Special Data Feature

Journal of Peace Research

Journal of Peace Research
2018, Vol. 55(5) 687–698

DOI: 10.1177/0022343318764254
journals.sagepub.com/home/jpr
SAGE

New spatial data on ethnicity: Introducing SIDE

Carl Müller-Crepon

Center for Comparative and International Studies, ETH Zurich

Philipp Hunziker

Lazerlab, Northeastern University & IQSS, Harvard University

Abstract

Research on ethnic politics and political violence has benefited substantially from the growing availability of cross-national, geo-coded data on ethnic settlement patterns. However, because existing datasets represent ethnic homelands using aggregate polygon features, they lack information on ethnic compositions at the local level. Addressing this gap, this article introduces the Spatially Interpolated Data on Ethnicity (SIDE) dataset, a collection of 253 near-continuous maps of local ethno-linguistic, religious and ethno-religious settlement patterns in 47 low- and middle-income countries. We create these data using spatial interpolation and machine learning methods to generalize the ethnicity-related information in the geo-coded Demographic and Health Surveys (DHS). For each DHS survey we provide the ethnic, religious and ethno-religious compositions of cells on a raster that covers the respective countries at a resolution of 30 arc-seconds. The resulting data are optimized for use with geographic information systems (GIS) software. Comparisons of SIDE with existing categorical datasets and district-level census data from Uganda and Senegal are used to assess the data's accuracy. Finally, we use the new data to study the effects of local polarization between politically relevant ethnic groups, finding a positive effect on the risk of local violence such as riots and protests. However, local ethno-political polarization is not statistically associated with violent events pertaining to larger-scale processes such as civil wars.

Keywords

ethnic groups, geographic information systems, settlement patterns

Introduction and motivation

Spatially disaggregated data are increasingly important for the study of ethnic politics and political violence. Subnational data on ethnicity allow investigating how ethnic settlement patterns influence the likelihood (Weidmann, 2009) and location (Buhaug & Rød, 2006) of civil wars, the occurrence of electoral violence (Wilkinson, 2004) and the prevalence of communal conflict (Fjelde & Østby, 2014). Moreover, combining spatial data on the location of ethnic communities with geo-coded covariates enables the investigation of important subnational determinants of political violence, such as intergroup economic inequalities (Cederman, Weidmann & Bormann, 2015), petroleum production (Asal et al., 2016) and climate variability (Fjelde & von Uexkull, 2012).

Many of these research designs are made possible by datasets that offer geo-coded information on ethnic settlement patterns for a large number of countries. In particular the GREG (Weidmann, Rød & Cederman, 2010) and GeoEPR (Wucherpfennig et al., 2011) projects provide data on ethnic settlement patterns across the globe in the form of geo-referenced polygon features.[1] While extremely valuable, these datasets exhibit a key

[1] See also the Ethnologue Atlas (Global Mapping International & SIL International, 2015) and Murdock's (Murdock, 1967; Nunn & Wantchekon, 2011) map of ethnic groups in Africa.

Corresponding author:
carl.mueller-crepon@icr.gess.ethz.ch

Figure 1. Ethnic maps contained in SIDE
Country borders from Weidmann & Gleditsch (2010).

shortcoming: they are ill-suited for capturing local ethnic diversity. The polygon-based coding of ethnic settlement patterns adopted by these datasets is only adequate where ethnic groups are spatially segregated. Local ethnic diversity is, however, a common phenomenon and ethnically segmented countries rarely feature complete segregation. Polygon-based datasets, even when letting polygons overlap, do not capture such local ethnic diversity due to their categorical nature.

This limitation entails two important caveats for applied research. First, ignoring local ethnic diversity can be problematic when settlement polygons are used to estimate group-level covariates. Specifically, if group members are not distributed exclusively and uniformly within 'their' polygon, estimates of group characteristics obtained via spatial operations may be associated with considerable error. Second, polygon-encoded ethnic settlement patterns all but prohibit investigating the consequences of local ethnic diversity.

We argue that redressing these issues requires the use of ethnic settlement data that reflect local ethnic mixing. Because geographically detailed ethnic census data are rare, relying on official statistics is an impractical strategy for most conflict-related research projects. As an alternative, this article introduces the SIDE – Spatially Interpolated Data on Ethnicity – dataset. With a total of 253 maps, SIDE provides high-resolution geo-referenced information on local ethnic (i.e. ethno-linguistic, religious, and ethno-religious) population shares for 47 low- and middle-income countries across the globe (see Figure 1). SIDE consists of raster data that encode estimates of local ethnic compositions at a resolution of 0.0083 degrees (ca. 1 km at the equator). Thus, when combined with suitable population data, SIDE

i. permits constructing maps of ethnic groups' settlement patterns that explicitly account for local variation in groups' population shares,
ii. allows calculating the extent and nature of local ethnic diversity for arbitrary spatial units,
iii. facilitates visualization.

All data are available at https://side.ethz.ch, or via the sidedata R package.[2]

The key innovation underlying SIDE is that it is estimated from geo-coded survey data via spatial interpolation methods. Specifically, SIDE is a statistical generalization of the ethnic information contained in the DHS data (DHS, 2015). Many DHS surveys are geo-coded, thus providing a set of spatial sampling points containing local ethnic composition estimates. We use methods from geo-statistics and machine learning to estimate the ethnic composition of areas in between these sampling points, thus producing a continuous map of ethnic compositions for each surveyed country. We also demonstrate that tuning spatial interpolation models through machine learning strategies is an effective

[2] See https://github.com/carl-mc/sidedata.

strategy for generalizing sparse, geo-coded survey data to locations not included in the original sample.

This article is structured as follows. The next section provides an overview of the DHS data underlying SIDE. We then outline our spatial interpolation procedure, followed by a presentation of the newly derived data. We evaluate the data's accuracy using information from GeoEPR, as well as district-level census statistics from Uganda and Senegal. Finally, we discuss SIDE's limitations, and offer an illustrative example of its use by exploring the violent effects of local polarization of politically mobilized ethnic groups.

