Supporting Information

Soares Magalhães et al. 10.1073/pnas.1106784108

SI Materials and Methods

General Framework of Analysis. The analysis was conducted in two phases (Fig. S1). First, we built geographical models of water, sanitation, and hygiene (WASH) indicators, using individual [demographic health survey (DHS) household-level] variables and one location-specific environmental variable [a rural/urban mask obtained from the Global Rural-Urban Mapping Project (GRUMP)], and performed spatial prediction for each indicator at the location of helminth surveys, using the best-fitting model (step 3.2). We also predicted the prevalence of WASH indicators at the nodes of a 5 × 5-km grid (step 3.3).

Second, we built geographical models of helminth prevalence, using individual-level data (child-level demographic information) and location-specific values for environmental covariates and for WASH indicators. Parameter estimation for each WASH indicator (step 5.1) was conducted by using the predicted values and associated uncertainty at the parasitological survey locations (step 3.2). These values were then used in population attributable fraction (PAF) estimation for each helminth infection due to individual WASH indicators (step 5.1.1).

Finally, we predicted the geographical prevalence of helminth infection (step 5.2), using the extracted values of the mean and variance of the posterior distributions of WASH indicators (step 3.3) and the extracted values of environmental covariates.

Hygiene and Sanitation Data. The DHS datasets used in this study are in the public domain and are available from the MACRO international Web site (1).

Questionnaires on household sanitation were given to mothers. Questions included their age, their education, and number of members in the household. The main floor material was recorded as a nominal variable with three categories: natural floor (earth, sand, and dung), rudimentary (wood, palm, and bamboo) and finished (parquet, vinyl, tiles, cement, and carpet). The source of drinking water was recorded as a nominal variable with three categories: piped water (piped into dwelling, piped into yard, and public tap), well water (public, in dwelling or yard, open or protected wells), and surface water (spring, river, pond, lake, dam, and rainwater). The type of toilet facility was recorded as a nominal variable with three categories: no facility, flush toilet, and pits (traditional pit and ventilated improved pit).

For the three countries in the analysis (Burkina Faso, Ghana, and Mali), the mean maternal age was 29.5 y [95% confidence interval (CI): 29.4, 29.6], and the mean number of members per household was 8.0 (95% CI: 7.9, 8.1). The percentage of mothers who were uneducated was highest in Burkina Faso, 87.4% (95% CI: 86.9, 88.0%), and lowest in Ghana, 47.6% (95% CI: 46.4, 48.7%). The percentage of households without piped water, toilet facilities, and finished floors in Burkina Faso, Ghana, and Mali is shown in Table S1.

Helminth Infection Data. We used parasitological data from national, school-based parasitological surveys conducted in Burkina Faso in 2006, Ghana in 2008, and Mali in 2007 with support from the Schistosomiasis Control Initiative (SCI). These surveys were originally designed and implemented with the objective of mapping urinary schistosomiasis, which is the most prevalent helminth infection in West Africa. To ensure good geographical coverage of the survey area, the number of schools selected from each district was proportional to the size of the district. The area of each district was calculated using a geographical information system (GIS). The sample size of 77 schools was calculated to give the same spatial density of schools across all surveys. The sampling frame for school selection was a list of all schools in the country, stored in a Microsoft Excel 2007spreadsheet. The geographic location of the school was determined using a handheld global positioning system device. In each school 60 children were selected at random (30 boys and 30 girls) when possible. Children were selected from within the selected schools using systematic random sampling of class lists. Selected children were assembled and asked to provide a stool and urine sample. Stool samples provided by each child were used to make two slides, which were examined microscopically using the semiquantitative Kato–Katz technique for the eggs of soil-transmitted helminthiases (STHs) (*Ascaris lumbricoides*, *Trichuris trichiura*, and hookworm) and *Schistosoma mansoni*. Stool samples were processed immediately and slides were prepared and examined in the field laboratory by experienced microscopists in diagnosing schistosomiasis and STHs, within 2 h of preparation to increase detection of the more labile hookworm eggs, by the Kato–Katz thick smear technique using a 41.7-mg template. The concentration of eggs was expressed as eggs per gram of feces (epg). From urine samples, up to 10 mL was filtered through a polycarbonate membrane and the number of eggs of *Schistosoma hematobium* was counted and expressed as eggs per 10 mL of urine.

