Seidel, André

Conference Paper

A global map of Amenities: Public Goods, Ethnic Divisions and Decentralization

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2020: Gender Economics

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Seidel, André (2020) : A global map of Amenities: Public Goods, Ethnic Divisions and Decentralization, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2020: Gender Economics, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/224555

A GLOBAL MAP OF AMENITIES: PUBLIC GOODS, ETHNIC DIVISIONS AND DECENTRALIZATION

André Seidel*

This version: 20.02.2020

Abstract:

I analyze the effects of ethnic divisions on the provision of public goods. Using OpenStreetMap data, I construct a new global dataset of locations of public amenities, such as schools, hospital and libraries. I allow for the possibility that the data may be systematically incomplete using two new proxies for mapping completeness. I provide strong evidence that more autonomous subnational regions with a high degree of ethnic fractionalization provide significantly fewer productive public goods. Therefore, my findings indicate that decentralization can lead to a failure in the provision of local public goods when it increases ethnical fractionalization among the policy makers responsible for collectively supplying public goods.

JEL: H41, H77, H75, D72, R53, C82

Keywords: public goods, amenities, decentralization, ethnic fractionalization, OpenStreetMap

*André Seidel: University of Bergen, Norway; andre.seidel@uib.no. I would like to thank David Weil, Maxim Pinkovskiy and Roland Hodler for their helpful comments at the second Workshop on Geodata and Economics at University in Hamburg. Special thanks go to Stelios Michalopoulos and Ferdinand Rauch for their constructive comments in an early stage of the project.

1. INTRODUCTION

Does ethnic heterogeneity prevent local governments from supplying public goods? A small but growing literature indicates that this might be the case at least in some countries. High ethnic fractionalization, which is the likelihood that a stranger does not belong to the same ethnicity as oneself, has most prominently been associated with lower subnational spending on education in the US (Alesina, Baqir, & Easterly, 1999). However, whether this phenomenon is global remains an open question that has become increasingly relevant as ethnic heterogeneity within countries continues to increase internationally.[1] Furthermore, there is an ongoing trend toward decentralization worldwide[2]. Countries increasingly delegate power from the central government to subnational regions without knowing the potential costs of decentralization. If decentralization leads to an increase in ethnical fractionalization among the policy makers responsible for collectively supplying public goods, decentralization may lead to a failure in the provision of local public goods.

In this paper, I study how ethnic heterogeneity and decentralization affect the supply of various public goods in first level administrative regions across almost all countries worldwide. I show that ethnic heterogeneity has a negative impact on the supply of regional public goods in subnational regions worldwide. However, this effect only emerges in subnational regions that are a part of decentralized countries. I do not find a similar effect on the supply of private goods. Both findings are consistent with the theory proposed by Alesina et. al. (1999), who predict that collective action to provide public goods is more likely to fail in the presence of high social heterogeneity. Additionally, consistent with the predictions by Alesina et. al. (1999), I find that only a subset of public goods is negatively impacted by the presence of high ethnic heterogeneity. Mainly productive public goods associated with education are affected, while other goods, such as public safety, are not affected.

The main policy implications of my findings are that decentralization should be accompanied by a careful evaluation of the resulting ethnic heterogeneity in the empowered subnational units. In some cases, it might be better to leave the public spending power to a central government to avoid the failure of local policymakers. In other cases, it might be possible to use administrative reforms to create more homogenous regions. However, the latter suggestion should be considered with a grain of salt as it may also increase the risk of separatism.

My analysis relies on a novel dataset that I assembled that contains the geocode locations of various amenities that are closely linked to some of the core public goods typically provided by government

[1] Measuring changes in ethnic heterogeneity over time is difficult; however, the existing data indicate an increasing trend toward more ethnic fractionalization across countries over the last 100 years, see for example Dražanová (2019).
[2] For a survey exploring the adoption of decentralization worldwide, see OECD (2019).

spending, such as schools, libraries, hospitals and police stations. I use the volunteered, crowd-sourced data collected by the OpenStreetMap (OSM) project. To study regional spending, I aggregate information at the regional level by counting the numbers of specific amenities in first-level administrative regions of countries. The new dataset covers 3342 regions in 204 countries. The number of public amenities is a simple but powerful proxy for spending on local public goods.[3] For example, in the case of the US, the number of primary and secondary schools per school district can explain between 68% and 74% of the variation in district-level educational spending.[4]

Using my new dataset, I estimate the effect of regional ethnic fractionalization on the number of different public amenities observed in subnational administrative regions across countries. The cross-country cross-regional setup allows me to use the variation in ethnic fractionalization among subnational regions of a country while controlling for country-level fixed effects. I show that regions in decentralized countries with high levels of social heterogeneity have a significantly lower supply of schools, libraries, and hospitals. The effect is large; for example, an increase in ethnic fractionalization by one standard deviation decreases the number of schools in a region by 7% to 14% if the region is a part of a federal country. Therefore, the average global effect is much larger than the effect reported by Alesina et. al. (1999) in the US, where an increase of ethnic fractionalization by one standard deviation decreases the share of educational spending by 2%.[5] I conduct placebo tests using non-public amenities and show that ethnic fractionalization does not impact the supply of restaurants differently in regions that are a part of a federal country.

My findings are robust to a large number of robustness tests, such as the use of different indicators of decentralization and ethnic heterogeneity. I further develop two new indicators to account for the degree of completeness of the OSM data. These indicators allow me to correct the data at the regional level or to control for the degree of completeness in cross-country analysis. When comparing the corrected data with official data of a subset of countries for which subnational data exist, I observe country-level correlations greater than 90%. Using official and OSM data to study the determinants of the degree of completeness, I find that completeness is primarily driven by national fixed effects and that regional development plays only a small role. The indicators of OSM completeness and national fixed effects explain between 85% and

[3]The number of public amenities per region is also a good proxy for welfare gains resulting from public goods. A greater local availability of public amenities usually results in higher welfare as the consumption of the associated public goods becomes less costly. The literature concerning school attainment indicates that a main driver of school attendance is distance to schools (e.g., Duflo (2001), Burde & Linden (2013), Kazianga et al. (2013) and Muralidharan & Prakash (2017)). The literature concerning other public goods, such as public safety (e.g., Blanes i Vidal & Kirchmaier (2018)) and emergency health care (e.g., Buchmueller et al. (2006) or Wilde (2013)), shows that the response time is a key issue. These data are correlated to the distance to the relevant amenity, which typically decreases as the number of amenities in a region increases.

[4] See section 3.2 for a related estimate.

[5] In section 3.2, I show that comparing increases in budget and the number of schools is reasonable as both are highly correlated. In the US, at the school district level, this correlation is approximately 70%.

95% of the variation in the observable completeness. Therefore, the risk of estimation bias arising from the OSM data collection seems small when using the indicators of OSM completeness and national fixed effects. The main findings of the paper however do not depend on the uses of these controls.

My findings contribute to economic literature on the political economy of government spending specifically on the effect of social heterogeneity. [6] The seminal paper of this literature by Alesina et. al. (1999) shows that for different levels of subnational units higher levels of ethnic fractionalization decreases spending on productive public goods, primarily education. I confirm this finding on a global scale. The existence of the effect only for productive public amenities and only in subnational regions within federal countries give further credit to the hypothesis that the underlining mechanism is indeed the collective action failure of local government.[7]

My findings further contribute to the economic literature concerning the costs and benefits of decentralization. There is a long standing discussion in the economics literature regarding the gains and costs of decentralization that dates back to the work conducted by Tiebout (1956), Musgrave (1959) and Oates (1972). Since their initial arguments, many scholars have identified various moderators that effect the outcomes of decentralization, such as the level of national development (Lessmann, 2012), the freedom of press (Lessmann & Markwardt, 2010), the level of government tiers (Fan, Lin, & Treisman, 2009) and the quality of the government (Neyapti, 2006).[8] My findings contribute to this literature by showing that the ethnic fractionalization of subnational regions determines whether decentralization leads to a reduction in the ability to provide public goods.

My study also contributes to the literature concerning the evaluation of volunteered geocoded data in geography. How to best measure the quality of volunteered geocoded data, such as the data available on Open street maps, Yelp, etc., remains an open question in the field of geography. Thus far, only a few studies evaluated the various aspects of the quality of OSM data. Senaratne et al. (2017) recently summarized this literature. I add to this line of research by providing the first globally available indicators of subnational mapping completeness. I also test the reliability of these indicators in a subset of countries at different stages of economic and OSM development.

The remainder of this paper is organized as follows. Section 2 presents the new data of the global location of public amenities and discusses the two new measures of subnational mapping completeness. This section also presents a test of the reliability of the new measures and dataset. Section 3 presents a replication of the

[6] See section 3.1 for more details regarding the general literature and section 4.1 for a summary of studies investigating the moderation of the effect of social heterogeneity on government spending by different institutions.

[7] With this finding, I also add to the general literature concerning heterogeneity and decision making; for a survey of this literature, see Ahn et. al. (2003).

[8] For a more detailed summary of the literature concerning the impact of decentralization, see Martinez-Vazquez et. al. (2017).

findings reported by Alesina et. al. (1999) using OSM data. Section 4 presents the main findings of the paper. Section 5 concludes.

2. A GLOBAL MAP OF PUBLIC AMENITIES

2.1. OPENSTREETMAP AS A SOURCE OF THE LOCATION OF PUBLIC AMENITIES

DATA COLLECTION

The data of the geographical location of the public amenities I use are extracted from the OSM project. The OSM project is a free, editable map of the whole world that is built by volunteers largely from scratch and released with an open-content license. By the end of 2017, the project had more than 4 million registered mappers, with an average of 40,000 people contributing data to the project per week.[9] The OSM project is the largest existing dataset of volunteered geographic information. The incredible success of the project is attributable to several factors, which have been well documented and discussed, such as by Senaratne et al. (2017). One factor is that untrained people, regardless of their expertise and background, have been able to add geographic information since the start of the project,[10] which is likely also the reason why, especially in less developed parts of the world, the OSM project has increased its coverage substantially in recent years. Different mappers and programmers associated with OSM have beautifully illustrated this point, for example, here[11] and here[12].

Data on the OSM project are provided by referencing with latitude/longitude nodes, lines, or polygons and attaching to these objects attributes in the form of tags (e.g., “amenity”=“yes” and “building”="pub”). The dataset is built using this information. I extract all polygons, multipolygon relationships, liens and points and their locations that carry tags associated with the various amenities under study from the OSM project until the end of 2017. For example, to identity schools, I use the tags “amenity”=“school” or “building”=“school”. Table 11 in the appendix summarizes all tags used. Section 6.1 in the appendix summarizes in greater detail how I extract and clean the raw OSM data.

GENERAL DATA QUALITY ISSUES AND INITIAL CLEANING OF THE RAW DATA

Using volunteered geocoded information generally has some drawbacks. Senaratne et al. (2017) summarized the current strand of the geography literature on the various quality issues associated with volunteered geocoded information. When examining economic geography, some issues are less important

[9] https://wiki.openstreetmap.org/wiki/Stats
[10] To see this point demonstrated, go to (https://wiki.openstreetmap.org/wiki/Beginners%27_guide) and see how easy it is to add something.
[11] http://tyrasd.github.io/osm-node-density/#2/19.1/21.4/latest
[12] https://www.youtube.com/watch?v=AM2fMJedqAc

than other issues. For example, topological consistency (e.g., whether objects overlap) and positional accuracy (e.g., whether objects are half a meter further south or north) are not of high importance for the applications in which economists are typically interested. However, there are other issues, such as thematic and semantic accuracy, that require discussion.

It is well known that tags are not consistently used in the OSM project since people are free to define new tags as they go. To address this problem, the OSM project has set guidelines on how and where to tag common objects, such as public amenities. The selection of tags I use to identify different amenities is based on these guidelines. Beyond the wording used in the different tags, they can be placed on different objects; for example, sometimes only the wall of a school is tagged with "building"="school", and sometimes the relationship between various objects that form the school is tagged with "amenity"="'school'". To avoid the resulting double counting (e.g., each school yard wall being counted as a separate school), I merge all objects with the same tag within a 100-meter radius into one observation.[13]

COMPLETENESS

A quality dimension critical for the application of OSM data in the study of economic geography is completeness. It is more than likely that, depending on the popularity of the OSM project, not all amenities that exist are recorded in the OSM data. Various issues could determine the magnitude of this effect, such as the lack of Internet access or legal boundaries. In the case of China, for example, mapping by private individuals is illegal.

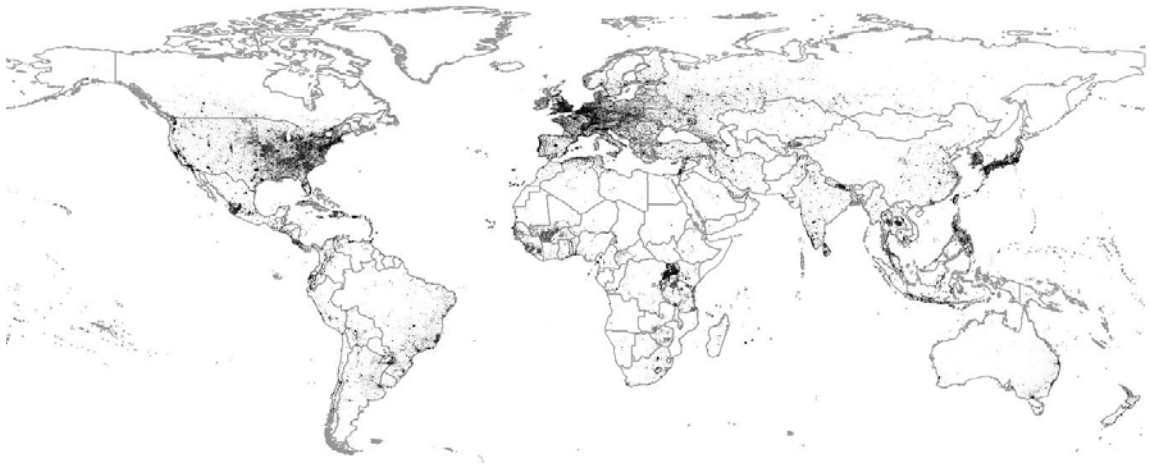
The descriptive statistics of the cleaned raw data can provide an initial impression of the data and the potential extent of missings. Figure 1 provides a first look at the data that I obtain after the initial cleaning. The figure displays all of the schools in the OSM project by the end of 2017 as a 50-m radius dot. At first glance, it is encouraging to see the close resemblance of Figure 1 to nightlight images and population density maps.

At a closer look, however, one might spot some unusual patterns, for example, the large numbers of schools in Uganda. An explanation for this finding might be that the Humanitarian OpenStreetMap Team (HOT)[14] has a large and successful project running in Uganda as a response to the ongoing refugee crises. As shown in the following section, despite the incredible increases in the number of OSM data volunteered in recent years, it seems that OSM data are incomplete in many dimensions. In this sense, Uganda is most likely an outlier at the top, with more data than other countries in Africa. An example of a possible outlier at the

[13] Obviously, this may create some error because in densely populated regions, public amenities could be in such close proximity that they are counted as one although there are actually two or more amenities. However, changing the radius to 50 meters does not change the results. For restaurants, the radius is reduced to 10 m.
[14] See https://www.hotosm.org/ for more details.

bottom might be China or North Korea, where the number of schools seems very small. This impression is reinforced when examining some simple descriptive statistics based on the cleaned raw data[15] that imply that there are 0.5 schools per 1000 citizens in the state of New York, whereas there are 0.02 schools per 1000 citizens in the province of Shanghai. The silver lining some might see in *Figure 1*, however, is that the distribution of schools across countries seems to be not dramatically distorted. A good example for this lack of distortion is China, where obviously many schools are missing, but the allocation still seems plausible. The greatest density of schools in the OSM data is observed in the heavily populated western regions of China. Therefore, it might be that missings are a mostly driven by country-level effect. Nevertheless, the following section discusses in detail how to account for the degree of completeness at least at the regional level.