The DHS data

The information used for estimating the SIDE data are the ethno-linguistic and religious identities of respondents enumerated in the Demographic and Health Surveys (DHS, 2015). The DHS is a demographics- and health-related survey that has been conducted regularly and in a growing number of countries since the 1980s. Beyond its primary focus, the survey also collects basic demographic information such as respondents' ethnicity. About 50% of all DHS rounds are geocoded, of which 119 include items on ethno-linguistic (74 surveys in 31 countries) and/or religious identities (114 surveys in countries in 45 countries). Combining the two ethnic identifiers, we can construct ethno-religious identities in 67 surveys (29 countries). An overview of the surveys used for generating SIDE maps is given in Figure 1 in the Online appendix.

In total, the DHS provides geocoded information on the ethnic identities of 1.83 million respondents. For the DHS sampling procedure, each country is divided into subregions out of which primary sampling units (PSU) are drawn with a probability proportional to their population (see Figure 2). Households are then sampled at random in each PSU (USAID, 2012). This procedure ensures that the DHS is representative on the national, subnational, and – on average – PSU-level. We exploit this fact by treating each PSU (also called cluster or point hereafter) as a geographic point associated with a certain ethnic composition. Beyond the PSU-level sampling error introduced by the relatively small size of local samples (10–60 respondents/PSU), the data provider randomly displaced the geocodes associated with the clusters by up to 10 km in rural and 2 km in urban areas to mitigate privacy concerns (Burgert et al., 2013). This displacement places a natural upper limit on the spatial precision of the SIDE data.

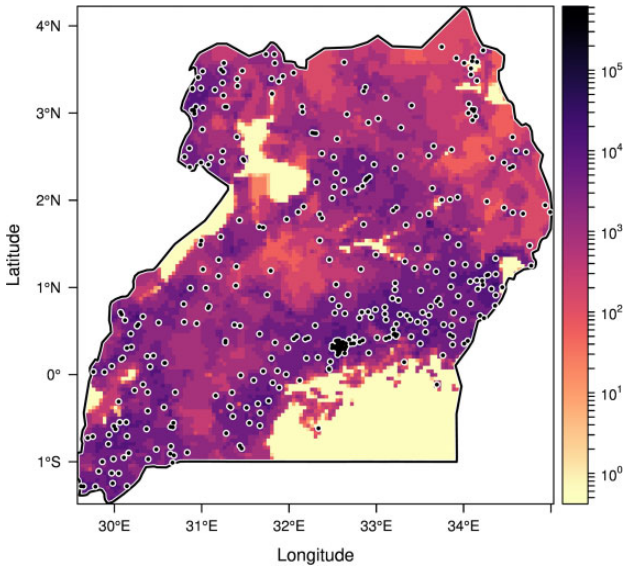

Figure 2. Geocoded DHS clusters (2011) and local population counts (2000) in Uganda

Data: USAID (2012) and CIESIN et al. (2011).

To limit the computational costs of the interpolation procedure discussed below, we only predict settlement patterns for ethnic groups that amount to at least 0.5% of a country's population. All other ethnic groups are added to those labelled 'others' by the DHS.[3] For most countries, multiple survey rounds from different years are available. These rounds vary with regard to the number of respondents, clusters and enumerated ethnic groups.[4] We therefore conduct separate estimation and prediction runs for each survey and provide separate maps for each survey round.

Methods

This section provides a brief overview of the spatial interpolation procedure we employ to generate the SIDE data. For a detailed discussion, we refer the reader to the Online appendix.

[3] To deal with the much more numerous ethno-religious groups, respondents of an ethno-linguistic group have been assigned the religious 'mixed' category if (1) their religious group constitutes less than 5% of the total population and less than 20% in the respective ethno-linguistic group, or (2) if their ethnic group is smaller than 5% of the total population and their religious group makes up less than 50% of their ethno-linguistic group.

[4] See the Online appendix.

Interpolating compositional data

The methodological challenge we face when generating the SIDE data is the following. Can we infer the ethnic composition at any location within a country given information on ethnic population shares for a finite sample of points? This task is a variant of spatial interpolation, which refers to the mapping of data sampled at discrete points onto a smooth surface (Mitas & Mitasova, 1999). In the two-dimensional case, spatial interpolation pursues the following goal. Given a sample of points $s_i, i = 1, 2, \ldots, N$, each associated with a coordinate pair (x_i, y_i) and a measured outcome value z_i, find a model that provides a prediction $\hat{z}_0$ for some target location s_0.

Most spatial interpolation methods are designed for continuous and unbounded response data. In this application, however, we are interested in estimating *compositions*. Modeling compositions is challenging, since they are bounded by 0 and 1, and must sum up to unity. To address this issue, we pre-process the compositions at each sample point by applying the additive log-ratio transform (Aitchison, 1986: 93). The transform maps a vector of compositions of length G, $[z_{i1}, z_{i2}, \ldots, z_{iG}]$, onto a vector of log-ratios of length $G-1$, $[z'_{i1}, z'_{i2}, \ldots, z'_{i(G-1)}]$, by dividing each of the first $G-1$ compositions by the final composition and taking the natural logarithm, that is,

$$z'_{ij} = \left[ln\left(\frac{z_{ij}}{z_{iG}}\right)\right]. \quad (1)$$

Because the transformed z'_i values are unbounded, we may now perform all estimation and prediction tasks on these individual log-ratios without worrying about the compositional nature of the data. The final output data are then generated by applying the inverse log-ratio transform to the vector of separately predicted log-ratios. This *guarantees* that the resulting values sum to 1.[5]

Interpolation

We use a three-step approach to generate a prediction $\hat{z}_0$ for some target location s_0:

i. **Sample selection:** We employ a *local modelling approach* for the interpolation task (cf. Lloyd, 2010). Thus, instead of generating predictions from a single global model fitted to all available data points from a given survey in a given country, we estimate a local model using only data points that are spatially proximate to s_0. We create local samples by selecting all observations that are either within distance D of the target point s_0, or belong to its K nearest neighbors.

ii. **Modelling:** We apply two types of models to generate predictions for a target point given its local sample. The first is the exponential distance decay (EDD) method. The EDD method assigns a given target point a weighted average of the response values of all other points in the (local) sample (Smith, 2016). Weights are constructed using an exponential distance decay function that reduces the influence of sample points that are further away from the target point. The second method we apply is the thin plate spline (TPS, see Lloyd, 2010: 158). The TPS is a smoothing spline that is used to fit a smooth surface to a sample of spatial points associated with a continuous outcome variable.

iii. **Model mixing and prediction:** We rely on both the EDD and the TPS methods because they have complementary properties. The EDD method is exceptionally robust against overfitting, whereas the TPS is able to recover complex response surfaces. To leverage this complementarity, we generate predictions using both modelling approaches, and then use a weighted average of the two as the final prediction. The weights used for averaging correspond to each model's out-of-sample predictive performance, the estimation of which is discussed below.

