WORKING PAPERS IN ECONOMICS

No. 05/19

ANDRÉ SEIDEL

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

Department of Economics
UNIVERSITY OF BERGEN

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

André Seidel*

This version: 20.09.2019

Abstract:

This paper presents a new global dataset on the geocode locations of public amenities, e.g., schools, hospitals or libraries, based on OpenStreetMap data. Volunteered geocoded information can be systematically incomplete; therefore, we develop and study two new proxies for the degree of completeness of OSM data in first-level administrative regions. Using our new data, we study the effects of decentralization and ethnic divisions on the provision of public amenities associated with various public goods. We find strong evidence for the existence of collective action failure at the subnational level worldwide. More autonomous regions with high degrees of ethnic fractionalization provide significantly fewer public amenities than others.

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

There is little debate over the idea that one of the central responsibilities of states is to secure the supply of essential public goods to its citizens. In many countries, the right to supply some of these essential public goods, e.g., border defense, lies solely with the state. The ratio of public to private supplies of goods, such as education and health care, can differ between countries; however, the elementary supply of such goods, e.g., basic education and emergency care is widely seen as a state responsibility. Several decades of economic research have discussed various reasons why states might fail to fulfill their responsibility to supply fundamental public goods. One of them, which is at the center of this study, is the collective action failure associated with social heterogeneity.

Existing country case studies have indicated that social heterogeneity increases the likelihood of a collective action failure, ultimately decreasing the provision of regional public goods. If this phenomenon is global, then there might exist a dark side of decentralization, which so far has received little attention in the economic research. Therefore, the main question that we empirically address in this paper is whether increasing local autonomy and decentralization decrease the provision of regional public goods in regions with high levels of ethnic division. To the best of our knowledge, this paper is the first to empirically study the consequences of decentralization in the presence of social heterogeneity for the subnational provision public goods across countries.

Our study relies on a new dataset collected by us that contains the geocode locations of various amenities that are closely linked to some of the core public goods typically provided via government spending, such as schools, libraries, hospitals and police stations. Aiming to study regional spending, we aggregate these data on a regional level by simply counting the numbers of these amenities in different first-level subnational administrative regions (GADM1). Our final dataset covers 3342 regions in 204 countries.

The main data source that we utilized in the construction of our new dataset consists of volunteered, crowd-sourced data collected by the OpenStreetMap (OSM) project until the end of 2017. To address the well-known problems associated with this type of data, we developed two different indicators that allow for accounting for the degree of completeness of OSM data. Our indicators allow us to correct the data at the regional level or control for the degree of completeness within estimations. Comparing the corrected data with official data for a subset of countries for which we could find regional data, we typically observe country-level correlations greater than 90%. Using official data and OSM data to study the determinants of the degree of completeness, we find that completeness is mainly driven by national fixed effects and only a little by regional development. Using our indicators of completeness and national fixed effects, we can explain between 85% and 95% of the variation in observed completeness.

Using the new data, we first replicate the findings of Alesina, Baqir, and Easterly (1999) for the US before asking our main question using the global dataset. Our estimates suggest that decentralization decreases the regional supplies of schools, libraries, and hospitals in regions with high levels of social heterogeneity. The effect is sizable; for example, 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 part of a federal country. This finding is robust to a large battery of robustness tests, for example, the use of different indicators for decentralization or social heterogeneity. We also run a placebo test using nonpublic amenities, suggesting that the effects can be attributed to a collective action failure.

2. A GLOBAL MAP OF PUBLIC AMENITIES

2.1. PUBLIC AMENITIES AND PUBLIC GOODS

The number of public amenities is a simple but powerful proxy for spending on local public goods. For example, using exceptionally good official data on US primary and secondary schools, we can explain between 68% and 74% of district-level educational spending purely with the number of schools in a school district.[1] The number of public amenities per region is also a simple but good proxy for the welfare gains resulting from specific public goods. Greater local availability of public amenities usually results in higher welfare since consumption of the associated public goods becomes less costly. Consulting, for example, the literature on school attainment, we find that one of the main drivers of school attendance is distance to school (e.g., Duflo (2001), Burde & Linden (2013), Kazianga et al. (2013) and Muralidharan & Prakash (2017)). Turning to the literature on 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)), we find that response times are a key issue. The main driver of response times is the distance to the relevant amenity, which typically decreases as the number of amenities in a region increases.

2.2. OPENSTREETMAP AS A SOURCE FOR THE LOCATION OF PUBLIC AMENITIES

DATA COLLECTION

The data that we use on geocoded public amenities are extracted from the OSM project. The OSM dataset is a free, editable map of the whole world that is being 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.[2] The OSM project is the largest existing dataset of volunteered geographic information. The incredible success of the project arises from

[1] See Table 3 for estimations on this point using the US Public Elementary-Secondary Education Finance Data from 2015.

[2] https://wiki.openstreetmap.org/wiki/Stats

several factors, which have been well documented and discussed, for example, 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,[3] 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[4] and here[5].

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"). Our dataset is built using this information. We extract all of the polygons, multipolygon relationships, liens and points and their locations from the OSM project till the end of 2017 that carry tags that we associate with the various amenities under study. For example, to identity schools, we use the tags "amenity" = "school"' or "building" = "school". Table 11 in the appendix summarizes all of the tags that we use. Section 6.1 in the appendix summarizes in more detail how we 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. Some of them are less important to us than to geographers. 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 that we 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" =

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

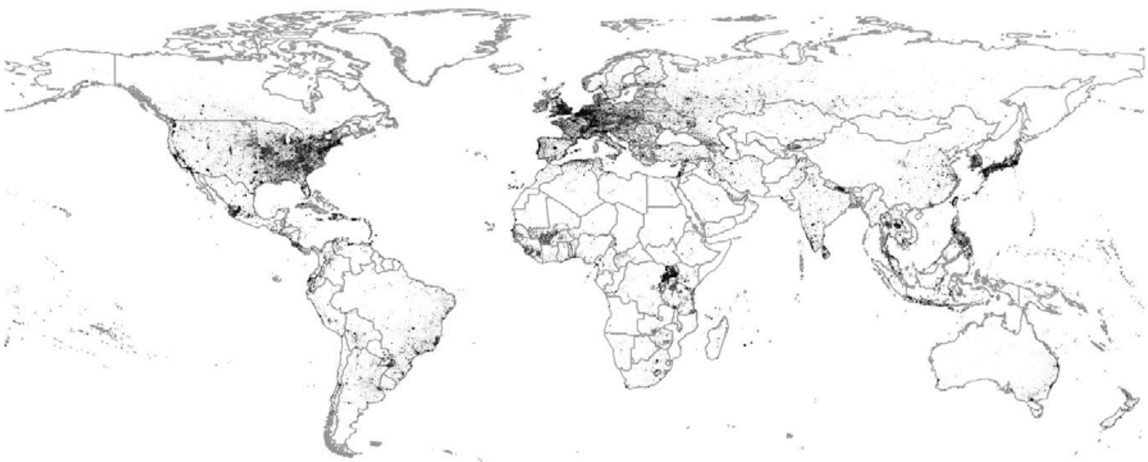
"'school"'. To avoid the resulting double counting (e.g., that each school yard wall is counted as a separate school), we merge all objects with the same tag within a 100-meter radius into one observation.[6]

COMPLETENESS

A quality dimension that is very important to us is completeness, hence the issue of 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. There various issues that could determine the magnitude of this effect, for example, 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 us with an initial impression of the data, as well as the potential extent of missing data. Figure 1 provides a first look at the data that we obtain after the initial cleaning, as described above. 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.