Geographical Risk Prediction of Helminth Infection in School-Aged Children. All predictive models had (*a*) the child covariates of age and sex, (*b*) environmental and sanitation covariates for obtaining the effect estimates and also to help with the helminth prediction, and (*c*) a geostatistical effect for modeling second-order variation (geographical variation). This geographical variation between locations was modeled using an exponentially decaying autocorrelation function. The variance of the posterior distribution for each WASH indicator was included to model the uncertainty in the geographic prediction of WASH. This involved modeling the distribution of probable values of WASH in a given location, using a beta distribution parameterized by the predicted posterior mean and variance for each parasitological survey location.

The number of children positive with a given parasite is a binomial variable Y_j. The models assume a conditional binomial model where the prevalence of infection p_j, given the location j of the sample survey, is given by

$$Y_j \sim \text{Binomial}\left(p_j, T_j\right)$$

$$\text{logit}\left(p_j\right) = \alpha + \sum_{k=1}^{p} \theta_k \times x_j + u_j,$$

where T_j is the total number of children tested in location j, α is the intercept, x_j is a matrix of the mean individual and environmental covariates, θ_k is a matrix of coefficients, and u_j is a geostatistical random effect defined by an isotropic powered exponential spatial correlation function,

$$f(d_{ab};\phi) = \exp[-(\phi d_{ab})],$$

where d_{ab} are the distances between pairs of points a and b, and ϕ is the rate of decline of spatial correlation per unit of distance.

In the case of WASH indicators, x_j was defined by

$$x_j \sim \text{Beta}(\alpha_i, \beta_i)$$

$$\alpha_i = x_j\left(\frac{x_j(1-x_j)}{\nu_j} - 1\right)$$
$$\beta_j = (1-x_j)\left(\frac{x_j(1-x_j)}{\nu_j} - 1\right).$$

So the mean WASH prevalence, x_j, was defined by a beta distribution with a shape α_i and a scale parameter β_j; these parameters were estimated using the mean posterior prediction x_j and the mean posterior variance ν_j from the WASH prediction models for each helminth survey location. This specification explicitly considers the uncertainty of WASH predictions into the spatial geostatistical models. Noninformative priors were used for the intercept α and the coefficients (normal prior with mean = 0 and precision = 1×10^{-4}). The precision of u_i was given a noninformative gamma distribution.

In all models, a burn-in of 5,000 iterations was allowed, followed by 10,000 iterations where values for the intercept, coefficients, and predicted probability of infection at the prediction locations were stored. Diagnostic tests for convergence were made, including visual examination of history and density plots; convergence was successfully achieved after 5,000 iterations.

Model Selection, Spatial Prediction, and Model Validation. Model selection for prediction was based on the deviance information criterion (DIC) (the lower the DIC was, the better the model fit) (2). The best-fitting model for each sanitation indicator and for each helminth species (per lowest DIC) was used for geographical prediction (Tables S2–S5). Predictions of WASH indicators were initially made to the helminth survey locations and then to a 5 × 5-km resolution grid of the study area, using ArcGIS version 10.0 (ESRI).

The predictions of the prevalence of helminth infections were made at the nodes of a 0.1 × 0.1 decimal degree grid (~11 km^2) by interpolating the geostatistical random effect and adding it to the sum of the products of the coefficients for the fixed effects and the values of the fixed effects at each prediction location. Values of predicted prevalence of helminth infections at unsampled locations were stored for all male children of 15–19 y of age (Fig. 3). The interpolation of the random effect was done using the *spatial.unipred* kriging function in WinBUGS; the spatial.unipred command implements Bayesian kriging (3) where the values of predicted prevalence at unsampled locations are estimated (interpolated) independently of neighboring values, as opposed to joint prediction that is conditional on the values of neighboring unsampled locations. Joint prediction was not considered feasible in this study as it is extremely computationally intensive due to the size of the dataset.

