of Connecticut in the U.S. Note that we abstract from any sub-national administrative boundaries, since these are man-made and endogenous.

Using a GIS system, we calculate the sum of DN values within a grid-cell and divide it by the number of land-covered pixels. Thus, we refer to a measure of light density as proxy of overall economic activity following Henderson et al. (2012) and others. Note that we do not refer to light per capita, since this would strongly depend on the population density, which is by definition a man-made factor resulting from second-nature agglomeration forces. In this part of the analysis, we want to stay in the first-nature world and investigate how physical geography affects overall economic activity, which is a function of productivity and population.

Data on first-nature determinants of regional development are taken from different sources (see table 7 in the appendix for data sources and definitions; table 5 provides summary statistics). We follow Henderson et al. (2017), considering measures of climate (temperature and precipitation), distance to water bodies (oceans and navigable rivers), terrain properties (soil constraints, elevation and ruggedness), and biomes. Overall, we consider 20 variables excluding fixed effects. The final data set consists of a cross-section of 15,664 cells (regions) covering the period 2008 – 2012.

2.1.3 Results

We use both methods described above – OLS regressions and random forest – to investigate the relationship between first-nature geography and grid-cell light emissions. We include country

dummy variables that capture unobserved country characteristics in both cases. Table 1 reports the estimation results. The dependent variable is light density at the grid-cell level. Note that we do not use a logarithmic transformation as in Henderson et al. (2017), since this would reduce variation in predicted values and affect the inequality measures we calculate later. Column (1) reports regression coefficients and t-statistics for the different first-nature variables. Column (2) reports the results we obtain by applying the machine learning algorithm. As mentioned above, this methodology assigns a value to every observation according to the final node it is in; therefore, regression coefficients for the different input variables are not available. As a measure of the importance of variables, we report the IncNodePurity and the IncMSE (in italics). Note that both importance measures have to be considered independent each other. In general, a higher value indicates higher importance.

Based on the regression analysis, we find that the average annual temperature has no significant effect on nighttime light densities.[5] Precipitation has an inverted-U shaped effect, i.e. more rainfall increases light up to a certain threshold. Being close to a river has no effect on light, while being far away from the coast has a negative effect. Ruggedness and latitude have no significant effect in our data. Soil constraints are negatively associated with night lights. Here, a low number indicates that there are no or only a few constraints for cultivating land, a high number reflects severe constraints. Thus, observed economic activity decreases with the number of constraints. Finally, elevation has a significant negative impact on lights. The coefficients of the biome variables indicate the difference to the base category *Desert*. The R^2 as a measure of the overall fit of the model is 0.49. If we alternatively use log light density as the independent variable, we get an R^2 of 0.56, which is the same value as in the baseline specification of Henderson et al. (2017). Note that we cluster standard errors at the country level.

Turning our attention to the results we obtain from the random forest algorithm, we find that the average temperature has a very high predictive power for night lights according to IncNodePurity, as do precipitation, distance to coast, latitude, etc. The results are quite different compared with the linear regression. For example, latitude is not statistically significant in the regression analysis, but it is a highly relevant factor in the random forest model. This may be a consequence of non-linearities and interdependencies of factors discussed above. Comparing the accuracy of the different models, random forest clearly outperforms regressions with an R^2 of 0.82.

In order to shed more light on the differences between the results of the two methods, we use both models for the prediction of grid-cell lights and compare them with observed lights. We aim to investigate whether the machine learning algorithm outperforms regressions in general or whether the improvement is only relevant for a subset of observations. Note that we set negative predicted values from the regressions to zero, since negative values are impossible by definition.[6]

Table 2 reports pairwise correlations for the different World Bank lending group regions. Column (1)

[5] We also tested a quadratic influence of temperature on light, but we could not find any significance for the quadratic term.

[6] Alternatively, we added the largest negative predicted value to all observations in order to get rid of negative values; this, however, yields a weaker relationship between observed lights and predicted lights.

Table 1: Results of OLS regression and importance of Random Forest determinants

	OLS	Random Forest
	(1)	(2)
Temperature (°C)	0.04769	17 744
	(1.479)	*22.10*
Precipitation (mm/month)	0.00811* *	12 850
	(1.662)	*32.69*
Precipitation (mm/month) squared	-0.00004**	
	(-2.542)	
Close to River (<25km) (dummy)	0.15140	944
	(1.006)	*10.86*
Distance to Coast (km)	-0.00106**	13 338
	(-2.130)	*38.61*
Latitude	-0.05779	11 612
	(-1.202)	*21.92*
Ruggedness	-0.20245	9 226
	(-1.528)	*33.03*
Soil Constraints	-0.25443***	8 475
	(-6.795)	*23.57*
Elevation (m)	-0.00061**	13 016
	(-2.413)	*33.39*
Tropical Moist Forest	0.00932	220
	(0.022)	*6.17*
Tropical Dry Forest	0.04262	205
	(0.101)	*5.19*
Temperate Broadleaf	3.52755***	19 292
	(4.275)	*38.17*
Temperate Conifer	0.65638	387
	(1.314)	*6.53*
Boreal Forest	0.08413	265
	(0.111)	*4.70*
Tropical Forest	0.05820	638
	(0.174)	*3.65*
Temperate Grassland	0.59878**	423
	(2.093)	*9.55*
Montane Grassland	1.78884***	57
	(2.944)	*3.93*
Tundra	-0.08401	202
	(-0.091)	*4.47*
Mediterranean Forest	2.95606***	2 955
	(2.660)	*17.23*
Mangroves	0.38374	824
	(0.453)	*9.37*
Desert	*Base category*	13
		-4.74
Constant	5.28631*	
	(1.911)	
Observations	15,664	15,664
R-squared	0.490	0.762
Country FE	YES	YES

Dep. var: light density. OLS (column 1) report beta coefficients and robust t-statistics in parentheses; RF (column 2) report IncNodePurity and *IncMSE* in italics. Note that due to the randomly chosen training sample the predictions, the importance measures and the R^2 vary every time random forest prediction is calculated.
*** p<0.01, ** p<0.05, * p<0.1

Table 2: Correlations observed and predicted light by World Bank Lending Group Regions

	L Obs - OLS Pred (s0) (1)	L Obs - RF Pred (2)	OLS Pred - RF Pred (3)
Not grouped; high developed	0.7514	0.8936	0.8577
East Asia and Pacific	0.5788	0.8192	0.7175
Europe and Central Asia	0.6805	0.8729	0.8212
Latin America and Caribbean	0.3092	0.7732	0.3309
Middle East and North Africa	0.4413	0.8508	0.5983
South Asia	0.2032	0.8579	0.6022
Sub-Saharan Africa	0.4709	0.7903	0.6044
World	0.7066	0.8803	0.8217

reports correlations between observed lights and lights predicted from the regression, column (2) reports correlations between observed lights and predictions from random forests, and column (3) reports correlations between the different prediction outcomes. Figure 6 in the appendix shows the corresponding scatter plots. The machine learning results outperform regression results in all parts of the world by means of a higher correlation with observed lights (column 1 versus 2). However, the improvement of the prediction is not distributed homogeneously across the globe. The differences are less pronounced in high income countries and Europe & Central Asia, while random forests generate considerably better predictions in the less developed parts of the world. The lowest correlations between predicted and observed lights are 0.20 for the regression results (South Asia) and 0.77 for random forests (Latin America and Caribbean), respectively. Overall, the correlation between Random Forest prediction and observed light is 0.88 and between the OLS prediction and observed light only 0.71. Thus, we are confident that the machine learning-based light predictions are much more informative than are the predictions based on regressions. Nevertheless, all main results of the forthcoming analysis hold for both methods.