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



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)[7] has a large and successful project running in Uganda as a response to the ongoing refugee crises. As we show in the next section, despite the incredible increases in the number of OSM data volunteered in recent years, it seems that OSM data are in many dimensions incomplete. In this sense, Uganda is most likely an

[6] Obviously, we create some error with this technique because, for example, in densely populated regions, public amenities could be in such close proximity that we count them as one when there are actually two or more. However, changing the radius to 50 meters does not change the results. For restaurants, we reduce the radius to 10 m.
[7] See https://www.hotosm.org/ for more details.

outlier at the top, with more data than other countries in Africa. An example of a possible outlier at the bottom might be China or North Korea, where the number of schools seems very small. This finding is confirmed by examining simple descriptive statistics from the cleaned raw data[8] 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. We observe the greatest density of schools in the OSM data in heavily populated western regions of China. Therefore, it might be that systematic absences are a mostly country-level effect. Nevertheless, the next section discusses in detail how we can account for the degree of completeness, at least at the regional level.

2.3. 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, let us first state the general problem. Let us assume that an existing amenity is only recorded in the OSM data with a certain probability. Let us furthermore assume that this probability depends on the type of amenity and is constant within subnational regions. We refer to this probability as $p_{i,r}$,where $i \in \{School,\ Library,\ Hospital,\ Police\ station \ldots\ \}$ 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

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

where $A_{i,r}$ is the true number of amenities within a region. Consequently, if we find a proxy for the amenity's specific completeness in the OSM data ($p_{i,r}$), we can 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 $p_{i,r}$ we must extend our theory and 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.[9] 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

[8] See Tables 13 and 14 for more descriptive statistics from the raw data.
[9] For a more elaborate discussion of the motivations of OSM volunteers, see, for example, Goodchild (2007).

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.

From this information, we can derive a simple theory on the process of mapping. Mapping happens in two stages. Let us assume that mapping in regions without any data in the OSM project starts by adding fundamental landmarks, e.g., roads. Only after the first stage is realized can the second stage start. 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.[10] Let us 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, we assume

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

In what follows, we show that we can obtain proxies for these two probabilities using a comparison between OSM data and satellite data.

A PROXY FOR THE FIRST STAGE OF MAPPING COMPLETENESS

We can approximate the completeness of the first stage in a region of the country by comparing satellite settlement data with OSM settlement indicators associated with the first stage of mapping. Let us define a settled area (a pixel\~one km^2) as an area with urban buildup and more than 100 inhabitants. We identify these areas using the Global Human Settlement Layer (GHSL).[11] Let us furthermore assume that, if in such a settled location, the first stage of mapping has taken place, we should then observe residential roads[12] in the OSM data. For simplicity, we refer below to these areas, where we observe settlement indicators in the GHSL layer and OSM project, as active OSM areas.

[10] Support for our model comes 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”.

[11] We use the Global Human Settlement Layer from 2015.

[12] We concentrate on road data since they were by far the most common data added to the OSM project in the early stages of mapping, added even before surface characteristics, such as mountains. We furthermore concentrate on residential roads (highway = residential roads or service or unknown) since they are a proxy for settlement structures and are usually not mapped by government institutions (unlike larger roads connecting towns, such as highways and motorways).

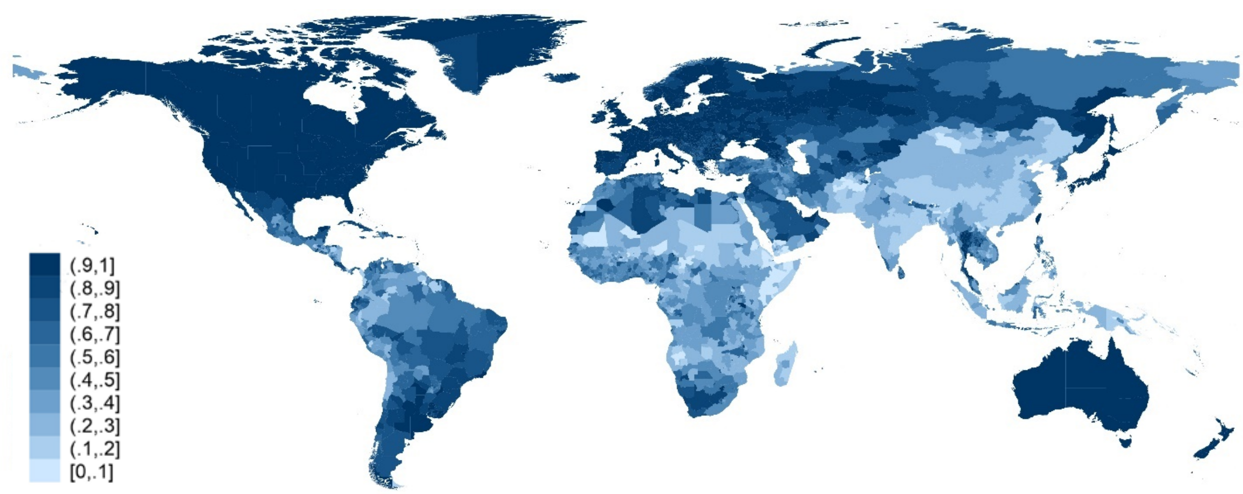
Building on these assumptions, we define our proxy $p_{i,r}^{I}$ for the completeness of the first stage of mapping in a region as

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

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, we assume that the share of the settled area in a region that contains residential roads is a good proxy for the degree to which the first stage of mapping has been realized in a in a region.

Figure 2 displays our proxy $p_{i,r}^{I}$ for the degree of completeness of fundamentals across the first-level administrative regions (GADM1) worldwide. The figure confirms the findings of our simple plausibility test on 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.