Figure 1 Schools in raw OSM data as 100-m dots



2.2. APPROXIMATING THE REGIONAL DEGREE OF OSM COMPLETENESS

A SMALL THEORY ON OSM DATA COLLECTION

To better understand and combat the issue of completeness in volunteered data, I first state the general problem. An existing amenity is only recorded in the OSM data with a certain probability. I assume that this probability depends on the type of amenity and is constant within subnational regions. I refer to this probability as $p_{i,r}$,where $i \in \{School,\ Library,\ Hospital,\ Police\ station \dots\}$ and $r \in [0, n]$, with n being the number of regions in a country. Given this assumption the expected number of amenities recorded in the OSM data $A_{OSM,i,r}$ can be calculated by

$$A_{OSM,i,r} = p_{i,r} \cdot A_{i,r} \quad [1]$$

[15] See Table 14 for more descriptive statistics from the raw data.

where $A_{i,r}$ is the true number of amenities within a region. Consequently, using a proxy for an amenity's specific completeness of the OSM data ($p_{i,r}$), it is possible to predict the total number of amenities in a region based on the number of amenities observed in the OSM data.

To find a proxy for $p_{i,r}$ I extend my theory to account for the process of mapping. Aside from large-scale organized group efforts, for example, by NGOs such as HOT, mapping for the OSM project usually starts with individuals interested in improving the availability of high-quality digital maps in the region where they live.[16] In many cases, these people do not have high-quality equipment for mapping. Without the availability of, for example, GPS-based mapping devices, it is difficult to add data to a blank map. This restriction changes when fundamental landmarks, such as roads, have already been added to the OSM project. Using these landmarks, mappers can add data even without having access to GPS devices. For example, they can simply use addresses or distances between road crossings as reference points.

Based on this approach, I derive a simple theory regarding the process of mapping. Mapping occurs in two stages. Mapping in regions without any data in the OSM project starts by adding fundamental landmarks, e.g., roads. The second stage can start only after the first stage is realized. In the second stage, detailed data, for example, social-economic features, such as schools, police stations, cinemas, and restaurants, are added. If so, then the probability that a specific amenity is recorded in the OSM project is the product of the probability that stages one and two have occurred.[17] I assume that the degree to which the first stage has been realized in a region $p_{i,r}^{I}$ is region specific and that the degree to which a specific type of amenity has been recorded in the second stage $\varphi_{i,r}$ is amenity and region specific. Hence, the probability $p_{i,r}$ that a specific amenity is recorded in a region can be calculated by

$$p_{i,r} = p_{i,r}^{I} \cdot \varphi_{i,r.} \quad [2]$$

Next, I argue that it is possible to derive proxies for these two probabilities by comparing the OSM data and satellite data.

[16] For a more elaborate discussion of the motivations of OSM volunteers, see Goodchild (2007).

[17] Support for this model is derived not only from observations of the evolution of OSM data over time but also from the guidelines provided by the OSM wiki. Under the rubric mapping techniques (https://wiki.openstreetmap.org/wiki/Mapping_techniques), there is text reading, "Mapping is done in two steps: First, you need to know where things are, mainly the streets and ways. Then you need to know what there is, namely the POIs, street names and types. You can do these one after another, or both at the same time, but you can hardly do the what before the where".

A PROXY FOR THE FIRST STAGE OF MAPPING COMPLETENESS

A simple indicator of the degree completeness of mapping is the number of recorded objects in OSM relative to the existing number of objects in a region. Obviously, obtaining this indicator is not possible as the number of existing objects is endless and may even be unknown in certain subsets. Therefore, obtaining a proxy for the completeness of mapping is only possible by focusing on a subset of objects that can be overserved and is associated with a specific stage of mapping.

Residential roads represent an essential subset of objects associated with the first stage of mapping. [18] To obtain an indicator of the existence of residential roads in the regions, I use the Global Human Settlement Layer (GHSL, 2015). I define each arear\pixel in the GHSL (~one km^2) with urban buildup and more than 100 inhabitants as a settled area. I assume that within such a settled area, at least one residential road must exist. Hence, if there are no residential roads recorded in the OSM data in a settled area, it is likely that the first stage of mapping has not occurred. Therefore, the proxy for the degree of completeness of the first stage of mapping $p_{i,r}^{I}$ in an region is:

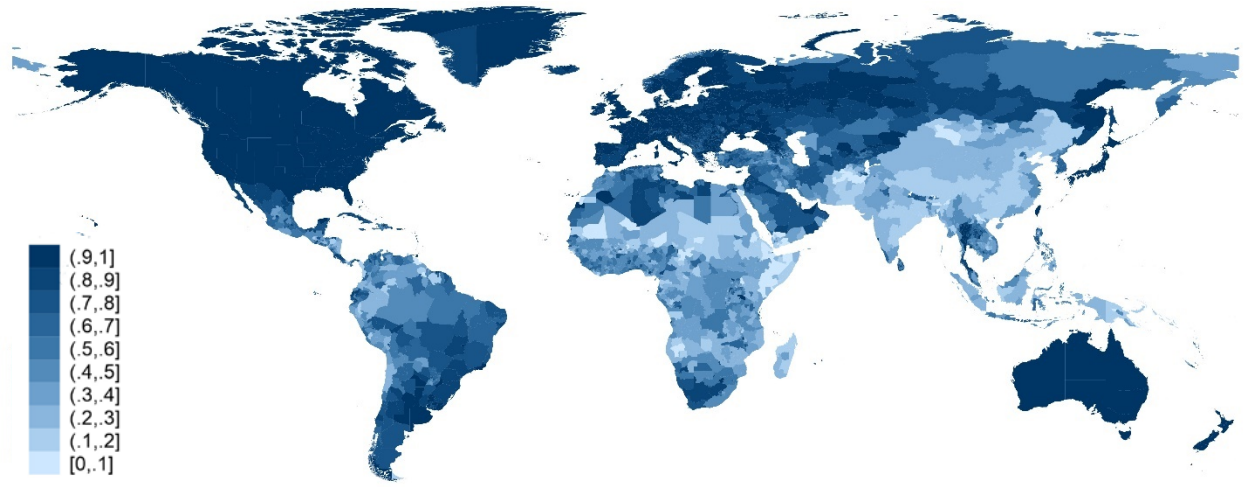
$$p_{i,r}^{I} = \frac{\#Pix_{S\cap R}}{\#Pix_{S}} \quad [3]$$

where $\#Pix_S$ is the number of settled pixels in a region (settled area), and $\#Pix_{S\cap R}$ is the number of settled pixels in a region that contains any residential roads in the OSM data (active OSM area). Hence, I assume that the share of the settled area in a region that contains residential roads in the OSM project is a good proxy for the degree to which the first stage of mapping has been realized in a region. For simplicity, below, I refer to $\#Pix_{S\cap R}$ as active OSM areas.

Figure 2 displays $p_{i,r}^{I}$ of the first-level administrative regions (GADM1) worldwide. This figure confirms the suspicions raised based on the simple plausibility test of the raw data in the previous section. First, in many African and Asian countries, the OSM data on the fundamentals are substantially incomplete. Second, the degree of completeness of fundamentals seems to be more heterogeneous between countries and less so within countries. Third, there is nevertheless heterogeneity within countries that should be considered when using OSM data in a scientific analysis.

[18] Road data are by far the most common data added to the OSM project in the early stages of mapping, and these data were added even before surface characteristics, such as mountains. Focusing only on residential roads (highway=residential roads or service or unknown) is advantageous as roads are a proxy for settlement structures and are usually not mapped by non-local mappers, such as government institutions (in contrast to larger roads connecting towns, such as highways and motorways).

Figure 2. Share of populated area with residential roads in the OSM data



A PROXY FOR THE FIRST AND SECOND STAGES OF MAPPING COMPLETENESS

A proxy for the extent to which the second stage of mapping has been realized in a region can be obtained using a logic similar to that used to derive the proxy for the first-stage realization. However, in the second stage, the subgroup of objects of interest is fixed as the aim is to obtain an amenity-specific indicator of the completeness of stage two of mapping. Consequently, finding an indicator of the likely existence of these objects is much more complicated.

I assume that each area that has undergone the first stage of mapping (active OSM areas) should contain at least one target amenity. Therefore, I obviously commit an error since it is unlikely that each square kilometer that is settled contains a specific amenity. However, I assume that this error is country specific. Therefore, the proposed proxy for the second stage of mapping completeness is as follows:

$$\varphi_{i,r} = \frac{\#Pix_{S\cap R\cap i}}{\#Pix_{S\cap R}} \cdot \varepsilon \qquad [4]$$

where $\#Pix_{S\cap R\cap i}$, is the number of pixels that contain a record of a specific amenity, and ε is the country-specific approximation arrow.

The proposed proxy for the completeness of the second stage of mapping is associated with an issue worthy of discussion. The assumption of a fixed country-specific error term ε is in some instances problematic. This assumption implies that the true share of the settled area that contains at least one amenity is fixed across regions countrywide. If there is a reason to believe that there is something effecting this type of amenity density and the number of amenities in the settled areas of a region, caution is advised when interpreting results that rely on $\varphi_{i,r}$ being a part of the proxy for the completeness of the OSM data. Circumventing this issue without knowing the true number of amenities in a region is impossible. Therefore,

as a robustness test, I recommend always controlling whether the results depend on the use of the indicator of the completeness of the second stage of mapping. Nevertheless, in the following section, I show that $\varphi_{i,r}$ despite its caveats still provides information worth utilizing when comparing OSM data with official data.

The indicator of the total completeness of mapping (e.g., stages one and two) can be obtained by inserting [3] and [4] into [2] to obtain the following:

$$p_{i,r} = \frac{\#Pix_{S \cap R \cap i}}{\#Pix_S} \cdot \varepsilon = p_{i,r}^{I+II} \cdot \varepsilon \qquad [5]$$

where $p_{i,r}^{I+II} = \#Pix_{S \cap R \cap i} / \#Pix_S$. For simplicity, I refer to $p_{i,r}^{I+II}$ as an indicator of the completeness of stages one and two of mapping.

Using the proxies for the completeness of mapping, I can predict the number of amenities based on the OSM data. After some algebraic computation, substituting [5] into [1] provides the number of amenities predicted by the OSM data.

$$A_{i,r} = \frac{A_{OSM,i,r}}{\#Pix_{S \cap R \cap i}} \cdot \#Pix_S \cdot \frac{1}{\varepsilon} = \frac{A_{OSM,i,r}}{p_{i,r}^{I+II}} \cdot \frac{1}{\varepsilon} \qquad [6]$$

Examining the middle of [4] shows that the proxy for the predicted number of amenities resulting from the use of both proposed proxies for completeness is equal to the average number of amenities within areas that contain at least one amenity of interest multiplied by the settled area in the region. Thus, the proposed correction of the OSM, data based on the indicator of the completeness of stages one and two of OSM mapping is to treat those areas that are active OSM areas that contain at least one amenity of interest as representative areas and to inflate their data to the settled area in a region. As discussed before, in some cases, this approximation could be problematic.

2.3. HOW TO USE THE PROXY FOR MAPPING COMPLETENESS

CROSS VALIDATION OF RESULTS

Based on the discussion in the previous section, I draw the conclusion that theoretically, the proposed proxy for the completeness of stages one and two could be biased in some cases. Therefore, I recommend always cross-validating the findings using the raw OSM data. Furthermore, I suggest cross-validating the findings using the proxy for stage one of mapping alone. This indicator partially accounts for the degree of completeness while avoiding the risk of bias by the assumption on which the indicator of completeness of stage two relies. To remain consistent with the theory underlying the approximation approach, I restrict the

observations to those within active OSM areas when using the proxies for the completeness of the OSM data.[19]

COUNTRY CASE STUDIES

When using the new amenity datasets and indicators of the completeness of stages one and two of mapping in a country case study, it is necessary to account for the country-specific approximation error ε. It is possible to calculate a proxy for the error if the true total number of amenities in country A_i is known. Given the assumption that bias is the same in all regions, ε can be obtain by summing both sides of [4] to obtain the following:

$$\varepsilon = \frac{\sum_r^n \frac{A_{OSM,i,r}}{p_{i,r}^{I+II}}}{A_i}. \quad [7]$$

Hence, the proxy $\tilde{A}_{i,r}$ for the true number of specific amenities in the region can be derived using [1] to [6] as follows:

$$A_{i,r} = \frac{A_{OSM,i,r}}{p_{i,r}^{I+II}} \cdot \frac{A_i}{\sum_r^n \frac{A_{OSM,i,r}}{p_{i,r}^{I+II}}} \quad [8]$$

CROSS COUNTRY ANALYSIS

When studying the regional determinants of the supply of amenities across countries, typically, the aim is to estimate the following:

$$\ln(A_{i,r,j}) = \alpha + \beta \boldsymbol{X} + \zeta \boldsymbol{Z} + \mu_{i,j} \quad [9]$$

where $\boldsymbol{X}$ is a vector of the explanatory variables of interest, $\boldsymbol{Z}$ is a vector of the controls, $\mu_{i,j}$ is the country fixed effect, and j is the country index. However, the problem is that the true number of amenities is unknown; thus, the need arises to approximate the number of amenities. The approximation approach described in the previous section suggests that the true number of amenities can be approximated by taking the log on both sides of [4], which yields the following:

$$\ln(A_{i,r,j}) = \ln(A_{OSM,i,r}) - \ln(p_{i,r,j}) - \ln(\varepsilon_{i,j}). \quad [10]$$

Substituting [10] into [9] gives the following estimation equation based on the OSM data:

[19] The indicator relies on the assumption that there should not be second-stage data in the OSM project if there are no first-stage data. This assumption is empirically not always true. However, the restriction has typically no large effect on the number of amenities within a region. The only noteworthy exception is the US, with its tendency to build school premises in more remote locations outside of towns. The results do not change when the US is excluded from the estimates.

$$\ln(A_{OSM,i,r}) = \alpha + \beta \boldsymbol{X} + \zeta \boldsymbol{Z} + \phi \ln(p_{i,r,j}) + \mu_{i,j} \quad [11]$$

Notably, $\log(\varepsilon_{i,j})$ becomes a part of the country fixed effect $\mu_{i,j}$, and based on the theory underlying the approximation approach, ϕ should be positive and close to 1.

2.4. TESTING THE RELIABILITY OF THE PROXIES OF OSM COMPLETENESS

COUNTRY-LEVEL RESULTS

As a first test of the reliability of the proxies for completeness, I use the assumptions described in the last section and [8] to calculate a proxy for the true number of amenities in a region. Then, compare this proxy with the true number of amenities per region in those cases for which I could obtain official data.[20] The focus of this analysis is on schools because it is possible to obtain the number of schools in first-level administrative regions for a decent set of countries. The scatterplots in Figure 3 show the number of schools for first-level administrative regions, as reported by government sources for various countries at different stages of development. The scatterplots always display the official data versus the raw OSM data with gray triangles and the adjusted OSM data with blue dots. For convenience, the 45° line is added in red.

