Modelling the impact of MAUP on environmental drivers for *Schistosoma japonicum* prevalence

Andrea L. Araujo Navas [1,*], Frank Osei [1], Ricardo J. Soares Magalhães [2,3], Lydia R. Leonardo [4], and Alfred Stein [1]

[1] Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, PO Box 217, 7500 AE, Enschede, The Netherlands.

[2] UQ Spatial Epidemiology Laboratory, School of Veterinary Science, The University of Queensland, Gatton 4343 QLD, Australia.

[3] Child Health and Environment Program, Child Health Research Centre, The University of Queensland, South Brisbane 4101 QLD, Australia.Institute of Biology, College of Science, University of the Philippines Diliman, the Philippines.

* Correspondence: a.l.araujonavas@gmail.com

E-mails:
ALAN: a.l.araujonavas@utwente.nl
FO: f.b.osei@utwente.nl
RJSM: r.magalhaes@uq.edu.au
LRL: lydialeonardo1152@gmail.com
AS: a.stein@utwente.nl

Abstract

Background: The modifiable areal unit problem (MAUP) arises when the support size of a spatial variable affects the relationship between prevalence and environmental risk factors. Its effect on schistosomiasis modelling studies could lead to unreliable parameter estimates. The present research aims to quantify MAUP effects on environmental drivers of *Schistosoma japonicum* infection by (i) bringing all covariates to the same spatial support, (ii) estimating individual-level regression parameters at 30 m, 90 m, 250 m, 500 m, and 1 km spatial supports, and (iii) quantifying the differences between parameter estimates using five models.

Methods: We modelled the prevalence of *Schistosoma japonicum* using sub-provinces health outcome data and pixel-level environmental data. We estimated and compared regression coefficients from convolution models using Bayesian statistics.

Results: Increasing the spatial support to 500 m gradually increased the parameter estimates and their associated uncertainties. Abrupt changes in the parameter estimates occur at 1 km spatial support, resulting in loss of significance of almost all the covariates. No significant differences were found between the predicted values and their uncertainties from the five models. We provide suggestions to define an appropriate spatial data structure for modelling that gives more reliable parameter estimates and a clear relationship between risk factors and the disease.

Conclusions: Inclusion of quantified MAUP effects was important in this study on schistosomiasis. This will support helminth control programs by providing reliable parameter estimates at the same spatial support, and suggesting the use of an adequate spatial data structure, to generate reliable maps that could guide efficient mass drug administration campaigns.

Keywords: schistosomiasis modelling; modifiable areal unit problem; uncertainty; Bayesian statistics; convolution model.

Background

Schistosomiasis (SCH) is a water-borne neglected tropical disease of public health significance [1] associated with important morbidity outcomes in school-aged children such as malnutrition, anaemia and stunted growth in school-aged children [2, 3]. Infection is caused by skin penetration of the cercariae, the larval infective stage of the parasite, also known as schistosome. Three schistosome species cause the infection: *Schistosoma japonicum, S.mansoni,* and *S. haematobium.* Due to its zoonotic life cycle [4], *Schistosoma japonicum* is the hardest to control; its infection life cycle includes the amphibious snail from the species *Oncomelania hupensis* as the intermediate host, and humans and other mammalians as definite hosts [5, 6]. SCH affects more than 252 million people worldwide [7] especially populations living in poor conditions, where access to clean water and sanitation is limited.

Traditionally, SCH is controlled by the use of anthelminthic drugs in at-risk populations [8]. Mass drug administration campaigns identify at-risk populations by using

SCH risk mapping. SCH mapping uses geographic information systems (GIS), global positioning systems and remotely sensed environmental data [9, 10]. Modelling those infections using various statistical methods have enabled the study of the distribution of populations at-risk [9, 10], and the role of the environmental variation on the geographical heterogeneity of infection burden (i.e. prevalence or intensity of infection) [11]. Statistical modelling of SCH quantifies empirical relationships between indirect morbidity indicators of public health significance and environmental risk factors. Those could be extracted from Earth Observation (EO) data such as monitor sites or satellite imagery. In addition, EO data help to interpolate the level of infection towards unsampled locations [12-14].