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



A PROXY FOR THE FIRST AND SECOND STAGES OF MAPPING COMPLETENESS

Our approach to finding a proxy for the extent to which the second stage of mapping has taken hold in a region follows a similar logic to our proxy for the first-stage realization. We count the number of square kilometers that contain at least one of the amenities of interest and divide by the number of square kilometers that have undergone the first stage of mapping (active OSM areas). With this technique, we obviously commit an error since we should not expect that every square kilometer that is settled contain a specific amenity. We assume, however, that this error is country specific. The proxy we suggest therefore can be calculated by

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

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.

Our proxy for the completeness of the second stage of mapping comes with an issue worth discussing upfront. The assumption of a fixed country-specific error term ε is in some instances problematic. We implicitly assume that the true share of settled area that contains at least one amenity is fixed across regions countrywide. If we suspect that there is an effect on this type of amenity density that is region specific, we must be careful when interpreting results that rely on $\varphi_{i,r}$ being part of a proxy for the completeness of the OSM data. We cannot circumvent this issue without knowing the true number of amenities within a region. We therefore recommend as a robustness test to always control whether the results depend on the use of our indicator for the completeness of the second stage of mapping. Nevertheless, we show in the next section that $\varphi_{i,r}$ still carries information worth utilizing when comparing OSM data with official data.

Plugging [3] and [4] into [2], we obtain our indicator for the completeness of mapping, that is,

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

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

Using our proxies for the completeness of mapping, we can predict the number of amenities based on the OSM data. Substituting [5] into [1], we obtain, after some algebra, the number of amenities as 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], we see that we ultimately calculate the average number of amenities within areas that contain at least one of the amenities of interest and multiply it by the settled area of the region. A possible interpretation of this step is that we treat those areas that are active OSM areas that, in addition, contain at least one amenity of interest, as representative of the region, and we inflate their data to the settled area of a region. As discussed before, in some cases, this approximation could be problematic.

2.4. How to Use the Proxy for Mapping Completeness

Cross Validation of Results

From the discussion of the previous section, we can draw the conclusion that, in theory, our proxy for the completeness of stages one and two could be biased in some cases. We therefore suggest always cross validating findings using the raw OSM data. We furthermore suggest also cross validating findings using our proxy for stage one of mapping alone. This indicator partly accounts for the degree of completeness while not being at risk of being biased by the assumption on which our indicator of completeness in stages one and two rests. To remain in line within the theory underlying the approximation approach, we furthermore suggest restricting observations to those within active OSM areas when using our proxies for the completeness of OSM data.[13]

Country Case Studies

When using our new amenity datasets and indicators of the completeness of stages one and two of mapping in a country case study, we must consider the country-specific approximation error ε. We can obtain a proxy for the error if we know the true total number of amenities in country A_i. Since we assume that the bias is the same in all regions, we can derive ε by totaling both sides of [4] and obtain

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

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

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

Cross Country Analysis

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

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

[13] Our indicator rests 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 if we exclude the US from our estimates.

where $\boldsymbol{X}$ is the vector of explanatory variables in which we are interested, $\boldsymbol{Z}$ is the vector of controls, $\mu_{i,j}$ is the country fixed effect, and j is the country index. The problem, however, is that we do not know the true number of amenities, so we must rely on the approximation described in the previous section. From [4], we can derive

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

Substituting [10] into [9], we obtain an estimation equation based on OSM data:

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

Note that $\log(\varepsilon_{i,j})$ is now part of the country fixed effect $\mu_{i,j}$ and that, based on our theory, we expect ϕ to be positive and close to 1.

2.5. TESTING THE RELIABILITY OF THE PROXIES OF OSM COMPLETENESS

COUNTRY-LEVEL RESULTS

As a first test of the reliability of our proxies for completeness, we use the assumptions made in the last section and [8] to calculate a proxy for the true number of amenities in a region. We compare this proxy with the true number of amenities per region for those cases in which we could obtain official data.[14] Since we obtain official data only on the locations of schools for a larger set of countries, we focus on schools in this analysis. 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 connivance, we added the 45° line in red.

Focusing first on the raw OSM data represented by gray triangles, we see that the perception that we derived from Figure 1, i.e., that the degree of completeness is entirely driven by country-level effects, was wrong. There is considerable heterogeneity in the missing data between regions of countries, for example, Malaysia or Mexico, which is something that we already suspected after studying the descriptive evidence provided by our indicator for the completeness of stage one of mapping. 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.[15]

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

[15] We examined US cases and found several reasons why we sometimes observe even more schools in the raw OSM data. Some of the reasons were related to tagging issues. For example, the OSM data include several historical schools in the Midwest that no longer exist. They are tagged as amenity = school with the Key = historic. We could not simply omit these schools since doing so might also mean dropping schools in historic buildings. Another reason is that our official data reflect the number of public schools, whereas the OSM data also contain private schools.

Looking finally at the adjusted data (blue dots), we see that the differences between the OSM and official data have decreased substantial. 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 from Table 1 reveals that adjusting the number of schools as proposed by our theory 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 in our favor. Clearly, the 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. Our approximation approach is furthermore also 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 of Table 1, we see that, in most cases, our approach is far superior to such a naïve approach.