Model tuning

The prediction approach outlined in the previous subsection involves a number of 'tuning' parameters that cannot be estimated directly during model fitting. These include the K and D parameters determining the size of local samples, the decay parameter determining the EDD weights, and a couple of parameters associated with the log-ratio transformation. We determine these parameters using a leave-one-out cross-validation (LOOCV) strategy on a subsample of the input point data. More precisely, for each of the two modelling approaches, we choose the set of parameters that maximize the model's out-of-sample predictive performance, as measured via LOOCV. Because sweeping the parameter space entirely is associated with prohibitively high

[5] One possible alternative to this log-ratio approach is to interpolate the compositional data directly using Kriging. However, Kriging has proven computationally too expensive for the task at hand; see Section 2.3.3 of the Online appendix.

computational costs, we employ the genetic optimization algorithm implemented by Mebane & Sekhon (2011) to determine appropriate tuning parameters.

Prediction

Finally, we use the tuned models to generate near-continuous maps of each country's ethnic composition at the local level. To this end, we divide each country into raster cells, choosing a resolution of 0.0083 decimal degrees (ca. 1 km at the equator).[6] After predicting the two maps of ethnic compositions on the basis of the EDD and TPS estimators, we mix the predictions for each point according to group-specific model fits of the optimized models.

Estimation results

As anticipated, we find that combining the EDD and TPS predictions consistently improves predictive performance. Moreover, we find that our models' predictive performance follows a number of expected regularities (see the Online appendix). First, unsurprisingly, we are typically better at predicting ethnic settlement patterns in countries featuring fewer ethnic groups, and covered by a higher density of DHS sampling units. Next, we appear to be slightly worse at predicting ethno-linguistic settlement patterns than at predicting ethno-religious and religious settlement patterns. This difference is likely driven by higher levels of spatial segregation among religious groups in many of the sampled countries. Finally, we find that model fit does not appear to be affected by the average number of DHS respondents per cluster, or the total number of respondents of a survey.

A look at the data

The SIDE data are organized as a collection of geo-coded raster grids with a resolution of 0.0083 decimal degrees. This format permits extracting the estimated ethno-linguistic, religious and ethno-religious composition of any point in a given country. As an example, Figure 3 shows the estimated ethno-religious composition of Ibadan in Nigeria in 2013. According to SIDE, in 2013, Ibadan was composed of approximately 45% Yoruba Christians, 40% Yoruba Muslims, and a range of smaller ethno-religious groups.

SIDE can also be easily combined with spatial population data, in particular the GRUMP data (CIESIN et al., 2011), which adhere to the same raster resolution as SIDE. As an illustration, Figure 4 displays the GRUMP population data for Nigeria in 2000, the SIDE estimate of the Yoruba settlement pattern in 2013 and the combined estimate of the absolute distribution of Yorubas in Nigeria. Combining SIDE with spatial population data also permits calculating measures of ethnic diversity for arbitrary spatial units. Figure 5, for instance, plots the ethno-linguistic, religious and ethno-religious fractionalization in Nigerian districts.

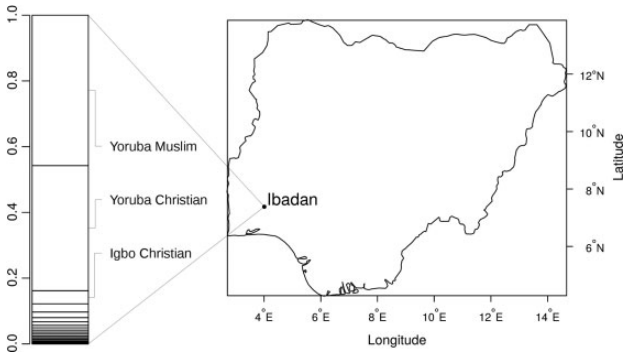

Figure 3. Point prediction: ethno-religious composition of Ibadan, Nigeria

Assessing the quality of SIDE

Our strategy for assessing the accuracy of SIDE is two-fold. First, we assess the face validity of the data by comparing them with existing polygon-encoded data on ethnic settlement patterns. Second, we use district-level census data from Uganda and Senegal to conduct a systematic test of the accuracy of SIDE.

Comparison with categorical maps

In the following, we compare the SIDE data for sub-Saharan Africa with the polygon data provided by the GeoEPR dataset (Wucherpfennig et al., 2011). To this end, we first match the ethnic groups enumerated in SIDE with those coded by GeoEPR.[7] For each SIDE-GeoEPR match, we calculate two quantities: (1) the proportion of the population within a group's GeoEPR polygon that SIDE codes as being part of that group, and (2) the proportion of the SIDE group-population located within the GeoEPR polygon. Figure 6 plots these two

[6] We choose this resolution in order to facilitate the combination of SIDE with the GRUMP population counts (CIESIN et al., 2011).

[7] For each country, we match the most recent SIDE data. Groups are matched primarily on the basis of equivalent names; in addition, we consulted online sources to match cases where a single EPR group corresponds to several SIDE groups.