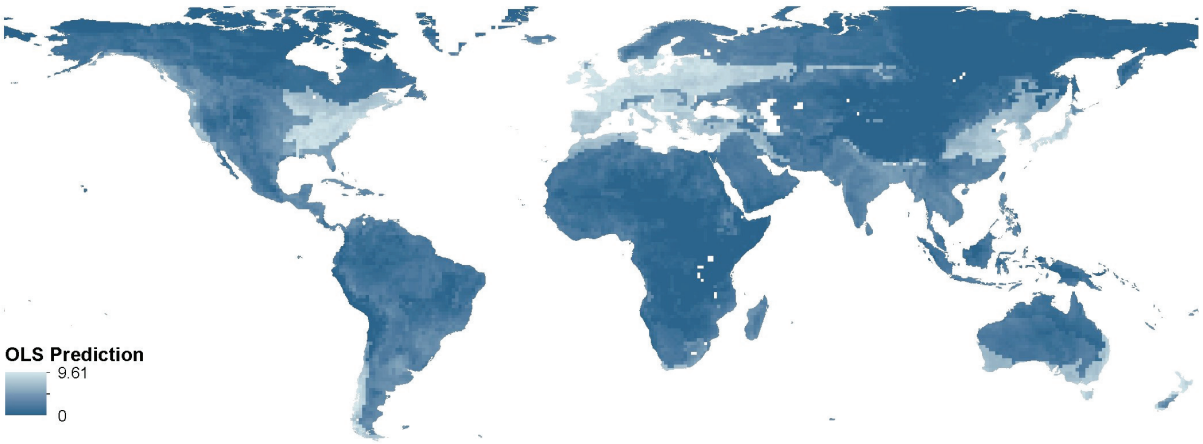
The differences between the grid-cell light prediction outcomes using the two approaches have important implications for measures of spatial inequality. A major advantage of the random forest results is the higher level of variation in the data, which could easily be seen by inspecting world maps of the prediction results. Figure 2 panel (a) shows the actual distribution of night-time light. Europe and the West Coast of the United States are the brightest areas, whereas the African continent is very dark. Especially in the Sahara Region and in the northern parts of Russia and Canada, almost no light is detected. Figure 2 panels (b) and (c) show predicted lights according to the OLS regression and Random Forest application, respectively. In both maps, light is distributed more smoothly than observed lights, particularly for the regression results. While the OLS predictions point to generally more light in Africa and South America, those areas are, according to the Random Forest Prediction, almost as dark as observed. This observation is in line with the correlations presented in Table 2 above.

The predictions yield the distribution of light across the world independent from man-made "second-nature" factors. Hence, they provide the initial geographic preconditions for economic

activity before any agglomeration process has occurred. According to the observed nighttime lights, economic activity seems to be concentrated in small regions (only occasionally lit cells), whereas according to first-nature characteristics, economic activity should be distributed more smoothly. The geographic starting conditions seem to be equitably allocated in many countries. Aligned with the new economic geography (Krugman (1991)), the concentration of economic activity in certain areas is a result of an agglomeration process that could not be explained solely by first-nature geography.

(a) Light Observed

(b) OLS Prediction



(c) Random Forest Prediction

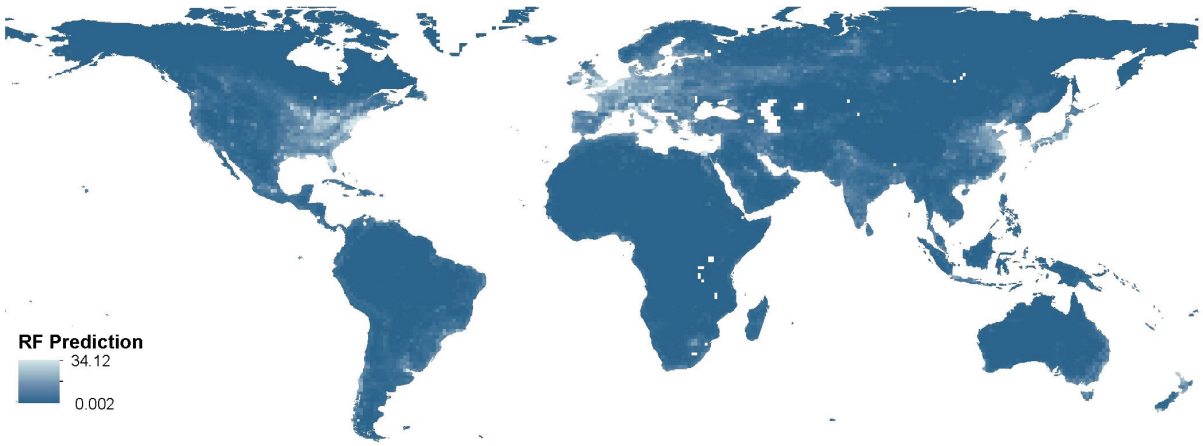


Figure 2: Maps of observed and predicted light

2.2 Spatial Inequality and First-Nature Geography

In this section, we calculate measures of spatial inequality within countries based on observed and predicted nightlights. The ultimate aim of our research is to estimate a causal effect of spatial inequality on economic activity. We will use an indicator of spatial inequality based on predicted light as an instrument for spatial inequality in observed light. Importantly, the variation in the instrumental variable comes only from differences in exogenous first-nature characteristics.

The literature on spatial inequality makes use of different inequality measures. Lessmann (2009) compares measures such as a coefficient of variation or a Gini coefficient with regard to their properties and different calculation outcomes. An important decision has to be made regarding the weights of the income data that enter the inequality formula. Commonly used are population weights, which reduce the effect of sparsely populated areas on the within-country inequality measure. Let us illustrate this issue based on the example of Canada: Among the industrial economies, Canada has one of the lowest levels of regional inequality based on population-weighted measures, while it has one of the highest inequality levels based on un-weighted income data. This is due to the extreme differences in per capita income between coastal regions and the giant – almost unpopulated – territories in the north. Researchers applying population-weighted measures focus on intergroup inequality, where groups are based on people's place of residence. This implies that the focus is still on inequality between human beings, or groups of human beings to be more precise. Taking this perspective is important if we are studying political economy determinants of regional inequality or the consequences of that inequality.

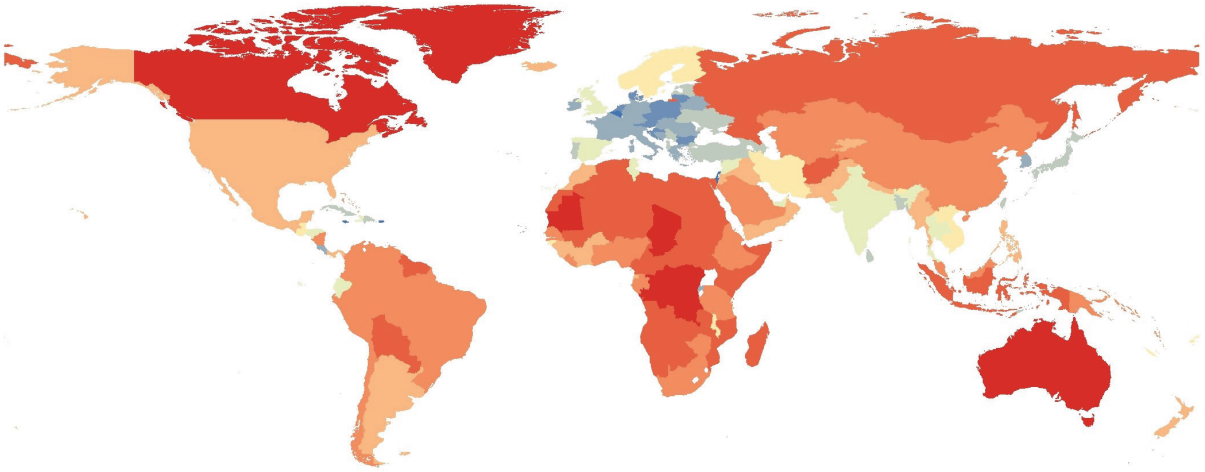
The perspective of our study is different. We aim to focus on spatial inequalities that are caused by first-nature geography prior to the agglomeration process. Consequently, we do not make use of population weights. Following Alesina et al. (2016) and Lessmann and Seidel (2017), we refer to the Gini index as an inequality measure, which is calculated as follows:

$$G_j = 1 + \frac{1}{n_j} - \frac{2}{y_j \cdot n_j^2} \sum_{i=1}^{n_j} (n_j + 1 - i) y_{i,j} \tag{2}$$

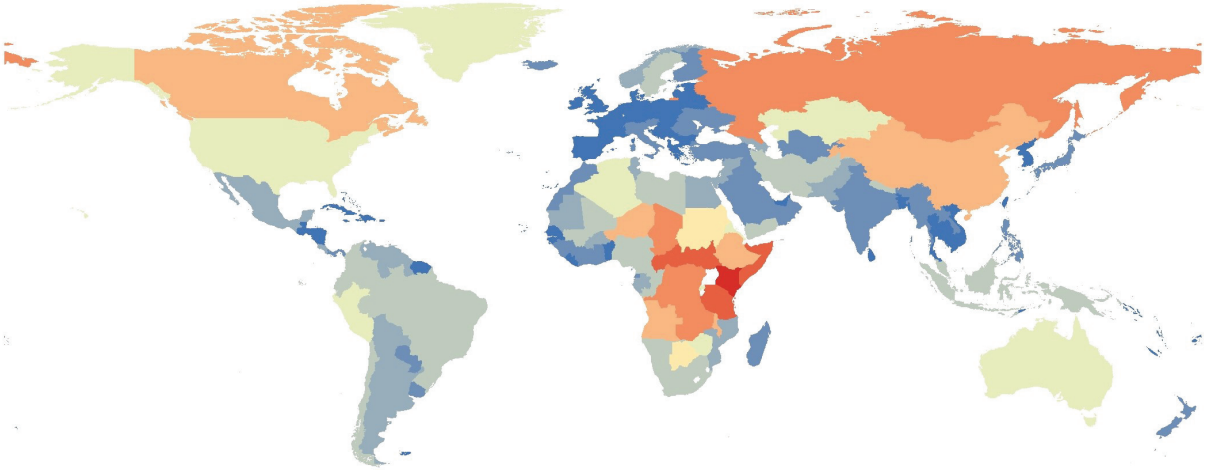
$y_{i,j}$ is the (predicted) light in cell i in country j and n_j is the number of cells attributed to country j. Weights are also important in our case, since not all grid cells have the same amount of land mass inside. Grid cells that contain abundant water area and that are therefore simply smaller in the underlying area contribute less to the country's output density. We therefore refer to an area-weighted version of the Gini coefficient.