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

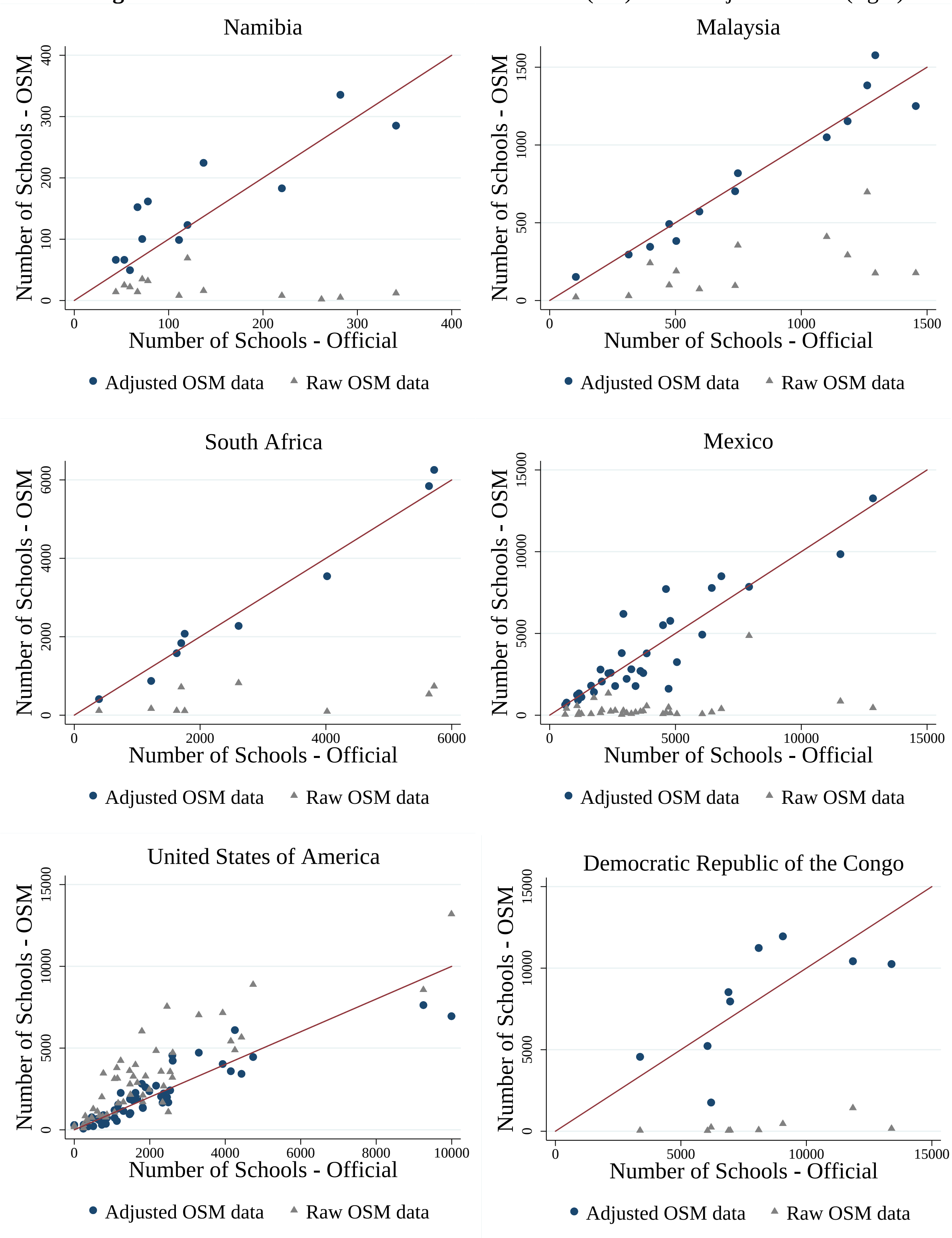


Table 1. Correlations of the official number of schools with different proxies

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 our proxies for completeness as described in section 2.4. # 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.

CROSS COUNTRY RESULTS

Table 2 presents estimates of the determinants of the true degree of completeness. Estimates are based on 124 regions in 6 countries for which we could obtain data on the official numbers of schools in the first-level administrative regions. As the determinant variable, we 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. We run estimates in the same fashion as we do when examining economic applications of the dataset. Hence, to account for potential bias in the OSM data, we run estimates including country fixed effects and our indicator of the degree of completeness. As discussed in section 2.4, it is suggested that we always should test the robustness of findings by considering both indicators of completeness separately. Hence, we present results for both completeness proxies $p_{i,r}^{I}$ and $p_{i,r}^{I+II}$.

Column 1 in Table 2 reports estimation statistics when we use only country fixed effects as the explanatory variable. The R^2 of 0.76 confirms our 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 average regional nightlight density as proxy for regional economic development, we 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 when we add our proxies for mapping completeness [column 4 and 6]. When we control for the log of $p_{i,r}^{I+II}$, light has no longer has any significant association with omissions [column 6].[16]

Considering the power of our proxies, both of which enter strongly significant and positive, we see that already our proxy for the first stage of mapping $p_{i,r}^{I}$ can explain, together with country fixed effects [column

[16] The link between development and the degree of missing data also becomes insignificant if we use the number schools in the raw data and not the number of schools in active OSM areas.

3], a considerable amount of variation in the data [R^2=0.848]. The fixed effects and our proxy for the completeness of stages one and two $p_{i,r}^{I+II}$ explain jointly [column 5] even more variation [R^2= 0.959].

Skipping already a bit ahead in our analysis, the results reported in Table 15 in the appendix suggest that ethnic fragmentation and decentralization seem not to impact the degree of completeness. Hence, even the raw data can be utilized to study the effects of both factors on the allocation of amenities.

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 deaminate variable of all of the 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. $\ln(p^I)$ 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 2.3. 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. PREVIOUS LITERATURE

Before applying our new data to a new question, we revisit a central result of 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. Using US regional data, they found 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, was confirmed in many subsequent studies. 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 what follows, we replicate the findings of Alesina, Baqir, and Easterly, (1999) using our new dataset. In doing so, we show that the number of amenities is linked to public expenditures and further that, despite the potential noisiness of our indicator for government spending, we can replicate the existing findings. Hence, we perform this exercise in part as an additional robustness test of our data, as well as an introduction to the discussion in the next 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, we focus on their findings on education spending. We perform the replication in stages: first we show that the number of schools is a good proxy of educational spending; and second, we show that the number of schools depends negatively on the degree of regional ethnic fractionalization. We do so with official government data on the number of schools, as well as our new data.

The most detailed data on educational spending in the US are available at the school district level. We were able to collect spending data for 7797 school districts[17] and matched them with our amenity data. Utilizing these data, we test the ability of the number of school district schools to predict total educational spending. To account for productivities of scale, we regress the number of schools in logs on the total expenditure on education in logs. Table 3, columns (1) to (3), summarizes the estimates using as an explanatory variable the official number of secondary and primary schools (1), the corrected number of OSM schools as defined

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

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, we also control in column (3) for the likely degree of completeness approximated by $\ln(\mathrm{p}_{i,r,j})$. 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 schools is associated with a 1% increase in educational spending. Overall, we observe a R^2 between 60% and 70%. The number of schools therefore seems to be a good proxy for educational expenditures.

Next, we test whether the number of schools depends negatively on the degree of ethnic fractionalization in US counties. For this purpose, we focus on the county level since it allows us to calculate the same fractionalization indicators as in Alesina et al. (1999). Hence, the indicators are based on the ethnicity definitions and population figures from the US Census of 2010.[18] Ultimately, we utilize data for 2131 US counties. In the first part of our replication analysis, we found a strong correlation between the log of the number of schools and the log of education expenditures. Consequently, we regress the level of ethnic fragmentation on the log number of schools. To account for size effects, we always control in all of the estimates for the log of area and population. 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 the potential distortion due to omissions from the raw OSM amenity data, we also control in column (6) for the likely degree of completeness approximated by $\ln(\mathrm{p}_{i,r,j})$. 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%.

In line with Alesina et al. (1999), we find mixed effects of ethnic fractionalization on the extent of public safety spending, measured by the number of police stations and health care spending, approximated by the number of hospitals. Furthermore, we find a weak, negative link between the number of libraries and the degree of ethnic fractionalization, fitting the theory of Alesina et al. (1999) that mostly productive public good should be affected.

