New Spatial Data on Ethnicity: Introducing SIDE

Online Appendix

1 The DHS data

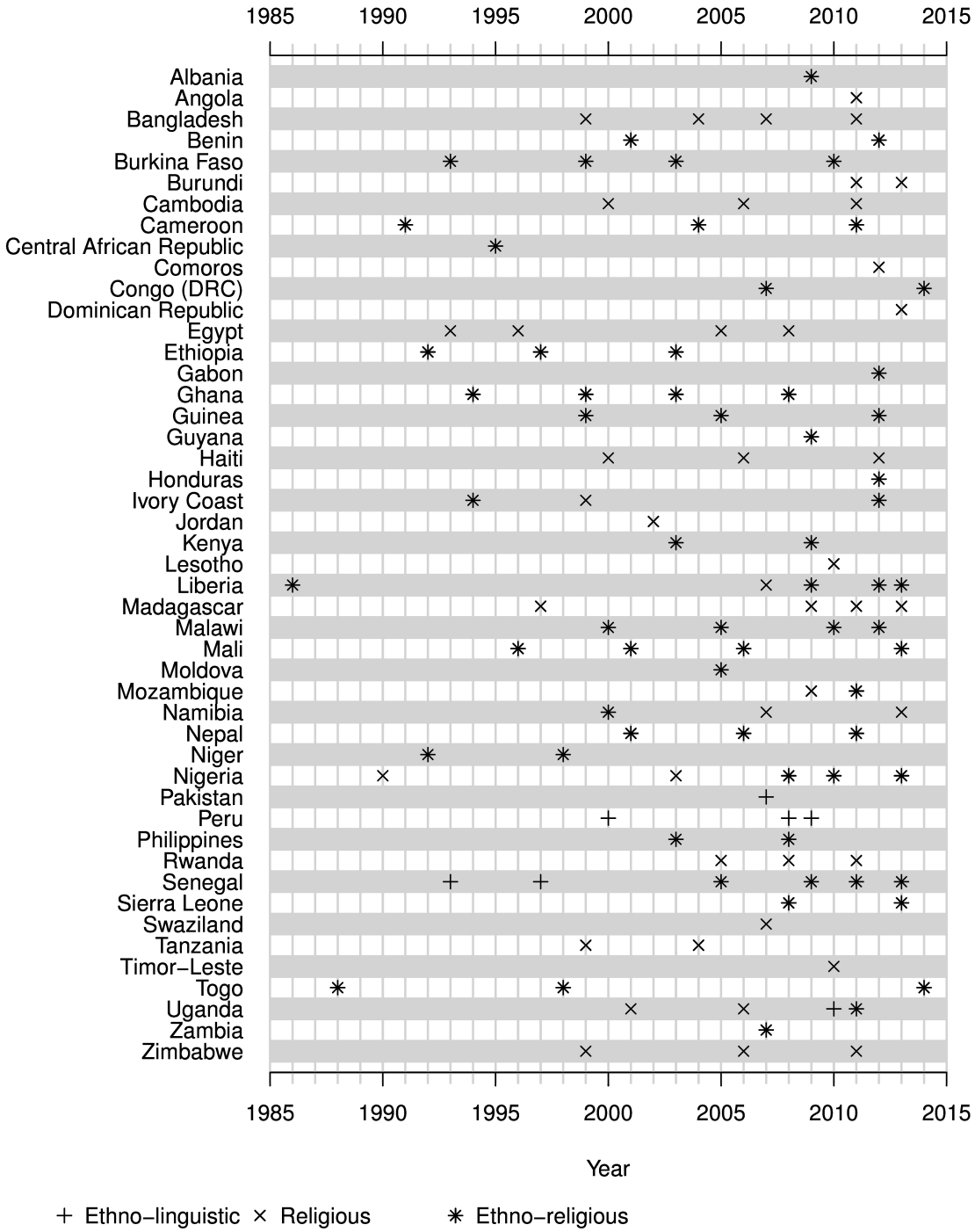


Figure 1. DHS surveys enumerating ethno-linguistic, religious, and ethno-religious identities that are used for generating the SIDE data.

Table I. Summary statistics: DHS surveys underlying SIDE

Statistic	N	Mean	St. Dev.	Min	Max
No. clusters	119	431.311	261.346	114	1,409
No. respondents	119	14,557.640	8,444.002	2,906	56,287
No. ethno-ling. groups	119	7.849	8.279	0	36
No. religious groups	119	5.429	2.510	0	12
No. ethno-rel. groups	119	19.580	20.829	0	75
No cluster / km^2	119	0.004	0.013	0.0001	0.140
Respondents / cluster	119	35.521	11.115	10.467	63.315

2 Compositional interpolation for ethnic survey data

The generalization of information sampled at discrete points in space is a well-known problem, and is usually referred to as spatial interpolation. More formally, in the 2-dimensional case, spatial interpolation pursues the following goal: Given a sample of points $s_i, \quad i = 1, 2, \ldots, N$, 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 the arbitrary target location s_0. These types of problems are commonly encountered in disciplines like bathymetry, topography, and geology, where spatial interpolation is routinely applied.

The suitability of any given spatial interpolation method is dependent on the type of problem and data at hand. Moreover, even once a method is chosen, most require additional, often non-trivial, and highly consequential modeling decisions. Since we do not have strong theoretically informed priors about the spatial distribution of ethnic compositions, we rely on a purely inductive approach for model tuning and selection. More precisely, we propose and implement a fairly generic model tuning algorithm that uses out-of-sample predictive performance to optimize and evaluate different spatial interpolation methods. In the following sections, we first introduce some basic concepts pertaining to spatial interpolation and compositional prediction, and then discuss our model tuning and selection scheme.

2.1 Global vs. local fitting

Spatial interpolation methods may be divided into local and global models (Lloyd, 2010). Global models use all available data to fit an interpolating equation that is applied uniformly across the plane to generate predictions.[1] In contrast, local models only use a subset of the available data to generate predictions for any particular point.

Though easier to implement and usually computationally less expensive, global models are often inadequate. They generally assume a globally uniform functional form and fixed parame-

[1] In the setup considered in this paper, "all available data" refers to all data points in a given survey.

ters, which may be too restrictive if the statistical properties of the outcome vary across space. Local models circumvent this problem by only considering local information for the generation of local predictions. Thus, given the target point s_0, we only use sample points in the vicinity of s_0 to estimate a model that yields a prediction $\hat{z}_0$. It follows that model fitting in the local case uses a restricted sample of size $N_0 < N$.

While local models are more flexible than global ones, they raise new difficulties. First, interpolation using local models is often computationally more expensive. Because in the local setting, the set of observations used for model fitting varies depending on the point for which a prediction is required, we usually have to estimate as many models as we want predictions. In the global case, only one model is needed. Second, local models require some method defining which observations should be employed for fitting the local model. There are essentially two options for this task:

- The *distance method* only considers observations that are within some distance D of the target point s_0.
- The *nearest neighbor method* only considers the K nearest neighbors of s_0.

In this paper, we employ local models, and use a combination of the above two methods to determine the composition of local samples. Specifically, we consider all observations that are either within distance D of the target point s_0, or belong to its K nearest neighbors. This approach ensures that regardless of the spatial clustering of observations in the neighborhood of s_0, we employ a local sample of at least size K, distributed across a sphere with a radius of at least D. Thus, varying the two parameters K and D jointly creates a very flexible set of possible local samples. Moreover, if D or K are sufficiently high, all sample observations are used, and we effectively estimate a global model. Because we have little a-priori information on what $D - K$ combination yields the best predictive performance, we use the model tuning algorithm introduced below to learn the parameters from the data.