Focusing first on the raw OSM data represented by the gray triangles shows that the naïve conclusion suggested by Figure 1, i.e., that the degree of completeness is entirely driven by country-level effects, is incorrect. There is considerable heterogeneity in the missing data across regions of countries. Nevertheless, it remains true that the average level of missing data seems to be country specific. It appears that, in Namibia and Mexico, almost all schools are missing, while in the US, there might even be too many.[21]

Finally, examining the adjusted data (blue dots) shows that the differences between the official data and the adjusted OSM data is considerably smaller than that compared with the raw OSM data. To put numbers to the magnitude of the adjustment effect, Table 1 summarizes the Pearson's correlation coefficients among the official number of schools and the raw and adjusted numbers of schools derived from the OSM data. Comparing rows one and two in Table 1 reveals that adjusting the number of schools as proposed by the approximation approach discussed earlier increases the correlation between the official data and the OSM data by a large margin. In most cases, the correlation with the adjusted data is greater than 90%. In the most extreme case of Namibia, even the sign of the correlation changes from negative to positive. Clearly, the

[20] To maximize the comparison dataset, data from various official sources from 2012 to 2017 are utilized. For more data sources, see Table 12 in the appendix.

[21] A closer examination of the US cases revealed several reasons why sometimes there are even more schools in the raw OSM data than the official data. Some reasons are related to tagging issues. For example, the OSM data include several historical schools in the Midwest that no longer exist. These schools are tagged as amenity=school with the Key=historic. Simply omitting these schools seems problematic since their removal might also imply the removal of schools in historic buildings. Another reason is that the official data reflect the number of public schools, whereas the OSM data also contain private schools.

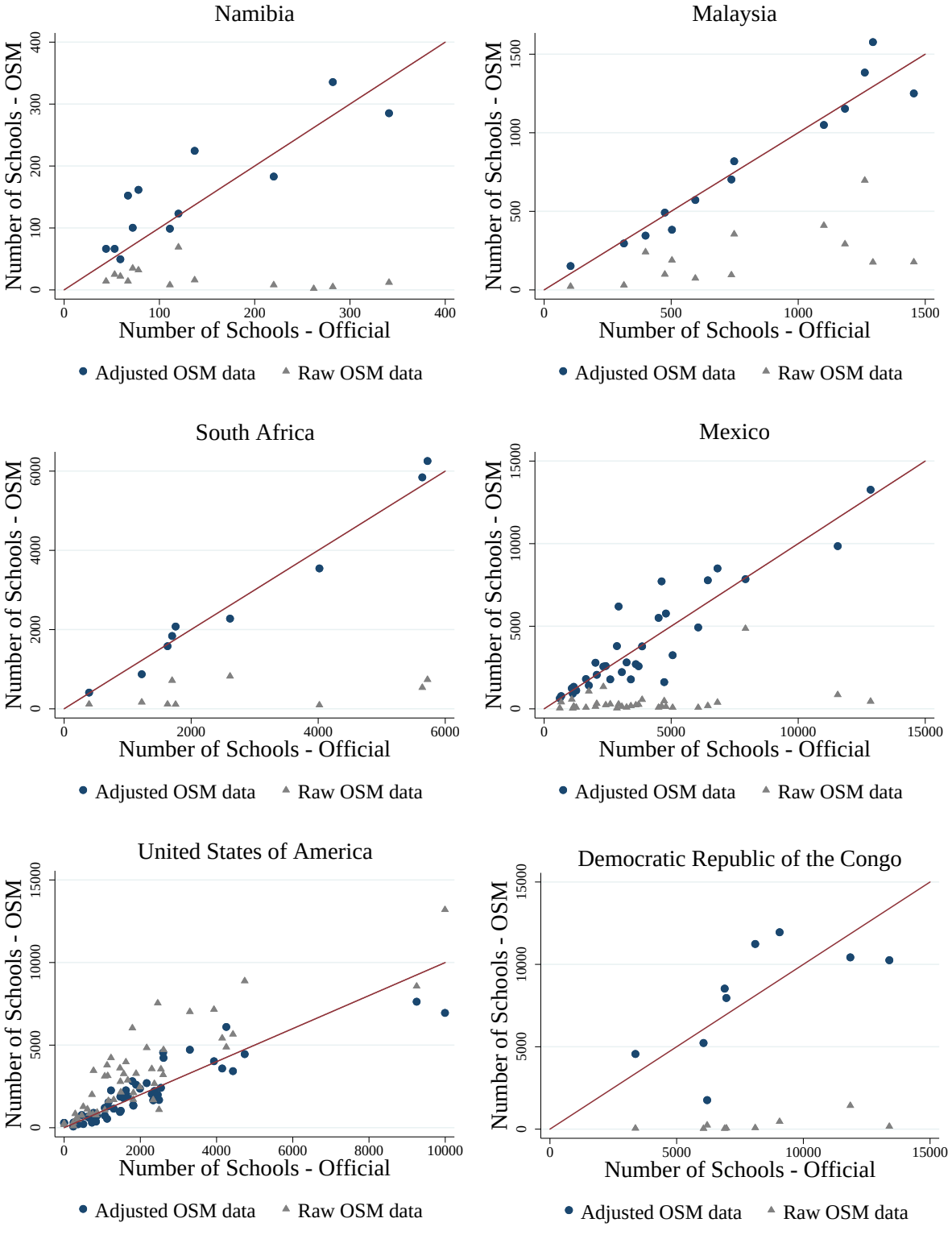
comparison shows that the correction is most important in less-developed countries but also helps to improve the correlation in advanced economies, such as the US. Furthermore, the proposed approximation approach is superior to a very simple and naïve approach in which one simply allocates the total number of amenities of a country to the different regions of the country depending on the regional population (Table 1, row 3). Comparing the first and last rows in Table 1 reveals that in most cases, the approximation approach derived in the previous sections is far superior to such a naïve approach.

Table 1. Correlations of the official number of schools with different proxies

Official # of schools in:	COD	MEX	MYS	NAM	USA	ZAF
# of schools OSM adjusted	0.6805	0.9082	0.9643	0.8555	0.9103	0.9871
# of schools OSM raw	0.5463	0.2987	0.5736	-0.4155	0.8605	0.4588
# of schools spread by pop.	0.5087	0.8382	0.7184	0.5709	0.9686	0.5921

Note: The table reports Pearson's correlation coefficients of the official number of schools in first-level subnational regions of individual countries with different proxies for the number of schools. # of schools OSM adjusted is the number of schools recorded in OSM in 2017 corrected using the proxies for completeness as described in section 2.3. # of schools OSM raw is the number of schools recorded in OSM in 2017. # of schools spread by Pop., is the population share-weighted total number of schools per country.

Figure 3. Number of schools observed vs raw OSM (left) and vs adjusted OSM (right)



CROSS COUNTRY RESULTS

Table 2 presents estimates of the determinants of the true degree of completeness. The estimates are based on 124 regions in 6 countries for which I could obtain data regarding the official numbers of schools in first-level administrative regions. As the determinant variable, I use the log of the share of the number of OSM schools in active OSM areas relative to the official number of schools, which measures the true degree of completeness. The estimation approach follows the cross-country study design suggested in the previous section. Hence, to account for potential bias in the OSM data, I conduct estimates including country fixed effects and indicators of the degree of completeness. Following the approach discussed in section 2.3, the robustness of the findings is tested by considering both indicators of completeness separately. Hence, the results of both completeness proxies $p_{i,r}^{I}$ and $p_{i,r}^{I+II}$ are presented.

Column 1 in Table 2 reports the estimation statistics obtained using only country fixed effects as the explanatory variable. The R^2 of 0.76 confirms the suspicion that the degree of completeness is mainly driven by country-level effects.

The estimates presented in column 2 in Table 2 support the very simple hypothesis that completeness is correlated with economic development. Using the average regional nightlight density as a proxy for regional economic development, I find a positive, significant correlation with completeness, which might be the case since with less income, the means of mapping are not available to most residents; hence, the number of contributors to the OSM project is smaller. Interestingly, income explains completeness less after adding the proxies for mapping completeness [columns 4 and 6]. After controlling for the log of $p_{i,r}^{I+II}$, light no longer has any significant association with omissions [column 6].[22]

Considering the power of these proxies, both of which are strongly significant and positive. The proxy for the first stage of mapping $p_{i,r}^{I}$ along with country fixed effects [column 3] can explain a considerable amount of variation in the data [R^2=0.848]. The fixed effects and the proxy for the completeness of stages one and two $p_{i,r}^{I+II}$ jointly explain [column 5] even more of the variation [R^2=0.959].

Skipping slightly ahead in the analysis, the results reported in Table 15 in the appendix suggest that ethnic fragmentation and decentralization do not seem to impact the degree of completeness. This finding suggests that even the raw data can be utilized to study the effects of both factors on the allocation of amenities.

[22] The link between development and the degree of missing data also becomes insignificant if the number schools in the raw data rather than the number of schools in active OSM areas is used.

Table 2. Determinants of the degree of completeness of OSM school data

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Var.	log(#School OSM / #School official)					
ln(light)		0.187***		0.106**		0.043
		(0.043)		(0.026)		(0.042)
$\ln(p^{I})$			1.781***	1.686***		
			(0.313)	(0.288)		
$\ln(p^{I+II})$					0.872***	0.857***
					(0.056)	(0.069)
Constant		-1.441***	-0.959***	-0.900***	0.564**	0.566**
		(0.038)	(0.114)	(0.095)	(0.140)	(0.146)
Observations	124	124	124	124	124	124
R-squared	0.755	0.774	0.848	0.854	0.959	0.960
Country FE	YES	YES	YES	YES	YES	YES

Note: The unit of observation is the first-level administrative regions. The dependent variable in all estimates is the log of the number of schools in the OSM data, recorded in active OSM areas in 2017 divided by the number of schools reported in official statistics (source years vary between 2012 and 2017). All of the estimates include country fixed effects that are not reported. ln(light) is the log of average nighttime light intensity extracted from the VIRS image of 2016. $\ln(p^{I})$ is the log of the proxy for OSM mapping completeness of stage one, and $\ln(p^{I+II})$ is the log of the proxy for completeness of stages one and two as defined in section 2.2. Standard errors are reported in parentheses and are clustered at the country level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

3. PUBLIC AMENITIES AND ETHNIC DIVISIONS

3.1. SOCIAL HETEROGENEITY AND COLLECTIVE ACTIONS - PREVIOUS FINDINGS

Before applying the new data to a new question, I revisit a central result reported in the previous literature. The provision of public goods depends on the cost of engaging in collective actions. With their seminal paper, Alesina, Baqir, and Easterly (1999) introduced the idea to the economic literature that these costs might depend on the social heterogeneity of the groups involved. Their hypothesis relies on two possible mechanisms. First, groups could simply differ in their preferences regarding different public goods; and second, the gains from using a public good could decrease if other groups also use it. The model built on these premises predicts that increasing social heterogeneity leads to a collective action failure, resulting, for example, in under-provision of public goods. Alesina, Baqir, and Easterly (1999) tested their theory by using US regional data. They reported the first empirical evidence of the under-provision of productive public goods in regions with high levels of social heterogeneity measured by ethnic fragmentation.

The link between ethnolinguistic fractionalization and the supply of public goods, such as education or health care, has also been found in countries other than the US. The vast majority of these studies relied on cross-regional data on specific countries and public goods (e.g., Alesina & La Ferrara (2000) (social

activities in the US); Dayton-Johnson (2000) (water supply in Mexico); Miguel & Gugerty (2005) (education in Kenya); Khwaja (2009) (infrastructure in Pakistan) or Díaz-Cayeros et al. (2014) (a range of public goods in Mexico)). Only a handful of studies adopted a cross-country perspective (e.g., Baqir (2002) or Alesina & Zhuravskaya (2011)). These studies, however, examined national-level outcomes, such as social sector spending or institutional quality. A small subset of studies has also attempted to approach the problem at the individual level using lab experiments and survey data, and they also confirmed that socially heterogenous groups have a greater tendency to mistrust one another and to fail in the provision of public goods (e.g., Glaeser et al. (2000), Bernhard et. al. (2006) or Habyarimana et al. (2007))

3.2. REPLICATING ALESINA ET. AL. (1999) WITH OSM DATA

In the following analysis, I replicate the findings reported by Alesina, Baqir, and Easterly, (1999) using my new dataset. I show that the number of amenities is linked to public expenditures and further reveal that despite the potential noisiness of the indicator of government spending, it is possible to replicate existing findings. Hence, this exercise is partially an additional robustness test of the data and an introduction to the discussion in the following section.

The main finding of Alesina et. al. (1999) is that, with increasing social heterogeneity, in US cities, metropolitan areas and counties, the spending on productive public goods decreases. To stay within reason, I focus on their findings regarding education spending. The replication is performed in the following two stages: first, I show that the number of schools is a good proxy for educational spending, and second, I show that the number of schools negatively depends on the degree of regional ethnic fractionalization. I obtain these results using official government data regarding the number of schools and the new OSM data.

The most detailed data on educational spending in the US are available at the school district level. Matching the spending data from the US Education Survey with the official and OSM data of the location of schools allows me to study spending in 7797 school districts[23]. Utilizing these data, I test the ability of the number of school district schools to predict the total educational spending. To account for productivities of scale, the estimates are biased on the number of schools in logs on the total expenditure on education in logs. In Table 3, columns (1) to (3) summarize the estimates using the official number of secondary and primary schools as an explanatory variable (1), the corrected number of OSM schools as defined in [8] (2), and the raw number of schools recorded in the OSM project (3). To account for potential distortion due to missing data in the raw OSM amenity data, in column (3), the likely degree of completeness approximated by $\ln(p_{i,r,j})$ is added. All three estimations reveal the same -- a strong correlation between educational expenditures and the number of schools in a school district. An average 1% increase in the number of

[23] Data are provided by the US Education Survey (2009).

schools is associated with a 1% increase in educational spending. Overall, the observed R^2 is between 60% and 70%. The number of schools therefore seems to be a good proxy for educational expenditures.

Next, I test whether the number of schools negatively depends on the degree of ethnic fractionalization in US counties. The focus in this analysis is on the county level since this approach enables the calculation of the same fractionalization indicators as those used by Alesina et al. (1999). Hence, the indicators are based on the ethnicity definitions and population figures from the US Census of 2010.[24] Ultimately, I can utilize data for 2131 US counties. The first part of the replication analysis indicates that a strong correlation exists between the log of the number of schools and the log of education expenditures. Consequently, the estimates are based on the log number of schools. To account for size effects, all estimates include the log of the area and population as controls. Table 3, columns (4) to (6), summarizes the estimates using as the dependent variable the official number of secondary and primary schools (4), the corrected number of OSM schools as defined in [8] (5), and the raw number of schools recorded in the OSM project (6). To account for potential distortion due to omissions from the raw OSM amenity data, in column (6), the likely degree of completeness approximated by $\ln(p_{i,r,j})$ is added to the set of controls. Despite a decrease in coefficient size, all three estimates show qualitatively the same effect that an increase of ethnic fractionalization by a standard deviation (0.060) decreases the number of schools by 1.5% to 2%. For example, an increase in the ethnic fractionalization of Starr County in Texas (0.01) to the level of Queens County in New York City (0.75) would decrease the number of schools by 20% to 25%.