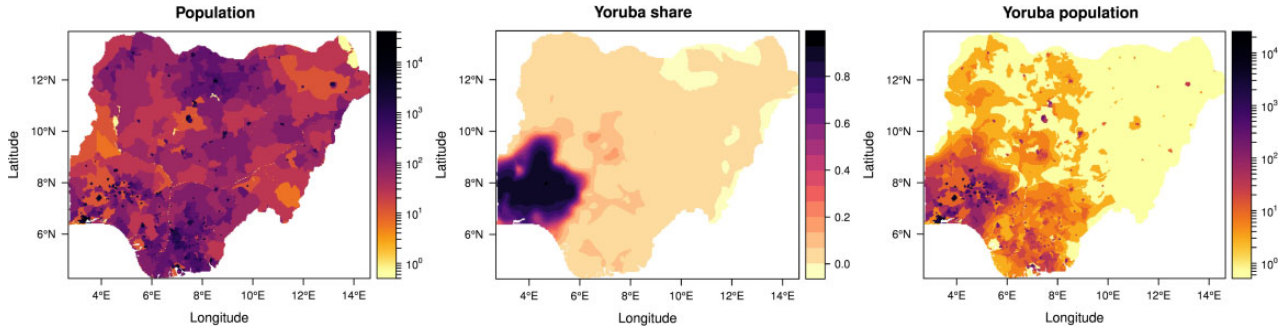

Figure 4. Constructing local ethnic population counts

The Yoruba population of each grid cell is computed as the product of its population count (CIESIN et al., 2011) and its Yoruba share from the SIDE data.

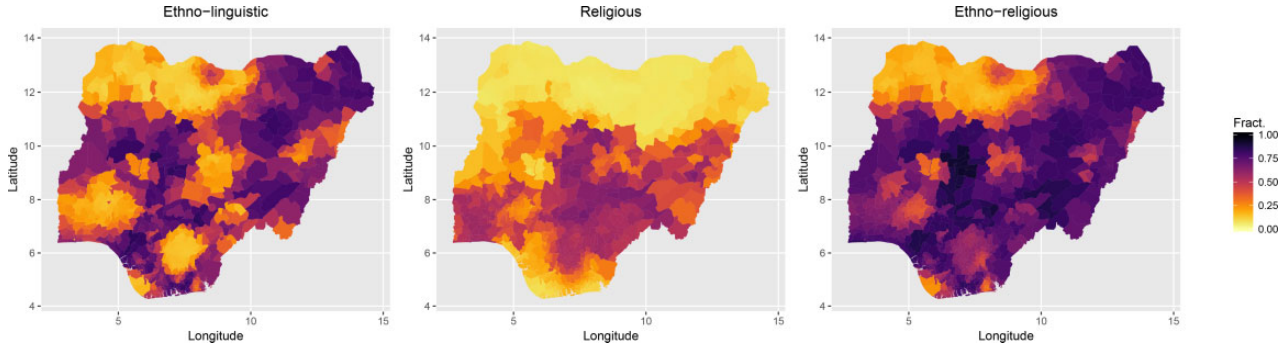

Figure 5. District-level ethno-linguistic, religious and ethno-religious fractionalization in Nigeria

Calculated based on SIDE, GRUMP (CIESIN et al., 2011), GAUL district borders (FAO, 2014) and the fractionalization formula in Alesina et al. (2003).

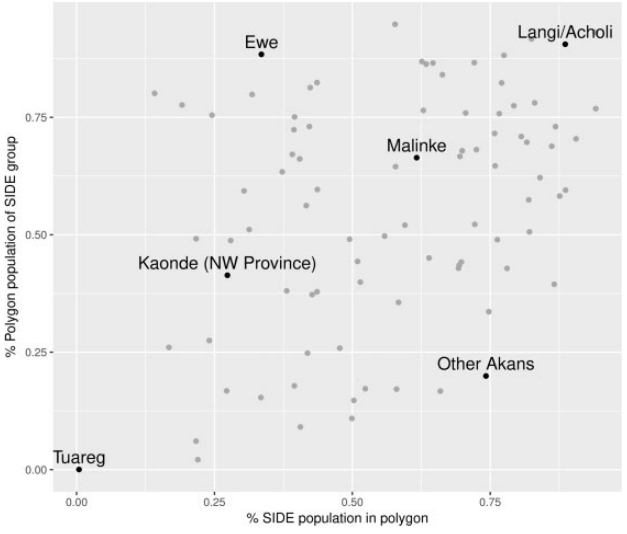

Figure 6. Comparison of SIDE with GeoEPR for sub-Saharan African ethnic groups contained in both datasets.

measures across all 104 matched groups. The upper-right quadrant of the plot thus contains those groups with the greatest overlap of the two data sources. It contains a significant share of groups (e.g. the Langi/Acholi of Uganda). However, those quadrants where there is less overlap between the two datasets are also populated: the lower-right corner of the plot contains groups for which the SIDE settlement pattern is contained within the GeoEPR polygon, but the latter also hosts a non-trivial number of individuals of other ethnicities. Conversely, the upper-left quadrant of the plot contains groups for which the GeoEPR polygon is populated almost exclusively by the respective SIDE group, but according to SIDE, the group's settlement area extends beyond its GeoEPR polygon. Finally, a few groups are located in the lower-left corner of the plot, where the SIDE and GeoEPR data exhibit little overlap.

To examine this variation in more detail, we select six representative groups from Figure 6 and map the