2.2 Interpolating compositional data

Additional modeling decisions are introduced by the fact that in this project, we aim to predict compositions. Compositions feature a number of notable properties:

i *Multidimensionality:* At any point in space, the response is a vector $z_i = [z_{i1}, z_{i2}, \ldots, z_{iG}]$, with each z_{ij} representing the proportion associated with group j in location i, and G indicating the total number of groups in the sample.

ii *Boundedness:* Each element of the response vector z_i is in the unit interval, $z_{ij} \in [0, 1]$.

iii *Unit sum:* For each location i, $\sum_{j=1}^{G} z_{ij} = 1$.

These properties introduce additional difficulties because most spatial interpolation methods are designed for univariate continuous outcomes. Fortunately, however, the boundedness and unit-sum requirements can easily be met by preprocessing the data. In particular, we conduct all estimation tasks after applying the additive log-ratio transform (Aitchison, 1986, 92ff.), which yields a response vector of length $G-1$ where each element is defined on the real line, and no additive requirement exists:

$$z_i' = \left[ln\left(\frac{z_{i1}}{z_{iG}}\right), ln\left(\frac{z_{i2}}{z_{iG}}\right), \ldots, ln\left(\frac{z_{i(G-1)}}{z_{iG}}\right)\right]. \tag{1}$$

Thus, for all estimation tasks, we work exclusively with the transformed responses z', which can easily be mapped back to true compositions.

Unfortunately, in practice, applying the additive log-ratio transform introduces a new challenge: dealing with observed zero-proportions, i.e., cases where $z_{ij} = 0$. Whenever zero-proportions occur, the log-ratio transformed values are non-finite. We address this issue in two ways. First, before applying the additive log-ratio transform, we preprocess the local sample of N_0 points by censoring groups that are likely irrelevant for predicting the outcome at s_0, the target point. Specifically, we determine the $L \leq N_0$ sampling points closest to s_0, and eliminate those groups from the local sample whose measured proportion never exceeds zero in any of these L neighbors. When predicting the composition for point s_0, these censored groups are simply assigned a predicted value of zero, without any further fitting or estimation. Not only does this procedure eliminate many zero-proportions in the sample, but it also eliminates those groups that are irrelevant for the compositional prediction at point s_0. L, the size of the neighborhood determining which groups to censor, is learned from the data using the optimization procedure discussed below (Section 2.4).

Second, to those zero-proportions that remain in the sample after this first censoring step, we assign positive placeholder values ξ_{ij}. Then, to ensure that the unit-sum restriction is still met at all locations, the remaining proportions are again adjusted accordingly. In principal, one could use a constant as a placeholder value, so that $\xi_{ij} = \xi$. However, we choose a more informed approach, and calculate a separate placeholder value for each observed zero-proportion.

More precisely, we choose the placeholder values in a two-stage procedure:

1. Calculate $r_i \in (0, 1]$, which is the total proportion assigned to groups with observed zero-proportion at location i. That is, if V_i is the set of all groups with an observed zero-

proportion in location i, then $\sum_{j \in V_i} \xi_{ij} = r_i$. r_i is defined as

$$r_i = 1 - q^{n_i^{-1}}, \tag{2}$$

whereas n_i is the number of individuals surveyed at location i, and $q \in (0, 1)$ is a sensitivity parameter.[2] This formulation has the desirable property that as the number of surveyed individuals in a particular location grows, the placeholder values assigned to groups with zero-proportions shrinks. Thus, with growing local survey size, we become ever more certain that groups with no members in the sample are indeed very small in reality. The sensitivity parameter q controls the shape of the function, with larger q values producing smaller r_i values, and vice-versa. We learn q from the data using the model tuning algorithm introduced below.

2. Allocate r_i to the groups with observed zero-proportions. This step is performed using the groups' relative sizes across all locations in the local sample. Specifically, let

$$v_j = \frac{\sum_{i=1}^{N_0} n_i * z_{ij}}{\sum_{i=1}^{N_0} n_i} \tag{3}$$

be group j's relative size across all locations in the local sample. Then, for each group $j \in V_i$, the placeholder value is set to

$$\xi_{ij} = \frac{v_j * r_i}{\sum_{j \in V_i} v_j}. \tag{4}$$

Thus, in summary, we deal with the boundedness and unit-sum aspects of the compositional data by preprocessing the local sample of N_0 points in three steps: (1) Remove irrelevant groups, (2) assign placeholder values to zero-proportions, and (3) apply the additive log-ratio transform to the compositional response vectors.

A final challenge is the fact that the log-ratio transformed z' response is still multidimensional. Some spatial interpolation methods support multivariate responses. In particular, Kriging is easily generalizable to multivariate outcomes, and even exists in a variant specifically targeted at compositional predictions (Walvoort & de Gruijter, 2001; Pawlowsky-Glahn & Olea, 2004). However, extensive tests by the authors have yielded that local compositional Kriging is associated with prohibitively high computational costs.[3] For this reason, we adopt the simpler

[2] Expression (2) also has an intuitive interpretation: r_i is the true relative size of a group that is *not* included in a random sample of n_i individuals with probability $1 - q$.

[3] The optimization and training of compositional Kriging models for a single DHS survey would have taken a full day, even if executed on a high-performance computing cluster.

approach of performing separate estimations for each dimension in z'. That is, we employ interpolation methods for univariate outcomes, and estimate a separate model for each log-ratio in z'. These separate models are then used for creating log-ratio predictions at the target point s_0, which are then translated back into true compositions using the inverse function of the additive log-ratio transform.

2.3 Candidate models

We consider two interpolation methods for predicting log-ratios: a simple exponential distance decay weighting scheme, and the more complex thin plate smoothing splines model.

2.3.1 Exponential distance decay

The exponential distance decay (EDD) method assigns a given target point a weighted average of the response values of all other points in the (local) sample (Smith, 2016, II.5-3). The weights are constructed using an exponential distance decay function that assigns sample points that are further away from the target point less influence. More precisely, the EDD prediction for the target point s_0 is

$$z'_0 = \frac{\sum_{i=1}^{N_0} e^{-\omega d_{i0}} z'_i}{\sum_{i=1}^{N_0} e^{-\omega d_{i0}}}, \tag{5}$$

whereas z' is the (univariate) log-ratio to be predicted, d_{i0} is the Euclidean distance between point s_i and the target point s_0, and $\omega > 0$ is a decay parameter determining how influential observations distant from the target point are. ω is an unknown free parameter which we learn from the data using the optimization scheme described below (Section 2.4). Note that this implies that we fit a single ω parameter for the entire sample of size N. Thus, the only "local" component of the EDD model, as employed here, is the size of the sample used for estimation, N_0.

The EDD method has two important advantages: First, because it necessitates no local optimization, predictions can be obtained *very* efficiently, even if the local sample is large. This is a desirable property because it leaves more computing power for global optimization, that is, selecting parameters like neighborhood sizes and those introduced in the previous section for preprocessing purposes. Second, EDD interpolations are conservative in the sense that the model only allows predictions within the range of the observed values. Thus, the model prevents bias due to the unwarranted extrapolation of local trends.

2.3.2 Thin plate splines

The thin plate spline (TPS, see e.g. Lloyd 2010, 158ff.) method is a multivariate smoothing approach that fits the spline function g to an observed sample of points by minimizing the objective function

$$\sum_{i=1}^{N_0}(z_i' - g(s_i))^2 + \lambda \int\int \left[\left(\frac{\partial^2 g}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 g}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 g}{\partial y^2} \right)^2 \right] dxdy. \tag{6}$$

The first term in expression (6) controls the model's goodness-of-fit and is equivalent to the well-known least-squares criterion. The second term is the 2-dimensional integral over the squared second derivative of the function g, and thus acts as a wiggliness-penalty. λ is a non-negative penalty parameter that determines the trade-off between fit and smoothness.