Consistent with Alesina et al. (1999), the effects of ethnic fractionalization on the extent of public safety spending as measured by the number of police stations and health care spending approximated by the number of hospitals are mixed. Furthermore, the data suggest that a weak negative link exists between the number of libraries and the degree of ethnic fractionalization, which is consistent with the theory proposed by Alesina et al. (1999) that mostly productive public goods should be affected.

[24] From 1990 to 2010, the number of ethnicities recorded in the US Census increased considerably as citizens of Hispanic or Latino origin, for example, became recognized as different ethnicities. However, the findings do not change when using the 1990 classification of ethnicities.

Table 3 Replication of Alesina, Baqir, and Easterly, (1999) using OSM data from 2017

	(1)	(2)	(3)		(4)	(5)	(6)
	ln(Educational expenditure)				ln(#of.S.)	ln($\widehat{\#S.}$)	ln(#S.)
ln(#of.S.)	1.000***			ln(pop)	0.788***	0.902***	0.921***
	(0.007)				(0.006)	(0.009)	(0.007)
ln($\widehat{\#S.}$)		0.951***		ln(area)	0.178***	0.020	0.016
		(0.009)			(0.009)	(0.013)	(0.010)
ln(#S.)			0.921***	Eth. Frac.	-0.41***	-0.40***	-0.257**
			(0.006)		(0.115)	(0.138)	(0.117)
ln(p^{I+II})			-0.39***	ln(p^{I+II})			0.741***
			(0.010)				(0.014)
Constant	1.611***	1.503***	1.175***	Constant	-6.60***	-6.67***	-5.78***
	(0.016)	(0.021)	(0.026)		(0.096)	(0.117)	(0.101)
# District	7,791	7,791	7,791	# Counties	2,131	2,131	2,131
R^2	0.684	0.623	0.740	R^2	0.905	0.912	0.948

Note: The unit of observation in columns (1)-(3) is consolidated US school districts and, in columns (4)-(5), US counties. The dependent variable in columns (1)-(3) is the log of educational expenditures as reported in the 2015 annual survey of school system finances. The dependent variable in column (4) is ln(#of.S.), i.e., the number of schools as reported in the National Center for Education Statistics Common Core of Data. The dependent variable in column (5) is the number of schools recorded in OSM in 2017 corrected using the proxies for completeness as described in section 2.3. The dependent variable in column (6) is the number of schools in OSM reported in active OSM areas in 2017. ln(p^{I+II}) is the proxy for completeness of stages one and two as defined in section 2.2. Robust standard errors are reported in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

4. PUBLIC AMENITIES, ETHNIC DIVISIONS AND DECENTRALIZATION

4.1. MODERATION BY POLITICAL INSTITUTIONS - PREVIOUS FINDINGS

Because of the lack of reliable cross-country, cross-regional data on the provision of public goods, studies shedding light on the effect of political institutions on the link between fragmentation and regional public goods supplies are rare. An alternative to utilizing cross-country variation in political institutions is to use variation within a country over time. Miguel (2004), for example, found a positive effect of nation building on regional education spending in ethnically heterogeneous regions in Kenya and Tanzania between 1996 and 2002. Glennerster et al. (2013) found no effect of ethnic fragmentation on regional public good supplies using data for regions in Sierra Leone before and after the civil war. Cinnirella and Schueler (2016) found a positive effect of centralization on educational spending in linguistically fragmented regions in the eastern border regions of Prussia between 1886 and 1896. Alesina et al. (2017) found a negative effect on

deforestation of administrative reforms that reduced the ethnic diversity of regions in Indonesia between 2000 and 2012. Despite focusing on very specific countries, time periods and public goods, these last two studies partially support the main hypotheses that decentralization can reduce the supply of regional public goods when power is allocated to socially heterogeneous administrative regions.

4.2. DATA

PUBLIC AMENITIES

For further details on the data on the allocation of public amenities, see section 2.

ETHNIC DIVISIONS OF FIRST SUBNATIONAL ADMINISTRATIVE REGIONS

Among the various dimensions of social heterogeneity, ethnic heterogeneity has been shown to be widely important to various economic outcomes, such as growth or the likelihood of civil conflicts (Montalvo & Reynal-Querol, 2005). Following the vast literature, I use the following two commonly used indicators: ethnic fractionalization and polarization. Both indicators rely on the number of people belonging to different ethnicities in a country or, in this study, regions of a country as a measure of ethnic fragmentation. The main difference between the two indicators is how the population weights contribute to the indicator. The general rule of thumb is that, in the case of the fractionalization indicator, large groups contribute more than their relative size to the indicator, while the opposite is the case for the polarization indicator.

Defining $\pi_{e,r}$ as the share of people belonging to group e in region r that hosts m ethnic groups, ethnic polarization is measured by

[1]
$$Ethnic\ Pola_{e,r} = 1 - \sum_{e=1}^{m} \left(\frac{1/2 - \pi_{e,r}}{1/2} \right)^2 \pi_{e,r} = 4 \sum_{e=1}^{m} \pi_{e,r}^2 (1 - \pi_{e,r}) \quad [12]$$

and ethnic fractionalization is measured by

[1]
$$Ethnic\ Frac_{e,r} = 1 - \sum_{e=1}^{m} \pi_{e,r}^2 = \sum_{e=1}^{m} \pi_{e,r} (1 - \pi_{e,r}) \quad [13]$$

Ethnic fractionalization has a very intuitive interpretation. The indicator measures the probability that two randomly selected individuals are not from the same ethnicity. In contrast, the polarization indicator measures how far the distribution of the ethnic groups is from a bipolar distribution. Hence, high values of the polarization indicator correspond to cases in which there is an ethnic majority that is challenged by a unified “large” minority. For an in-depth discussion of the origin and uses of both indicators, see Montalvo and Reynal-Querol (2005).

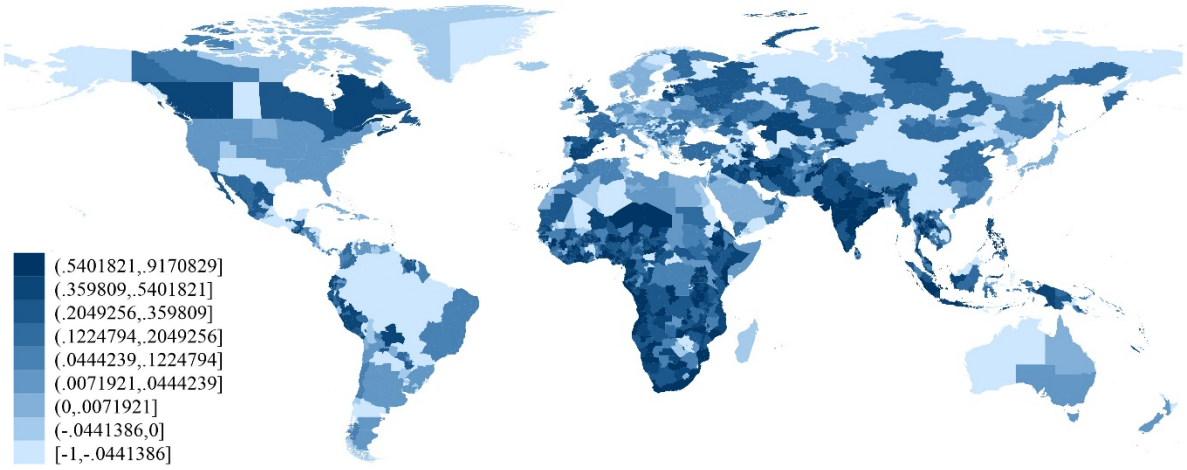
In the existing literature, ethnic fractionalization is the indicator of social heterogeneity most commonly used when studying collective action failure. Hence the main analysis focuses primarily on ethnic fractionalization as a measure of social heterogeneity.[25] Higher fractionalization is associated with a lower likelihood of collective action. A shift in the distribution of ethnicities toward a system with an ethnic majority should therefore decrease the failure of collective action. This outcome might not be the case if a simultaneous shift also "unifies" minorities into an opposing political force. The latter effect is more likely to be detected by the polarization indicator. Therefore, the robustness of the findings is tested using the indicator of ethnic polarization as an alternative indicator of social heterogeneity.

The data of the population belonging to different ethnicities are derived from a combination of gridded population data from the 2015 GHSL and the ethnic homeland data provide by GREG, which were reported by Weidmann et al. (2010). The GREG database reflects the distribution of ethnic groups worldwide in the 1960s and is based on a digitized version of the classical Soviet Atlas Narodov Mira. GREG documents the location of 928 ethnic groups in 8969 homelands. These homelands are projected to the current political boundaries of the first-level subnational administrative regions as defined by ADM. This approach creates 23874 regional homelands within 3219 regions.[26] For 2658 of these regional homelands, GREG reports more than one ethnicity residing in the area. For these regions, it is not possible to contribute their population to a specific ethnicity[27]. These multigroup homelands are spread across 1044 of the 3219 regions for which OSM data are available. Applying a strict exclusion criterion would therefore ultimately decrease the sample size by 1/3. Furthermore, it is likely that regions that contain homelands in which multiple ethnicities reside are also regions with higher levels of ethnic heterogeneity. Excluding these regions from an analysis, therefore, might induce a sample selection effect. To mitigate this issue while simultaneously reducing measurement error, regions that have more than 1% of the regional population living in homelands with multiple ethnicities are excluded from the main analysis, leading to the omission of 845 observations. The main results are robust to this exclusion criterion and extending the cut-off to the 10% level. Furthermore, the results do not depend on how the population residing in the multigroup homelands is allocated to the different ethnicities when calculating the social heterogeneity indicators.[28]

[25] This choice was most likely driven by data availability problems at the beginning of the literature since Alesina, Baqir, and Easterly (1999), in their seminal paper, already discussed the effect of polarization. Given the available data, however, they only tested for the effect of fractionalization.
[26] To minimize measurement error, regional homelands with a population smaller than one are excluded.
[27] Gridded population data are taken from GHSL (2015), 1000-m resolution image.
[28] For the main specification, the assumption is that the first named group in a multiple group homeland is the dominant one, and the population of the homeland is added to the total population of this group. The results do not depend on whether the population of multigroup homelands is allocated equally among the named groups or the same shares as in the rest of the region.

Ethnic heterogeneity has thus far mostly been studied at the national level or the regional level within selected countries. Therefore, the question arises of whether there is a meaningful difference between regional and national ethnic heterogeneity. To visualize this difference, Figure 4 displays the difference between national and regional ethnic fractionalization.[29] It is clear from Figure 4 that there are substantial differences in the degree of regional ethnic fractionalization within countries. These differences can go in both directions in Brazil; for example, most of the regions are more fractionalized than the overall country, and the opposite is the case for India, where the regions are much more homogeneous than the overall country.

Figure 4. National - Regional ethnic fractionalization



FEDERALISM AND DECENTRALIZATION

I use the following two different types of measures of decentralization: de facto and de jure measures. The de jure measures include the commonly used federalism indicator described by Treisman (2008), which indicates whether a federal constitution exists (1) or not (0), and a new federalism indicator derived from the CIA World Fact Book, which states whether the government type is federal (1) or not (0). The Treisman indicator is available for 155 countries, and the CIA World Fact Book indicator covers 199 countries that are a part of the amenity dataset. However, the CIA World Fact Book indicator is built based on only one very simple source of information, whereas the Treisman indicator is built based on multiple sources and, therefore, might be more accurate in some cases. This difference might also explain why the two indicators are highly correlated, at 0.92, but not perfectly correlated. Since the Treisman indicator is the standard

[29] The picture does not change when examining the level of regional fractionalization or polarization or the difference between national and regional polarization; see Figure 6 to Figure 8 and Figure 7 in the appendix. Note that, in the figures, regions with ethnic homelands that have residents belonging to multiple ethnicities are not omitted.

indicator used in the literature, the main analysis is performed using this indicator, and the CIA World Fact Book indicator is used as a robustness test.

The de facto measures are obtained from the IMF Government Financial Statistics. The three commonly used measures of fiscal decentralization include the share of subnational expenditures of the total expenditures, the share of subnational revenue of the total revenue and the subnational transferee share. The first two measures aim to directly proxy fiscal autonomy; however, they are not without problems. Neither indicator necessarily reflects autonomous decision making. The central government might still determine large parts of regional spending through its own legislation. A possible solution to this shortcoming is to use the third indicator. This indicator measures the share of subnational revenue provided by grants from other parts of the government. Hence, it proxies the fiscal dependence of subnational governments. The measure is also referred to as "vertical imbalance".[30] The main analysis focuses on vertical imbalance since it has the additional advantage of maximizing the number of available observations.

4.3. HYPOTHESIS AND ESTIMATIONS APPROACH

The previous literature indicates that social heterogeneity hinders the provision of public goods at the local level because of the increased risk of a collective action failure. This effect should increase with increasing local power and autonomy of regions within a country; hence, it should increase with increasing decentralization. Extending [11], the specific estimation equation used to test this prediction is as follows:

$$\ln\left(A_{OSM,i,r,j}\right) = \alpha + \beta_1 Hete. + \beta_2 Hete \times Auto + \zeta \boldsymbol{Z} + \phi \ln\left(p_{i,r,j}\right) + \mu_{i,j} \quad [14]$$

where $Hete$ is a measure of social heterogeneity, $Auto$ is a measure of the degree of local autonomy, $\mu_{i,j}$ are country fixed effects, and $\boldsymbol{Z}$ is a vector of the controls. The main prediction is that β_2 is negative. The previous literature on growth and social heterogeneity would indicate that, if β_1 is significant, it is most likely negative. The idea here is that social heterogeneity can decrease growth, which in turn reduces the ability to finance public amenities.[31]

4.4. IDENTIFICATION

There are considerable omissions in the OSM data as revealed by the analysis in section 2. Therefore, using OSM amenity data to study the allocation of amenities across regions must be performed with caution. The

[30] For a more in-depth discussion of the various approaches used in the literature on decentralization, see, for example, Lessmann (2009).

[31] Indeed, ethnic fractionalization has a significant, negative effect on the level of nightlight intensity in a region. For further details see Table 17 in the appendix and the discussion in section 4.6.

descriptive analysis and the analysis in section 2.4 indicate that the omissions seem to be mostly associated with country-specific factors and to a small degree with regional development. There is no evidence suggesting that regional ethnic fragmentation or the degree of decentralization impacts mapping completeness in countries for which official data of the allocation of schools across regions are available.[32] Nevertheless, to decrease the risk of omitted variable bias from the selection processes of OSM data, it is advisable to test the robustness of the findings by controlling for the degree of completeness of the OSM data using the proxies discussed in section 2. To show that the findings do not depend on the assumptions associated with the proxy for the completeness of the second stage of mapping, the findings obtained after controlling for the completeness of the first stage of mapping alone $\ln(p^{I})$ (see [2] and [3]) and those obtained after controlling for the completeness of the first stage and second stage of mapping jointly $\ln(p^{I+II})$ (see [2], [3] and [5]) are presented. It is important to note that the dependent variables differ between the estimates that include the proxies for completeness and those that do not. As suggested in section 2.3, when conducting estimates containing the proxies for completeness, the dependent variable is the number of amenities in active OSM areas[33]; otherwise, the number of all OSM amenities within a region is used. Given the theoretical argument presented in sections 2.2 and 2.3 and the empirical findings presented in section 2.4, the expectation is that the coefficients of the proxies for completeness are positive and close to one.

To reduce the likelihood of omitted variable bias further, the set of controls always includes country-level fixed effects $\mu_{i,j}$ and the regional log of the population and log area. The number of amenities is expected to increase as the number of regional residents increases, while the expectations of the effect of area are ambivalent.