Figure 3 panels (a)-(c) show the Gini coefficients of observed lights and predicted lights according to the OLS regression and Random Forest application, respectively. Comparing the different panels, it is obvious that inequality based on the observed light distribution (highest value: 0.95) is higher than the inequality measure based on the first-nature characteristics, predicted either by OLS or RF (highest value 0.85). This is not surprising, since spatial inequality in observed lights is a result of all determinants of regional development: first-nature geography, second-nature geography and all other man-made factors such as political institutions and history. By contrast, spatial inequality

(a) Gini Log Light Observed



(b) Gini Log OLS Prediction



(c) Gini Log RF Prediction

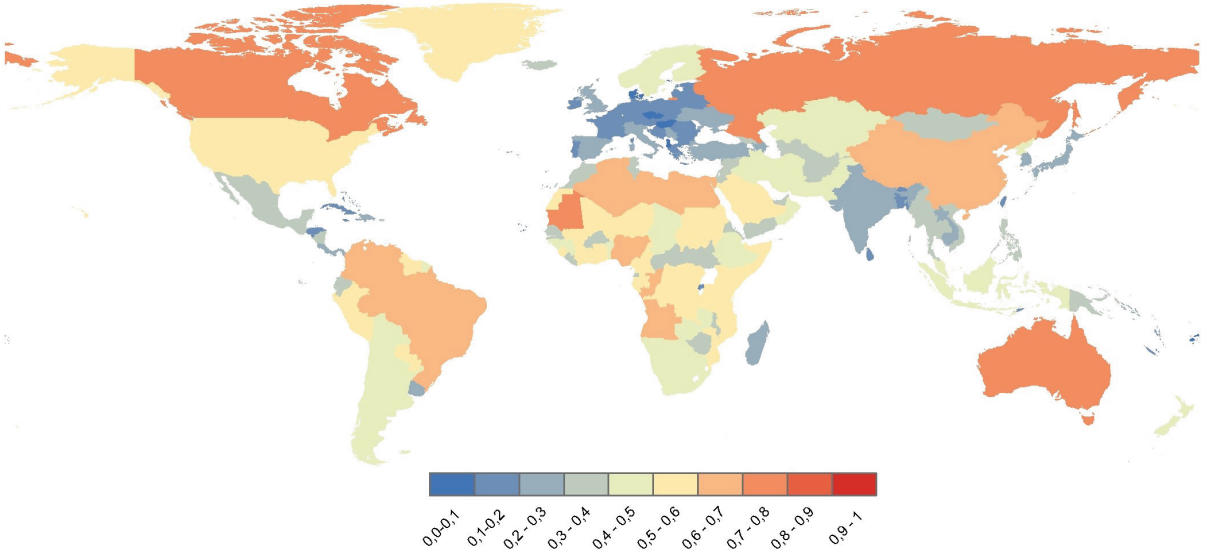


Figure 3: Maps of Ginis

in predicted lights captures only exogenous first-nature determinants.

When we consider different continents, the data reveal that Central Europe is the region with the lowest inequality in the case of observed economic activity and in the case of geographic fundamentals. Large differences between both observed inequality and predicted inequality occur in Africa and Latin America. In these parts of the world, economic activity might be distributed – according to the first-nature geographic factors – more equally than we actually observe.

There are a number of possible explanations for this result – apart from arguments from the new economic geography. In Sub-Saharan Africa, artificial borders drawn in the process of decolonization may have created large ethnic heterogeneity within countries, causing spatial income inequality today although geographic pre-conditions are relatively homogeneous (Alesina et al. (2011), Alesina et al. (2016)). Also politics determine agglomeration patterns, causing urban primacy in developing countries, particularly if weak democratic institutions are in place (Ades and Glaeser (1995)). In addition, the timing of early urbanization might be important, since economies of scale were not as important in history compared to more recent times (Puga (1998), Henderson et al. (2017)). Although the observations in our data are consistent with these arguments, this is not what we are ultimately concerned with.

In the next section, we investigate the causal impact of spatial inequality on economic development using inequality based on predicted lights as an instrument for observed spatial inequality. Figure 4 shows a scatter plot for the Gini coefficient in observed lights compared to the predictions.

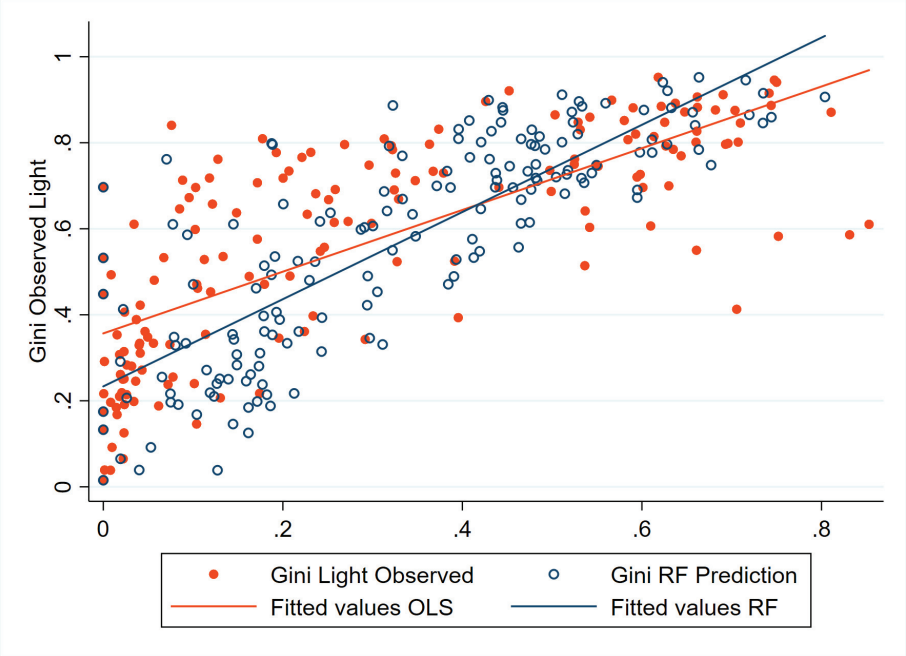


Figure 4: Correlation of Ginis

Although the inequality measures differ in mean levels, for the reasons mentioned above, the correl-

ation between them is fairly high, especially between the Gini from the Random Forest prediction and the observed light (0.82) The correlation between the Gini of observed light and with OLS predicted light is smaller (0.74). Again, the Random Forest predictions are more closely related to observed inequality compared with the OLS approach. Altogether, first-nature characteristics explain about $\frac{3}{4}$ of spatial inequality within countries.

3 Spatial Inequality and Development

3.1 Baseline estimates

In this section, we turn our attention to the investigation of a causal link between economic activity and spatial inequality using instrumental variable regressions. Several potential sources of endogeneity may bias estimations using simple OLS regressions.

Potential sources of endogeneity are reverse causality and omitted variables. For example, an increased level of national economic development may influence the spatial income distribution. Rich countries have more scope for redistribution policies that affect the regional income distribution. These policies could be direct instruments such as vertical or horizontal equalization transfers, or indirect transfers through the social security system, where relatively poor regions benefit overproportionally. Moreover, a country's wealth is positively related to investments in infrastructure that stimulate factor mobility, which also affects the spatial income distribution. Note that from a theoretical point of view, the directions of these effects are ambiguous. In a neoclassical framework, lower mobility costs for capital and labour cause a more equal spatial income distribution, since factor returns equalize across units. However, in models of the new economic geography, decreasing factor costs in combination with scale economies may facilitate agglomeration and therefore spatial income disparities. Also, institutions affect and are affected by both economic activity and spatial inequality. This list could be extended in several directions. Importantly, coefficients from OLS regressions are likely to be biased if observed levels of spatial inequality are considered on the right-hand side of the regression.

In order to deal with the endogeneity issue, we run instrumental variable regressions. We use constructed spatial inequality based on the first-nature characteristics as IV for observed spatial income inequality. Second-nature man-made factors amplify the agglomeration of economic activities in areas with favourable geographic starting conditions. An increased level of economic activity only affects those intensifying forces and not the physical geographic pre-conditions. Thus, we can claim that spatial inequality based on first-nature predicted incomes is strictly exogenous to observed spatial inequality, i.e., the instrument is uncorrelated with the error term. The relevance assumption will be proven later by showing the first-stage F-statistics.