[18] 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. Our findings, however, do not change when we use 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 deaminate variable of columns (1)-(3) is the log of educational expenditures as reported in the 2015 annual survey of school system finances. The deaminate variable of column (4) is ln(#of.S.), that is, the number of schools as reported in the National Center for Education Statistics Common Core of Data. The deaminate variable of column (5) is the number of schools recorded in OSM in 2017, corrected using our proxies for completeness as described in section 2.4. The deaminate variable of column (6) 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 2.3. 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. These last two studies, despite focusing on very specific countries, time periods and public goods, deliver partial support for our 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, we use 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 our cases, 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, we can write the ethnic polarization indicator as

[1]

$$Ethnic\ Pola_{e,r} = 1 - \sum_{e=1}^{m} \left(\frac{1/_2 - \pi_{e,r}}{1/_2} \right)^2 \pi_{e,r} = 4 \sum_{e=1}^{m} \pi_{e,r}^2 \left(1 - \pi_{e,r}\right) \quad [12]$$

and the ethnic fractionalization indicator as

[1]

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

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, which is why we focus on it in the main part of our analysis.[19] 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, we test the robustness of or findings using the indicators of ethnic polarization.

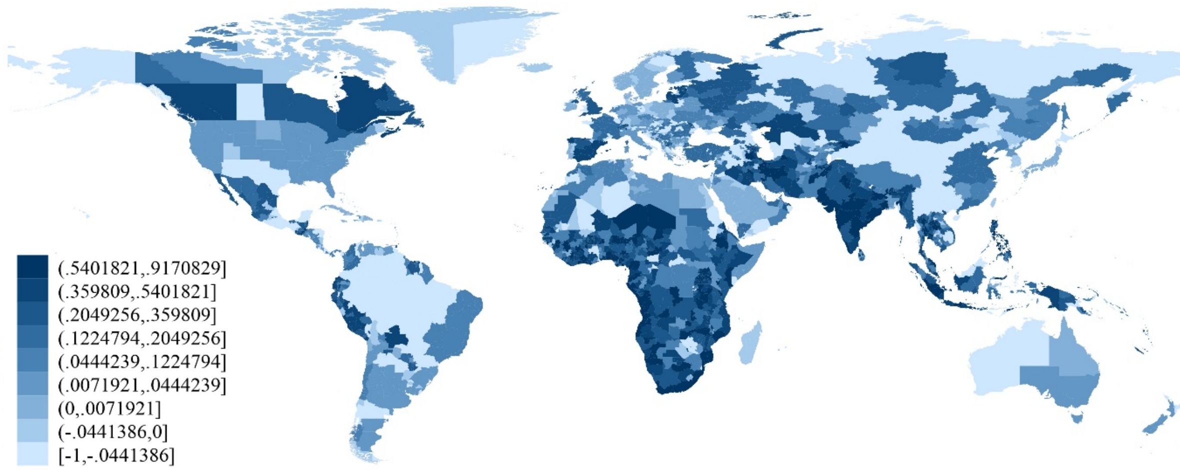
We measure the population belonging to different ethnicities by combining gridded population data from the 2015 GHSL with the ethnic homeland data provide by GREG, which go back to Weidmann et al. (2010). The GREG database reflects the distribution of ethnic groups worldwide in the 1960s and is based on a digitized version of the classical Soviet Atlas Narodov Mira. GREG documents the location of 928 ethnic groups in 8969 homelands. We project these homelands to the current political boundaries' first-level subnational administrative regions defined by ADM. Doing so, we obtain 23,874 regional homelands within 3219 regions.[20] 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[21]. These multigroup homelands are spread across 1044 of our 3219 regions for which we have OSM data. 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 at the same time reducing measurement error, we exclude regions from our main analysis that have more than 1% of the regional population living in homelands with multiple ethnicities, leading to the omission of 845 observations. Our main results are robust to dropping this exclusion criterion, as well as extending the cut-off to a 10% level. The results furthermore do not depend on how we treat the population residing in the multigroup homelands when calculating our social heterogeneity indicators.[22]

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

[19] 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.
[20] To minimize measurement error, we exclude regions with a population smaller than one.
[21] Gridded population data are taken from GHSL (2015), 1000-m resolution image.
[22] 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 we allocate the population of multigroup homelands equally among the named groups or with the same shares as in the rest of the region.

regional and national ethnic heterogeneity. To visualize this difference, Figure 4 displays the difference between national and regional ethnic fractionalization.[23] 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

We use two different types of measures for decentralization: de facto and de jure measures. The de jure measures that we use are the commonly used federalism indicator by Treisman (2008), which indicates whether a federal constitution exists (1) or not (0), and a new federalism indicator that we derive from the CIA World Fact Book that states whether the government type is federal (1) or not (0). The Treisman indicator is available for 155 countries in our dataset, and our CIA World Fact Book indicator covers 199 countries in our dataset. However, our indicator builds on only one very simple source of information, whereas the Treisman indicator builds 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 indicator used in the literature, we rely on it in our main analysis, and use our CIA World Fact Book indicator as a robustness test.

We derive our de facto measures from the IMF Government Financial Statistics. The three commonly used measures of fiscal decentralization that we use are the share of subnational expenditures in total

[23] 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, we did not omit regions with ethnic homelands that have residents belonging to multiple ethnicities.

expenditures, the share of subnational revenue in 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".[24] We focus in our analysis on vertical imbalance since it has the additional advantage of maximizing the number of available observations.

4.3. HYPOTHESIS AND ESTIMATIONS APPROACH

From the previous literature, we draw the conclusion that social heterogeneity hinders the provision of public goods at the local level because of the increasing 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. We can derive for [11] a specific estimation equation to test this prediction, that is,

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

where $Hete$ is our measure for social heterogeneity, $Auto$ is our measure of the degree of local autonomy, $\mu_{i,j}$ are the country fixed effects, and $\boldsymbol{Z}$ is the vector of controls. Our 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.[25]

4.4. IDENTIFICATION

There are considerable omissions with the OSM project, as we discussed in detail in section 2. Therefore, we must be careful when using OSM amenity data to study the allocation of amenities across regions. Our descriptive analysis, as well as our analysis in section 2.5, indicates that the omissions seem to be mostly associated with country-specific factors and a little bit with regional development. We do not find evidence that regional ethnic fragmentation or the degree of decentralization impact mapping completeness with countries for which we could obtain official data on the allocation of schools across regions.[26] Nevertheless, to decrease the risk of an omitted variable bias from the selection processes of OSM data, we always present