The function g is given by

$$g(s) = a_0 + a_1 x + a_2 y + \sum_{i=1}^{N_0} c_i \phi(s_i, s), \tag{7}$$

whereas a_0, a_1, a_2, and $c_1, c_2, \ldots, c_{N_0}$ are the $N_0 + 3$ coefficients to be estimated, and

$$\phi(s_i, s) = d_i^2 ln(d_i) \tag{8}$$

is the TPS radial basis function. Note that d_i denotes the Euclidean distance between s_i and s. Further, note that the TPS fit is a combination of a linear trend, parametrized by a_0, a_1, and a_2, and the weighted spline components, $\phi(s_i, s)$, which may distort the linear fit locally. The penalty parameter λ determines how much local distortion is allowed.

Given λ, there exists a closed-form solution for the a and c parameters that minimizes (6). The λ penalty is obtained via generalized cross-validation on the local sample N_0.

The key advantage of the TPS approach is its flexibility: it allows recovering both completely linear as well as highly complex response surfaces, while local cross-validation prevents overfitting. Its primary shortcoming is that it is computationally expensive, as for every new target point s_0 with its associated local sample, a new TPS fit needs to be estimated, which is a notoriously slow process (Wood, 2003).

2.3.3 Alternative models

Among the many models for spatial interpolation discussed in the literature, two stand out as possible alternatives to the approach followed here:

1. *Inverse distance weighting:* A popular alternative to the EDD method is the closely related *inverse* distance weighting (IDW) scheme, whereby the weights are constructed as $d_{i0}^{-\omega}$ rather than $e^{-\omega d_{i0}}$ (Lloyd, 2010, 154ff.). We employ EDD instead of the IDW approach mainly for computational reasons: we have found that the two weighting methods produce almost equivalent results, but the exponential variant used in the paper prevents numerical overflow issues that arise with the IDW method when $d_{i0} \approx 0$.

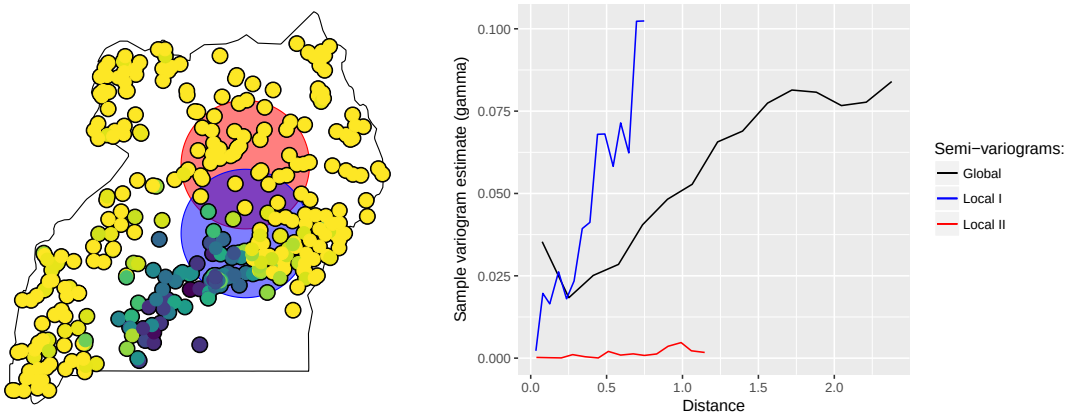
2. *Kriging:* Kriging involves the fitting of a global variogram that captures the structure of spatial covariance between the input data points. In principle, Kriging is a suitable alternative to our approach. In fact, scholars have proposed Kriging methods specifically targeted at compositional data (Walvoort & de Gruijter, 2001; Pawlowsky-Glahn & Olea, 2004). However, by relying on a set of global parameters, traditional Kriging is not able to recover potential non-stationarity in the spatial covariance of the data (see Figure 2 for an illustration). To address this problem, some authors have proposed local moving window Kriging (LMWK), whereby variograms are fitted only on local data within some window of size D_{Krig} (Atkinson & Lloyd, 2007; Lloyd, 2010, 2012, 2015). Because we do not know D_{Krig} ex-ante, however, we would have to learn it using some search heuristic, similar to the strategy we discuss below. Unfortunately, however, the computational costs associated with this strategy are simply prohibitive, even if computed in parallel. Our current approach, relying on computationally efficient interpolation methods, already takes between 30min and 3h per map if run in parallel on 114 CPU cores on a high-performance computing cluster. Compositional Kriging would take several orders of magnitude longer. One of the reasons why compositional Kriging is so costly is that computation time increases exponentially in the number of groups. The complexity of our approach, in contrast, increases linearly in the number of groups.

2.4 Model tuning

Before the EDD and TPS methods can be applied for interpolation, a number of global parameters need to be determined:

- The D and K parameters, which regulate the composition and size of the local samples used for estimation.

- The L parameter, which determines which groups are retained in the local sample, and which are deemed locally irrelevant.



(a) % Baganda in survey clusters (b) Local and global semivariograms

Figure 2. Comparison of global and local semivariograms for the ethnic Baganda in Uganda. Red (blue) area in (a) corresponds to red (blue) local semivariogram in (b). Data comes from DHS round 6.1 in 2011. Proportion of ethnic Baganda in DHS clusters in (a) range from 0 (light) to 100% (dark).

- The q parameter, which moderates the size of the placeholder values assigned to observed zero-proportions prior to applying the additive log-ratio transform.
- For the EDD model: the ω decay parameter, which governs how much influence is assigned to observations that are remote from the target point.

We determine these parameters by optimizing their out-of-sample predictive performance, estimated using leave-one-out cross-validation (LOOCV). In other words, we attempt to find the set of global parameters that are associated with the most accurate local EDD and TPS predictions, respectively.

In a first step, we randomly draw an evaluation sample of size $N_t < N$ from the full sample of points.[4] We only assess predictive performance for these evaluation points in order to decrease the computational burden associated with optimization. Next, let θ_m be the global parameter vector required to perform interpolation using model m, whereas m refers to the TPS and EDD models, respectively.[5] Then, the global parameters are chosen as follows,

$$\hat{\theta}_m = \underset{\theta_m}{\operatorname{argmin}}\, \epsilon_m, \tag{9a}$$

$$\epsilon_m = N_t^{-1} \sum_{l=1}^{N_t} \sum_{j=1}^{G} (z_{lj} - \hat{z}_{lj,m}^{CV}(\theta_m))^2, \tag{9b}$$

whereas $\hat{z}_{lj,m}^{CV}(\theta_m)$ is the LOOCV prediction for group j in point s_l. Hence, it is derived from

[4]Note that in practice, the "full sample" refers to a single DHS survey. That is, we estimate separate sets of global parameters for each survey.

[5]Thus, $\theta_{EDD} = [D, K, L, q, \omega]$ and $\theta_{TPS} = [D, K, L, q]$.

a sample where s_l is *not* used for estimation, thus ensuring that we evaluate *out-of-sample* predictive performance. Note that the error ϵ_m in (9) evaluates predicted proportions $\hat{z}$, rather than the log-ratio transformed values. Hence, before assessing performance, we retransform the $\hat{z}'$ values yielded by the EDD/TPS models back into proper proportions, simply because we are interested in how well we are able to predict actual compositions, not the log-transformed values that are introduced for purely methodological reasons. Further, as is evident from expression (9), to quantify performance, we use the mean-squared-error loss function, applied to each proportion for every point of the evaluation sample.

(9) is a non-trivial optimization problem. The L and K parameters are discrete, all parameters are constrained, and the objective function is likely irregular, thus prohibiting derivative-based optimization routines. For this reason, we employ the genetic optimization algorithm implemented by Mebane & Sekhon (2011) in the R programming language. This approach serves our purposes because it supports discrete parameter spaces, allows optimizing arbitrarily irregular functions, and permits utilizing parallel computing infrastructure with minimal overhead.