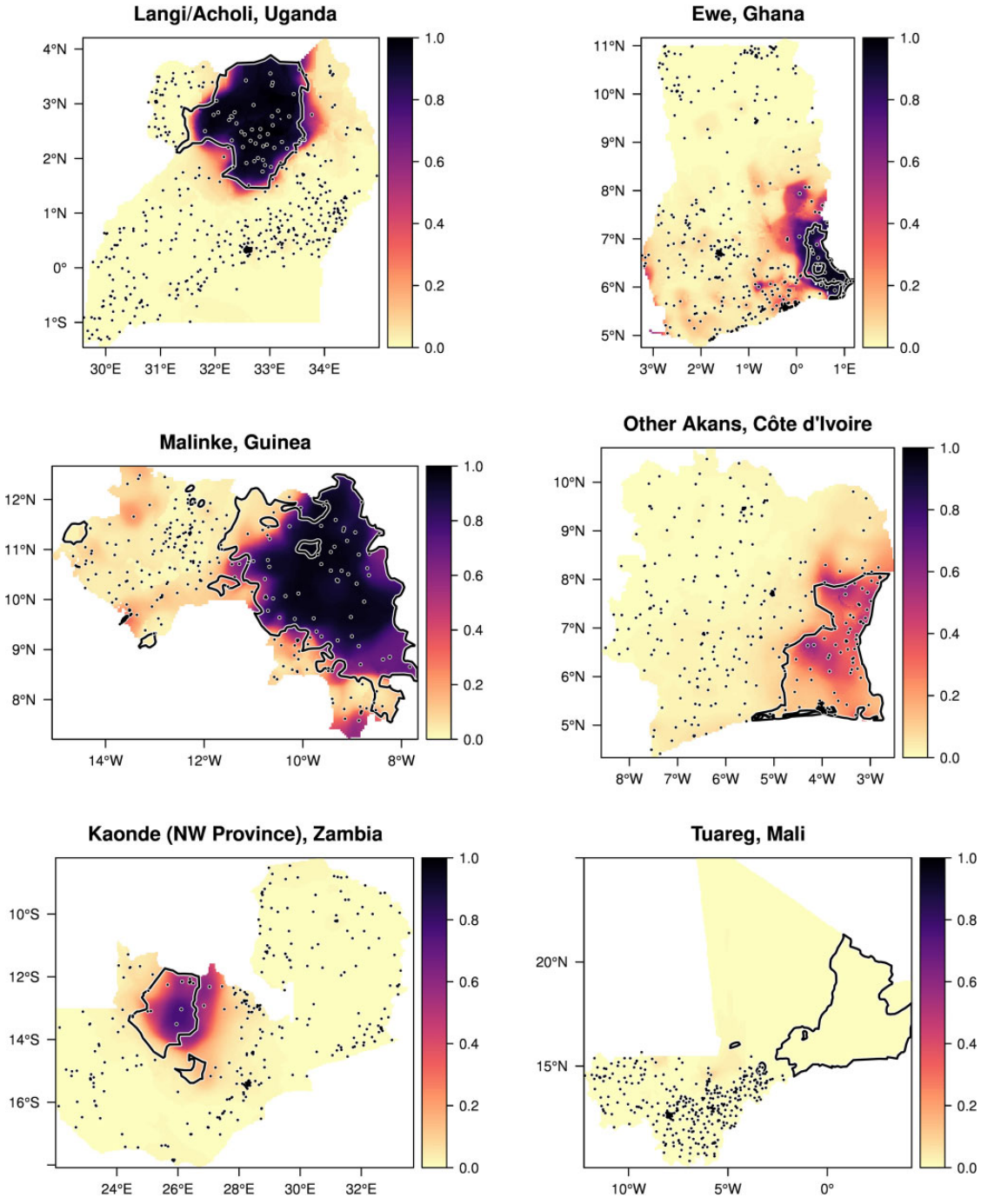

Figure 7. Comparison of SIDE estimates with GeoEPR polygons for selected ethnic groups
Lines indicate GeoEPR polygons; points indicate locations of DHS PSUs.

respective SIDE and GeoEPR data in Figure 7. Visual inspection suggests that, with the notable exception of the Tuareg in Mali, the SIDE estimates overlap heavily with the respective GeoEPR polygons, indicating that the two datasets convey the same underlying signal. The variation exhibited in Figure 6 originates primarily from cases like Ghana's Ewe, where the SIDE settlement area 'spills over' the respective GeoEPR polygon, or the 'Other Akan' group in the Ivory Coast, where GeoEPR and SIDE essentially agree, but where the group represents less than 100% of the local population. This pattern is precisely what we would expect if SIDE indeed constitutes an improvement over the GeoEPR data: the two datasets agree on the general location of groups, but

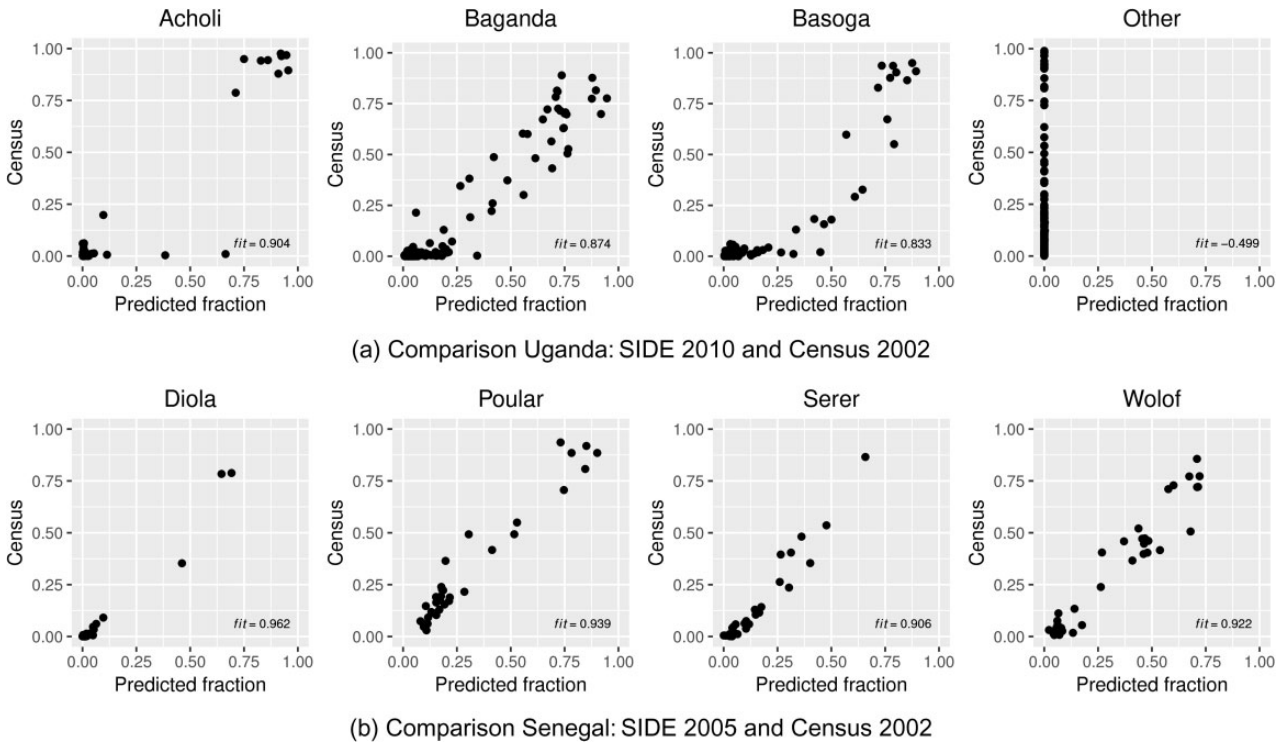

Figure 8. Comparison of SIDE data with ground truth census data from Uganda and Senegal
Fits are identical to the R^2 of a regression of the census data on the SIDE predictions with an intercept of 0 and a slope of 1.

the SIDE data suggest that ethnic settlement patterns are not as homogeneous or segregated as polygon data would suggest.