[24] For a more in-depth discussion of the various approaches used in the literature on decentralization, see, for example, Lessmann (2009).
[25] Indeed, when estimating the effect of ethnic fractionalization on the level of nightlight intensity in a region, we find a significant, negative effect. For further details see Table 17 in the appendix and the discussion in section 4.6.
[26] See Table 15 in the appendix.

estimates accounting for the degree of completeness of the OSM amenity data using the proxies discussed in section 2 alongside estimates based on the raw data alone. To show that findings do not depend on the assumptions associated with our proxy for the completeness of the second stage of mapping, we present findings controlling for the completeness of the first stage of mapping alone $\ln(p^{I})$ (see [2] and [3]) and when controlling for completeness of the first stage and second stage of mapping jointly $\ln(p^{I+II})$ (see [2], [3] and [5]). It is important to note that the dependent variables differ for estimates including our proxies for completeness and those that do not. As suggested in section 2.4, we use as the dependent variable the number of amenities in active OSM areas[27] when running estimates containing our proxies for completeness and otherwise the number of all OSM amenities within a region. Given the theoretical argument presented in sections 2.3 and 2.4 and the empirical findings presented in section 2.5, we expect that the coefficients of our proxies for completeness are positive and close to one.

To reduce the likelihood of further omitted variable bias, we always control for country-level fixed effects $\mu_{i,j}$, as well as the regional log of population and log area. We expect that the number of amenities increases with the number of regional residents. Our 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 we study the phenomenon at the regional level, the endogeneity of institutions seems not to be of greater relevance, given that regional ethnic fractionalization and all of the measures of decentralization that we use 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).

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

[27] E.g. areas with urban buildings, more than 100 residents and residential roads in the OSM data

4.5. MAIN RESULTS

Table 5 reports our main results. The determinant variable in our baseline is the log number of schools within first-level subnational regions. All of the estimates are based on a consistent dataset that is restricted by the availability of our indicator.[28] We have data for 1965 subnational regions within 155 countries.[29]

We started in column (1) with a specification that only includes our country fixed effects, the log of population and the log of area. Our standard controls explain 85% of the variation within our observations. The coefficient of log of population is positive and strongly significantthe coefficient for the local area is positive but not significant. These findings confirm the reasonable expectation that the main determinant for the number of schools within a region is the population of a region.

In column (2), we add the level of regional ethnic fractionalization, which enters with a significant, negative effect coefficient. The effect becomes insignificant if we add to column (3) the interaction of ethnic fractionalization and our indicator for decentralization, which enters with a strongly 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 in column (3) suggest that ethnic fractionalization decreases the supply of schools in regions that are part of a decentralized country by a considerable margin. We see this outcome as our main finding.

We next ensure that our findings in columns (2) and (3) are not affected by the regional degree of completeness of the OSM amenity data. In columns (4) and (5), we include the log of our indicator for the completeness of the first stage of mapping ($\ln(p^{I})$) into the estimation. In columns (6) and (7), we add our 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, we observe a decrease in effect size most notably when controlling for the completeness of stages one and two in columns (6) and (7). The coefficient suggests that a reduction of 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 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 in column (3) are

[28] Note that we omit regions where we do not observe any data in the OSM project, which leaves us with 2956 observations. We also omit regions where more than 1% of the total population lives in ethnic homelands in which multiple ethnicities reside, leaving us with 2226 observations. The availability of our decentralization indicator decreases the number of observations finally to 1965.

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

overestimated if we do not control for the degree of completeness, which could be the case if the degree of completeness is negatively affected by ethnic fractionalization and decentralization. Our findings in section 2.5 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 we underestimate effect sizes in column (7), which might be the case since our proxy for the completeness of mapping of stages one and two rests implicitly on the assumption that we can treat the cells that contain amenities as representative for the region. Hence, we might miss an effect of ethnic fragmentation and decentralization on the number of amenities outside of these cells. If this effect goes in the same direction as in the representative cells, then we underestimate the total effect. 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), we only control for the completeness of the first stage of mapping. Remember that our indicator for the first stage of mapping is essentially the share of populated cells that have any data on rule roads. Hence, this indicator decreases the potential bias of systematic mapping that could inflate the estimates in column (3) without making the restrictive assumptions of our indicator for the completeness of stages one and two of mapping, which could downplay the effects in column (7).

We run a large a set of robustness test.[30] It is possible that our indicator for decentralization also proxies for the level of general country development (correlation 0.29). In Table 18 in the appendix, columns (1,3,5), we add the interaction of ethnic fractionalization with the log of national GDP per capita[31], without any changes to our main finding. Larger regions might have a greater likelihood of being an ethnically fractionalized regions (correlation 0.23). If so, we might simply detect a size effect of regions that are part of a federal state. In Table 18, columns (2,4,6), we therefore add the interaction of the federalism indicator with the log of area, and it does not change our findings. Capital regions might be special for various reasons; hence, we run estimates that include a dummy for capital regions, or we exclude capital regions and do not find different results. 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 our findings.

[30] Some might recall at this point that some potential omitted variables are already addressed by the court fixed effect included in all of our estimates.

[31] Using the interaction of national GDP per capita not in logs does not change the result.

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 deaminate variable in columns (1)-(3) is the log of the number of schools reported in OSM and, in columns (4)-(7), 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 2.3. 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

As was the case in previous studies, we cannot observe directly the failing of collective actions; we observe only the outcomes of successful actions. Therefore, we must be careful when contributing our findings to collective action failure. We must hedge against the risk that there are indirect effects of ethnic fractionalization and decentralization on the supply of public amenities.

The most prominent one is that regional development can be affected by ethnic fractionalization and decentralization and, in turn, affect the capacity to finance public amenities. To determine whether we simply pick up a regional development-level effect, we add in Table 9, columns (1, 3 and 5), the log of average regional nightlight and the interaction of our indicator of decentralization with nightlight intensity in columns (2,4 and 6).[32] Controlling for regional economic development does not change our results. It seems that the effect of ethnic fractionalization does not arise from the indirect effect of ethnic fractionalization on development.[33] 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 if we do not control for the completeness of stages one and two of mapping. Our findings in section 2.5 indicate that the degree of completeness of OSM data is positively associated with regional development, which might explain why the effect becomes insignificant if we control for the degree of completeness of stages one and two 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 narrow down further that our findings can be attributed to collective action failure, we can perform a placebo test. To do so, we extract the number of restaurants in a region from the OSM Project. Restaurants are amenities that are not provided by the government and therefore should not be directly influenced by the political economy of regional government spending. Hence, we expect to see no differences between ethnically fractionalized regions in decentralized and non-decentralized countries.

In Table 7, we present the findings for our bassline specification when using the log of the number of restaurants per regions as a dependent variable. Only when we do not control for degree of completeness of the OSM data do we find a significant, negative effect of ethnic fragmentation. If we control for regional development, even this effect becomes insignificant. Most importantly, we never see that decentralization 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).