2.5 Model selection and prediction

Once optimal global parameters are established, we are able to generate predictions for any given point on the plain. The only remaining question, then, is whether to use the EDD or TPS predictions. One viable approach would be to perform model selection based on out-of-sample predictive performance. That is, rely exclusively on the predictions of the model yielding the smallest out-of-sample error, as established by the optimization routine described in the previous section. However, a substantial body of machine learning literature suggests model mixing as an alternative and superior approach (see e.g. Hastie, Tibshirani & Friedman, 2005, 288ff.). Here, instead of relying on one model exclusively, the predictions of numerous models are averaged.

We adopt this model mixing approach and follow the standard practice of using the two candidate models' estimated out-of-sample performance as weights when averaging their predictions. Moreover, instead of using the same weights across all predicted proportions, we employ group-level weights. Thus, when averaging the proportional prediction for a given group at a given location, we assign more influence to the prediction of the model that performs best for the group in question. This approach accounts for the possibility that the TPS and EDD models may perform differently for specific types of groups. This would be the case, for instance, if the more conservative EDD model systematically produced better predictions for groups that are consistently small across space.

More formally, our model mixing scheme proceeds in three steps. First, based on a second evaluation sample, we calculate group-level out-of-sample goodness-of-fit estimates introduced in the previous section:

$$u_{j,m} = 1 - \frac{\sum_{l=1}^{N_t}(z_{lj} - \hat{z}_{lj,m}^{CV}(\hat{\theta}_m))^2}{\sum_{l=1}^{N_t}(z_{lj} - \bar{z}_j)^2}, \tag{10}$$

whereas $\bar{z}_j$ is the mean proportion associated with group j across the evaluation sample. Thus, $u_{j,m}$ is similar to the well-known R^2 metric. The numerator of the second term is the sum of squared errors associated with model m's predictions for group j, whereas denominator is the variance in the true proportions for group j across all locations in the evaluation sample. Thus, $u_{j,m}$ measures the reduction in observed group-level variance associated with the predictions of model m. It follows that if the predictions are perfectly accurate, $u_{j,m} = 1$, whereas $u_{j,m} < 0$ if our models systematically perform worse than if we had simply used group j's mean-proportion as a prediction.

Second, given these group-level goodness-of-fit metrics, we calculate the weights

$$\zeta_{j,m} = \frac{u_{j,m}}{u_{j,EDD} + u_{j,TPS}}. \tag{11}$$

Finally, we calculate the final mixed prediction for group j at the arbitrary point s_0 as

$$\hat{z}_{0j}^{MI} = \frac{\zeta_{j,TPS} * \hat{z}_{0j,TPS} + \zeta_{j,EDD} * \hat{z}_{0j,EDD}}{\sum_{j=1}^{G}(\zeta_{j,TPS} * \hat{z}_{0j,TPS} + \zeta_{j,EDD} * \hat{z}_{0j,EDD})}, \tag{12}$$

whereas $\hat{z}_{0j,m}$ is the proportional prediction for group j at point s_0 obtained with model $m = \{\text{TPS, EDD}\}$ and the optimized set of global parameters $\hat{\theta}_m$. The denominator of (12) is necessary to ensure that the unit-sum requirement is met after model mixing.

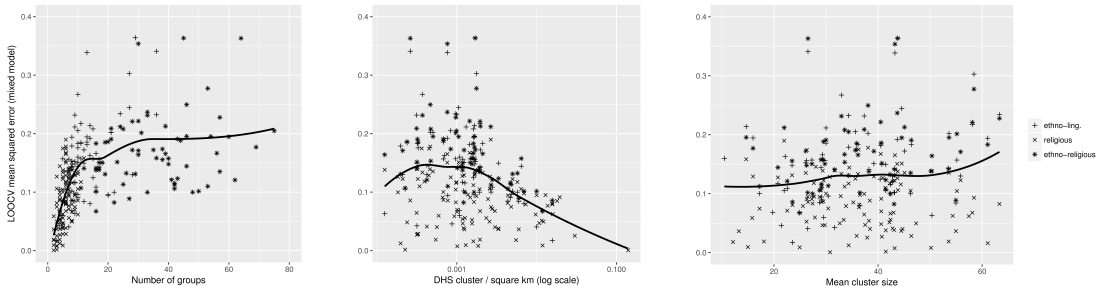
2.6 Model optimization

The global parameters of the two spatial interpolation models – exponential distance decay and thin plate spline estimation – are tuned following the above presented optimization procedure.[6] Out of each DHS cluster sample, we randomly draw a sample of evaluation points of size $N_t = 0.25 * N$ to perform the leave-one-out cross-validation (LOOCV, see Subsection 2.4).[7] Solving

[6]The genetic optimization procedure is carried out with the following parameters:

- Initial population size: 40 (EDD) and 20 (TPS)
- Hard maximum number of generations: 50
- Convergence threshold: no improvement of over $1 * 10^{-3}$ in 5 consecutive generations

[7]Since we parallelize the leave-one-out predictions over 114 CPUs on a high-performance computing cluster, the number of training points is rounded up to multiples of the number of CPUs in order use all available computing power.



(a) No. of groups (b) Cluster density per km^2 (c) Mean respondents per cluster

Figure 3. Correlates of LOOCV mean squared error (mixed model). Points indicate fitted models, lines result from a smoothing local polynomial estimation.

the optimization problem brings about major computational challenges, in particular in the case of the thin plate splines (TPS). Even with a reduced number of evaluation points, the time it takes to find the optimal set of global parameters for *one* DHS sample is prohibitive if the procedure is run on an office computer (between 60 and 360 hours). Hence, we parallelize the entire optimization procedure using 120 CPUs on a high-performance computing cluster. Thereby, we significantly reduce the time needed to tune the models to between 0.5 and 3 hours.

2.7 Estimation results

2.7.1 Determinants of SIDE model fits

We analyze the impact of likely determinants of the LOOCS mean squared error in order to better understand which characteristics of the DHS input data are likely to affect the quality of the SIDE data. The three most important factors are likely to be (1) the number of ethnic groups in a DHS survey, (2) the spatial density of the DHS data, and (3) the average number of respondents per DHS cluster. The correlations plotted Figure 3 in are in line with the intuition that predictions of local ethnic compositions are more error-prone if there (1) are many ethnic groups to be predicted and (2) if the spatial density of DHS clusters is low. Importantly however, the average number of respondents per DHS cluster does not strongly influence the model fit. This corroborates the assumption that we can treat DHS clusters as being on average representative of a locality's ethnic population and thus providing valuable information to the model, even if the number of respondents is very small.

Statistically more rigorous, Table II explores the correlates of the LOOCV mean squared errors using a simple linear model. The results reaffirm that the fit of the models decreases in the number of ethnic groups and increases in the density of DHS point data fed to the model. Furthermore, models of ethno-linguistic compositions have a, ceteris paribus, lower fit than

models of religious or ethno-religious compositions. All other potential correlates, such as the number of respondents per survey cluster, the total area of a country or the simple number of survey clusters or respondents turn out to have negligible effects.

Table II. Explaining LOOCV model fits across models

	Dependent variable:		
	ϵ_{EDD}	ϵ_{TPS}	ϵ_{MI}
	(1)	(2)	(3)
Ethno-religious map	-0.038^{**}	-0.039^{**}	-0.035^{**}
	(0.012)	(0.013)	(0.012)
Religious map	-0.055^{**}	-0.053^{**}	-0.052^{**}
	(0.008)	(0.009)	(0.008)
No. ethnic groups	0.002^{**}	0.002^{**}	0.002^{**}
	(0.0004)	(0.0004)	(0.0004)
No cluster / km^2	-0.987^{**}	-1.003^{**}	-0.961^{**}
	(0.361)	(0.374)	(0.356)
Respondents / cluster	-0.0004	-0.0002	-0.0003
	(0.001)	(0.001)	(0.001)
Area (km^2)	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
No. clusters	-0.0001^{**}	-0.0001^{*}	-0.0001^{**}
	(0.00004)	(0.00004)	(0.00004)
No. respondents	0.00000	0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)
Constant	0.169^{**}	0.163^{**}	0.161^{**}
	(0.024)	(0.025)	(0.024)
Observations	250	251	253
R^2	0.465	0.457	0.455
F Statistic	26.234^{**} (df = 8; 241)	25.508^{**} (df = 8; 242)	25.491^{**} (df = 8; 244)
Note:			†p<0.1; *p<0.05; **p<0.01.