The ability of the final model to predict household floor type, water supply, and sanitation and helminth infection thresholds was assessed using the area under the curve (AUC) of the receiver operating characteristic (4). An AUC value of <0.7 indicates poor discriminatory performance, 0.7–0.8 indicates acceptable, 0.8–0.9 indicates excellent, and >0.9 indicates outstanding. The predicted prevalence of household floor type, water supply, and sanitation was compared with the observed prevalence, dichotomized at 75%. The predictive ability of the helminth models was assessed by comparing the predicted prevalence of *S. hematobium*, *S. mansoni*, and hookworm to the observed prevalence, dichotomized at 50% for *S. hematobium*, 3% for *S. mansoni*, and 5% for hookworm using the AUC. A 50% threshold is of operational relevance for urinary schistosomiasis control in that, on the basis of World Health Organization mass-drug administration (MDA) guidelines communities over a 50% threshold require annual MDA (5); thresholds of 3% and 5% were used for *S. mansoni* and hookworm, respectively, to assess the discriminatory performance of these models at predicting values above the mean endemicity class. Using a cutoff of 75% the final WASH models had AUCs of 0.78 (95% CI: 0.71, 0.84) for floor type, 0.83 (95% CI: 0.78, 0.86) for water supply, and 0.79 (95% CI: 0.73, 0.82) for sanitation. Using cutoffs of 50%, 3%, and 5% for *S. hematobium*, *S. mansoni*, and hookworm, each model was able to discriminate between these prevalence thresholds with AUCs of 0.75 (95% CI: 0.70, 0.80), 0.78 (95% CI: 0.73, 0.82), and 0.72 (95% CI: 0.66, 0.78), respectively.

The complete results for the tested models are in Tables S3–S5. The sizes of geographical clusters were ~128 km for *S. hematobium*, 112 km for *S. mansoni*, and 167 km for hookworm, and the tendency for clustering was strongest for *S. mansoni*.

Estimation of the PAF of Helminth Infection Due to Floor Type, Water Supply, and Sanitation. Estimates of the PAF for specific predictors are used to guide policy makers in planning public health interventions (6).

We estimated the PAF of anemia due to a specific helminth infection, using the standard equation (6)

$$\text{PAF}_1 = \frac{P_1(\text{OR}_1 - 1)}{P_1(\text{OR}_1 - 1) + 1},$$

where P_1 is the mean prevalence of one parasite in the 15- to 19-y age group, and OR_1 is the prevalence-specific odds ratio (OR). The OR for the prevalence of infection with one parasite was estimated by exponentiating the mean posterior estimate of the coefficient (obtained from the spatial prediction model).

1. Demographic Health Survey (2011) Measure DHS. Available at http://www.measuredhs.com/start.cfm. Accessed September 22, 2011.
2. Diggle PJ, Moyeed RA, Tawn JA (1998) Model-based geostatistics. *Appl Stat* 47:299–350.
3. Thomas A, Best N, Lunn D, Arnold R, Spiegelhalter D (2004) *GeoBUGS User Manual Version 1.2* (Medical Research Council Biostatistics Unit, Cambridge, UK).
4. Brooker S, Hay SI, Bundy DA (2002) Tools from ecology: Useful for evaluating infection risk models? *Trends Parasitol* 18:70–74.
5. WHO (2002) Prevention and control of schistosomiasis and soil-transmitted helminthiasis: Report of a WHO expert committee. *WHO Technical Report Series 912* (World Health Organization, Geneva), pp 1–57.
6. Rothman KJ, Greenland S, Lash TL (2008) *Modern Epidemiology* (Lippincott, Williams, & Wilkins, Philadelphia), 3rd Ed.