[32] 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)). We use the VIIRS global nightlight images from 2015, the latest year for which cleaned high-resolution images are available. The data are provided by Earth Observation Group at NOAA/NCEI.

[33] In fact, when estimating the effect of ethnic heterogeneity and its interaction with decentralization, we find the opposite effect. 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. x Federal state	-1.360***	-1.457***	-1.162***	-1.275***	-0.349**	-0.510***
	(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. x ln(light)		0.073		0.085		0.122***
		(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 deaminate variable in columns (1)-(3) is the log of the number of schools reported in OSM and, in columns (4)-(7), 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(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 2.3. 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. x Federal state			-0.077		-0.055		0.029
			(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 deaminate variable in columns (1)-(3) is the log of the number of restaurants reported in OSM and, in columns (4)-(7), 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). $\ln(p^{I})$ 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 2.3. 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 we can extend our findings to a broader set of public amenities, we utilize the number of other public amenities that are part of our new dataset.

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 our main estimates using the log of the number of libraries as the dependent variable. Note that we refer to the amenity specific proxy for the completeness of stages one and two, when referring to $\ln(p^{I+II})$ in Table 8. As with schools, we see a negative effect of ethnic fractionalization that mainly comes from regions that are part of a federal country. The effect, however, is only significant if we control for our indicators of mapping completeness in a region. The effect is also significant when using the raw data if we control 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. In our cases, however, the results are less mixed (see Table 8, columns 2, 5, 8). We find that fractionalization has a significant, negative effect on the number of hospitals within 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. Using the log of the number of police stations as a dependent variable, we find that ethnic fractionalization has a significant, negative effect. 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 deaminate 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 log number of hospitals and, in columns (3), (6) and (9), 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 2.3. Standard errors are reported in parentheses and are clustered at the country level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

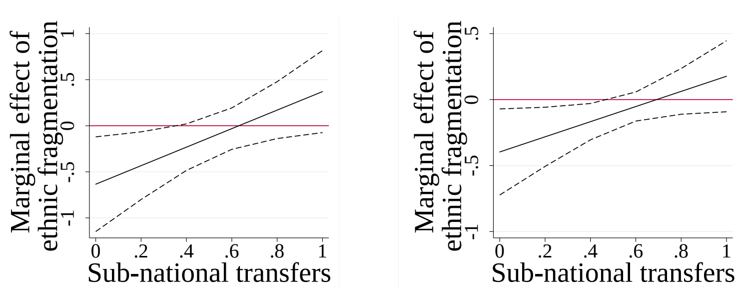
4.8. ALTERNATIVE MEASURES OF THE DETERMINANTS

ALTERNATIVE MEASURES OF DECENTRALIZATION

To ensure that our measure of decentralization does not bias our findings, we replicate our main findings using alternative indicators. The federalism indicator that we use in our main analysis is well established in the literature but is not available for all countries in our dataset; therefore, using the CIA World Fact Book, we constructed a new indicator that allows us to increase the number of countries to 199.[34] Adding the additional countries or estimations (Table 9, columns (1-3)) does not change our findings, and the effect size remains roughly the same as that in our bassline specification.

We can derive for at least a subset of countries de facto measures of fiscal decentralization using the IMF government finance statistics. To maximize the number of observations, we focus on the share of transfers, and we use 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. The first thing to note is that using the IMF data drastically reduces our sample to almost a half of its original size. Second, the coefficient of ethnic fragmentation enters negatively. Third, in line with the idea that higher shares of transfers reflect decreasing local autonomy, we find a positive interaction effect with ethnic fractionalization. The effect is significant if we control 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% when we control for $\ln(p^{I})$ (Figure 5, left) and 50% when we control for $\ln(p^{I+II})$ (Figure 5, right) and is otherwise negative. Hence, ethnic fragmentation negatively affects the supply of public amenities in regions that are more financially independent.

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



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

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 deaminate variable in columns (1)-(3) is the log of the number of schools reported in OSM and, in columns (4)-(7), 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 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 2.3. 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

We next examine whether our findings depend on the measure of social heterogeneity that we are applying. In Table 10, we reperform the estimates presented in Table 5 using ethnic polarization as an indicator of social heterogeneity. Comparing both sets of results, we see 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.			-0.753***		-0.655***		-0.190*
x Federal state			(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 deaminate variable in columns (1)-(3) is the log of the number of schools reported in OSM and, in columns (4)-(7), 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 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 2.3. 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. We find 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 that we find 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 part of a federal country.

The analysis is based on a new dataset that we derive from the OSM project, which contains global location of various public amenities associated with public goods that are typically provided to a large extent by the state. We address well-known accuracy problems associated with using volunteered geocode data by developing a new method that allows us to account for the completeness of OSM data within first level subnational regions, by cross-referencing of OSM settlement indicators with indicators derived from satellite data. Our new approach allows us to minimize any bias in our estimations stemming from omitted variables creating a systematic bias in the OSM data. We test the quality of our approach by correcting the OSM raw data and comparing the corrected data with official data for a subset of countries where such data exist. The correlation between our final data and the official data on these countries is typically greater than 90%. Our results also hold when we use the original data and when we alter the different technical details of the algorithms used to clean the raw data or to account for the possibility of systematically missing data. Our findings are robust to a large set of robustness tests based on a large set of controls, as well as alternative indicators for public goods, social heterogeneity and decentralization. We also run a placebo test and find that the supply of regional non-public amenities, such as restaurants, is not affected by the joint effect of social heterogeneity and decentralization. Examining our data, we do not see any 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, we are confident that we find a causal effect of social heterogeneity and decentralization on the provision of regional public goods.

Our findings elucidate the dark side of decentralization, which has received little 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 us to understand the mechanism underlying the existing finding that favoritism impacts growth (Hodler & Raschky, 2014; De Luca, Hodler, Raschky, & Valsecchi, 2018). We leave such questions open for further research since addressing them would go beyond the scope of this paper.

REFERENCES

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.

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.

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

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.

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.

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, we created a SpatiaLite Link to the OSM data. 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, we deleted all objects for which tags were empty. 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. For the final cleaning step, we merged all of the point layers from the OSM Geofabric dump, as well as those obtained from the Overpass API (multipolygon).
6. When an amenity consisted of multiple buildings and was not labelled as a multipolygon, we might have overestimated the presence of amenities, as we, for example, counted a hospital complex with two buildings as two hospitals. To account for this fact, we merged close-by observations.