2.7.2 Determinants of optimized parameters

Next, we assess the face-validity of the optimized parameters used to generate each SIDE map to ensure that our optimization procedure, though in principle well founded in geo-statistical theory, does not produce unreasonable results. Figure 4 plots the distributions of all optimized parameters across ethnic, religious, and ethno-religious maps. The distributions show that the optimized parameters can only in a few cases be characterized as outliers. In some cases, these outliers constitute corner-solutions to the parameters (e.g. $D_{TPS} = 2490km$ for the ethnic

SIDE in Mozambique covers the entire country). Importantly, the distributions also show that optimized parameters vary between SIDE maps. This vindicates our initial assumption that there is no global set of parameters that is locally optimal.

But are the locally optimal parameters determined by a set of characteristics of the DHS data itself? Table III presents the results of simple OLS models that regress each parameter on the most important characteristics of DHS samples, the unit of analysis being the SIDE map. As one would expect, different characteristics of the data co-determine different parameters. Intuitively, for example, the minimal number of clusters used to determine the local neighborhood (K_{EDD} and K_{TPS}) increases in the number of clusters enumerated in a survey. Also, the results show that the size of local windows used in the EDD models (K_{EDD} and D_{EDD}) is significantly bigger for estimating religious than ethnic maps. This probably relates to the fact that there are most often less religious than ethnic groups in a country, so that points distant from each other carry more religious than ethnic information for the prediction. In general, the fit of the regression models ranges between an R^2 of .2 and .5. The models thus explain substantial parts of the variation among the optimized parameters.

It must however be remembered, that these models can only be constructed with the information on the optimized parameters. It seems unlikely that any of the regression coefficients could have been 'guessed' ex ante to circumvent the optimization procedure.

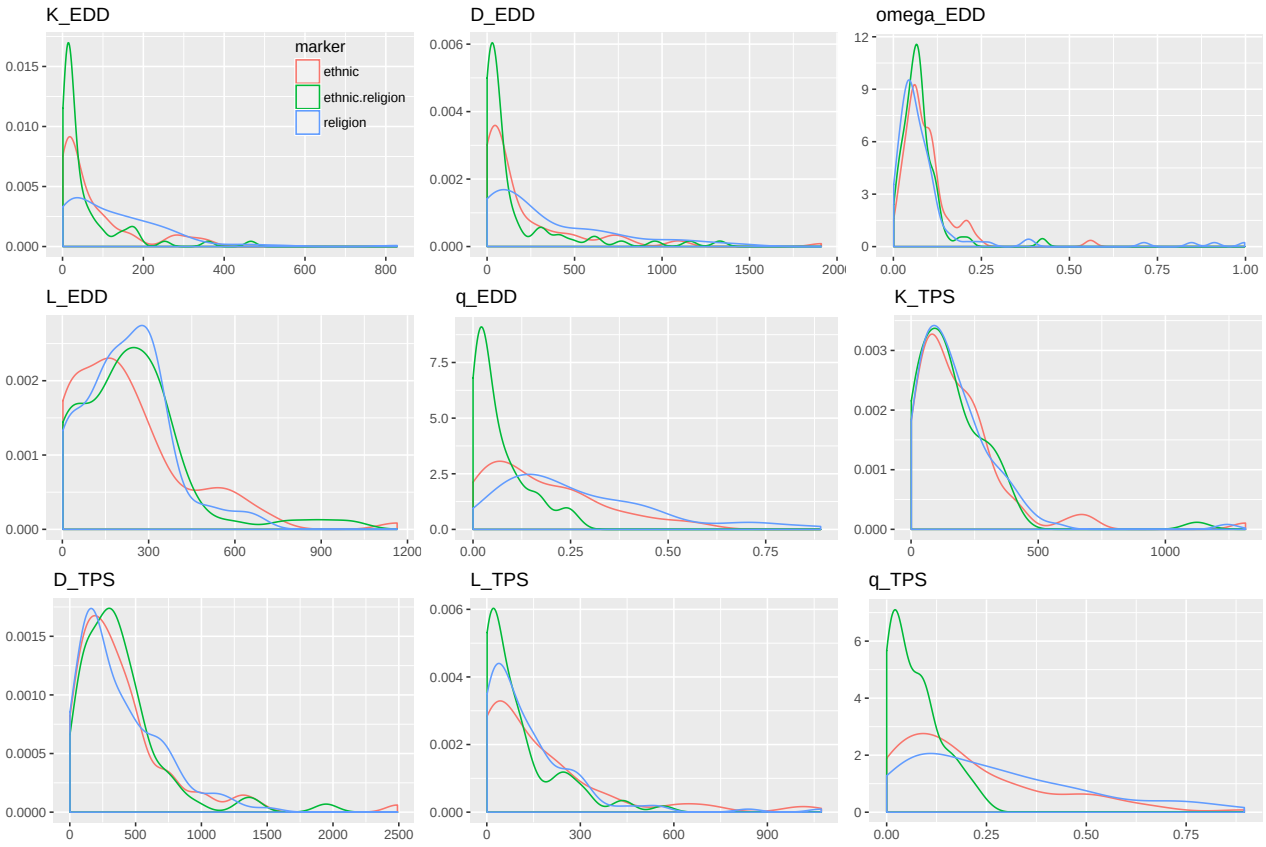


Figure 4. Distribution of optimized parameters over estimated models.

Table III. Explaining optimized parameter values

	Dependent variable:								
	K_{EDD}	D_{EDD}	ω_{EDD}	L_{EDD}	q_{EDD}	K_{TPS}	D_{TPS}	L_{TPS}	q_{TPS}
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Ethno-religious map	28.334	8.681	−0.001	−4.313	−0.015	−18.189	−53.400	−33.256	−0.057
	(23.813)	(70.258)	(0.024)	(33.903)	(0.030)	(35.098)	(65.161)	(32.247)	(0.039)
Religious map	44.984**	109.541*	−0.011	20.165	0.062**	−2.957	18.036	−31.437	0.048†
	(16.265)	(47.988)	(0.016)	(23.157)	(0.020)	(23.973)	(44.507)	(22.026)	(0.026)
No. groups	−1.949**	−2.839	−0.001	1.424	−0.004**	0.203	2.140	−0.839	−0.003*
	(0.740)	(2.182)	(0.001)	(1.053)	(0.001)	(1.090)	(2.024)	(1.002)	(0.001)
No cluster / km^2	−121.193	−2,118.400	6.283**	−1.109	4.686**	487.963	−1,291.107	155.632	4.469**
	(710.299)	(2,095.666)	(0.710)	(1,011.282)	(0.890)	(1,046.908)	(1,943.651)	(961.876)	(1.154)
Respondents / cluster	2.377†	−0.997	0.001	−1.253	0.001	1.711	2.582	3.491*	0.001
	(1.234)	(3.642)	(0.001)	(1.758)	(0.002)	(1.819)	(3.378)	(1.672)	(0.002)
Area (km^2)	0.00003†	0.0003**	0.000	−0.00003	0.00000	0.00001	0.0004**	−0.00000	0.00000†
	(0.00001)	(0.00004)	(0.000)	(0.00002)	(0.00000)	(0.00002)	(0.00004)	(0.00002)	(0.00000)
No. clusters	0.170*	−0.270	0.0001*	0.373**	0.0005**	0.426**	−0.111	0.763**	0.0004**
	(0.073)	(0.216)	(0.0001)	(0.104)	(0.0001)	(0.108)	(0.201)	(0.099)	(0.0001)
No. respondents	−0.002	0.002	−0.00000	0.004	−0.00001**	−0.002	0.002	−0.012**	−0.00001**
	(0.002)	(0.006)	(0.00000)	(0.003)	(0.00000)	(0.003)	(0.006)	(0.003)	(0.00000)
Constant	−43.639	241.325†	0.050	48.971	0.122*	−37.238	110.309	−88.273	0.160*
	(47.407)	(139.871)	(0.047)	(67.496)	(0.059)	(69.874)	(129.725)	(64.198)	(0.077)
Observations	253	253	253	253	253	253	253	253	253
R^2	0.197	0.209	0.286	0.373	0.491	0.249	0.346	0.383	0.333
F Statistic (df = 8; 244)	7.485**	8.075**	12.207**	18.145**	29.478**	10.121**	16.107**	18.925**	15.244**
Note:									†p<0.1; *p<0.05; **p<0.01.