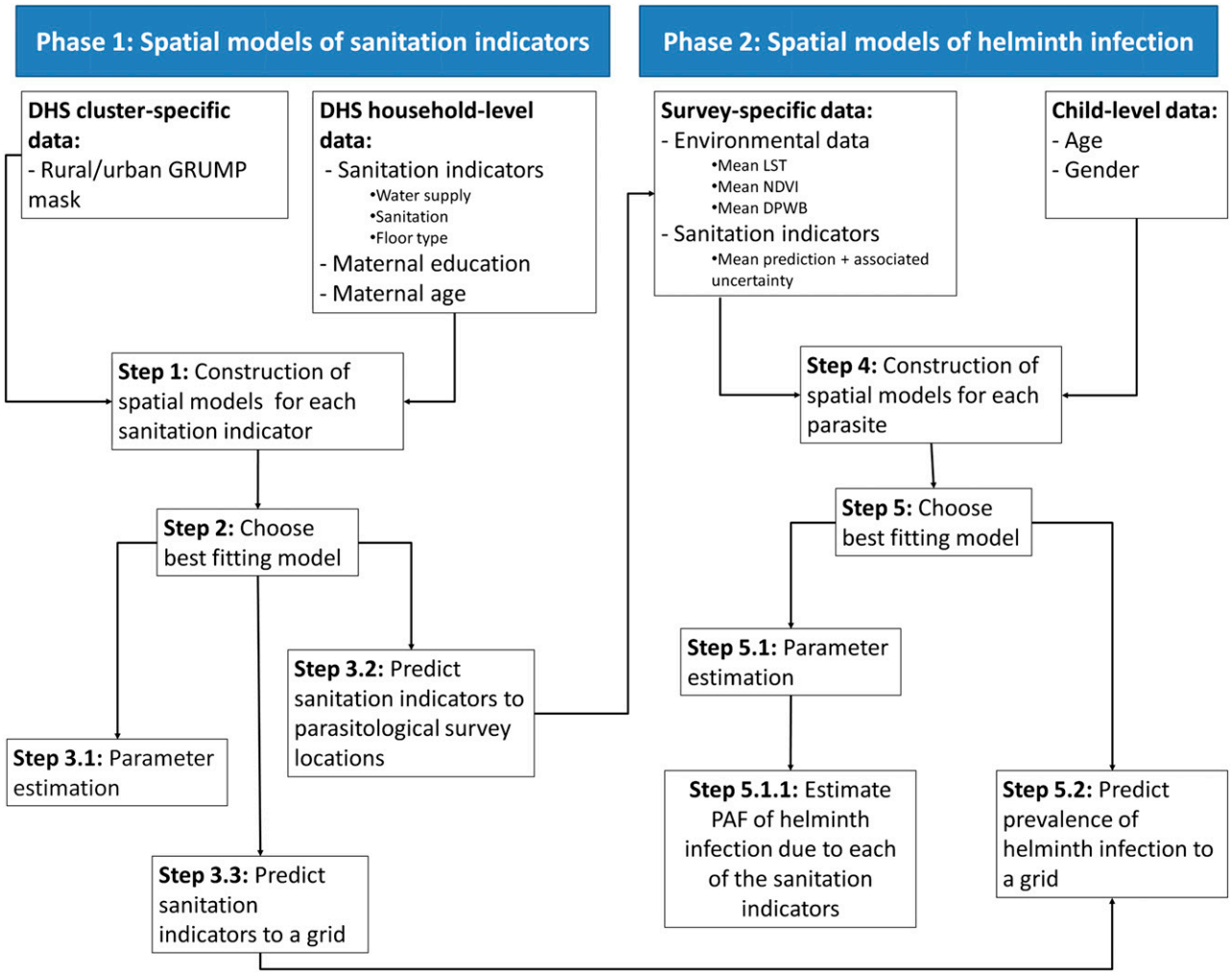

Fig. S1. Flow diagram showing the data sources and analytical steps of the analysis.