The structure of the IV regression is the following:

First stage:

$$\log(G_j) = \beta_0 + \beta_1 \cdot GIV_j + \beta_2 \cdot GEO_j + \beta_i \cdot \sum_i X_i j + \gamma + \epsilon_j \tag{3}$$

Second stage:

$$\log(Y_j) = \tilde{\beta}_0 + \tilde{\beta}_1 \cdot \hat{G}_j + \tilde{\beta}_2 \cdot GEO_j + \tilde{\beta}_i \cdot \sum_i X_{ij} + \gamma + \eta_j \tag{4}$$

In the first stage regression (eq. 3), G is the observed spatial inequality of economic activity in country j and GIV is our instrument based on predicted incomes from first-nature geography. GEO are the first-nature geographic characteristics averaged on the country level and X is a matrix of i control variables we use in some of our specifications. We include world region fixed effects, γ, and ϵ is the error term. In the second stage, we regress the national log light density Y on spatial inequality $\hat{G}_j$ predicted from the first stage regression with error η_j.

As dependent variable for economic activity we refer to log light densities Y (the sum of lights within a country divided by the land surface). Alternative indicators will be used in the robustness section. Note that we refrain from using the GDP per capita, since the population of a country is endogenous to economic activity, geography and spatial inequality.[7] We control for a country's geographic patterns, since geography affects both national and regional economic outcomes. Including the national means of the first-nature variables used to predict regional incomes allows us to separate the direct effect from geography on economic activity from indirect effects through spatial inequality.

All regressions include world region fixed effects (γ). Given that first-nature geographic variables do not vary over time – except precipitation and temperature – our instrument is time-invariant, too. Therefore, we cannot use country fixed effects and have to concentrate on a cross section of countries. The world region fixed effects aim to reduce an omitted variable bias coming from unobserved heterogeneity between different parts of the world (Europe, Latin America, etc.). In order to reduce a bias from unobserved heterogeneity between countries, we include a number of country-level control variables reflecting the main determinants of economic growth, i.e. schooling, institutional quality, fertility, and investments (Barro (1991)).

Table 3 reports our main findings. Columns (1) - (4) report results of OLS regressions. Columns (5) - (7) and (8) - (10) report results of IV regressions using the random forest prediction and the predictions of the linear regression as instruments, respectively.

We first examine the relationship between a country's first-nature geography and economic activity. The results are reported in column (1). The average temperature and precipitation have no significant effect on national average light density. The only significant variables are ruggedness, elevation and soil constraints. A rough terrain as measured by the ruggedness variable is positively related to development (Nunn and Puga (2012)), while soil constraints and elevation have negative

[7] Note that GDP per area is highly correlated with the average light per country. According to our calculations, the correlation coefficient is 0.86.

effects. The first-nature geographic variables explain about 0.52% of the variation in lights per area between countries.

Next, we regress economic activity on the observed level of spatial inequality in a simple bivariate model. The results reported in column (2) suggest a strong negative relationship. This is in line with Alesina et al. (2016). Note that the coefficient indicates the value of the Gini multiplied by 100. Thus a one-unit increase in this variable corresponds to a 0.01-unit increase in the Gini coefficient defined in the range of zero to one. So, in column (1), economic activity decreases by 2.4% if the spatial inequality, measured by the Gini coefficient, increases by 0.01 units. In column (3), we combine regressions (1) and (2). Again, spatial income inequalities have a negative impact on economic development, but the coefficient decreases by approximately 0.004 units using the country-level averages of the first-nature controls. This specification explains about 0.67% of the variation in economic activity. In column (4), we add political controls. The coefficient of inequality increases compared to columns (2) and (3) but remains negative and statistically significant. The R^2 increases up to 0.76. Note that the sample size decreases significantly; therefore, we cannot compare the regression coefficients directly. Still, it is quite intuitive that the coefficient of the inequality measure decreases, since politics affect both development and inequality simultaneously.

Turning to the IV regressions, we run three different specifications for each prediction method. The results reported in columns (5) and (8) are similar to those reported in column (2) and contain inequality as the only explanatory variable. In columns (6) and (9), we add geographical controls, and in column (7) and (10), political controls. In every specification, the coefficient of spatial inequality is negative and statistically significant at conventional confidence levels. In absolute terms, the coefficients using the random forest prediction instrument (columns 5 – 7) are smaller than the coefficients of the OLS regression. The results suggest that the OLS coefficients are inflated due to endogeneity. According to these specifications, the true negative effect of spatial inequality on economic activity is less pronounced. The coefficients of spatial inequality instrumented by the linear prediction (columns 8 – 10) are in absolute terms higher than the OLS coefficients. These results contradict our previous conclusion. However, the country group fixed effects smooth the light prediction strongly. If we exclude country fixed effects, the coefficients of spatial inequality are smaller for both prediction methods. The results are reported in table 8 in the Appendix.

Turning our attention to the first-stage regressions, we find a positive and significant relationship between the instrument for spatial inequality and observed inequality. Note that to save space, we do not report the coefficients of the included instruments. The F-statistic of the first stage is larger than 36 for the specification using the linear prediction and even above 120 for the specifications using the random forest prediction. Therefore, these results comply with the suggested rule of thumb that the first-stage F-statistic should be larger than 10 if the used instrument is relevant (Staiger and Stock (1997), Stock and Yogo (2005)). We therefore conclude that our instrument is meaningful.

Table 3: IV Results

					Instrumental variable regressions					
		OLS			Random Forest			Linear Prediction		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	Geo	Ineq	Ineq+Geo	Ineq+Geo+Pol	Ineq	Ineq+Geo	Ineq+Geo+Pol	Ineq	Ineq+Geo	Ineq+Geo+Pol
First stage										
Gini Light Predicted$^{\diamond}$					0.941***	0.878***	0.926***	0.652***	0.488***	0.514***
					(14.98)	(11.12)	(11.76)	(10.06)	(7.39)	(6.01)
Second stage										
Gini Light Observed $^{\diamond}$		-0.024***	-0.028***	-0.028***	-0.017***	-0.020***	-0.025***	-0.029***	-0.038***	-0.035***
		(-9.006)	(-8.915)	(-6.589)	(-5.968)	(-5.425)	(-5.621)	(-7.741)	(-6.935)	(-5.231)
Temperature (°)	0.030		0.009	-0.000		0.015	0.005		0.001	-0.012
	(1.633)		(0.583)	(-0.012)		(1.039)	(0.282)		(0.056)	(-0.517)
Precipitation (mm/month)	-0.000		-0.002**	-0.003***		-0.002	-0.003***		-0.003***	-0.004***
	(-0.075)		(-2.118)	(-3.130)		(-1.546)	(-3.189)		(-2.779)	(-3.785)
Distance to Coast	0.015		0.022	0.018		0.020	0.018		0.025	0.017
	(0.856)		(1.470)	(0.846)		(1.511)	(1.002)		(1.415)	(0.760)
Latitude	0.016		-0.005	-0.015		0.001	-0.013		-0.012	-0.019
	(1.477)		(-0.501)	(-1.424)		(0.130)	(-1.397)		(-1.103)	(-1.560)
Ruggedness	0.364***		0.080	-0.006		0.158***	0.022		-0.018	-0.070
	(5.402)		(1.535)	(-0.074)		(2.808)	(0.298)		(-0.328)	(-0.948)
Soil Constraints	-0.116*		0.162***	0.066		0.085	0.038		0.259***	0.131**
	(-1.924)		(3.562)	(1.110)		(1.609)	(0.646)		(4.800)	(2.198)
Elevation	-0.001***		-0.000**	-0.000		-0.000***	-0.000*		-0.000	-0.000
	(-3.634)		(-2.239)	(-1.385)		(-3.030)	(-1.673)		(-0.915)	(-1.026)
Years of schooling				0.035			0.037			0.030
				(0.942)			(1.117)			(0.763)
Rule of Law				0.022			0.035			-0.009
				(0.247)			(0.426)			(-0.106)
Fertility Rate				-0.064			-0.077			-0.034
				(-0.950)			(-1.268)			(-0.554)
Gross Capital Formation				0.004			0.003			0.007
				(0.665)			(0.506)			(1.180)
Constant	-0.050	2.506***	2.109***	3.248***	2.186***	1.514**	3.006***	2.735***	2.862***	3.817***
	(-0.061)	(12.913)	(2.986)	(2.732)	(9.997)	(2.162)	(2.795)	(13.128)	(3.435)	(2.683)
Observations	164	164	164	117	164	164	117	164	164	117
R-squared	0.516	0.675	0.763	0.837						
WB Region FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
First Stage F-Statistic.					224.36	123.66	138.38	101.27	54.63	36.12

t-statistics in parentheses, clustered SE; Dependent variable: Log Light; $^{\diamond}$Ginis = Ginihttps://www.overleaf.com/project · 100

*** p<0.01, ** p<0.05, * p<0.1. Coefficients of included instruments are not reported to save space.

Overall, the findings of our three-step analysis indicate that countries with unequally distributed economic activity at the regional level have a lower level of economic activity at the national level. Using the random forest prediction with geographical and political control variables, we find that an 0.01 unit increase in the Gini coefficient decreases economic activity by about 2.5%. Thus, geography affects national economic activity through an unequal distribution of first-nature factors at the sub-national level. Geographic heterogeneity is important for national development.