The robustness of SCH geographical modelling efforts is affected by uncertainties propagated from the use of EO data at various spatial and temporal scales of analysis [15]. EO data are generally constrained by their spatial and temporal scale of sampling [16]. In this study, we focus on spatial scale. Scale is a major concern in spatial epidemiology [17, 18] since it determines the significance of the various environmental risk factors on the disease distribution [19]. Spatial scale encompasses the spatial support and the spatial extent of analysis [20]. The spatial support refers to the area that each individual observation occupies in space. In the case of a raster grid, the spatial support is the spatial resolution (e.g. a 30 x 30 m-resolution Landsat pixel). The spatial extent is the spatial coverage of a set of observations (e.g. administrative units) and is gathered following a sampling scheme [20]. For a given extent, the support size and shape of spatial units may affect the patterns identified in the survey and environmental data [21, 22] and the relationship between the disease morbidity indicators and the environmental risk factors. This is known as the modifiable areal unit problem (MAUP) [23, 24]. The MAUP arises because spatial units of analysis are often created using different *ad hoc* shapes and sizes. Statistical analyses of data performed according to these varying spatial units may lead to different results (e.g. correlation and regression coefficients) [24].

Various studies investigated the consequences of ignoring MAUP effects in spatial epidemiological modelling. For instance, Hellsten et al. [25] studied the influence of using aggregated covariate data to model ammonia emissions at the farm level. They showed that the size and shape of spatial aggregation areas strongly affect the location of the emissions estimated by the model, e.g. too small areas resulting in false emission “hot spots”. Schur et al [21] and Schur et al [22] aggregated SCH prevalence maps to estimate endemicity for various administrative units [26]. Such aggregation showed different patterns of endemicity and

intervention approaches. As a consequence, localized areas of high endemicity may not be addressed properly. In a recent study [27] we quantified the effect of pure specification bias, that originates when using group-level (i.e. aggregated) survey data at an administrative level for individual-level inferences. Equation 1 shows the common method used to model schistosomiasis. Data on the human *Schistosoma japonicum* infection variable y are commonly aggregated at barangay k level, y_k has a binomial distribution with parameters N_k and p_k corresponding to the number of sampled individuals and the probability of infection, respectively. Parameters for this distribution are obtained from the mean of various environmental risk factors within barangay k as predictors, denoted as $\bar{\boldsymbol{x}}_{\boldsymbol{k}}$, where γ are the barangay-level coefficients, γ_0 being the intercept, and $\gamma_{(1\ldots n)}$ the regression coefficients for n environmental covariates (equation 1).

$$y_k|\bar{\boldsymbol{x}}_{\boldsymbol{k}}, \boldsymbol{\gamma} \sim Binomial(N_k, \hat{p}_k)$$

$$logit(\hat{p}_k) = \gamma_0 + \gamma_1 \cdot \bar{x}_{1_k} + \gamma_2 \cdot \bar{x}_{2_k} + \cdots + \gamma_n \cdot \bar{x}_{n_k}. \quad (1)$$

We calculated individual-level regression parameters by modifying equation 1 into a convolution model. We observed differences ranging from -0.19 to 0.28 between individual (i.e. γ coefficients) and group level parameter estimates and their uncertainties. High differences were observed for NDWI (0.28), LSTN (-0.19) and LSTD (0.16). Although some covariates got a less significant effect on schistosomiasis, uncertainties in their individual level coefficients were lower than the group-level regression coefficients (e.g. LSTD and elevation). We concluded that the choice of spatial support affects the model parameter estimates and their associated uncertainties by changing the within covariates variability in exposure areas. The selection of spatial support should be further investigated as it might represent a significant source of uncertainty in SCH modelling [15].

Up to date, the majority of SCH studies have put little attention to the size of spatial support. They use EO data at various spatial supports with misaligned grids ignoring the possible consequences on the observed patterns of the data [21, 22]. Moreover, MAUP effects on the various environmental risk factors used as drivers for SCH infection have not been quantified. This is important as the relevance of the environmental risk factors on SCH depends on their scale of analysis [7, 19]. Ignoring MAUP effects might produce unreliable predictions of at-risk populations, and consequently, wrong decisions based upon inefficient mass drug administration campaigns.

The purpose of this research is to quantify MAUP effects on environmental drivers of *Schistosoma japonicum* infection. To achieve this objective we aim to: (i) aggregate and disaggregate EO data in order to bring all covariates to a the same spatial support of analysis, (ii) estimate individual-level covariate regression parameters at 30 m, 90 m, 250 m, 500 m, and 1 km spatial supports, by using a convolutional model that accounts for pure specification bias; and (iii) quantify the differences between parameter estimates using five different models.

Methods

Study Area and Data on *Human Schistosoma japonicum* infection

We use *Schistosoma japonicum* infection data collected as part of the 2008 Nationwide schistosomiasis Survey in the Philippines. Here, *Schistosoma japonicum* is endemic in 28 of its 81 provinces [28], with approximately 1.8 million estimated infected people [29]. The disease affects children, adolescents, and individuals with high-risk occupations, such as farmers and fishermen [29, 30]. The area of study is the region of Mindanao in the Philippines (Figure 1). This area was selected due to the high response rate of 70.9 percent of the individuals to the 2008 survey [31, 32], and the good spatial coverage of the sampling.