An identification threat that one might see is that decentralization might be triggered by high levels of ethnic fractionalization. Since the focus is on studying a phenomenon at the regional level, the endogeneity of institutions does not seem to be of greater relevance given that regional ethnic fractionalization and all measures of decentralization used are only weakly correlated [see Table 4, column one]. One explanation for this observation might be that, at least in developing countries, decentralization was often pushed from international organizations and aid donors, rather than country forces. This fact might also explain why the correlation is slightly stronger in wealthier countries, but even among them, the correlation is very weak (see Table 4, column (2)). If ethnic fragmentation drives the decision to decentralize, then it seems that fragmentation at the national level and not within regions might play a role; however, even then, the correlation is very weak (see Table 4, column 3).

[32] See Table 15 in the appendix.
[33] E.g. areas with urban buildings, more than 100 residents and residential roads in the OSM data

Table 4. Correlations between decentralization and ethnic fractionalization

	Ethnic frag. ADM 1	Ethnic frag. ADM 1 Gdp p.c > 9000 $	Ethnic frag. ADM 0
Federal in Treismann	0.0242	0.1666	-0.0060
Federal in CIA World Factbook	0.0422	0.1683	0.0121
Share of subnational revenue mean 90-18	-0.1012	0.0380	-0.3597

4.5. MAIN RESULTS

Table 5 reports the main results of the paper. The determinant variable in the baseline estimates is the log number of schools within first-level subnational regions. All estimates are based on a consistent dataset that is restricted by the availability of the main indicator[34] and consists of observations in 1965 subnational regions in 155 countries.[35]

Column (1) presents estimates that only include the country fixed effects, the log of the population and the log of area. The standard controls explain 85% of the variation within the data. The coefficient of the log of the population is positive and strongly significant, while the coefficient of the local area is positive but not significant. These findings confirm the reasonable expectation that the main determinant of the number of schools within a region is the population of the region.

In column (2), the level of regional ethnic fractionalization is added and exhibits a significant, negative effect coefficient. The effect becomes insignificant in column (3) after adding the interaction between ethnic fractionalization and the indicator of decentralization, which exhibits a strong negative coefficient. The coefficient suggests that an increase in ethnic fractionalization by a standard deviation (0.19) is associated with a decrease in the number of schools by 3% in a non-federal state and by 14.2% in a federal state. The results shown in column (3) suggest that ethnic fractionalization decreases the supply of schools in regions that are a part of a decentralized country by a considerable margin. I consider this outcome my main finding.

It is important to test whether the findings shown in columns (2) and (3) are affected by the regional degree of completeness of the OSM amenity data. The estimates presented in columns (4) and (5) include the log of the indicator of the completeness of the first stage of mapping ($\ln(p^{I})$). The estimates presented in columns (6) and (7) include the indicator of the total degree of completeness of mapping ($\ln(p^{I+II})$). The inclusion of these controls does not change the quality of the main findings. However, the effect size

[34] Note that regions without any data in the OSM project are omitted, leaving 2956 observations. Regions where more than 1% of the total population lives in ethnic homelands in which multiple ethnicities reside are also omitted, leaving 2226 observations. The availability of the decentralization indicator decreases the number of observations finally to 1965.

[35] Table 16 in the appendix summarizes the main descriptive statistics for the baseline dataset.

decreases, particularly after controlling for the completeness of stages one and two in columns (6) and (7). The coefficient suggests that a reduction in ethnic fragmentation by a standard deviation is associated with a 2% decrease in the number of schools in regions in non-federal countries and a 6.7% decrease in regions that are a part of federal countries.

There are two possible reasons for the difference in effect magnitude between the coefficients of interest (the interaction effect) in columns (3) and (7). First, it is possible that the effect sizes shown in column (3) are overestimated if the degree of completeness is not considered, which could be the case if the degree of completeness is negatively affected by ethnic fractionalization and decentralization. The findings in section 2.4 indicate that this case is not true. These findings, however, rest on a dataset limited by the availability of official data on the number of amenities in subnational regions of different countries. Second, it is possible that the effect sizes shown in column (7) are underestimated, which might be the case since the proxy for the completeness of mapping of stages one and two implicitly relies on the assumption that areas that contain at least one amenity are representative of the region. Hence, the effect of ethnic fragmentation and decentralization on the number of amenities outside of these cells may not be accounted for. If this effect proceeds in the same direction as that in the representative cells, the total effect is underestimated. This interpretation is in line with the effect magnitude in column (5), which is somewhere between the estimates of columns (3) and (7). In column (5), only the completeness of the first stage of mapping is used as a control. The indicator of the first stage of mapping is essentially the share of populated cells that have at least basic OSM data. Hence, this indicator decreases the potential bias of systematic mapping that could inflate the estimates in column (3) without making the restrictive assumptions of the indicator for the completeness of stages one and two of mapping, which could downplay the effects in column (7).

The findings are subjected to a large a set of additional robustness tests.[36] It is possible that the indicator of decentralization also proxies for the level of general country development (correlation 0.29). In Table 18 in the appendix, in columns (1,3,5), the interaction between ethnic fractionalization and the log of the national GDP per capita is added to the main specification[37] without any changes to the main finding. Larger regions might have a greater likelihood of being an ethnically fractionalized regions (correlation 0.23). If so, the findings may reflect a simple size effect in regions that are a part of a federal state. In Table 18, in columns (2,4,6), the interaction between the federalism indicator and the log of area is added to the main specification without any change to the main findings. Capital regions might be special for various reasons. The estimates that include a dummy for capital regions or those that exclude capital regions confirm the

[36] Notably, some potential omitted variables are already addressed by the country fixed effect included in all estimates.
[37] Using the interaction of national GDP per capita not in logs does not change the result.

main findings. Excluding all regions with ethnic homelands where multiple ethnicities live or including those where more than 1% of the population lives in such homelands does not change the findings.

Table 5. Public amenities, decentralization and ethnic fragmentation

Dep. Var.:	(1) ln(#S)	(2) ln(#S.)	(3) ln(#S.)	(4) ln(#S.)	(5) ln(#S.)	(6) ln(#S.)	(7) ln(#S.)
ln(pop)	0.882***	0.874***	0.869***	0.926***	0.922***	0.837***	0.836***
	(0.026)	(0.027)	(0.027)	(0.025)	(0.025)	(0.020)	(0.020)
ln(area)	0.002	0.014	0.014	0.010	0.010	0.116***	0.116***
	(0.023)	(0.023)	(0.022)	(0.021)	(0.020)	(0.020)	(0.020)
Ethnic Frac.		-0.357**	-0.173	-0.335**	-0.180	-0.162**	-0.113
		(0.166)	(0.156)	(0.145)	(0.137)	(0.072)	(0.077)
Ethnic Frac. x Federal state			-1.182***		-0.997***		-0.318**
			(0.389)		(0.281)		(0.140)
$\ln(p^{I})$				0.920***	0.918***		
				(0.087)	(0.085)		
$\ln(p^{I+II})$						0.806***	0.805***
						(0.026)	(0.026)
Constant	-7.635***	-7.590***	-7.518***	-8.179***	-8.120***	-6.166***	-6.152***
	(0.399)	(0.402)	(0.399)	(0.352)	(0.350)	(0.218)	(0.221)
# Countries	155	155	155	155	155	155	155
# Regions	1,965	1,965	1,965	1,965	1,965	1,965	1,965
R-squared	0.841	0.842	0.843	0.876	0.877	0.957	0.957
Country FE	YES	YES	YES	YES	YES	YES	YES

Note: The unit of observation is the first-level administrative region. The dependent variable in columns (1)-(3) is the log of the number of schools reported in OSM, and the dependent variable in columns (4)-(7) is the log of the number of schools in OSM in active OSM areas. All estimates include country fixed effects that are not reported. ln(pop) and ln(area) are the log of regional population and land area, respectively. Ethnic Frac. is regional ethnic fragmentation biased on GREG and GHSL data. Federal state is a dummy for being a federal country, as defined by Treisman (2008). $\ln(p^{I})$ is the log of the proxy for OSM mapping completeness of stage one, and $\ln(p^{I+II})$ is the log of the proxy for completeness of stages one and two as defined in section 2.2. Standard errors are reported in parentheses and are clustered at the country level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

4.6. COLLECTIVE ACTION FAILURE OR IMPAIRED DEVELOPMENT

The available data do not allow the direct observation of the failing of collective actions; the data only document the outcomes of successful actions. Hence, to provide findings related to collective action failure with confidence, further robustness tests are needed to minimize the risk that other effects of ethnic fractionalization and decentralization on the supply of public amenities drive the results.

The most prominent alternative mechanism that could drive the results is that regional development can be affected by ethnic fractionalization and decentralization, which, in turn, can affect the capacity to finance public amenities. To determine whether the results are driven by a simple regional development-level effect, in Table 9, in columns (1, 3 and 5), the log of the average regional nightlights is added, and the interaction between the indicator of decentralization and nightlight intensity is added to the main specification in columns (2,4 and 6).[38] Controlling for regional economic development does not change the results. It seems that the effect of ethnic fractionalization does not arise from the indirect effect of ethnic fractionalization on development.[39] The relationship between nightlight and the number of schools is positive, in line with what some might expect, i.e., more prosperous regions can afford larger numbers of schools. However, the effect is only significant without the control for the completeness of stages one and two of mapping. The findings reported in section 2.4 indicate that the degree of completeness of the OSM data is positively associated with regional development, which might explain why the effect becomes insignificant after controlling for the degree of completeness of stages one and two as shown in columns (5 and 6). The effect in columns (1-4) might simply be attributed to the increases in the recording of schools associated with higher income levels.

To further examine whether the findings can be attributed to collective action failure, it is possible to perform a placebo test. Thus, the number of restaurants in a region is extracted from the OSM Project. Restaurants are amenities that are not provided by the government and should not be directly influenced by the political economy of regional government spending. Hence, the expectation is to observe no differences between the effect of ethnically fractionalization in decentralized and non-decentralized countries.

Table 7 presents the results of the bassline specification when using the log of the number of restaurants per region as the dependent variable. The effect of ethnic fragmentation is negative and significant only when the controls for the degree of completeness of the OSM data are omitted. If the control for regional development is included, this effect becomes insignificant. Most importantly, decentralization never has a significant impact on the effect of ethnic fragmentation on the number of restaurants in a region. Controlling for the regional level of development does not change this finding (see Table 19 in the appendix).

[38] The previous literature indicates that nightlight data are currently the most reliable globally available proxy for economic development (e.g., Henderson et.al. (2012), Lessmann & Seidel (2017) or Henderson et.al. (2018)). Lights are extracted from VIIRS global nightlight images from 2015, which is the latest year for which clean high-resolution images are available. The data are provided by Earth Observation Group at NOAA/NCEI.

[39] In fact, when estimating the effect of ethnic heterogeneity and its interaction with decentralization, the opposite effect is observed. Ethnic heterogeneity decreases growth less in regions that are part of a decentralized country; see Table 17 in the appendix.

Table 6. Public amenities, decentralization, ethnic fragmentation and regional development

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	ln(#S.)	ln(#S.)	ln(#S.)	ln(#S.)	ln(#S.)	ln(#S.)
ln(pop)	0.675***	0.678***	0.756***	0.759***	0.811***	0.815***
	(0.050)	(0.051)	(0.046)	(0.046)	(0.028)	(0.028)
ln(area)	0.208***	0.205***	0.174***	0.171***	0.141***	0.136***
	(0.041)	(0.042)	(0.036)	(0.036)	(0.032)	(0.032)
Ethnic Frac.	-0.066	0.148	-0.102	0.145	-0.101	0.254**
	(0.157)	(0.222)	(0.135)	(0.211)	(0.082)	(0.100)
Ethnic Frac.	-1.360***	-1.457***	-1.162***	-1.275***	-0.349**	-0.510***
x Federal state	(0.416)	(0.384)	(0.313)	(0.295)	(0.150)	(0.143)
ln(light)	0.218***	0.211***	0.195***	0.187***	0.030	0.018
	(0.042)	(0.044)	(0.039)	(0.041)	(0.029)	(0.031)
Ethnic Frac.		0.073		0.085		0.122***
x ln(light)		(0.073)		(0.064)		(0.037)
$\ln(p^{I})$			0.821***	0.820***		
			(0.089)	(0.089)		
$\ln(p^{I+II})$					0.798***	0.798***
					(0.028)	(0.028)
Constant	-6.359***	-6.377***	-7.142***	-7.163***	-6.013***	-6.042***
	(0.498)	(0.501)	(0.448)	(0.450)	(0.219)	(0.216)
# Countries	155	155	155	155	155	155
# Regions	1,965	1,965	1,965	1,965	1,965	1,965
R-squared	0.848	0.848	0.881	0.881	0.957	0.958
Country FE	YES	YES	YES	YES	YES	YES

Note: The unit of observation is the first-level administrative regions. The dependent variable in columns (1)-(3) is the log of the number of schools reported in OSM, and in columns (4)-(7), the dependent variable is the log of the number of schools in OSM in active OSM areas. All estimates include unreported country-fixed effects. ln(pop) and ln(area) are the log of regional population and land area, respectively. Ethnic Frac. is regional ethnic fragmentation biased on GREG and GHSL data. Federal state is a dummy for being a federal country, as defined by Treisman (2008). ln(light) is the log of average nighttime light intensity extracted from the VIRS image of 2016. $\ln(p^{I})$ is the log of the proxy for OSM mapping completeness of stage one, and $\ln(p^{I+II})$ is the log of the proxy for completeness of stages one and two as defined in section 2.2. Standard errors are reported in parentheses and are clustered at the country level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 7. Non-public amenities, decentralization and ethnic fragmentation: A placebo test

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dep. Var.:	ln(#R.)	ln(#R.)	ln(#R.)	ln(#R.)	ln(#R.)	ln(#R.)	ln(#R.)
ln(pop)	0.947***	0.940***	0.940***	1.017***	1.017***	0.899***	0.899***
	(0.055)	(0.055)	(0.055)	(0.035)	(0.035)	(0.026)	(0.026)
ln(area)	-0.179***	-0.168***	-0.168***	-0.100***	-0.100***	0.098***	0.098***
	(0.035)	(0.035)	(0.035)	(0.028)	(0.028)	(0.026)	(0.026)
Ethnic Frac.		-0.292*	-0.281*	-0.156	-0.148	-0.115	-0.119
		(0.151)	(0.160)	(0.148)	(0.154)	(0.091)	(0.099)
Ethnic Frac.			-0.077		-0.055		0.029
x Federal state			(0.478)		(0.470)		(0.266)
$\ln(p^{I})$				0.809***	0.810***		
				(0.135)	(0.135)		
$\ln(p^{I+II})$						0.911***	0.911***
						(0.036)	(0.036)
Constant	-7.234***	-7.204***	-7.201***	-8.646***	-8.643***	-6.065***	-6.066***
	(0.851)	(0.851)	(0.852)	(0.452)	(0.453)	(0.314)	(0.314)
# Countries	155	155	155	155	155	155	155
# Regions	1,694	1,694	1,694	1,638	1,638	1,635	1,635
R-squared	0.809	0.809	0.809	0.833	0.833	0.941	0.941
Country FE	YES	YES	YES	YES	YES	YES	YES

Note: The unit of observation is the first-level administrative region. The dependent variable in columns (1)-(3) is the log of the number of restaurants reported in OSM, and in columns (4)-(7), the dependent variable is the log of the number of restaurants in OSM in active OSM areas. All estimates include unreported country-fixed effects. ln(pop) and ln(area) are the log of regional population and land area, respectively. Ethnic Frac. is regional ethnic fragmentation biased on GREG and GHSL data. Federal state is a dummy for being a federal country, as defined by Treisman (2008). $\ln(p^{I})$ is the log of the proxy for OSM mapping completeness of stage one, and $\ln(p^{I+II})$ is the log of the proxy for completeness of stage ones and two as defined in section 2.2. Standard errors are reported in parentheses and are clustered at the country level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

4.7. A UNIVERSAL EFFECT ON PUBLIC AMENITIES

The collective action failure associated with social heterogeneity is suspected to be more relevant for specific types of public goods. The theoretical argument presented by Alesina et. al. (1999) indicates that the supply of productive public goods is mainly diminished by social heterogeneity. To determine whether this argument remains true from a global perspective and whether the findings can be extended to a broader set of public amenities, the number of other public amenities that are a part of the new dataset are examined.