Finally, the idiosyncratic Tuareg case serves to highlight a risk coming with the SIDE data. As illustrated in Figure 7, the 2013 SIDE data for Mali fail to capture the Tuareg settlements because the respective DHS survey does not cover the respective area, which was controlled by insurgents at that time. We note, however, that we have not found any other case where a lack of DHS sampling points has led to similarly drastic errors in the SIDE data.

Comparison with Ugandan and Senegalese census data

In a next step, we compare the SIDE data to district-level census statistics from Uganda (2002) and Senegal (1988 and 2002).[8] To this end, we aggregate the population-weighted ethno-linguistic proportions from SIDE to second-level administrative units.[9] Next, we match the ethno-linguistic groups enumerated in the censuses to those of the SIDE data. The definitions of ethno-linguistic groups differ at times and lead to unavoidable inconsistencies in the matching.[10] The results reported below are therefore conservative estimates of SIDE's accuracy.

Comparing the SIDE estimates for Uganda with the census, we find that SIDE explains 81.4% of district-level variation in ethno-linguistic group sizes.[11] In Senegal, SIDE explains 96.1% (1988 census) and 97.8% (2002 census) of district-level variation. The lower accuracy of SIDE in Uganda is due to the country's greater number of ethno-linguistic groups and districts. The mismatch of the 'others' category between SIDE and the census introduces additional error (see Figure 8a).

Figure 8 illustrates that SIDE is most accurate for demographically large groups, such as the Acholi and

[8] For the comparison, we used the 2010 SIDE data for Uganda, and the 1992/2005 maps for Senegal. Census data were obtained from the Minnesota Population Center (2015).

[9] Population data were obtained from CIESIN et al. (2011), and administrative boundaries from FAO (2014).

[10] For instance, we match those groups present in the censuses, but not in the SIDE data, to the SIDE category of 'others'.

[11] This number is derived from averaging the ethnic composition explained at the district-level: $1 - \sum_{d=0}^{D}\left(\sum_{g=0}^{G}\frac{(z_{d,g}-\hat{z}_{d,g})^2}{(z_{d,g}-\bar{z}_g)^2}\right) * \frac{1}{D}$, where d indexes districts, g ethno-linguistic groups in district d of size $z_{d,g}$.

Baganda in Uganda, and the Poular and Serer in Senegal.[12] In each of these cases, the SIDE estimates account for between 75% and 96% of the district-level variation in group-sizes. In contrast, SIDE performs poorly on the 'Other' category, primarily because the group's definitions in the DHS and the Ugandan census do not coincide.

The census data also permit evaluating how SIDE compares to a district-level aggregation of the raw DHS data from 2010. Performing this comparison for Uganda, we first note that the sparseness of DHS PSUs leads to missing values for 47 of the country's 162 districts. Naturally, the near-continuous SIDE data avoid this issue. For the remaining 115 districts, the aggregated DHS data perform *worse* than the SIDE data, with 78.0% vs. 81.4% of district-level variation in the census data explained. This difference likely originates in that SIDE makes use of out-of-district information to predict local ethnic compositions.

In sum, we highlight two qualities of SIDE: first, SIDE captures considerable local variation in ethnic settlement patterns that polygon-based data cannot reflect. Second, SIDE provides reasonably accurate estimates of true, local-level ethnic group proportions as enumerated in census data.

Limitations

SIDE exhibits important limitations. These originate primarily from the characteristics of the DHS data that SIDE is based upon. First, DHS sampling may be not always be representative due to social phenomena, such as political violence. We test this possibility in the Online appendix but find only very weak evidence for undersampling in areas that experienced conflict in the year prior to a survey. However, we cannot conclude that sampling is always unbiased. To enable users of SIDE to single out potentially problematic cases, we provide the convex hull of DHS sampling clusters underlying each SIDE map. Second, even though for many countries SIDE covers multiple years, we caution against relying on this *temporal variation* for inferential purposes. Since DHS sampling units and ethnic groups' definitions vary over time, a substantial share of intertemporal variance in the SIDE data is random noise. Third, although the SIDE data are provided as high-resolution rasters, *very local variation* in the data may not be meaningful. This depends on (1) the local density of DHS clusters and (2) their random displacement of up to 2 km (10 km) in urban (rural) areas. Although enhancing the precision of SIDE may be possible by incorporating additional covariates during estimation, we refrain from doing so to avoid potential endogeneity issues in subsequent applications of the data.

Given these limitations, we encourage the use of SIDE for cross-national analyses that require consistent, cross-national data on local ethnic diversity, rather than single-country studies that call for high-precision data on ethnic demographies.

The effect of ethno-political polarization on local violence

By offering high resolution geographical data on ethnic compositions, SIDE contributes to spatially disaggregated conflict research (Cederman & Gleditsch, 2009). In particular and as explored in the following, the SIDE dataset adds an explicitly local dimension to the literature on the link between ethnic polarization and political violence (Montalvo & Reynal-Querol, 2005; Wilkinson, 2004).

The main theoretical argument to explain the impact of ethnic polarization on violent conflict focuses on increased interethnic competition for political power and economic resources. However, not all ethnic groups are politically mobilized in a uniform manner – some are politically irrelevant, while others form coalitions (Vogt et al., 2015). We therefore expect the degree of polarization between politically relevant ethnic groups to have a greater effect on conflict risk than pure ethnic polarization.[13] Furthermore, local ethno-political polarization is unlikely to have a uniform impact across all types of political violence. Rather, we hypothesize that local ethno-political polarization increases the risk of local conflict, such as riots, militia violence and one-sided violence (e.g. Wilkinson, 2004). We do not expect a strong association between local polarization and the occurrence of civil-war related events which are most likely related to macro-cleavages.