3 Assessing the quality of SIDE

3.1 Comparison with Ugandan and Senegalese census data

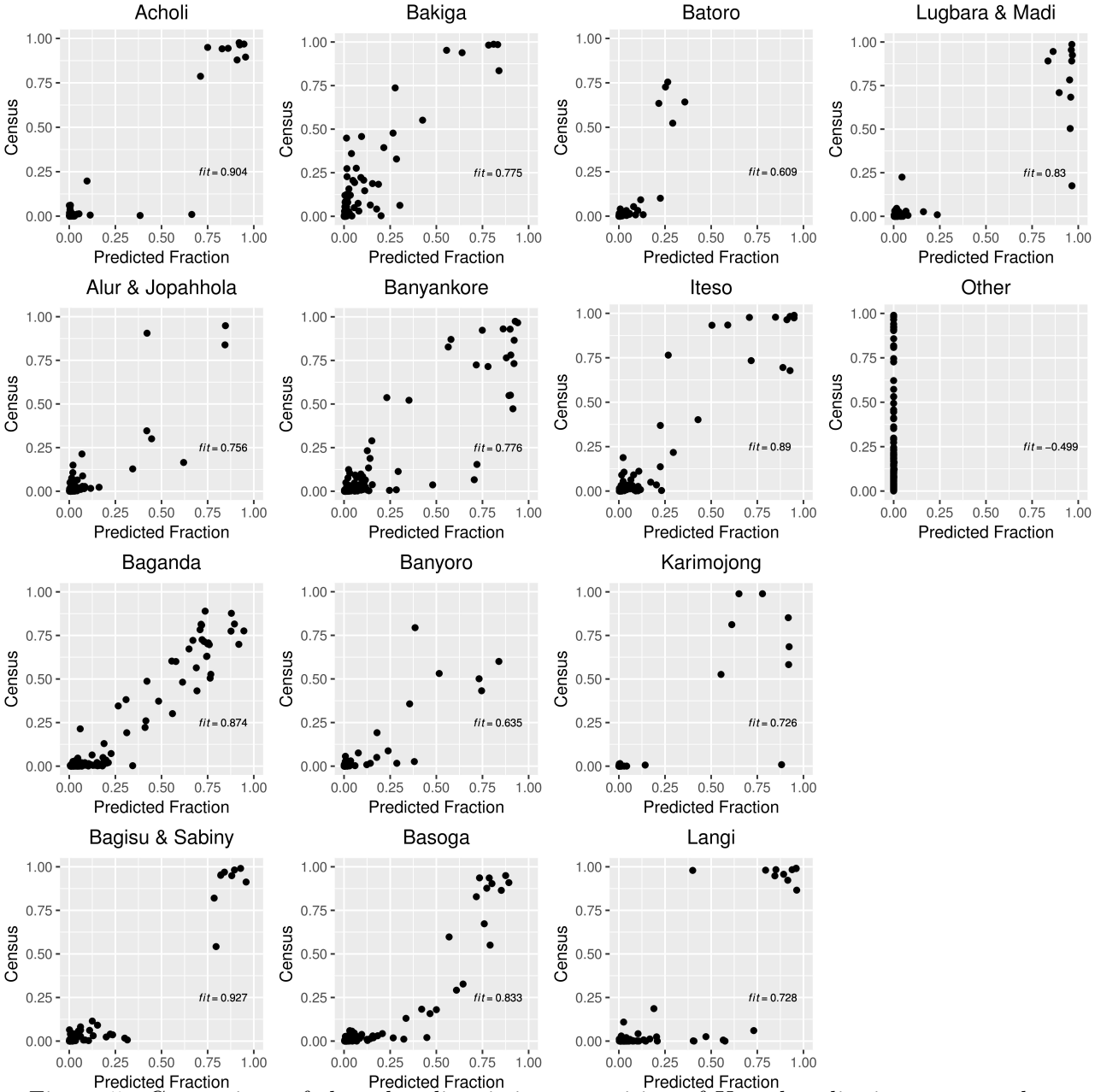
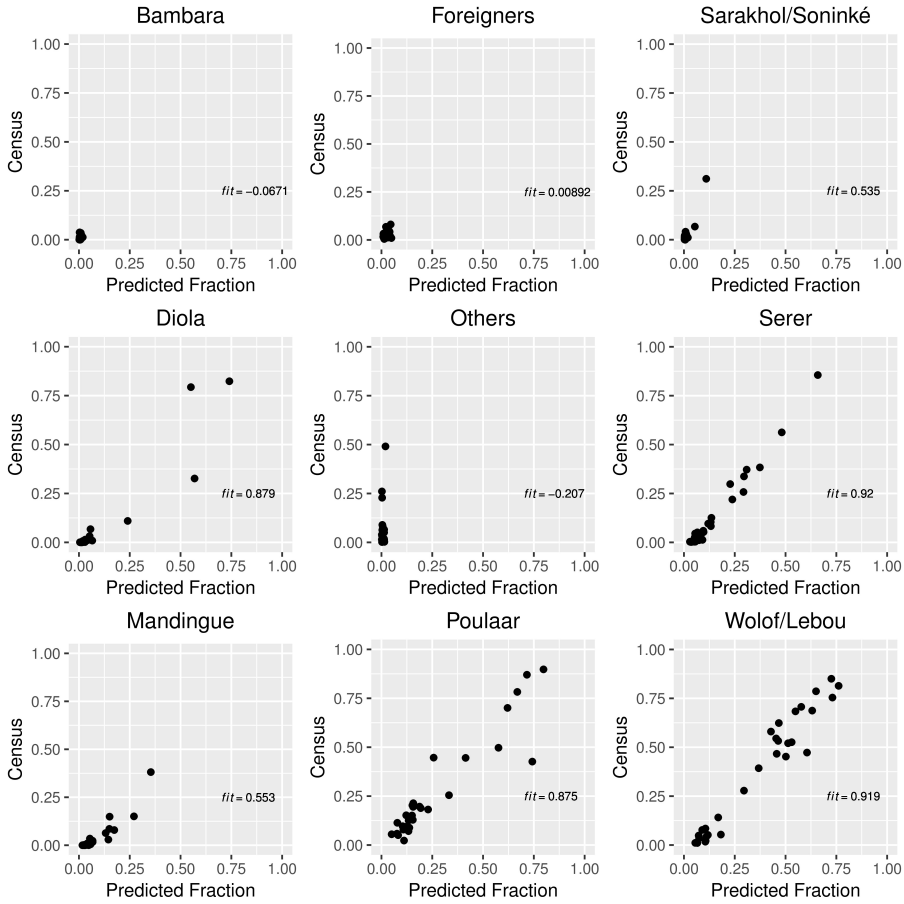
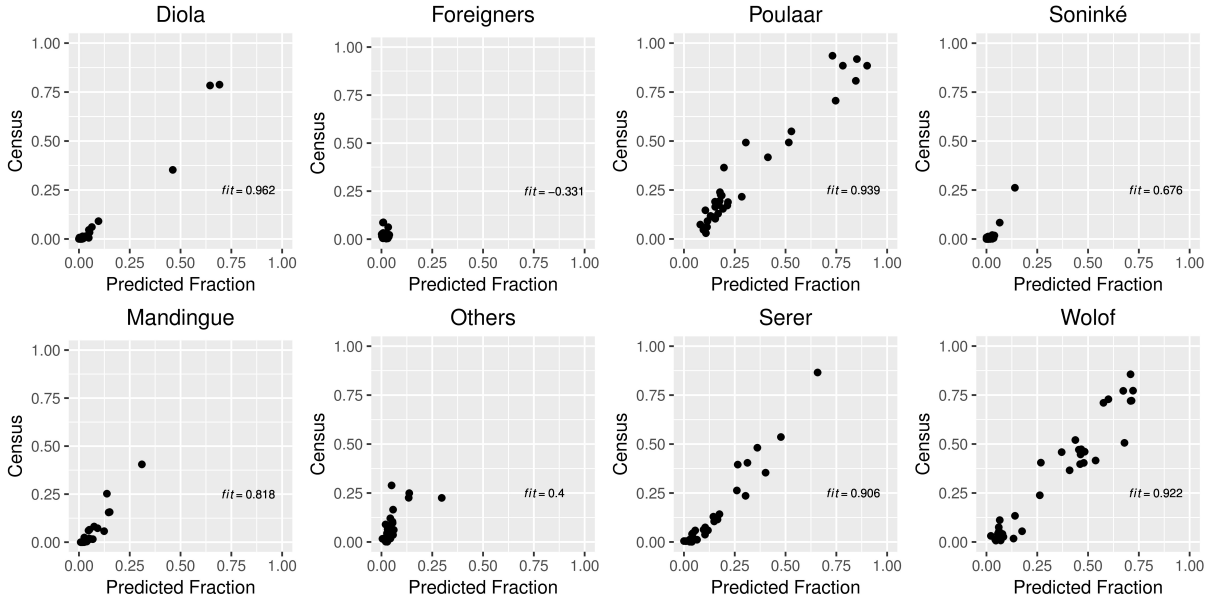