An alternative measure of educational spending that can, by the definition of Alesina et. al. (1999), be classified as a productive public good is the number of libraries within a region. Columns (1, 4 and 7) in Table 8 report the findings of the main specification using the log of the number of libraries as the dependent variable. Notably, the proxy for the completeness of stages one and two when referring to $\ln(p^{I+II})$ in Table 8 is amenity specific. Similar to the school analysis, the negative effect of ethnic fractionalization mainly occurs in regions that are a part of a federal country. However, the effect is only significant after controlling for mapping completeness in a region. The effect is also significant when using the raw data after controlling for regional development.

The number of hospitals in a region can be interpreted as a proxy for health care spending. Obviously, this measure is not without problems since hospitals in many countries are at least partly private. In many countries, governments nevertheless subsidize hospitals for their provision of ambulance services with the aim of securing a country-wide emergency health care provision. Given this issue, Alesina et. al. (1999) was not completely clear on whether spending on hospitals is a productive public good. Their empirical findings on the link between health care spending and ethnic fragmentation were mixed. However, from a global perspective, the results are less mixed (see Table 8 columns 2, 5, 8). Fractionalization has a significant, negative effect on the number of hospitals in a region. The effect is larger in regions that are part of decartelized countries. The effect is significant even when only utilizing the raw data.

Public safety is an alternative public good, the provision of which might be affected by social heterogeneity and decentralization. Spending on law and order should be positively associated with the number of police stations in regions. The argument here is that a higher police station density decreases response times. Alesina et. al. (1999) argued that, in contrast to educational spending, the effect of social heterogeneity on spending on public safety is theoretically ambiguous. Their empirical results are, if significant, positive. In contrast, the number of police stations is significantly smaller in ethnic fractionalization regions worldwide. This effect, however, is not significantly different in regions that are part of federal countries (see Table 8, columns 3, 6 and 9). Hence the effect is most likely not associated with a collective action failure triggered by social heterogeneity among local policy makers.

Table 8. Public amenities, decentralization and ethnic fragmentation: Alternative output measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dep. Var.:	ln(#L.)	ln(#H.)	ln(#P.)	ln(#L.)	ln(#H.)	ln(#P.)	ln(#L.)	ln(#H.)	ln(#P.)
ln(pop)	0.706***	0.753***	0.645***	0.782***	0.845***	0.729***	0.749***	0.811***	0.745***
	(0.060)	(0.043)	(0.047)	(0.029)	(0.025)	(0.025)	(0.024)	(0.018)	(0.020)
ln(area)	-0.053*	-0.019	-0.013	-0.033	-0.001	-0.019	0.133***	0.105***	0.111***
	(0.031)	(0.028)	(0.027)	(0.025)	(0.024)	(0.026)	(0.025)	(0.019)	(0.023)
Ethnic Frac.	-0.397**	-0.301**	-0.514***	-0.267	-0.303**	-0.453***	-0.113	-0.175**	-0.288***
	(0.175)	(0.151)	(0.123)	(0.195)	(0.125)	(0.112)	(0.111)	(0.077)	(0.077)
Ethnic Frac.	-0.961	-1.100**	-0.520	-1.047*	-0.657*	-0.424	-0.452*	-0.345*	-0.201
x Federal state	(0.620)	(0.491)	(0.366)	(0.566)	(0.392)	(0.385)	(0.272)	(0.177)	(0.209)
$\ln(p^{I})$				0.416***	0.615***	0.656***			
				(0.120)	(0.065)	(0.092)			
$\ln(p^{I+II})$							0.687***	0.683***	0.690***
							(0.023)	(0.026)	(0.024)
Constant	-7.006***	-7.302***	-6.086***	-8.140***	-8.573***	-7.015***	-6.202***	-6.616***	-5.823***
	(0.961)	(0.687)	(0.722)	(0.450)	(0.375)	(0.345)	(0.283)	(0.225)	(0.233)
Observations	1,383	1,967	1,866	1,339	1,900	1,767	1,331	1,897	1,758
R-squared	0.872	0.826	0.819	0.885	0.859	0.857	0.960	0.943	0.947
Country FE	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: The unit of observation is the first-level administrative regions. The dependent variable in columns (1), (4) and (7) is the log of the number of libraries reported in OSM; in columns (2), (5) and (8), the dependent variable is the log number of hospitals; and, in columns (3), (6) and (9), the dependent variable is the log number of police stations. In columns (1)-(3) amenity number refers to total observations and, in columns (4)-(9), to observations within active OSM areas. All of the estimates include country fixed effects that are not reported. ln(pop) and ln(area) are the log of regional population and land area, respectively. Ethnic Frac. is regional ethnic fragmentation biased on GREG and GHSL data. Federal state is a dummy for being a federal country, as defined by Treisman (2008). $\ln(p^{I})$ is the log of the proxy for OSM mapping completeness of stage one, and $\ln(p^{I+II})$ is the log of the amenity specific proxy for completeness of stages one and two as defined in section 2.2. Standard errors are reported in parentheses and are clustered at the country level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

4.8. ALTERNATIVE MEASURES OF THE DETERMINANTS

ALTERNATIVE MEASURES OF DECENTRALIZATION

To ensure that the results are not driven by the specifics of the measure of decentralization used, the findings are replicated using alternative indicators. The federalism indicator used in the main analysis is well established in the literature but is not available for all countries. Utilizing the CIA World Fact Book allows the construction of a new alternative indicator that covers 199 counties, e.g., almost all countries worldwide.[40] Adding the additional countries to the estimates does not change the findings, and the effect size remains approximately the same as that in the baseline specification (Table 9, columns (1-3)).

For at least a subset of countries, it is possible to derive de facto measures of fiscal decentralization using the IMF government finance statistics. The number of observations is maximized by focusing on the share of transfers and using average data from 1990 to 2018. Considering past spending abilities also seems plausible since the construction of public amenities usually requires time. Hence, it is not likely that changes in current local sovereignty regarding spending have an immediate effect on the existence of publicly financed amenities, such as schools. Table 9, column (4-6), presents the estimation results when interacting the share of transfers with the degree of regional ethnic fractionalization. First, notably, using the IMF data drastically reduces the sample to almost half of its original size despise all efforts to maximize the number of observations. Second, the coefficient of ethnic fragmentation enters negatively. Third, consistent with the idea that higher shares of transfers reflect decreasing local autonomy, the interaction effect with ethnic fractionalization is positive. The effect is significant after controlling for the degree of regional completeness. The marginal effect plots indicate that ethnic fragmentation has no effect on the number of schools per region if the share of transfers is greater than 30% after controlling for $\ln(p^{I})$ (Figure 5, left) and 50% after controlling for $\ln(p^{I+II})$ (Figure 5, right); otherwise, the effect is negative. Hence, ethnic fragmentation negatively affects the supply of public amenities in regions that are more financially independent.

[40] The indicator is equal to one if the government type description contains the word “federal” in the CIA World Fact Book and zero otherwise.

Figure 5 Marginal effect of ethnic fragmentation (90% confidence interval)

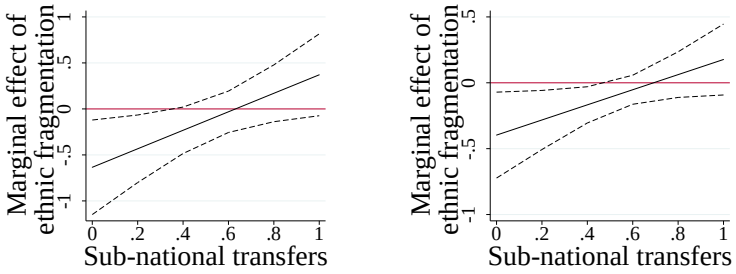


Table 9. Public amenities, decentralization and ethnic fragmentation: Alternative decentralization measures

Dep. Var.:	(1) ln(#S.)	(2) ln(#S.)	(3) ln(#S.)	(4) ln(#S.)	(5) ln(#S.)	(6) ln(#S.)
ln(pop)	0.852***	0.904***	0.820***	0.901***	0.950***	0.830***
	(0.025)	(0.024)	(0.020)	(0.023)	(0.023)	(0.025)
ln(area)	0.013	0.009	0.115***	0.041	0.019	0.121***
	(0.022)	(0.020)	(0.020)	(0.026)	(0.025)	(0.025)
Ethnic Frac.	-0.175	-0.172	-0.101	-0.726*	-0.634**	-0.397**
	(0.151)	(0.133)	(0.075)	(0.424)	(0.309)	(0.196)
Ethnic Frac.	-1.193***	-1.042***	-0.361**			
x Federal state CIA	(0.384)	(0.274)	(0.140)			
Ethnic Frac.				0.962	1.005*	0.574*
x Subn. trans. 90-18				(0.689)	(0.512)	(0.334)
$\ln(p^{I})$		0.911***			0.961***	
		(0.080)			(0.109)	
$\ln(p^{I+II})$			0.798***			0.813***
			(0.026)			(0.040)
Constant	-7.208***	-7.795***	-5.902***	-7.806***	-8.274***	-5.919***
	(0.365)	(0.323)	(0.211)	(0.368)	(0.349)	(0.284)
# Countries	197	197	197	88	88	88
# Regions	2,222	2,222	2,222	1,316	1,316	1,316
R-squared	0.861	0.888	0.959	0.864	0.892	0.960
Country FE	YES	YES	YES	YES	YES	YES

Note: The unit of observation is the first-level administrative region. The dependent variable in columns (1)-(3) is the log of the number of schools reported in OSM, and in columns (4)-(7), the dependent variable is the log of the number of schools in OSM in active OSM areas. All estimates include unreported country-fixed effects. ln(pop) and ln(area) are the log of regional population and land area, respectively. Ethnic Frac. is regional ethnic fragmentation biased on GREG and GHSL data. Federal state CIA is a dummy for being a federal country, as defined by the CIA world fact book 2018. Subn. trans. 90-18 is the mean of subnational transfers between 1990 and 2018 reported by the IMF Government Financial Statistics. $\ln(p^{I})$ is the log of the proxy for OSM mapping completeness of stage one, and $\ln(p^{I+II})$ is the log of the proxy for completeness of stages one and two as defined in section 2.2. Standard errors are reported in parentheses and are clustered at the country level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

ALTERNATIVE MEASURES OF SOCIAL HETEROGENEITY

Finally, the question of whether the findings depend on the measure of social heterogeneity used remains. Table 10 reports the baseline estimates analogous to Table 5 using ethnic polarization as an indicator of social heterogeneity. Comparing both sets of results reveals very few differences. The results based on ethnic polarization are slightly weaker, and the coefficients are a bit smaller, but otherwise, the results are very similar.

Table 10. Public amenities, decentralization and ethnic polarization

Dep. Var.:	(1) ln(#S.)	(2) ln(#S.)	(3) ln(#S.)	(4) ln(#S.)	(5) ln(#S.)	(6) ln(#S.)	(7) ln(#S.)
ln(pop)	0.882***	0.875***	0.869***	0.927***	0.921***	0.837***	0.836***
	(0.026)	(0.027)	(0.027)	(0.026)	(0.026)	(0.020)	(0.020)
ln(area)	0.002	0.012	0.012	0.008	0.008	0.116***	0.116***
	(0.023)	(0.023)	(0.022)	(0.021)	(0.020)	(0.020)	(0.020)
Ethnic Pola.		-0.178*	-0.064	-0.169*	-0.070	-0.096**	-0.067
		(0.101)	(0.092)	(0.092)	(0.084)	(0.044)	(0.046)
Ethnic Pola. x Federal state			-0.753***		-0.655***		-0.190*
			(0.248)		(0.189)		(0.100)
$\ln(p^{I})$				0.923***	0.923***		
				(0.088)	(0.086)		
$\ln(p^{I+II})$						0.807***	0.805***
						(0.026)	(0.026)
Constant	-7.635***	-7.589***	-7.499***	-8.176***	-8.097***	-6.161***	-6.143***
	(0.399)	(0.404)	(0.404)	(0.355)	(0.356)	(0.219)	(0.222)
# Countries							
# Regions	1,965	1,965	1,965	1,965	1,965	1,965	1,965
R-squared	0.841	0.841	0.843	0.876	0.877	0.957	0.957
Country FE	YES	YES	YES	YES	YES	YES	YES

Note: The unit of observation is the first-level administrative region. The dependent variable in columns (1)-(3) is the log of the number of schools reported in OSM, and in columns (4)-(7), the dependent variable is the log of the number of schools in OSM in active OSM areas. All estimates include unreported country-fixed effects. ln(pop) and ln(area) are the log of regional population and land area, respectively. Ethnic Frac. is regional ethnic polarization biased on GREG and GHSL data. Federal state is a dummy for being a federal country, as defined by Treisman (2008). $\ln(p^{I})$ is the log of the proxy for OSM mapping completeness of stage one, and $\ln(p^{I+II})$ is the log of the proxy for completeness of stages one and two as defined in section 2.2. Standard errors are reported in parentheses and are clustered at the country level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

5. CONCLUSION

This paper provides a first global view of the effects of decentralization and social heterogeneity on the provision of regional public goods. The estimates indicate that increasing local autonomy hampers the provision of public goods in regions that face high levels of social heterogeneity. This finding is in line with the theory of collective action failure and social heterogeneity. The effect is also sizable since it implies that an increase in ethnic fractionalization by a standard deviation decreases the supply of schools in a region by 7% to 14% if the region is a part of a federal country.

The analysis is based on a new dataset that I derived from the OSM project, which contains the global locations of various public amenities associated with public goods that are typically provided to a large extent by the state. Well-known accuracy problems associated with using volunteered geocode data are addressed by developing a new method that accounts for the completeness of the data within first-level subnational regions by cross-referencing of OSM settlement indicators with indicators derived from satellite data. The new approach minimizes the risk of potential biases due to omitted variables creating systematic missing data in the OSM data. The quality of the approach is tested by correcting the OSM raw data and comparing the corrected data with official data of a subset of countries for which such data exist. The observed correlation between the corrected OSM data and the official data is typically greater than 90%. The main findings of the paper also hold when using only the raw OSM data and when different technical details of the algorithms used to clean the raw data or account for the possibility of systematically missing data are altered.

The findings are robust to a large set of robustness tests based on a large set of controls and alternative indicators of public goods, social heterogeneity and decentralization. The placebo test indicates that the supply of regional non-public amenities, such as restaurants, is not affected by the joint effect of social heterogeneity and decentralization. Examining the data shows no indication that regional social heterogeneity might be the driver of decentralization or that the provision of public goods might induce social heterogeneity or decentralization; hence, it is likely that the findings document the causal effect of social heterogeneity and decentralization on the provision of regional public goods.