3.2 Robustness Checks

To check the stability of our results, we perform a number of robustness checks, namely considering only a specific subsample of countries according to their size, income or time of development, and including additional controls like country size or ethnic inequality. In addition, we construct artificial countries instead of using current national borders.

3.2.1 Sample variations and additional controls

It is conceivable that in developed countries, inequality may affect economic activity less because dissimilarities can be compensated more easily. Therefore, we take a closer look at subsamples depending on development status and income. The results are reported in table 4. Note that we show only the main coefficient of interest, that is, the coefficient of the inequality measure from the second stage using the random forest prediction. Here, we include political controls just as in table 3 column (10).

In row (1), column (a), we exclude every country that is not categorized by the World Bank. Thus, 27 highly developed countries are excluded from the sample. The coefficient of inequality remains significantly negative. In addition, we divide the sample into low- and high-income countries. As the threshold, we take the mean of the logarithmic GDP per pixel. The results are shown in columns (b) and (c). While the main result for high income countries is very similar to our previous result for the whole sample, the coefficient is in absolute terms smaller if only low income countries are observed. Note that our sample shrinks significantly. However, columns (a) and (c) point to a lower effect of spatial inequality on economic activity in developing countries.

Especially in large countries, geographic characteristics vary substantially within the country. Thus, large countries may be outliers and bias our results. To deal with this issue, we estimate our model for different subsamples. First, we exclude the three largest countries, i.e. the USA, Canada and Russia. In a second step, we exclude the six largest countries, resulting in a supplementary exclusion of China, Brazil and Australia. The results of these estimations are shown in row (2), columns (a) and (b) of table 4. Even if we exclude those large countries, the effect of inequality is still significant and negative. As in the case of large countries, very small countries may influence our results since they are so small that geographic inequality can hardly arise. In row (2), column (c) of table 4, we exclude all countries that are smaller than or equal to the size of Belgium ($< 30,000$ km^2). Consequently, we exclude seven countries. Note that small island

states have already been excluded because they contain too few grid cells to calculate the Gini coefficients. Again, our main result is robust to this sample change. To account for the size of a country, we also include country size as an additional control variable. In row (3), column (a), we see that the effect of inequality on economic activity is still negative and significant. Not shown in the table is that the coefficient of area is negative, almost zero and only significant using the random forest prediction. We conclude that our results are not driven by either very small or very large countries.

Next, we control for the degree of ethnolinguistic fractionalization. Michalopoulos (2012) shows that ethnic fractionalization is determined by geographic fundamentals. Alesina et al. (2016) claim that geographic endowments are connected to ethnic homelands and show a negative correlation between economic development and geographic inequality across ethnic homelands. To check whether our effect of pure spatial inequality and economic activity diminishes when accounting for ethnic differences, we use ethnicity as additional control. The results are shown in row (3), column (b) of table 4. The effect of inequality is still negative and highly significant. The coefficient of ethnic fractionalization is not significant in any specification. Note that the correlation between geographic inequality and ethnic fractionalization may cause a multicollinearity problem, which inflates the standard error of the ethnicity coefficient. Still, there may be a direct effect of ethnic fractionalization on economic activity if we do not include the other geographic factors.

3.2.2 History and time of development

History may also play an important role in spatial inequality. Following Henderson et al. (2017), the differences in regional development depend on the time when a country completed the structural transformation from the agricultural to the manufacturing sector, i.e. before or after the decrease in transportation costs. In countries that developed early– which are today's high-income countries – the increase in agricultural productivity happened before the 1950s. Consequently, less labour was necessary for food production. Given the labour surplus in the agricultural sector, workers moved into cities and engaged in manufacturing (Desmet and Henderson, 2015). Since transportation costs were still high, several cities close to areas with good agricultural suitability evolved to guarantee food supply. This structure of cities is highly persistent over time. In contrast, in developing countries, the transportation costs were already low when the structural transformation occurred. This has lead to a concentration of manufacturing in a few regions that have locational advantages concerning access to trade. Proximity to farms was no longer necessary. According to this line of reasoning, the levels of spatial inequalities are larger in (late) developing countries.

To test whether early development influences the effect of spatial inequality on economic activity, we include three different variables that are indicators for early-developed countries. Following Henderson et al. (2017), the indicator variables are education, urbanization and GDP per capita in 1950. We run robustness checks including every variable in addition to our empirical model, which considers geographic and political control variables. Neither of the indicators of early development has a statistically significant influence. We show the main regression results in row (4) of table 4.

Table 4: Robustness checks

	Differentiation marks	(a)	(b)	(c)
(1)	**INCOME**	Without WB Reg 0	High Income	Low Income
		-0.011***	-0.024***	-0.011***
		(-3.05)	(-5.37)	(-2.78)
		89.74	*88.12*	*15.77*
		n=90	n=66	n=51
(2)	**SIZE**	Without big 3	Without big 6	Without small countries
		-0.028***	-0.020***	-0.025***
		(-5.86)	(-4.94)	(-6.95)
		108.74	*106.14*	*112.69*
		n=114	n=111	n=110
(3)	**FURTHER CONTROLS**	Country Size	Ethnic Inequality	
		-0.022***	-0.024***	
		(-5.44)	(-6.57)	
		104.29	*121.97*	
		n=117	n=115	
(4)	**HISTORY CONTROLS**	Urbanization	Education	GDP
		-0.023***	-0.023***	-0.027***
		(-6.81)	(-7.33)	(-8.46)
		143.61	*141.02*	*112.06*
		n=110	n=110	n=95
(5)	**HISTORY - URBANIZATION**	Low	High	
		-0.015***	-0.037***	
		(-4.02)	(-8.80)	
		65.68	*81.73*	
		n=78	n=46	
(6)	**HISTORY - EDUCATION**	Low	High	
		-0.012**	-0.39***	
		(-2.92)	(-9.99)	
		46.96	*80.63*	
		n=69	n=55	
(7)	**HISTORY - GDP**	Low	High	
		-0.014***	-0.036***	
		(-3.86)	(-10.01)	
		100.45	*88.29*	
		n=87	n=52	
(8)	**ARTIFICIAL COUNTRIES**	Grid 16°	Grid 16° right shift	Grid 16° upper shift
		-0.019***	-0.023***	-0.008**
		(-4.39)	(-7.48)	(-2.66)
		60.01	*64.80*	*65.83*
		n=123	n=122	n=130
(9)	**OTHER DEP VAR**	GDP per Area		
		-0.027***		
		(-3.60)		
		126.14		
		n=114		

t-statistics in parentheses, clustered SE, First stage F-statistics statistics in italics;
Dependent variable (if not defined differently): Light density
*** p<0.01, ** p<0.05, * p<0.1

The coefficient of spatial inequality on economic activity remains negative and significant. In addition, we divide our sample into two subgroups, late and early developers, according to historic urbanization rates, education and GDP. Rows (5) – (7) show the regression coefficients for every subgroup. The negative effect of spatial inequality on economic development remains in every subgroup. However, the effect of spatial inequality on economic activity is larger in early-developed countries.

3.2.3 Artificial countries

National borders are, of course, also man-made, but they do not change for the majority of countries. In order to rule out a potential bias from national borders, we also investigate artificial countries of a $16 \times 16°$ grid, which corresponds to the average size of a country of about 748 thousand square kilometres (e.g. Turkey). We double-check this by creating two independent large grids where the $16 \times 16°$ grid is shifted by $8°$ to the right or up north, respectively. Thereby, we can create new artificial states. Based on the new data set, we start our analysis from the beginning. We again predict light using two different methods, linear prediction and random forest, both with $16°$ grid cell fixed effects. On the basis of this predicted light and the observed light, we calculate the Gini coefficients and use them to conduct our instrumental variable estimation as in section 3. Since we divide the world into grids independent of any political borders, we cannot include country-specific variables such as schooling or investment. Thus, we can only show the results of the specifications using country-specific geographic controls as in columns (9) of table 3. The results of the regressions using artificial countries are shown in table 4, rows (8), columns (a), (b) and (c). The coefficient of inequality is negative and significant. We conclude that the effect of inequality on growth is independent of political borders.

3.2.4 Alternative dependent variable

We measure economic activity by national light densities. In a final robustness check, we use the GDP per area as alternative dependent variable. The results are shown in row (9) of table 4. The effect of spatial inequality on economic activity, this time measured by GDP, is still highly significant and negative.