Figure 5. Comparison of the ethno-linguistic composition of Ugandan districts: census data (2002) vs. prediction (2011)



(a) Senegal SIDE 1992 and Census 1988



(b) Senegal SIDE 2005 and Census 2002

Figure 6. Comparison of the ethno-linguistic composition of Senegalese districts: census data vs. SIDE prediction.

4 Limitations: DHS sampling bias

This section briefly addresses the potential for sampling bias in the DHS data which might, if extant, affect the quality of SIDE. In particular, we aim to provide insights into the question whether local violence affects the sampling probability of individuals and districts within a given DHS survey. To this intent we use the data set described in Section 5 which aggregates conflict-event data from SCAD and ACLED (Raleigh et al., 2010; Salehyan et al., 2012) on the district-year level. These data are combined with geocoded DHS respondents sampled in a given year. Based on the DHS data, we construct two measures of district-level sampling. The first encodes the number of DHS respondents per 1 million inhabitants of a district.[8] The second measure is a dummy for whether any person in a district has been sampled in a given year. These two measures are then used as the dependent variables in straightforward OLS models, where we regress them on dummies for whether a district has seen different forms of political violence in the preceding year. We also include basic co-variates, as well as country-year fixed effects.

The results are encouraging insofar as we do not find robust and consistent signs of violence-induced sampling bias in the DHS. Except for a negative, but hardly significant effect of prior civil war on the sampling dummy in Model 3 in Table 4, all point estimates are associated with substantive standard errors, and mostly coming with a positive, rather than negative sign.[9] Although we do only find a very weak signal that the DHS sampling might be biased in the case of civil wars, this result does not preclude the bias to be substantial in single cases. Furthermore, users of the SIDE data should be aware that, beyond a potential within-country sampling bias of the DHS, there is substantial variance in the number of survey conducted across countries, some of which is very likely to be explained by civil wars that prohibit the carrying out of an extensive and elaborate DHS survey.

[8] Data on the population size of a district comes from CIESIN (2013).

[9] This positive sign, in particular in the analysis of DHS respondents per 1 million inhabitants, might stem from the fact that the DHS uses different population for sampling than we use for analysis.

Table IV. DHS respondents per 1'000'000 inhabitants and local violence (dummy)

	Dependent variable:		
	Respondents per 1'000'000		
	(1)	(2)	(3)
Population$\times 1e-6$	$-1,597.51^{**}$	$-1,835.04^{**}$	$-1,887.74^{**}$
	(271.50)	(310.80)	(327.61)
(Population$\times 1e-6)^2$	319.63^{**}	363.19^{**}	369.64^{**}
	(84.43)	(91.28)	(98.16)
Urban share (%)	5.94^{**}	5.93^{**}	5.83^{**}
	(2.03)	(2.23)	(2.24)
Area (ln)	$90.70^{\dagger}$	95.89	$107.22^{\dagger}$
	(53.21)	(58.61)	(61.20)
Any$_{t-1}$>0 SCAD	304.82		
	(210.06)		
Any$_{t-1}$>0 ACLED		237.96	
		(152.85)	
non-organ.$_{t-1}$>0 SCAD			394.04
			(278.21)
militia$_{t-1}$>0 SCAD			−74.08
			(120.75)
non-organ.$_{t-1}$>0 ACLED			330.73
			(305.78)
one-sided$_{t-1}$>0 ACLED			44.52
			(115.66)
civil war$_{t-1}$>0 ACLED			−96.45
			(98.60)
Country-year FE	yes	yes	yes
District FE	no	no	no
Observations	10,755	9,611	9,611
R^2	0.33	0.34	0.34

Note: †p<0.1; *p<0.05; **p<0.01. Standard errors are clustered on the district-level.

Table V. Local DHS sampling (dummy) and local violence (dummy)

	Dependent variable:		
	District is sampled (dummy)		
	(1)	(2)	(3)
Population$\times 1e-6$	0.33**	0.32**	0.32**
	(0.04)	(0.04)	(0.04)
(Population$\times 1e-6)^2$	−0.07**	−0.07**	−0.07**
	(0.02)	(0.01)	(0.01)
Urban share (%)	0.001**	0.001**	0.001**
	(0.0002)	(0.0002)	(0.0002)
Area (ln)	0.04**	0.03**	0.03**
	(0.004)	(0.004)	(0.004)
Any$_{t-1}$>0 SCAD	0.01		
	(0.02)		
Any$_{t-1}$>0 ACLED		0.01	
		(0.01)	
non-organ.$_{t-1}$>0 SCAD			0.01
			(0.02)
militia$_{t-1}$>0 SCAD			0.01
			(0.03)
non-organ.$_{t-1}$>0 ACLED			0.01
			(0.02)
one-sided$_{t-1}$>0 ACLED			0.01
			(0.01)
civil war$_{t-1}$>0 ACLED			−0.03†
			(0.02)
Country-year FE	yes	yes	yes
District FE	no	no	no
Observations	10,755	9,611	9,611
R^2	0.55	0.57	0.57

Note: †p<0.1; *p<0.05; **p<0.01. Standard errors are clustered on the district-level.

5 SIDE in conflict research

5.1 Data and empirical strategy

Unit of analysis: The unit of analysis is the district-year between 1990 and 2013 – the union of the temporal coverage of the SCAD (Salehyan et al., 2012) and EPR-ETH data (Vogt et al., 2015). Data on district boundaries comes from FAO's GAUL data (FAO, 2014), taking the earliest available year, 1992, as the baseline.