The findings shed light on a further dark side of decentralization, which has received limited attention to date. Increasing local autonomy might increase, on average, the effectiveness of government spending within the regions of a country. However, in some cases, the opposite might be the case since power is given to a layer of government that is too socially heterogeneous to execute collective actions. This finding might explain how decentralization can lead to increases in regional disparities (e.g., Rodríguez-Pose & Ezcurra (2009) or Lessmann (2012)). One possible conclusion that some might draw from this finding is that decentralization should be accompanied by administrative reforms that decrease social heterogeneity

within regions. However, such a policy might increase separatist tendencies and hence should be further studied before being enacted.

The dataset generated for this study, along with the proposed approach to account for missing OSM data, offers a variety of opportunities for possible further research. For example, studies could examine other aspects of the political economy driving the provision of public goods via public amenities. Some might examine favoritism and whether political leaders use the provision of public goods to pamper their favorite regions. Such examinations might, for example, help to understand the mechanism underlying the existing finding that favoritism impacts growth (Hodler & Raschky, 2014; De Luca, Hodler, Raschky, & Valsecchi, 2018). These questions remain open for further research since addressing these issues is beyond the scope of this paper.

REFERENCES

Ahn, T. K., Ostrom, E., & Walker, J. M. (2003). Heterogeneous preferences and collective action. *Public choice, 117*(3-4), 295-314.

Alesina, A., & La Ferrara, E. (2000). Participation in heterogeneous communities. *The quarterly journal of economics, 115*(3), 847-904.

Alesina, A., & Zhuravskaya, E. (2011). Segregation and the Quality of Government in a Cross Section of Countries. *The American Economic Review, 101*(5), 1872-1911.

Alesina, A., Baqir, R., & Easterly, W. (1999). Public Goods and Ethnic Divisions. *The Quarterly Journal of Economics, 4*(1), 1243–1284.

Alesina, A., Gennaioli, C., & Lovo, S. (2017). Public Goods and Ethnic Diversity: Evidence from Deforestation in Indonesia. *NBER Working Paper, No. w20504*.

Baqir, R. (2002). Social Sector Spending in a Panel of Countries. *Workingpaper International Monetary Fund., 2*(35).

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature, 442*(7105), 912-915.

Blanes i Vidal, J., & Kirchmaier, T. (2018). The Effect of Police Response Time on Crime Clearance Rates. *The Review of Economic Studies, 85*(2), 855–891.

Buchmueller, T. C., Jacobson, M., & Wold, C. (2006). How far to the hospital?: The effect of hospital closures on access to care. *Journal of health economics, 25*(4), 740-761.

Burde, D., & Linden., L. L. (2013). Bringing Education to Afghan Girls: A Randomized ControlledTrial of Village-Based Schools. *American Economic Journal: Applied Economics, 5*(3), 27–40.

Cinnirella, F., & Schueler, R. (2016). The cost of decentralization: Linguistic polarization and the provision of education. *CESifo Working Paper, No. 5894*.

Dayton-Johnson, J. (2000). Determinants of collective action on the local commons: a model with evidence from Mexico. *Journal of Development Economics, 62*(1), 181-208.

De Luca, G., Hodler, R., Raschky, P. A., & Valsecchi, M. (2018). Ethnic favoritism: An axiom of politics? *Journal of Development Economics, 132*, 115-129.

Díaz-Cayeros, A., Magaloni, B., & Ruiz-Euler, A. (2014). Traditional governance, citizen engagement, and local public goods: evidence from Mexico. *World Development, 53*, 80-93.

Dražanová, L. (2019). Historical Index of Ethnic Fractionalization Dataset (HIEF). *Harvard Dataverse, V1.0*. doi:10.7910/DVN/4JQRCL

Duflo, E. (2001). Schooling and Labor Market Consequences of School Construction in Indonesia:Evidence from an Unusual Policy Experiment. *American Economic Review, 91*(4), 795–813.

Fan, C., Lin, C., & Treisman, D. (2009). Political decentralization and corruption: Evidence from around the world. *Journal of Public Economics, 93*(1-2), pp. 14-34.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *The Quarterly Journal of Economics, 115*(3), 811-846.

Glennerster, R., Miguel, E., & Rothenberg, A. D. (2013). Collective action in diverse Sierra Leone communities. *The Economic Journal, 123*(568), 285-316.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal, 69*(4), 211-221.

Habyarimana, J., Humphreys, M., Posner, D. N., & Weinstein, J. M. (2007). Why does ethnic diversity undermine public goods provision? *American Political Science Review, 101*(04), 709-725.

Henderson, J. V., Squires, T. L., Storeygard, A., & Weil, D. N. (2018). The global spatial 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.

Hodler, R., & Raschky, P. A. (2014). Regional favoritism. *The Quarterly Journal of Economics, 129*(2), 995-1033.

Kazianga, H., Levy, D., Linden, L. L., & Sloan., M. (2013). The Effects of‘ Girl-Friendly’ Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso. *American Economic Journal: Applied Economics, 5*(3), 41–62.

Khwaja, A. I. (2009). Can good projects succeed in bad communities? *Journal of public Economics, 93*(7), 899-916.

Lessmann, C. (2009). Fiscal Decentralization and Regional Disparity: Evidence from Cross-Section and Panel Data. *Environment and Planning A: Economy and Space, 41*(10), 2455–2473.

Lessmann, C. (2012). Regional inequality and decentralization: an empirical analysis. *Environment and Planning A, 44*(6), 1363-1388.

Lessmann, C., & Markwardt, G. (2010). One size fits all? Decentralization, corruption, and the monitoring of bureaucrats. *World Development, 38*(4), 631-646.

Lessmann, C., & Seidel, A. (2017). Regional inequality, convergence, and its determinants – a view from outer space. *European Economic Review, 92*(1), 110-132.

Martinez-Vazquez, J., Lago-Peñas, S., & Sacchi, A. (2017). The impact of fiscal decentralization: A survey. *Journal of Economic Surveys, 31*(4), 1095-1129.

Miguel, E. (2004). Tribe or nation? Nation building and public goods in Kenya versus Tanzania. *World politics, 56*(3), 327-362.

Miguel, E., & Gugerty, M. K. (2005). Ethnic diversity, social sanctions, and public goods in Kenya. *Journal of public Economics, 89*(11), 2325-2368.

Montalvo, J. G., & Reynal-Querol, M. (2005). Ethnic Polarization, Potential Conflict, and Civil Wars. *American Economic Review, 95*(3), 796-816.

Muralidharan, K., & Prakash, N. (2017). Cycling to school: increasing secondary school enrollment for girls in India. *American Economic Journal: Applied Economics, 9*(3), 321-50.

Musgrave, R. (1959). *The Theory of Public Finance: A Study in Public Economy*. New York: McGraw-Hill.

Neyapti, B. (2006). Revenue decentralization and income distribution. *Economics Letters, 92*, 409–166.

Oates, W. (1972). *Fiscal Federalism.* New York:: Harcourt Brace Jovanovich.

OECD. (2019). Current trends in decentralisation. In *Making Decentralisation Work: A Handbook for Policy-Makers.* Paris: OECD Publishing. doi:10.1787/f9117991-en.

Rodríguez-Pose, A., & Ezcurra, R. (2009). Does decentralization matter for regional disparities? A cross-country analysis. *Journal of Economic Geography, 10*(5), 619-644.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (2017). A review of volunteered geographic information quality assessment methods. International Journal of Geographical Information Science, 31(1), 139-167. *International Journal of Geographical Information Science, 31*(1), 139-167.

Tiebout, C. (1956). A pure theory of local expenditures. *The Journal of Political Economy, 64*, 416–424.

Treisman, D. (2008). *Decentralization Dataset.* Retrieved from https://www.danieltreisman.org/workinprogress/

Weidmann, N. B., Jan Ketil Rød, & Cederman., L.-E. (2010). Representing Ethnic Groups in Space: A New Dataset. *Journal of Peace Research, 47*(4), 491–99.

Wilde, E. T. (2013). Do emergency medical system response times matter for health outcomes? *Health economics, 22*(7), 790-806.

6. APPENDICES

6.1. APPENDIX TO THE PUBLIC AMENITY DATASET

DATA SOURCE

The bulk of OSM data come from the planet/continent dumps provident by Geofabrike (http://download.geofabrik.de/). The QGIS OSM data converter cannot handle multipolygon relationships; these observations are manually obtained from the Web-based Overpass API (http://overpass-turbo.eu/).

DATA CLEANING

The following steps were undertaken to clean these raw data.

1. After downloading the OSM data, a SpatiaLite Link to the OSM data is created. The QGIS importer for SpatiaLite data was used to create point and polygon layers that only contained objects with the keys/tags that were later used to identify different amenities, e.g., amenities and buildings. For details, see step 3.
2. For ease of calculation, all objects with empty tags were deleted. The following query was used: “amenity” IS NULL AND “building” IS NULL AND “religion” IS NULL AND “denomination” IS NULL.
3. In the next step, specific amenity shape files were created from the main files. The following query was used to extract all of the libraries ("amenity" IS 'library' OR "building" IS 'library'). The queries were designed to account for the problem that not all tags were always in the field as required by the tagging guidelines of OSM. Staying with the previous example, in some cases, a building was tagged as building=library, while the value of amenity was null and vice versa. Table 11 in the appendix summarizes all of the tags used.
4. The polygon layers were converted into point layers by calculating the centroids of the polygons.
5. In the final cleaning step, all point layers from the OSM Geofabric dump and those obtained from the Overpass API (multipolygon) were merged.

6. When an amenity consisted of multiple buildings and was not labelled a multipolygon, the presence of amenities may be overestimated, e.g., a hospital complex with two buildings may be counted as two hospitals. To account for this issue, close-by observations were merged.

Table 11. Tags used to identify public amenities within the OSM data

Amenity	OSM Tags
Kindergarten	amenity=kindergarten or building=kindergarten
School	amenity=school or building=school
College	amenity=college or building=college
University	amenity=university or building=university
Library	amenity=library or building=library
Police station	amenity=police or building=police
Prison	amenity=prison or building=prison
Hospital	amenity=hospital or building=hospital or clinic=hospital or building=hospital
Restaurant	amenity=restaurant or building=restaurant
Road	highway=residential

6.2. SUPPLEMENTARY FIGURES AND TABLES

Table 12. Data sources

Variables	Description and data source
#S, (#L., #H., #P., #R.)	The number schools (libraries, hospitals, police stations and restaurants) in first-level administrative regions reported in OpenStreetMap (OSM) by the end of 2017. Depending on specification, the number refers either to total observations or observations within active OSM areas (areas with more than 100 residents, urban buildup and residential roads in OSM) **Source:** OSM data are from the planet/continent dumps provided by Geofabrike and Overpass API. Settlement indicators are extracted on a one-km^2 grid from the Global Human Settlement Layer (GHSL) from 2015. Boundary data of first level administrative regions are taken from GADM
p^I	Proxy for the completeness of the first stage of OSM mapping **Source:** Own calculations biased on OSM and GHSL; for details, see section 2.2.
p^{I+II}	Proxy for the completeness of the first and second stages of OSM mapping **Source:** Own calculations based on OSM and GHSL data; for details, see section 2.2.
Ethnic Frac.	Regional ethnic fragmentation **Source:** GHSL and GREG provided by Weidmann et al. (2010)
Ethnic Pola.	Regional ethnic polarization **Source:** GHSL and GREG provided by Weidmann et al. (2010)
Federal state	Dummy for being a federal country, **Source:** Treisman (2008)
Federal state CIA	Dummy for being a federal country **Source:** CIA World Fact Book 2018
Subn. trans. 90-18	Mean of subnational transfers between 1990 and 2018 **Source:** IMF Government Financial Statistics
pop	Population **Source:** GHSL 2015
area	Area **Source:** GADM
light	Average nighttime light intensity **Source:** VIIRS global nightlight images from 2015
Educational expenditure	Total educational expenditures **Source:** Annual survey of school system finances 2015
#School official	Official number schools in first-level administrative regions reported **Source:** USA: National Center for Education Statistics Common Core database 2012 via SABINS; Malaysia: Government statistics 2017 retrieved from https://www.moe.gov.my/en/statistik-menu; Mexico: INEGI-SEP. Censo de Escuelas, Maestros y Alumnos de Educación Básica y Especial, CEMABE 2013; South Africa: EMIS Program 2016; Democratic Republic of the Congo: Ministry of Education 2014 via Education Policy and Data Center (EPDC); Namibia: Fifteenth School Day Report for 2017

Table 13. Number of schools per 1000 citizens in raw OSM data 2017 by country income level

Income Level	Observations	Mean	Std. Dev.	Min	Max
Low	608	0.128	0.209	0	1.312
Middle	837	0.174	0.320	0	3.424
High	1,869	0.328	1.176	0	45.122

Note: The definition of low-, middle- and high-income countries follows the World Bank definition for 2015, where LIC: $y_j < 4.086\ US\$pc$; MIC: $4.086\ US\$pc \leq y_j < 12.615 US\$pc$; HIC: $y_j \geq 12.615\ US\$pc$

Table 14. Number of schools per 1000 citizens in capital regions of countries in raw OSM data 2017

ISO	S.p.c.	ISO	S.p.c.	ISO	S.p.c.	ISO	S.p.c.	ISO	S.p.c.	ISO	S.p.c.
AFG	0.00	CHL	0.26	GRD	0.30	MDG	0.02	PHL	0.08	SWE	0.39
ALB	0.15	CHN	0.02	GUM	0.85	MWI	0.02	POL	0.26	CHE	0.67
DZA	0.19	COL	0.09	GTM	0.07	MYS	0.09	PRT	0.24	SYR	0.06
ASM	0.50	CRI	0.26	GIN	0.04	MLI	0.14	PRI	0.43	TJK	0.04
AGO	0.02	CIV	0.03	GNB	0.01	MRT	0.18	QAT	0.14	TZA	0.09
ATG	0.41	HRV	0.20	GUY	0.29	MUS	0.21	COG	0.01	THA	0.03
ARG	0.26	CUB	0.18	HTI	0.22	MEX	0.06	ROU	0.14	TGO	0.08
ARM	0.18	CYP	0.30	HND	0.09	FSM	0.24	RUS	0.14	TON	0.94
AUS	0.39	CZE	0.24	HKG	0.06	MDA	0.22	RWA	0.03	TTO	0.23
AUT	0.23	COD	0.04	HUN	0.22	MNG	0.10	KNA	0.42	TUN	0.12
AZE	0.16	DNK	0.26	ISL	0.47	MNE	0.14	LCA	0.57	TUR	0.14
BGD	0.01	DJI	0.04	IND	0.03	MAR	0.05	VCT	0.40	TKM	0.11
BRB	0.38	DMA	0.51	IDN	0.18	MOZ	0.03	WSM	0.07	TCA	0.00
BLR	0.21	DOM	0.14	IRN	0.05	MMR	0.05	SMR	0.00	UGA	0.16
BLZ	0.63	ECU	0.20	IRQ	0.08	NAM	0.16	STP	0.23	UKR	0.14
BEN	0.16	EGY	0.01	IRL	0.38	NPL	0.49	SAU	0.04	ARE	0.08
BTN	0.30	SLV	0.08	ITA	0.18	NLD	0.24	SEN	0.11	GBR	0.42
BOL	0.20	GNQ	0.02	JAM	0.25	NCL	0.57	SRB	0.15	USA	0.52
BIH	0.17	ERI	0.01	JPN	0.17	NZL	0.48	SLE	0.16	URY	0.14
BWA	0.18	EST	0.22	JOR	0.02	NIC	0.16	SVK	0.39	UZB	0.11
BRA	0.25	ETH	0.05	KAZ	0.16	NER	0.11	SVN	0.31	VUT	0.38
BRN	0.33	FRO	0.50	KEN	0.05	NGA	0.00	SLB	0.05	VEN	0.04
BGR	0.20	FIN	0.45	KGZ	0.22	PRK	0.01	SOM	0.01	VNM	0.01
BFA	0.15	FRA	0.31	LAO	0.15	MNP	0.40	ZAF	0.06	VIR	0.33
BDI	0.09	GAB	0.09	LVA	0.32	NOR	0.28	KOR	0.10	YEM	0.03
KHM	0.12	GMB	0.04	LSO	0.19	OMN	0.08	SSD	0.01	ZMB	0.07
CMR	0.14	GEO	0.15	LBR	0.03	PAK	0.09	ESP	0.29	ZWE	0.07
CAN	0.37	DEU	0.27	LBY	0.19	PLW	0.49	LKA	0.13		
CPV	0.93	GHA	0.07	LIE	1.18	PAN	0.08	SDN	0.01		
CAF	0.11	GRC	0.20	LTU	0.24	PRY	0.35	SUR	0.14		
TCD	0.04	GRL	1.06	MKD	0.18	PER	0.14	SWZ	0.03		