4 Summary and Conclusion

We analyse the effect of spatial inequality on economic activity measured by nighttime light data. Our general approach relies on the idea that geographic pre-conditions affect economic activity. Krugman (1993) distinguishes between first-nature and second-nature determinants. First-nature determinants are exogenous to economic activity, while second-nature variables are all man-made factors, which depend on economic activity itself. Given the endogeneity of observed spatial inequality to economic activity, we propose an instrumental variable approach.

In order to construct our instrumental variable, we investigate the relationship between grid-cell lights (1×1 degree) and first-nature geography (temperature, precipitation, biomes, etc.). Note that we neglect sub-national administrative borders; therefore, the final instrument depends only on exogenous first-nature geographic factors. We use two different approaches: linear regressions comparable to Henderson et al. (2017) and a machine learning algorithm (random forest) that accounts for potential non-linear relationships between physical geography and local economic activity. The latter approach turns out to provide much better prediction results.

Based on predicted grid-cell lights, we calculate measures of country-level spatial inequality depending only on geographic pre-conditions. These variables are then used as instruments of observed spatial inequality in standard linear IV regressions. We find a robust negative relationship between spatial inequality and economic activity. Note that the time-variation in the geographic variables is very low; therefore, we have to rely on a cross-section of countries. This opens up the door for an omitted variable bias. We aim to reduce the bias by including world regions fixed effects and additional country-level controls.

Our study reveals that geography affects economic activity and thus societies not only through their country-level characteristics but also through the variation at the sub-national level. A more heterogeneous geography yields a more heterogeneous spatial distribution of economic activities within countries which, in turn, harms overall national outcomes.

Given that this study is the first of its type, we do not investigate in detail the channels through which spatial inequality harms national economic activity. There are many factors that might affect the relationship, e.g. infrastructure and equalization payments. Countries might counteract the disadvantages arising from heterogeneity in first-nature geographic characteristics with a dense infrastructure network. One might therefore expect that the relationship between spatial inequality and development is less pronounced in countries with high-quality infrastructure. However, infrastructure investments are highly endogenous to economic activity; therefore, a cautious empirical analysis should account for this. Just to provide some initial correlations, we interact road densities with spatial inequality. The results concerning the marginal effect of spatial inequality on economic activity depending on the density of the road network are illustrated in figure 8. As expected, the relationship is negative in countries with a low road density, while the effect is insignificant in countries with a high road density. Although we cannot make a causal claim on this finding, it supports the argument that infrastructure investments may be an effective instrument to counteract the negative effects of spatial inequality. A detailed causal investigation would be a promising topic for future research.

References

Ades, A. F., Glaeser, E. L., 1995. Trade and circuses: Explaining urban giants. Quarterly Journal of Economics 110 (1), 195–227.

Alesina, A., Easterly, W., Matuszeski, J., 2011. Artificial states. Journal of the European Economic Association 9, 246–77.

Alesina, A., Michalopoulos, S., Papaioannou, E., 2016. Ethnic inequality. Journal of Political Economy 124 (2), 428–488.

Alesina, A., Perotti, R., 1996. Income distribution, political instability, and investment. European Economic Review 40 (6), 1203–1228.

Alesina, A., Rodrik, D., 1994. Distributive politics and economic growth. The Quarterly Journal of Economics 109 (2), 465–490.

Amos, O. M., 1988. Unbalanced regional growth and regional income inequality in the latter stages of development. Regional Science and Urban Economics 18 (4), 549–566.

Barrios, S., Strobl, E., 2009. The dynamics of regional inequalities. Regional Science and Urban Economics 39 (5), 575–591.

Barro, R., 1991. Economic growth in a cross section of countries. The Quarterly Journal of Economics 106 (2), 407–443.

Barro, R. J., 2000. Inequality and growth in a panel of countries. Journal of Economic Growth 5 (1), 5–32.

Barro, R. J., Lee, J. W., 2013. A new data set of educational attainment in the world, 1950–2010. Journal of Development Economics 104, 184–198.

Benabou, R., 2000. Unequal societies: Income distribution and the social contract. American Economic Review, 96–129.

Chen, X., Nordhaus, W. D., 2011. Using luminosity data as a proxy for economic statistics. Proceedings of the National Academy of Sciences 108 (21), 8589–8594.

Desmet, K., Henderson, J. V., 2015. The Geography of Development Within Countries. Vol. 5 of Handbook of Regional and Urban Economics. Elsevier, Ch. 22, pp. 1457–1517.

Dhingra, S., Ottaviano, G., Sampson, T., van Reenen, J., 2016. The consequences of brexit for uk trade and living standards. LSE Research Online Documents on Economics, London School of Economics and Political Science.

Diamond, J., 1997. Guns, Germs, and Steel. The Fates of Human Societies. W.W. Norton and Co, New York.

Donaldson, D., Storeygard, A., 2016. The view from above: Applications of satellite data in economics. Journal of Economic Perspectives 30 (4), 171–198.

Easterly, W., 2007. Inequality does cause underdevelopment: Insights from a new instrument. Journal of Development Economics 84 (2), 755–776.

Einav, L., Levin, J., 2014. Economics in the age of big data. Science 346 (6210).

Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., Davis, E. R., 1997. Mapping city lights with nighttime data from the DMSP operational linescan system. Photogrammetric Engineering and Remote Sensing 63 (6), 727–734.

Fischer, G., Van Velthuizen, H., Shah, M., Nachtergaele, F. O., 2002. Global agro-ecological assessment for agriculture in the 21st century: Methodology and results. IIASA Research Report 02–02, International Institute for Applied Systems Analysis.

Forbes, K. J., 2000. A reassessment of the relationship between inequality and growth. American Economic Review, 869–887.

Fujita, M., Krugman, P. R., Venables, A. J., Fujita, M., 1999. The Spatial Economy: Cities, Regions and International Trade. Vol. 213. MIT Press, Cambridge.

Galor, O., 2009. Inequality and Economic Development: The Modern Perspective. Edward Elgar Publishing.

Galor, O., Moav, O., 2004. From physical to human capital accumulation: Inequality and the process of development. The Review of Economic Studies 71 (4), 1001–1026.

Henderson, J. V., Squires, T., Storeygard, A., Weil, D., 2017. The global distribution of economic activity: Nature, history, and the role of trade. The Quarterly Journal of Economics 133 (1), 357–406.

Henderson, J. V., Storeygard, A., Weil, D. N., 2012. Measuring economic growth from outer space. American Economic Review 102 (2), 994–1028.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Vol. 112. Springer.

Kanbur, R., Venables, A. J., 2005. Spatial Inequality and Development. Oxford University Press, Oxford.

Kaufmann, D., Kraay, A., Mastruzzi, M., 2011. The worldwide governance indicators: Methodology and analytical issues. Hague Journal on the Rule of Law 3 (2), 220–246.

Krause, M., Bluhm, R., 2016. Top lights - bright spots and their contribution to economic development. Annual Conference 2016 (Augsburg): Demographic Change 145773, Verein fuer Socialpolitik / German Economic Association.

Krugman, P., 1991. Increasing returns and economic geography. Journal of Political Economy 99 (3), 483–499.

Krugman, P., 1993. First nature, second nature, and metropolitan location. Journal of Regional Science 33 (2), 129–144.

Krugman, P. R., 1997. Development, Geography, and Economic Theory. Vol. 6. MIT Press.

Kuznets, S., 1955. Economic growth and income inequality. The American Economic Review, 1–28.

Lessmann, C., 2009. Fiscal decentralization and regional disparity: Evidence from cross-section and panel data. Environment and Planning A 41 (10), 2455–2473.

Lessmann, C., 2014. Spatial inequality and development – is there an inverted-u relationship? Journal of Development Economics 106, 35–51.

Lessmann, C., 2016. Regional inequality and internal conflict. German Economic Review 17 (2), 157–191.

Lessmann, C., Seidel, A., 2017. Regional inequality,convergence, and its determinants – a view from outer space. European Economic Review 92, 110–132.

Li, H., Zou, H.-f., 1998. Income inequality is not harmful for growth: Theory and evidence. Review of Development Economics 2 (3), 318–334.

Matsuura, K., Willmott, C. J., 2015a. Terrestrial air temperature: 1900–2014 gridded monthly time series, version 4.01.

Matsuura, K., Willmott, C. J., 2015b. Terrestrial precipitation: 1900-2014 gridded monthly time series, version 4.01.

Michalopoulos, S., June 2012. The origins of ethnolinguistic diversity. American Economic Review 102 (4), 1508–1539.

Neves, P. C., Afonso, Ó., Silva, S. T., 2016. A meta-analytic reassessment of the effects of inequality on growth. World Development 78, 386–400.