A two-stage systematic cluster sampling was used where stratification was done using high, medium and low prevalence levels, obtained from the 1994 World Bank-assisted Philippine Health Development Program. Provinces and sub-municipalities called barangays were the primary and secondary sampling units respectively. A barangay is the smallest administrative division in the Philippines, numbering from 58 to1158 within a single province. In total 11 provinces with high ($\geq 2\%$) and medium (0.091 -1.99%) prevalence rates were included, while 9 low-prevalence (0.04-0.09%) provinces were randomly selected. Within the selected provinces, barangays with high prevalence rates were surveyed. In total, between 2 and 10 barangays were surveyed per province, resulting in 108 out of 10021 barangays that were surveyed in Mindanao.

For *Schistosoma japonicum* diagnosis, a Kato-Katz thick smear examination [32] was used based on a two-sample stool collection. Due to inconsistencies in the second stool sample submission, however, only the results of the first sample were available [8]. Samples were taken from people aged two years and above and were analysed using a microscope. Active infection was indicated by the presence of *Schistosoma japonicum* eggs.

Data such as age and gender were recorded for 19763 individuals. Barangay and province information for each individual was recorded but not geo-referenced. For this reason, individual-level survey data were aggregated and geo-located to the centroids of the 108 barangays. We used a probability of infection p in barangay k as our disease outcome variable. We obtained an up-to-date barangay centroids shape file from DIVA geographic information system [33]. More details about the sampling design and surveyed information can be found in Leonardo et al [31, 34].

Environmental risk factors

We included in our analysis six relevant environmental risk factors for SCH transmission [35, 36]. These are the nearest distance to water bodies (NDWB), the normalized difference vegetation index (NDVI), the normalized difference water index (NDWI), land surface temperature at day (LSTD) and at night (LSTN) and elevation (E). NDWB shows the accessibility of people to water bodies that represent potential infection foci as they may contain contaminated snail hosts that release the infective larval stages of the parasite [8]. NDVI is and indicator of flooded vegetation [8], particularly rice paddy fields, and environmental moisture [37, 38]. Both are an important risk factor for Asian SCH [39]. NDWI was used as a proxy indicator of flooding [37, 40] showing potentially hidden water bodies. LSTD and LSTN are determinant for the survival of larval stages of snails [41, 42] and are used as proxies for water temperature given that the thermal condition of shallow waters usually reflects the ambient temperature of the air [8]. Elevation is relevant for SCH transmission as the local topography of the area determines the presence of snails [43, 44]. For instance, at lower altitudes the risk of finding snails increases.

NDWB values range from 0.17 to 26.2 km and were calculated using the closest facility network analysis tool from ArcGIS [45]. We used the river and road network, and the cities and hamlets locations as input for the network. Rivers and roads were extracted from the Open Street Map Project in the Philippines [46]. Cities and hamlet locations were obtained from the National Mapping and Resource Information Authority from The Philippines [47] data base from 2010. We calculated the nearest distance from each city and hamlet to a water body following a road and interpolated those values within all surveyed barangays towards a spatial support of 30 m.

NDVI values range from 0 to 0.84 and were obtained from two sources of information, i.e. a series of Landsat 5 images from 2008 with a spatial support of 30 m and

the MODIS MOD13Q1 product with a spatial support of 250 m. NDWI values range from 0.06 to 0.61 and were also obtained from two sources of information, i.e. a Landsat 5 imagery product from 2008 with a spatial support of 30 m and the annual composite from Landsat 7 from 2008 derived from Google Earth Engine with a spatial support of 500 m. LSTD values range from 297.77 to 309.52 °K and LSTN ranges from 289.73 to 297.29 °K. LSTD and LSTN values were derived from MODIS MOD11A2_LST product with a spatial support of 1 km. Finally, Elevation values range from 0 to 969.57 m was obtained from ASTER GDEM version 2 from USGS [48] with a spatial support of 30 m. All covariates were set to a common coordinate system UTM zone 51N and were standardized to have mean = 0 and standard deviation = 1 before being used. Table 1 summarizes all sources of information.

Modifying the areal units of analysis

From now onwards we will refer to an areal unit as the spatial support of analysis (SSA). We used five SSAs, with a spatial support equal to 30 m, 90 m, 250 m, 500 m, and 1 km, respectively. These spatial supports increase when going from low to high data aggregation. These values were selected based upon the commonly used spatial supports at which the environmental information is originally provided.