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
$\mathbf{p^{I}}$	Proxy for the completeness of the first stage of OSM mapping **Source:** Own calculations biased on OSM and GHSL; for details, see section 2.3
$\mathbf{p^{I+II}}$	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 2.3
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. x Federal state		-1.100		-0.508		-0.052
		(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 deaminate variable of 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 2.3. 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 our main indicators for 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

	(1)	(2)	(3)	(4)	(5)
Dependent Variable	ln(light)	ln(light)	ln(light)	ln(light)	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 deaminate variable in all of the estimates is the ln(light) the log of average nighttime light intensity extracted from the VIRS image of 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 2.3. 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

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	ln(#S.)	ln(#S.)	ln(#S.)	ln(#S.)	ln(#S.)	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 deaminate variable in columns (1)-(3) is the log of the number of schools reported in OSM and, in columns (4)-(7), 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 2.3. 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 deaminate variable in columns (1)-(3) is the log of the number of restaurants reported in OSM and, in columns (4)-(7), 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. $\ln(p^{I})$ 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 2.3. 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

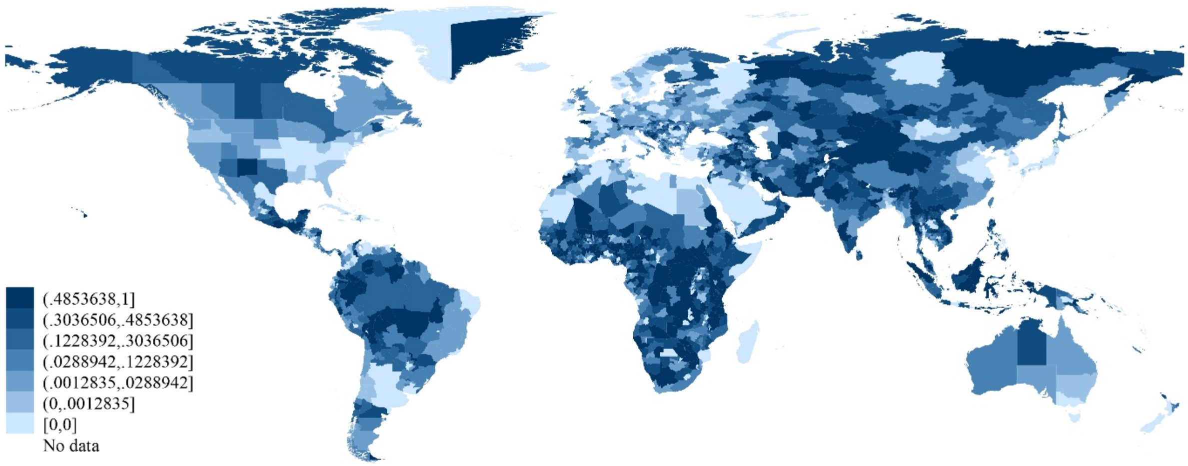


Figure 7. Regional ethnic polarization

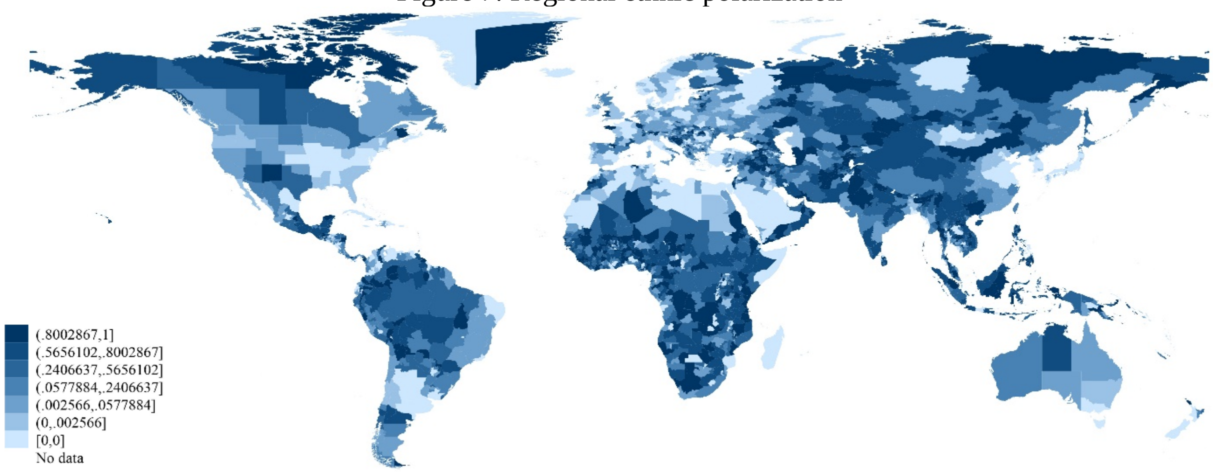
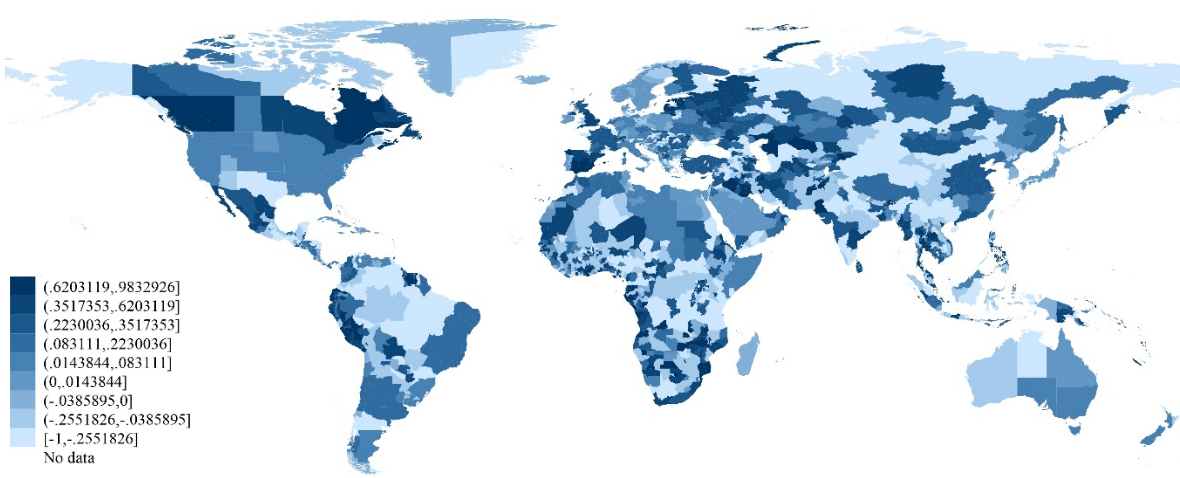


Figure 8. National-regional ethnic polarization



Department of Economics
University of Bergen
PO BOX 7802
5020 Bergen
Visitor address: Fosswinckels gate 14
Phone: +47 5558 9200
www.uib.no/econ/