Nunn, N., Puga, D., 2012. Ruggedness: The blessing of bad geography in africa. Review of Economics and Statistics 94 (1), 20–36.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., et al., 2001. Terrestrial ecoregions of the world: A new map of life on earth: A new global map of terrestrial rcoregions provides an innovative tool for conserving biodiversity. BioScience 51 (11), 933–938.

Østby, G., 2008. Polarization, horizontal inequalities and violent civil conflict. Journal of Peace Research 45 (2), 143–162.

Perotti, R., 1996. Growth, income distribution, and democracy: What the data say. Journal of Economic growth 1 (2), 149–187.

Persson, T., Tabellini, G., 1994. Is inequality harmful for growth? American Economic Review 84 (3), 600–621.

Puga, D., 1998. Urbanization patterns: European versus less developed countries. Journal of Regional Science 38 (2), 231–252.

Rodríguez-Pose, A., 2018. The revenge of the places that don't matter (and what to do about it). Cambridge Journal of Regions, Economy and Society 11 (1), 189–209.

Staiger, D., Stock, J. H., May 1997. Instrumental variables regression with weak instruments. Econometrica 65 (3), 557–586.

Stewart, F., 2000. Crisis prevention: Tackling horizontal inequalities. Oxford Development Studies 28 (3), 245–262.

Stewart, F., 2005. Horizontal Inequalities: A Neglected Dimension of Development. Palgrave Macmillan UK, London, Ch. 5, pp. 101–135.

Stock, J. H., Yogo, M., 2005. Testing for Weak Instruments in Linear IV Regression. Cambridge University Press, New York, Ch. 5, pp. 80–108.

Varian, H. R., 2014. Big data: New tricks for econometrics. Journal of Economic Perspectives 28 (2), 3–28.

Williamson, J. G., 1965. Regional inequality and the process of national development: A description of the patterns. Economic Development and Cultural Change 13 (4, Part 2), 1–84.

Table 5: Summary statistics on grid cell level

VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) max
Light	17,686	2.000	4.871	0	61.66
Log Light	17,686	0.424	0.821	0.00139	4.122
Light - OLS prediction	15,664	2.017	2.998	0	32.37
Light - RF prediction	15,664	1.887	3.386	0.000150	40.79
Temperature (°C)	15,693	10.53	13.42	-20.04	31.32
Precipitation (mm/month)	17,354	67.88	67.78	0	599.2
Distance to Coast (km)	16,940	439.3	466.8	0	2,387
Latitude	17,669	37.87	21.57	0	74.88
Ruggedness	17,013	3.431	1.455	0	8
Soil Quality	17,013	5.163	1.582	0	8
Elevation (m)	17,391	540.0	730.5	-46.48	5,405

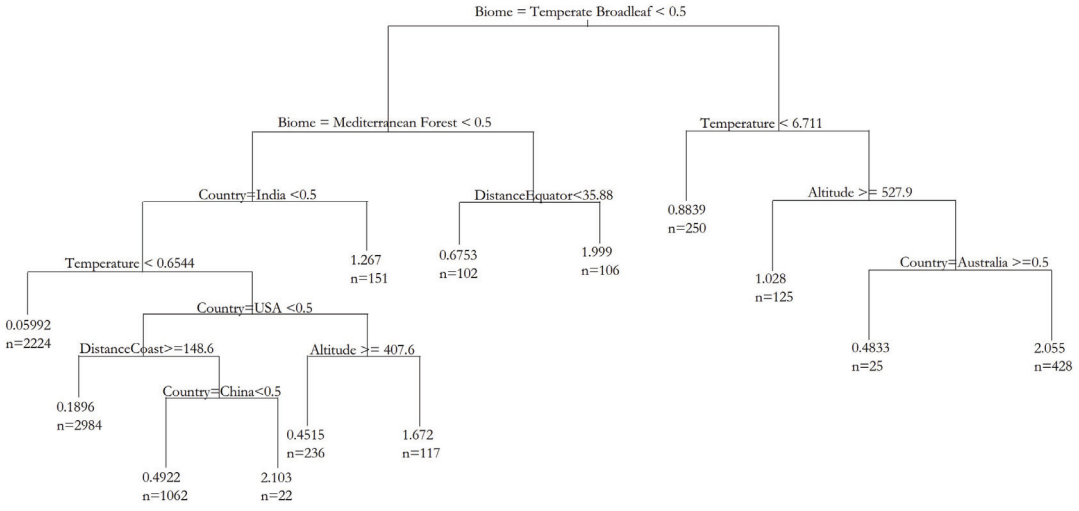


Figure 5: Importance of variables in Random Forest application

Table 6: Summary Statistics for variables on country level

VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) max
Light	157	3.923	6.107	0.00854	38.14
Gini Light Observed * 100	157	58.00	24.57	3.856	95.21
Gini Light - OLS Prediction * 100	157	31.03	25.62	0.0500	85.31
Gini Light - RF Prediction * 100	157	34.59	19.36	1.898	80.39
Temperature (°C)	157	18.36	8.836	-14.84	29.17
Precipitation (mm/month)	157	92.73	66.45	1.413	277.3
Distance to Coast	157	2.835	3.231	0.0371	16.36
Latitude	157	27.53	18.00	1.350	70.08
Ruggedness	157	3.491	1.132	0	6.884
Soil Quality	157	4.589	0.974	0	6.721
Elevation	157	505.2	508.7	5.830	3,115
Years of schooling	121	8.330	2.890	1.880	13.18
Gross Capital Formation	139	24.72	8.539	6.199	56.56
Rule of Law	149	-0.157	1.028	-2.486	1.951
Fertility Rate	151	3.016	1.520	1.222	7.664
GDP per Pixel (i.Th.)	145	1,106	2,200	5.636	11,983

Table 7: Data Sources

Variable	Source
Geographical	
Light	DMPS, NOAA and NGDC
Temperature	Matsuura and Willmott (2015a)
Precipitation	Matsuura and Willmott (2015b)
Distance to Coast	Euclidian Distances based on coastline borders taken from VMap0
Distance to River	River and lake centerlines database from Natural Earth
Latitude	"World Latitude and Longitide Grid" by esri online
Ruggedness	Fischer et al. (2002)
Soil Constraints	Fischer et al. (2002)
Elevation	GTOPO30, U.S. Geological Service
Biome	Olson et al. (2001)
Political	
GDP	World Bank
Years of schooling	Barro and Lee (2013)
Rule of Law	World Bank (WGI), Kaufmann et al. (2011)
Fertility Rate	World Bank
Gross Capital Formation	World Bank
Road Density	FAO