Dependent variables: We use the geo-coded event data available from SCAD (1990-2012; Salehyan et al., 2012) and ACLED (1996-2010; Raleigh et al., 2010) to construct dummy variables for the occurrence of several types of violence within each district-year:

1. **Riot/demo** SCAD: Combines events encoded as organized or spontaneous strikes, demonstrations, and riots.
2. **Militia** SCAD: Combines events encoded as pro-, anti-, intra-, and extra-government violence by militias.
3. **Riot/demo** ACLED: Comprises all events encoded as 'Riots/Protests'.
4. **One-sided** ACLED: Comprises all events encoded as 'Violence against civilians'.
5. **Civil war** ACLED: Combines all other events in ACLED, in particular all battle-related events, including remote violence.

All dummy variables are multiplied by a factor of 100 to facilitate the interpretation of all coefficients as percentage points.

Ethno-political polarization: The measure of ethno-political polarization for each district-year is constructed in three steps.

- First, each district-year is matched to the most recent SIDE data[10] and corresponding population counts (CIESIN, 2013), to enable the computation of district-level ethnic proportions (see Section 4 in the main paper).
- Second, the ethnic groups encoded in SIDE are matched to the EPR-ETH data (Vogt et al., 2015) to distinguish politically relevant from non-relevant ethnic groups and to identify those SIDE groups which together form a politically relevant ethnic cluster. The matching

[10] Where no SIDE data based on DHS surveys prior to the district-year is available, the most proximate SIDE data from 'the future' is taken.

is based on (1) a simple string matching of ethnic groups names and their synonyms, and (2) information available from online encyclopedias such as `ethnologue.com` and `joshuaproject.net`.

- Third, the district-level ethno-political polarization index is calculated using the formula derived by Esteban & Ray (1994).[11]

Control variables: We control for the logged population size of a district and the share of its urban population (CIESIN, 2013), as well as the area of a district. Finally, we include the temporal and spatio-temporal lags for $t-1$ and $t-2$ of the respective dependent variable for each district year. Spatio-temporal lags are derived on the basis of past conflict in a district's neighboring districts in the same country.

Estimation strategy: We specify a linear probability model and add country-year fixed effects to account for heterogeneity in the data that is constant within country-years. We thus ensure that the coefficients do only pick-up cross-sectional variation in the dependent variable.

Table VI. Summary statistics: Ethno-political polarization and local violence

Statistic	N	Mean	St. Dev.	Min	Max
Riot/demo SCAD	62,025	2.39	15.27	0	100
Militia SCAD	62,025	1.56	12.39	0	100
Riot/demo ACLED	44,658	5.78	23.33	0	100
One-sided ACLED	44,658	7.82	26.85	0	100
Civil war ACLED	44,658	8.56	27.98	0	100
Ethno-pol. polar.	53,250	0.41	0.32	0.00	1.00
Population (log)	61,544	11.23	1.39	5.52	15.44
Urban pop. (%)	61,544	16.90	28.04	0.00	100.00
Area (log)	62,025	−2.19	1.94	−21.34	3.36

5.2 Robustness checks:

[11] $Ethno-pol.polar. = 4 * \sum_{i=1}^{I}(size_i^2 * (1 - size_i))$, where $size_i$ is the size of politically relevant ethnic group i relative to the size of all politically relevant groups in a district.

Table VII. District-level ethno-political polarization & violence: Pure ethnic polarization

	Dependent variable:				
	Riot/demo SCAD	Militia SCAD	Riot/demo ACLED	One-sided ACLED	Civil war ACLED
	(1)	(2)	(3)	(4)	(5)
Ethno-pol. polar.	1.77**	0.54	2.01**	1.80*	1.02
	(0.43)	(0.44)	(0.70)	(0.88)	(0.98)
Ethnic polarization	−0.48	−0.55	0.68	−0.95	−2.80*
	(0.60)	(0.47)	(1.02)	(1.19)	(1.28)
Population (log)	1.26**	0.49**	2.41**	1.71**	1.33**
	(0.19)	(0.15)	(0.29)	(0.33)	(0.35)
Urban pop. (%)	0.05**	0.02**	0.10**	0.09**	0.09**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Area (log)	−0.27†	0.29*	−0.27	1.05**	1.65**
	(0.14)	(0.13)	(0.20)	(0.27)	(0.27)
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

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

Table VIII. District-level ethno-political polarization & violence: Weighting by country

	Dependent variable:				
	Riot/demo SCAD	Militia SCAD	Riot/demo ACLED	One-sided ACLED	Civil war ACLED
	(1)	(2)	(3)	(4)	(5)
Ethno-pol. polar.	2.32**	0.74	1.90*	1.98†	0.50
	(0.69)	(0.63)	(0.92)	(1.12)	(1.08)
Population (log)	1.82**	0.60**	2.83**	1.90**	1.11**
	(0.29)	(0.22)	(0.37)	(0.41)	(0.42)
Urban pop. (%)	0.08**	0.04**	0.12**	0.09**	0.09**
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
Area (log)	−0.46*	0.34*	−0.39	0.87**	1.51**
	(0.20)	(0.16)	(0.26)	(0.31)	(0.31)
Weighted by	country	country	country	country	country
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	0.26	0.16	0.33	0.32	0.36

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

Table IX. District-level ethno-political polarization & violence: Drop districts outside convex hull

	Dependent variable:				
	Riot/demo SCAD	Militia SCAD	Riot/demo ACLED	One-sided ACLED	Civil war ACLED
	(1)	(2)	(3)	(4)	(5)
Ethno-pol. polar.	2.01**	0.38	2.67**	1.80†	0.75
	(0.54)	(0.43)	(0.75)	(1.06)	(1.10)
Population (log)	1.39**	0.54**	2.59**	2.05**	1.97**
	(0.24)	(0.18)	(0.33)	(0.35)	(0.37)
Urban pop. (%)	0.04**	0.02**	0.10**	0.08**	0.08**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Area (log)	−0.38*	0.12	−0.43†	0.66*	1.07**
	(0.16)	(0.14)	(0.23)	(0.29)	(0.28)
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	39,027	39,027	27,226	27,226	27,226
R^2	0.23	0.14	0.31	0.27	0.27

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

References

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

Atkinson, Peter M & Christopher D Lloyd (2007) Non-stationary variogram models for geostatistical sampling optimisation: An empirical investigation using elevation data. *Computers and Geosciences* 33(10): 1285–1300.

CIESIN (2013) Low Elevation Coastal Zone (LECZ) Urban-Rural Population and Land Area Estimates, Version 2. *Center for International Earth Science Information Network, Columbia University*.

Esteban, Joan-Maria & Debraj Ray (1994) On the Measurement of Polarization. *Econometrica* 62(4): 819–851.

FAO (2014) Global Administrative Unit Layers. *Accessed from http://data.fao.org/map?entryId=f7e7adb0-88fd-11da-a88f-000d939bc5d8&tab=metadata on 2015/07/28*.

Hastie, Trevor; Robert Tibshirani & Jerome Friedman (2005) The Elements of Statistical Learning: Data Mining , Inference and Prediction. *The Mathematical Intelligencer* 27(2): 83–85.

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

Lloyd, Christopher D (2012) Analysing the spatial scale of population concentrations by religion in Northern Ireland using global and local variograms. *International Journal of Geographical Information Science* 26(1): 57–73.

Lloyd, Christopher D (2015) Assessing the spatial structure of population variables in England and Wales. *Transactions of the Institute of British Geographers* 40(1): 28–43.

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

Pawlowsky-Glahn, Vera & Ricardo A Olea (2004) *Geostatistical Analysis of Compositional Data*. Oxford: Oxford University Press.

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. *Retrieved on 10/04/2017 from http://www.seas.upenn.edu/ ese502/notebook*.

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.

Walvoort, Dennis J. J & Jaap J de Gruijter (2001) Compositional Kriging: A Spatial Interpolation Method for Compositional Data. *Mathematical Geology* 33(8): 951–966.

Wood, Simon N (2003) Thin plate regression splines. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 65(1): 95–114.