H1: Local polarization between politically relevant ethnic groups increases the risk of local conflict.

H2: Local polarization between politically relevant ethnic groups does not increase the risk of civil war at the local level.

[12] For a full comparison of all groups see the Online appendix.

[13] A politically relevant group is either (1) politically mobilized at the national level or (2) discriminated by the state (Vogt et al., 2015).

Table I. District-level ethno-political polarization & violence, linear probability models

	Dependent variable				
	Riot/demo SCAD (1)	*Milita SCAD* (2)	*Riot/demo ACLED* (3)	*One-sided ACLED* (4)	*Civil war ACLED* (5)
Ethno-pol. polar.	1.66**	0.41	2.16**	1.58†	0.39
	(0.45)	(0.43)	(0.67)	(0.93)	(0.96)
Population (log)	1.26**	0.50**	2.40**	1.72**	1.36**
	(0.19)	(0.15)	(0.29)	(0.33)	(0.35)
Urban pop. (%)	0.05**	0.02**	0.10**	0.08**	0.09**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Area (log)	–0.28*	0.28*	–0.25	1.03**	1.59**
	(0.13)	(0.13)	(0.19)	(0.26)	(0.26)
Country-year FE	yes	yes	yes	yes	yes
$spat.lag_{t-1,t-2}$	yes	yes	yes	yes	yes
$temp.lag_{t-1,t-2}$	yes	yes	yes	yes	yes
Observations	48,524	48,524	33,756	33,756	33,756
R^2	0.21	0.14	0.29	0.27	0.29

†$p < 0.1$; *$p < 0.05$; **$p < 0.01$. Standard errors are clustered on the district- and country-year-level.

Our empirical strategy for assessing the impact of local ethno-political polarization on violence consists of the following four elements (see Online appendix 5.1 for a detailed discussion):

1. The *unit of analysis* is the district-year in 22 African countries.[14]
2. The *dependent variable* is a dummy (0/100), indicating whether a district-year saw an event encoded in the SCAD (Salehyan et al., 2012) or ACLED (Raleigh et al., 2010) datasets. Events are recoded into the following categories: (1) riots/demonstrations, (2) militia violence, (3) one-sided violence and (4) civil war events.
3. The main *independent variable* is the ethno-political polarization index of a district-year. It is calculated on the basis of the most recent SIDE map matched to the EPR-ETH data that encode the political relevance of ethnic groups (Vogt et al., 2015). We add a vector of control variables and spatio-temporal lags.
4. Our *empirical model* consists in a linear probability model with country-year fixed-effects so that all variation constant within country-years is taken into account.

The results partly confirm our initial expectations (see Table I). The two indicators of riots and demonstrations from SCAD and ACLED are robustly associated with districts' degree of ethno-political polarization. Substantively, a one standard-deviation rise in a district's ethno-political polarization increases the probability of it experiencing a riot/demonstration in a given year by between .53 (SCAD) and .69 (ACLED) percentage points, with the average probability being 2.4% and 5.8%, respectively. The equivalent effect on one-sided violence is an imprecisely estimated increase of .5 percentage points that compares to an average probability of 7.8%.

Consistent with the argument that larger-scale political violence is not related to *local* dynamics in a mechanical manner, the occurrence of civil war- and militia-related events is not significantly associated with local ethno-political polarization.

These results are robust to a number of alternative specifications (see Online appendix 5.2). In particular, the degree of pure ethnic polarization does not drive the pattern. Also, the results are not caused by a single country with many and/or small districts. The described patterns also hold if districts (partly) outside the convex hull of DHS samples are excluded from the analysis to account for potential bias of the SIDE data.

Conclusion

This article introduces the SIDE data, a collection of 253 maps of local ethno-linguistic, religious and

[14] These are all countries covered by the ethno-linguistic maps in SIDE from 1990 to 2013, with the exception of Burkina Faso which has no 'politically relevant ethnic groups' according to EPR-ETH.

ethno-religious compositions in 47 low- and middle-income countries. SIDE improves on existing data projects by providing estimates of local ethnic diversity, rather than encoding ethnic settlement patterns in a binary manner. In applied research, this continuous representation of ethnic geographies yields three tangible benefits. First, it facilitates the measurement of group-characteristics via spatial operations, as it explicitly takes into account variation in groups' local population shares. Second, it permits calculating the extent and nature of local ethnic diversity for arbitrary spatial units. Third, the dataset's raster format makes it particularly easy to visualize.

Given these advantages, the SIDE dataset creates opportunities for expanding the discipline's understanding of the local dynamics of violence and other socio-economic phenomena. With its high-resolution data for many countries, SIDE contributes significantly to the possibilities of comparative cross-country research rooted in the micro level. Potential applications of the new data include analyses of the role of local ethnic identities for the occurrence of political violence, as well as research on the effects of local ethnic compositions on (ethnic) party politics, ethnic favoritism and interethnic trust. The data may also find applications in forecasting models of political violence. For practitioners, SIDE may prove useful as a source of politically neutral map material, for example to determine optimal sites for development programs.[15]

This article also introduces and evaluates spatial interpolation methodology as a tool for transforming geo-referenced survey data into smooth maps. We demonstrate that by combining very simple spatial interpolation models with a model tuning setup, we are able to produce near-continuous estimates of ethnic geographies from point-like survey clusters with surprisingly little error. We believe that this method has considerable potential beyond the scope of this article. Specifically, as geo-coded surveys are becoming increasingly common, spatial interpolation methods provide a powerful toolbox for making the most of this type of data.

Replication data

The dataset and R-code for the empirical analysis in this article, along with the Online appendix, can be found at http://www.prio.org/jpr/datasets and https://side.ethz.ch. All analyses have been conducted using R 3.1.

[15] However, keeping in mind that SIDE relies on *estimates*, we caution against using SIDE in cases where precise ethnic compositions are required at singular locations.