Figure 6: Correlation of observed and predicted light for each World Bank Region

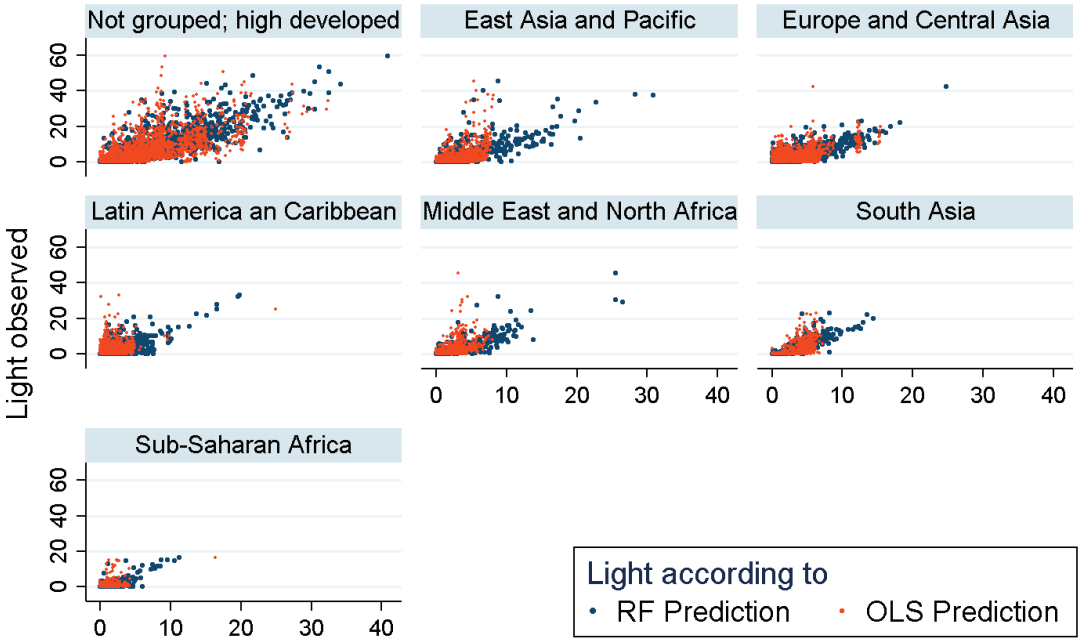


Table 8: IV Results without country fixed effects

	OLS				Random Forest			Linear Prediction		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	Geo	Ineq	Ineq+Geo	Ineq+Geo+Pol	Ineq	Ineq+Geo	Ineq+Geo+Pol	Ineq	Ineq+Geo	Ineq+Geo+Pol
First stage										
Gini Light Predicted$^{\Diamond}$					0.954***	0.790***	0.936***	0.521***	0.344***	0.409***
					(13.70)	(8.59)	(10.99)	(8.07)	(3.61)	(3.05)
Second stage										
Gini Light Observed $^{\Diamond}$		-0.025***	-0.028***	-0.027***	-0.019***	-0.020***	-0.023***	-0.020***	-0.016**	-0.028***
		(-11.177)	(-8.546)	(-6.206)	(-7.420)	(-4.477)	(-5.231)	(-7.899)	(-2.188)	(-3.272)
Temperature (°)	0.033*		0.007	0.003		0.015	0.009		0.019	0.002
	(1.666)		(0.459)	(0.147)		(0.985)	(0.541)		(1.063)	(0.067)
Precipitation (mm/month)	0.001		-0.002	-0.003***		-0.001	-0.003***		-0.001	-0.003***
	(0.375)		(-1.538)	(-2.936)		(-0.901)	(-2.843)		(-0.522)	(-2.914)
Distance to Coast	0.017		0.024	0.014		0.022*	0.014		0.021	0.014
	(1.024)		(1.567)	(0.639)		(1.649)	(0.748)		(1.545)	(0.687)
Latitude	0.017		-0.003	-0.013		0.004	-0.010		0.006	-0.013
	(1.522)		(-0.263)	(-1.193)		(0.385)	(-1.070)		(0.572)	(-1.054)
Ruggedness	0.318***		0.070	-0.010		0.145**	0.022		0.180**	-0.015
	(4.586)		(1.311)	(-0.128)		(2.486)	(0.292)		(2.369)	(-0.207)
Soil Constraints	-0.150**		0.133***	0.083		0.047	0.051		0.008	0.089
	(-2.279)		(2.633)	(1.417)		(0.790)	(0.875)		(0.100)	(1.267)
Elevation	-0.001***		-0.000**	-0.000		-0.000***	-0.000		-0.000***	-0.000
	(-3.271)		(-2.019)	(-1.170)		(-2.813)	(-1.455)		(-2.812)	(-1.288)
Years of schooling				0.031			0.033			0.031
				(0.846)			(1.000)			(0.876)
Rule of Law				0.017			0.031			0.014
				(0.180)			(0.360)			(0.171)
Fertility Rate				-0.065			-0.082			-0.062
				(-0.961)			(-1.379)			(-0.990)
Gross Capital Formation				0.005			0.004			0.006
				(0.977)			(0.860)			(1.070)
Constant	0.150	2.670***	2.164***	3.013**	2.404***	1.554**	2.697**	2.421***	1.272	3.068**
	(0.166)	(16.728)	(2.823)	(2.575)	(13.455)	(2.023)	(2.533)	(13.421)	(1.309)	(2.234)
Observations	157	157	157	115	157	157	115	157	157	115
R-squared	0.544	0.742	0.783	0.835						
WB Region FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
First Stage F-Statistic.					187.74	73.81	120.87	65.15	13.01	9.31

t-statistics in parentheses, clustered SE; Dependent variable: Log Light; $^{\Diamond}$Ginis = Gini · 100

*** p<0.01, ** p<0.05, * p<0.1. Coefficients of included instruments are not reported to save space.

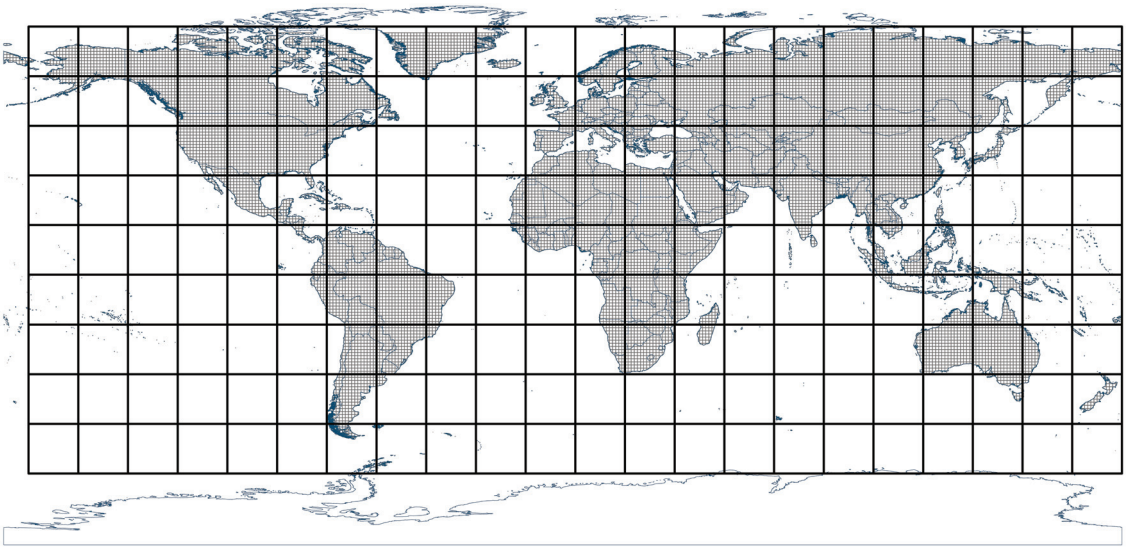


Figure 7: Observation units - Grid cells and administrative borders

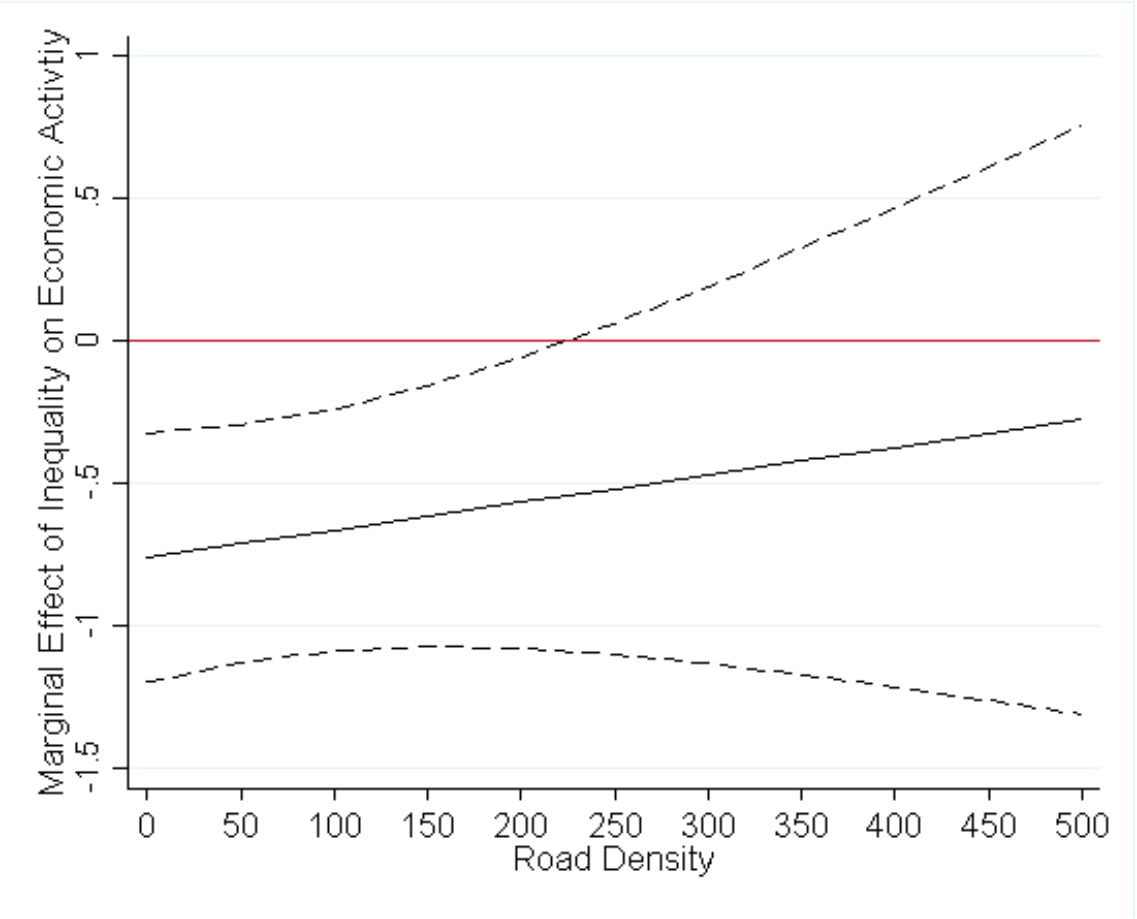


Figure 8: Marginal effects with interaction with road density