For NDVI, SSA = 30 m, we obtained NDVI values from Landsat 5 images. Many of these images presented gaps due to the presence of clouds. These gaps were covered using disaggregated NDVI MODIS images at the Landsat resolution. Disaggregation was performed using a linear model that predicted NDVI Landsat values based on NDVI MODIS values. NDVI values were obtained by merging the original and predicted Landsat NDVI values. For SSA = 90 m, we aggregated the previously merged NDVI values using their mean. For SSA = 250 m, we used the NDVI MODIS product directly. Finally, for SSA = 0.5 and 1 km, we aggregated the NDVI mean values from MODIS.

NDWI values were obtained from the Landsat 5 images. Gaps in some of these images were covered using disaggregated NDWI composite images at the Landsat resolution. Disaggregation towards SSA = 30 m was done by interpolating NDWI values using ordinary Kriging. For SSA = 90 m and 250 m, we aggregated the combined 30 m NDWI using its mean. For SSA = 500m, we directly used the Landsat 7 composite. Finally, for SSA = 1km, we aggregated the mean of the original Landsat 7 composite.

To obtain LSTD and LSTN values for SSA = 30 m we disaggregated the original MODIS values by using Ordinary Kriging interpolation. For SSA = 90 m, 250 m, and 500 m, we aggregated the previously interpolated values using their mean. For SSA =1 km, we used directly LSTD and LSTN from MODIS.

The interpolated NDWB values for SSA = 30 m were used to obtain NDWB for SSA = 90 m, 250 m, 500 m, and 1 km by aggregating the mean values. For elevation, we directly used the original 30 m SSA Aster images. For SSA = 90 m, 250 m, 500 m, and 1 km, we aggregated the mean values of the original Aster images.

Modelling *Schistosoma japonicum* infection under the MAUP

Convolution model

We modelled human *Schistosoma japonicum* infection at the five increasing SSAs using a convolution model that accounts for pure specification bias [27]. Pure specification is a source of uncertainty [11, 49] that produces loss of information on the real relationship between the disease and the environmental covariate data, when using aggregated survey data in a non-linear model, for example for individual-level inferences [50]. It is called 'pure' because it specifically addresses model specification bias [51], and it biases the estimates because any direct link between exposure and health outcomes is imperfectly measured [52]. This is because the regression function does not approximate the real relationship between the affected population and their exposure [27]. Pure specification bias can be reduced as the within area exposure is more homogenous [50]. This could be done by having a finer partition of space at which environmental risk factors are available [50, 53].

In this study we propose to minimize and quantify pure specification bias by extracting covariate information from cities within barangays (Figure 2) and by modelling the disease using a convolution model [53]. The city level is the finest available extent of analysis. Cities thus serve as a proxy for individual-level exposure locations. We identified all cities within the surveyed barangays using Google Earth Images. Available cities were extracted from the 2010 build-up data base from the National Mapping and Resource Information Authority from the Philippines [54]. We completed unavailable cities using Google Earth Images.

For the convolution model, we used the aggregate data method proposed by Prentice and Sheppard [55]. For each SSA, we obtained covariate information $\boldsymbol{x}$ for image pixel i belonging to a city j within a specific barangay k (Figure 2). Let $n = 6$ be the number of covariates $\boldsymbol{x_{ijk}}$ measured at locations s_{ijk} $i = 1, \dots, m_k$ where m_k denotes the number of city pixels within barangay k. Note that with an increasing resolution the possibility increases that there are no pixel points falling in cities of the within-pixel sizes. Data on the human *Schistosoma japonicum* infection are available at individual-level recorded within a barangay k. Because the exact response locations of the individual-level data are unknown, we aggregated them to their corresponding barangay centroid, denoted by y_k. To estimate the average probability of infection of the individuals in barangay k, and the individual level coefficients $\boldsymbol{\beta}$, we obtained the mean risk function $\hat{\hat{p}}_k$ over the total number of city pixels or exposure locations (equation 2). We accounted for the spatial variability at barangay-level by adding a spatial structure random effects term s_k. Pure specification bias results as $\gamma \neq \beta$, and is then minimized by using the individual-level regression coefficients $\boldsymbol{\beta}$ instead of the group-level coefficients $\boldsymbol{\gamma}$. The accompanying uncertainties are quantified by the difference between the group an individual-level credible intervals [27] for each SSA. The convolution model used is of the form:

$$y_k|\boldsymbol{x_{ijk}}, \boldsymbol{\beta} \sim Binomial\left(N_k, \hat{\hat{p}}_k\right)$$

$$\hat{\hat{p}}_k = \frac{1}{m_k} \cdot \sum_{i=1}^{m_k} \frac{1}{1+\exp(-(\beta_0+\beta_1 \cdot x_{1ijk}+\beta_2 \cdot x_{2ijk}+\cdots+\beta_n \cdot x_{nijk}+s_k))}. \quad (2)$$