Acknowledgements

We thank the editor and three anonymous referees, John McCauley, and participants of the 2015 ENCoRe Meeting in Barcelona and the 2016 ISA convention for helpful comments and suggestions.

Funding

This research was supported by the Swiss National Science Foundation under grants P0EZP1_165233 and P2EZP1_168826 to Carl Müller-Crepon and Philipp Hunziker, respectively.

References

Aitchison, John (1986) *The Statistical Analysis of Compositional Data*. London: Chapman & Hall.

Alesina, Alberto; Arnaud Devleeschauwer, William Easterly & Sergio Kurlat (2003) Fractionalization. *Journal of Economic Growth* 8(2): 155–194.

Asal, Victor; Michael Findley, James A Piazza & James Igoe Walsh (2016) Political exclusion, oil, and ethnic armed conflict. *Journal of Conflict Resolution* 60(8): 1343–1367.

Buhaug, Halvard & Jan Ketil Rød (2006) Local determinants of African civil wars, 1970–2001. *Political Geography* 25(3): 315–335.

Burgert, Clara R; Josh Colston, Thea Roy & Blake Zachary (2013) Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys. *DHS Spatial Analysis Report* 7.

Cederman, Lars-Erik & Kristian Skrede Gleditsch (2009) Introduction to special issue on 'disaggregating civil war'. *Journal of Conflict Resolution* 53(4): 487–495.

Cederman, Lars-Erik; Nils B Weidmann & Nils-Christian Bormann (2015) Triangulating horizontal inequality: Toward improved conflict analysis. *Journal of Peace Research* 52(6): 806–821.

CIESIN, IFPRI, World Bank & CIAT (2011) Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Population Density Grid. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC).

DHS (2015) Demographic and Health Surveys. Integrated Demographic and Health Series (IDHS), version 2.0. Minnesota Population Center and ICF International (http://idhsdata.org, accessed on 28 July 2015).

FAO (2014) Global Administrative Unit Layers (http://data.fao.org/map?entryId=f7e7adb0-88fd-11da-a88f-000d939bc5d8&tab=metadata, accessed on 28 July 2015).

Fjelde, Hanne & Gudrun Østby (2014) Socioeconomic inequality and communal conflict: A disaggregated analysis of sub-Saharan Africa, 1990–2008. *International Interactions* 40(5): 737–762.

Fjelde, Hanne & Nina von Uexkull (2012) Climate triggers: Rainfall anomalies, vulnerability and communal conflict

in sub-Saharan Africa. *Political Geography* 31(7): 444–453.

Global Mapping International & SIL International (2015) World Language Mapping System. Version 17 (https://www.ethnologue.com/about/language-maps, accessed on 24 January 2018).

Lloyd, Christopher D (2010) *Local Models for Spatial Analysis*. Boca Raton, FL: CRC.

Mebane, Walter R & Jasjeet S Sekhon (2011) Genetic optimization using derivatives: The rgenoud package for R. *Journal of Statistical Software* 42(11): 1–26.

Minnesota Population Center (2015) Integrated Public Use Microdata Series, International: Version 6.4. Minneapolis, MN: University of Minnesota.

Mitas, Lubos & Helena Mitasova (1999) Spatial interpolation. In: Paul A Longley, Michael F Goodchild, David J Maguire & David W Rhind (eds) *Geographical Information Systems: Principles, Techniques, Management and Applications. Volume 1*. New York: Wiley, 481–492.

Montalvo, José G & Marta Reynal-Querol (2005) Ethnic polarization, potential conflict, and civil wars. *American Economic Review* 95(3): 796–816.

Murdock, George Peter (1967) *Ethnographic Atlas*. Pittsburgh, PA: University of Pittsburgh Press.

Nunn, Nathan & Leonard Wantchekon (2011) The slave trade and the origins of mistrust in Africa. *American Economic Review* 101(7): 3221–3252.

Raleigh, Clionadh; Andrew Linke, Håvard Hegre & Joakim Karlsen (2010) Introducing ACLED: An armed conflict location and event dataset. *Journal of Peace Research* 47(5): 651–660.

Salehyan, Idean; Cullen S Hendrix, Jesse Hamner, Christina Case, Christopher Linebarger, Emily Stull & Jennifer Williams (2012) Social conflict in Africa: A new database. *International Interactions* 38(4): 503–511.

Smith, Tony (2016) Notebook on spatial data analysis (http://www.seas.upenn.edu/ese502/#notebook, accessed on 10 April 2017).

USAID (2012) DHS Sampling and Household Listing Manual (http://dhsprogram.com/pubs/pdf/DHSM4/DHS6SamplingManualSept2012DHSM4.pdf, accessed on 15 July 2015).

Vogt, Manuel; Nils-Christian Bormann, Seraina Rüegger, Lars-Erik Cederman, Philipp M Hunziker & Luc Girardin (2015) Integrating data on ethnicity, geography, and conflict: The Ethnic Power Relations dataset family. *Journal of Conflict Resolution* 59(7): 1327–1342.

Weidmann, Nils B (2009) Geography as motivation and opportunity: Group concentration and ethnic conflict. *Journal of Conflict Resolution* 53(4): 526–543.

Weidmann, Nils & Kristian Skrede Gleditsch (2010) Mapping and measuring country shapes: The cshapes package. *R Journal* 2(1): 18–24.

Weidmann, Nils B; Jan Ketil Rød & Lars-Erik Cederman (2010) Representing ethnic groups in space: A new dataset. *Journal of Peace Research* 47(4): 491–499.

Wilkinson, Steven I (2004) *Votes and Violence*. Cambridge: Cambridge University Press.

Wucherpfennig, Julian; Nils B Weidmann, Luc Girardin, Lars-Erik Cederman & Andreas Wimmer (2011) Politically relevant ethnic groups across space and time: Introducing the GeoEPR dataset. *Conflict Management and Peace Science* 28(5): 423–437.

CARL MÜLLER-CREPON, b. 1989; PhD Candidate in Political Science at ETH Zurich (2015–).

PHILIPP HUNZIKER, b. 1987; PhD in Political Science (ETH Zurich); Postdoctoral Researcher at Northeastern University and Harvard University (2016–).