Table 15. Determinates of the degree of completeness: decentralization and ethnic fragmentation

	(1)	(2)	(3)	(4)	(5)	(6)
Depen.Var.	log(#School OSM / #School official)					
ln(light)	0.147**	0.146**	0.083*	0.083*	0.044	0.044
	(0.042)	(0.040)	(0.037)	(0.037)	(0.041)	(0.041)
Ethnic Frac.	-0.998	-0.378	-0.633	-0.352	0.033	0.061
	(0.572)	(0.430)	(0.411)	(0.555)	(0.123)	(0.213)
Ethnic Frac.		-1.100		-0.508		-0.052
x Federal state		(0.710)		(0.726)		(0.290)
$\ln(p^{I})$			1.625***	1.603***		
			(0.221)	(0.210)		
$\ln(p^{I+II})$					0.859***	0.858***
					(0.074)	(0.077)
Constant	-1.304***	-1.315***	-0.832***	-0.843***	0.566**	0.563**
	(0.073)	(0.053)	(0.110)	(0.099)	(0.148)	(0.158)
Observations	124	124	124	124	124	124
R-squared	0.786	0.790	0.859	0.860	0.960	0.960
Country FE	YES	YES	YES	YES	YES	YES

Note: The unit of observation is the first level administrative region. The dependent variable in all estimates is the log of the number of schools in OSM reported in active OSM areas in 2017 divided by the number of schools reported in official statistics (source years vary between 2012 and 2017). All of the estimates include country fixed effects that are not reported. ln(light) is the log of average nighttime light intensity extracted from the VIRS image of 2016. Ethnic Frac. is regional ethnic fragmentation biased on GREG and GHSL data. Federal state is a dummy for being a federal country, as defined by Treisman (2008). $\ln(p^{I})$ is the log of the proxy for OSM mapping completeness of stage one, and $\ln(p^{I+II})$ is the log of the proxy for completeness of stages one and two as defined in section 2.2. Standard errors are reported in parentheses and are clustered at the country level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 16. Descriptive statistics on the main dataset in section 4

Variable	Obs	Mean	Std. Dev.	Min	Max
#School OSM raw	1,965	276.759	883.876	1	22470
#School OSM	1,965	211.779	727.401	1	21073
Ethnic Frac.	1,965	0.114	0.191	0	0.831
Federal state	1,965	0.152	0.359	0	1
$\ln(p^{I})$	1,965	0.690	0.263	0.019	1
$\ln(p^{I+II})$	1,965	0.117	0.122	0.000	1

Note: Descriptive statistics refer to the dataset used in the main estimations in section 4. The dataset is limited by the availability of the main indicators of decentralization and ethnic fragmentation. #School OSM raw refers to the raw number of schools in the OSM data, and #School OSM refers to the number of schools in active OSM areas.

Table 17. Regional development, decentralization and ethnic fragmentation

Dependent Variable	(1) ln(light)	(2) ln(light)	(3) ln(light)	(4) ln(light)	(5) ln(light)
ln(pop)	0.893***	0.885***	0.889***	0.890***	0.898***
	(0.043)	(0.044)	(0.044)	(0.044)	(0.030)
ln(area)	-0.903***	-0.891***	-0.891***	-0.893***	-0.894***
	(0.025)	(0.025)	(0.025)	(0.025)	(0.030)
Ethnic Frac.		-0.365**	-0.492***		-0.851***
		(0.143)	(0.140)		(0.183)
Ethnic Frac. x Federal state			0.812**		
			(0.401)		
Ethnic Pola.				-0.275***	
				(0.084)	
Ethnic Pola. x Federal state				0.495**	
				(0.215)	
Ethnic Frac. x Subn. exp. 90-18					2.365***
					(0.585)
Constant	-5.301***	-5.255***	-5.304***	-5.309***	-4.974***
	(0.546)	(0.561)	(0.568)	(0.566)	(0.337)
# Countries	155	155	155	155	155
# Regions	1,965	1,965	1,965	1,965	1,156
R-squared	0.917	0.918	0.918	0.918	0.935
Country FE	YES	YES	YES	YES	YES

Note: The unit of observation is first-level administrative region. The dependent variable in all estimates is ln(light), i.e., the log of the average nighttime light intensity extracted from VIRS images in 2016. All of the estimates include country fixed effects that are not reported. ln(pop) and ln(area) are the logs of regional population and land area, respectively. Ethnic Frac. (Pola.) is regional ethnic fragmentation (polarization) biased on GREG and GHSL data. Federal state is a dummy for being a federal country, as defined by Treisman (2008). Subn. trans. 90-18 is the mean of subnational transfers between 1990 and 2018 reported by the IMF Government Financial Statistics. $\ln(p^{I})$ is the log of the proxy for OSM mapping completeness of stage one, and $\ln(p^{I+II})$ is the log of the proxy for completeness of stages one and two as defined in section 2.2. Standard errors are reported in parentheses and are clustered at the country level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 18. Public amenities, decentralization and ethnic fragmentation: Additional controls 1

Dep. Var.:	(1) ln(#S.)	(2) ln(#S.)	(3) ln(#S.)	(4) ln(#S.)	(5) ln(#S.)	(6) ln(#S.)
ln(pop)	0.879***	0.872***	0.932***	0.922***	0.833***	0.835***
	(0.027)	(0.027)	(0.026)	(0.026)	(0.021)	(0.019)
ln(area)	0.021	0.024	0.008	0.011	0.115***	0.113***
	(0.023)	(0.028)	(0.021)	(0.026)	(0.021)	(0.025)
Ethnic Frac.	-2.435*	-0.188	-2.447**	-0.182	-0.504	-0.109
	(1.416)	(0.157)	(1.117)	(0.136)	(0.574)	(0.078)
Ethnic Frac.	-1.448***	-1.132***	-1.268***	-0.992***	-0.369***	-0.333**
x Federal state	(0.376)	(0.383)	(0.261)	(0.275)	(0.134)	(0.137)
Ethnic Frac.	0.261*		0.264**		0.045	
x ln(GDP p.c. national)	(0.158)		(0.125)		(0.063)	
Federal state		-0.035		-0.004		0.011
x ln(area)		(0.048)		(0.034)		(0.032)
$\ln(p^{I})$			0.918***	0.918***		
			(0.087)	(0.086)		
$\ln(p^{I+II})$					0.805***	0.805***
					(0.027)	(0.026)
Constant	-7.672***	-7.591***	-8.221***	-8.127***	-6.091***	-6.131***
	(0.386)	(0.421)	(0.347)	(0.376)	(0.227)	(0.221)
# Countries	148	155	148	155	148	155
# Regions	1,888	1,965	1,888	1,965	1,888	1,965
R-squared	0.843	0.843	0.877	0.877	0.957	0.957
Country FE	YES	YES	YES	YES	YES	YES

Note: The unit of observation is the first level administrative region. The dependent variable in columns (1)-(3) is the log of the number of schools reported in OSM, and in columns (4)-(7), the dependent variable is the log of the number of schools in OSM in active OSM areas. All of the estimates include country fixed effects that are not reported. ln(pop) and ln(area) are the log of regional population and land area, respectively. Ethnic Frac. is regional ethnic fragmentation biased on GREG and GHSL data. Federal state is a dummy for being a federal country, as defined by Treisman (2008). ln(GDP p.c. national) is the log of the PPP GDP per capita taken from the WDI 2017. $\ln(p^{I})$ is the log of the proxy for OSM mapping completeness of stage one, and $\ln(p^{I+II})$ is the log of the proxy for completeness of stages one and two as defined in section 2.2. Standard errors are reported in parentheses and are clustered at the country level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table 19. Restaurants, decentralization and ethnic fragmentation, regional development: A placebo test

Dep. Var.:	(1) ln(#R.)	(2) ln(#R.)	(3) ln(#R.)	(4) ln(#R.)	(5) ln(#R.)	(6) ln(#R.)	(7) ln(#R.)
ln(pop)	0.451***	0.450***	0.444***	0.606***	0.601***	0.863***	0.863***
	(0.073)	(0.073)	(0.073)	(0.058)	(0.059)	(0.041)	(0.042)
ln(area)	0.340***	0.342***	0.347***	0.302***	0.305***	0.132***	0.132***
	(0.061)	(0.061)	(0.061)	(0.051)	(0.052)	(0.040)	(0.040)
ln(light)	0.557***	0.555***	0.560***	0.470***	0.474***	0.042	0.042
	(0.055)	(0.055)	(0.055)	(0.050)	(0.051)	(0.041)	(0.041)
Ethnic Frac.		-0.101	-0.006	-0.051	0.019	-0.103	-0.102
		(0.152)	(0.160)	(0.136)	(0.139)	(0.090)	(0.098)
Ethnic Frac. x Federal state			-0.623		-0.477		-0.011
			(0.434)		(0.397)		(0.270)
$\ln(p^{I})$				0.540***	0.540***		
				(0.133)	(0.132)		
$\ln(p^{I+II})$						0.898***	0.898***
						(0.042)	(0.042)
Constant	-4.540***	-4.540***	-4.493***	-6.296***	-6.254***	-5.887***	-5.886***
	(0.717)	(0.716)	(0.716)	(0.537)	(0.546)	(0.345)	(0.347)
# Countries	155	155	155	155	155	155	155
# Regions	1,694	1,694	1,694	1,638	1,638	1,635	1,635
R-squared	0.835	0.835	0.835	0.850	0.850	0.941	0.941
Country FE	YES	YES	YES	YES	YES	YES	YES

Note: The unit of observation is the first level administrative region. The dependent variable in columns (1)-(3) is the log of the number of restaurants reported in OSM, and in columns (4)-(7), the dependent variable is the log of the number of restaurants in OSM in active OSM areas. All of the estimates include country fixed effects that are not reported. ln(pop) and ln(area) are the log of regional population and land area, respectively. Ethnic Frac. is regional ethnic fragmentation biased on GREG and GHSL data. Federal state is a dummy for being a federal country, as defined by Treisman (2008). ln(light) is the log of average nighttime light intensity extracted from the VIRS image of 2016. ln(pI) is the log of the proxy for OSM mapping completeness of stage one, and ln(pI+II) is the log of the proxy for completeness of stages one and two as defined in section 2.2. Standard errors are reported in parentheses and are clustered at the country level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Figure 6. Regional ethnic fractionalization

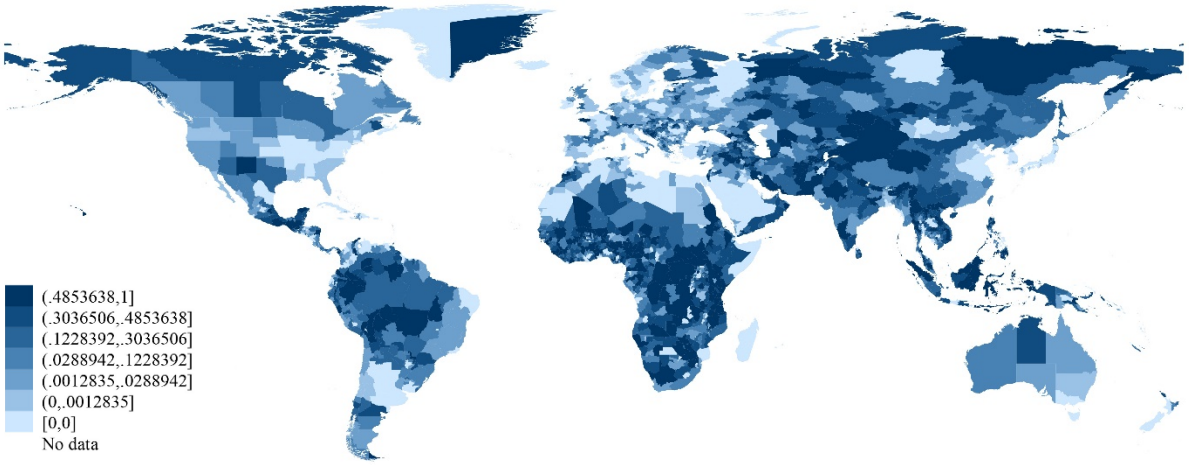


Figure 7. Regional ethnic polarization

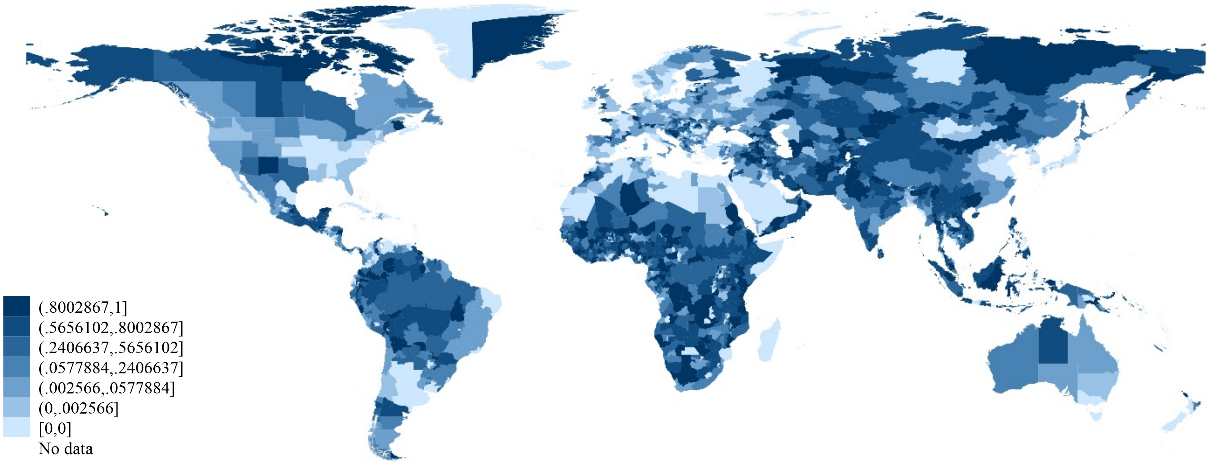


Figure 8. National-regional ethnic polarization