Table S1. Percentage (95% confidence intervals) of household indicators in the West Africa region, from the Demographic Health Surveys for Burkina Faso 2003, Ghana 2003, and Mali 2006

Household indicator	Burkina Faso	Ghana	Mali
Maternal education			
No education	87.4 (86.9, 88.0)	47.6 (46.4, 48.7)	84.9 (84.4, 85.4)
Primary	8.9 (8.4, 9.4)	21.5 (20.3, 22.7)	10.0 (9.5, 10.5)
Secondary	3.5 (3.2, 3.8)	29.9 (28.7, 31.1)	4.8 (4.5, 5.2)
Higher	0.2 (0.1, 0.2)	1.00 (0.7, 1.3)	0.3 (0.2, 0.4)
Main floor material			
Natural floor	63.9 (63.2, 64.5)	19.8 (18.8, 20.8)	77.6 (77.1, 78.1)
Rudimentary	—	0.2 (0.04, 0.3)	—
Finished	36.1 (35.5, 36.8)	80.0 (79.0, 81.1)	22.4 (21.9, 22.9)
Source of drinking water			
Piped water	14.8 (14.4, 15.1)	24.7 (24.1, 25.3)	22.5 (22.1, 22.8)
Well water	70.0 (69.5, 70.6)	49.6 (48.7, 50.6)	72.2 (71.7, 72.6)
Surface water	15.0 (14.7, 15.6)	25.65 (24.81, 26.50)	5.4 (5.1, 5.6)
Type of toilet facility			
No toilet facility	73.8 (73.3, 74.4)	40.7 (40.0, 41.5)	23.5 (23.0, 24.0)
Flush toilet	0.9 (0.8, 1.1)	5.5 (5.0, 6.0)	1.7 (1.5, 1.9)
Pits	25.2 (24.7, 25.8)	53.7 (52.9, 54.6)	74.8 (74.2, 75.2)

Table S2. DIC and effective number of parameters for the competing models of sanitary indicators in West Africa

	Natural floor		No piped water		No toilet facility	
Model	pD	DIC	pD	DIC	pD	DIC
Model 1, quadratic	8.2	9,372.5	8.50	7,733	8.1	9179.7
Model 2, cubic	9.4	9,346.8	9.96	7,736.1	9.2	9298.7
Model 3, spline	24.2	9,826.8	24.72	8,532.1	23.93	9,947.8
Model 4, geostatistical	243.8	10,117.9	234.8	9,342.3	231.1	10,748.6

DIC, deviance information criterion; pD, effective number of parameters.

Table S3. Associations with prevalence of *S. hematobium* infection in school-aged children in Burkina Faso, Ghana, and Mali, estimated using model-based geostatistical models considering estimation uncertainty of sanitary indicators at parasitological survey locations

Variable	Model A, posterior mean (95% CrI)	Model B, posterior mean (95% CrI)	Model C, posterior mean (95% CrI)
Female vs. male	−0.29 (−0.36, −0.22)	−0.29 (−0.37, −0.22)	−0.29 (−0.37, −0.22)
Age 10–19 y vs. 5–9 y	0.39 (0.30, 0.48)	0.39 (0.30, 0.48)	0.39 (0.30, 0.48)
DPWB	−1.22 (−1.72, −0.73)	−1.40 (−1.86, −0.98)	—
LST	−0.03 (−0.64, 0.53)	0.02 (−0.36, 0.41)	—
NDVI	−0.37 (−0.76, −0.04)	−0.32 (−0.64, 0.01)	—
Natural floor vs. other floor type	0.36 (−1.24, 2.10)	—	0.56 (−0.90, 1.69)
No piped water vs. with piped water	1.43 (0.20, 2.88)	—	0.80 (−0.11, 1.64)
No toilet facility vs. with toilet	−0.94 (−2.38, 0.46)	—	−1.14 (−2.46, 0.03)
ϕ, rate of decay of spatial correlation	2.60 (1.86, 3.53)	2.51 (1.82, 3.35)	1.93 (1.37, 2.56)
σ^2, variance of spatial random effect	5.20 (3.97, 7.82)	7.50 (5.87, 9.82)	6.29 (5.02, 7.95)
DIC	4,806.53	4,813.43	4,811.71

Posterior means and credible intervals are for the log odds ratios. CrI, Bayesian credible interval; DIC, deviance information criterion; DPWB, distance to perennial water body; LST, land surface temperature; NDVI, normalized difference vegetation index.

Table S4. Associations with prevalence of *S. mansoni* infection in school-aged children in Burkina Faso, Ghana, and Mali, estimated using model-based geostatistical models considering estimation uncertainty of sanitary indicators at parasitological survey locations

Variable	Model A, posterior mean (95% CrI)	Model B, posterior mean (95% CrI)	Model C, posterior mean (95% CrI)
Female vs. male	−0.46 (−0.66, −0.27)	−0.46 (−0.65, −0.26)	−0.46 (−0.65, −0.27)
Age 10–19 y vs. 5–9 y	0.44 (0.25, 0.62)	0.42 (0.22, 0.62)	0.43 (0.23, 0.62)
DPWB	−4.28 (−7.98, −0.77)	7.35 (3.18, 11.06)	—
LST	3.55 (−0.18, 7.13)	−8.39 (−11.58, −4.49)	—
NDVI	−0.09 (−0.70, 0.57)	0.07 (−0.66, 0.64)	—
Natural floor vs. other floor type	0.27 (−1.76, 3.53)	—	0.08 (−5.08, 4.07)
No piped water vs. with piped water	0.77 (−2.78, 4.06)	—	3.30 (0.25, 6.44)
No toilet facility vs. with toilet	−2.92 (−5.74, -0.36)	—	−4.55 (−6.81, −2.32)
ϕ, rate of decay of spatial correlation	2.98 (1.26, 8.29)	2.52 (1.02,6.58)	3.24 (1.23, 10.16)
σ^2, variance of spatial random effect	13.37 (8.14, 21.87)	14.41 (8.48,23.04)	13.90 (8.34, 23.34)
DIC	1,065.5	1,072.26	1,073.71

Posterior means and credible intervals are for the log odds ratios. CrI, Bayesian credible interval; DIC, deviance information criterion; DPWB, distance to perennial water body; LST, land surface temperature; NDVI, normalized difference vegetation index.

Table S5. Associations with prevalence of hookworm infection in school-aged children in Burkina Faso, Ghana, and Mali, estimated using model-based geostatistical models considering estimation uncertainty of sanitary indicators at parasitological survey locations

Variable	Model A, posterior mean (95% CrI)	Model B, posterior mean (95% CrI)	Model C, posterior mean (95% CrI)
Female vs. male	−0.55 (–0.68, –0.41)	−0.54 (–0.67, –0.42)	−0.55 (–0.69, –0.43)
Age 10–14 y vs. 5–9 y	0.38 (0.23, 0.55)	0.38 (0.23, 0.54)	0.38 (0.21, 0.54)
Age 15–19 y vs. 5–9 y	0.67 (0.37, 0.98)	0.66 (0.36, 0.95)	0.65 (0.35, 0.95)
DPWB	0.20 (–0.49, 1.14)	0.07 (–0.63, 0.78)	—
LST	−2.08 (–2.82, –1.28)	−0.49 (–1.11, 0.07)	—
NDVI	0.64 (0.24, 1.02)	0.44 (0.12, 0.76)	—
Natural floor vs. other floor type	2.25 (1.82, 3.44)	—	−1.02 (–3.13, 0.80)
No piped water vs. with piped water	−3.76 (–6.54, –1.19)	—	2.15 (–0.19, 4.80)
No toilet facility vs. with toilet	0.10 (–1.98, 1.88)	—	−0.65 (–3.32, 2.35)
ϕ, rate of decay of spatial correlation	2.00 (0.81, 4.42)	1.45 (0.76, 2.38)	1.48 (0.75, 2.45)
σ^2, variance of spatial random effect	5.42 (3.64, 8.92)	6.27 (4.10, 9.60)	5.90 (3.86, 9.41)
DIC	2,264.34	2,268.05	2,268.13

Posterior means and credible intervals are for the log odds ratios. CrI, Bayesian credible interval; DIC, deviance information criterion; DPWB, distance to perennial water body; LST, land surface temperature; NDVI, normalized difference vegetation index.