Model implementation

Five models were implemented, all including an intercept (β_0), pixel-level environmental variables ($\boldsymbol{x_{ijk}}$= NDVI, NDWI, LSTD, LSTN, E, NDWB) and their corresponding individual-level coefficients $\boldsymbol{\beta}$**.** Collinearity between covariates was assessed with the Pearson correlation coefficient. All covariates were standardized to have mean = 0 and standard deviation = 1.

The intercept β_0 was given a diffuse uniform prior distribution with wide bounds $\beta_0 \sim U[-100,100]$. The other $\boldsymbol{\beta}$ parameters were given a diffuse normal distribution $\boldsymbol{\beta} \sim N[0, \frac{1}{\sigma^2}]$, with σ uniformly distributed on a wide range of $\sigma \sim U[0, 100]$. These distributions avoid overestimating the parameters [56] and allow a good sequences mixing

used for Markov Chain Monte Carlo (MCMC) simulations, contributing to a fast convergence [57].

Prior information for the spatially structured random effects was based upon a geo-statistical model that can be used as a sampling distribution for continuous spatial data [58]. The vector of random variables $\boldsymbol{s}$ associated with point locations (x_k, y_k), $k = 1, \ldots, K$, was modelled with a multivariate normal distribution $s \sim MVN_K[\mu, \Sigma_{ab}]$ with mean μ = 0 and a covariance matrix $\Sigma_{ab} = \sigma^2 \cdot \exp[-(\phi \cdot d_{ab})^{\kappa}]$ defined by a powered exponential spatial decaying correlation function.

The covariance matrix Σ_{ab} is specified as a function of the distances d_{ab} between barangay centroids a and b, with the rate of decline of spatial correlation per unit of distance ϕ, the scalar parameter representing the overall variance σ^2 and the scalar parameter κ controlling the amount of spatial smoothing. Because extreme values of κ (0 and 2) could lead to undesirable smoothing, we used $\kappa = 1$. Prior information for ϕ was set to be uniform: $\phi \sim U[2\text{E}10^{-7}, 3\text{E}10^{-3}]$. These values give a diffuse but plausible prior range of correlations between 0.1 and 0.99 at the minimum distance between points (575 m), and between 0 and 0.3 at the maximum distance between points (< 552 km), assisting identifiability [59]. For σ^2, a half-normal distribution was selected: $\sigma^2 \sim HN[0,1]$ to restrict the prior σ^2 to positive values and avoid problems with convergence [56, 60].

To run the model we used three sequences or chains with 50000 iterations. This number of iterations ensured that the simulations were representative of target distributions and a stable convergence [57]. In order to diminish the influence of starting values, we discarded the first half of each sequence [57] using a burn-in of 25000 iterations. Convergence was monitored visually and statistically by inspecting the trace plots, and by checking the $\widehat{R}$ statistic [61, 62] also called the potential scale reduction factor. This potential scale reduction factor assesses sequences mixing by comparing the between and within variation. An $\hat{R}$ value < 1.1 indicates evidence that sequences have converged [61], while higher values suggest that an increase in the number of simulations may improve the inferences [57].

Survey and environmental data were structured in a rectangular format where columns are headed by the array name. Survey data and the codes in BUGS for the various SSA are provided in the Additional files 1 and 2 respectively.

Model validation

The five models were validated using two methods. The first method compared the data generated from the simulations of the predictive distribution to the observed data using a test statistic. A test statistic is a value derived from the sampled data and is used to perform hypothesis testing. This test statistic is the posterior predictive *p*-value (pp*p*-value) generated by calculating the proportion of the predicted values which are more extreme than the observed maximum, minimum and mean prevalence observed value. We calculated (i) the proportion of simulations of the data from the model for which the maximum prevalence across simulated barangays is greater than or equal to the maximum observed value, (ii) the proportion of simulations of the data from the model for which the minimum prevalence across simulated barangays is greater than or equal to the minimum observed value, and (iii) the proportion of simulations of the data from the model for which the mean prevalence across simulated barangays is greater than or equal to the mean observed value. If the model fits the data, the simulated values distribute closely around the observed values, thus, we expect a pp*p*-value of around 0.5. Otherwise, for a biased model, the pp*p*-value will be close to 0 or 1. Our aim was to test whether the model predicts a similar number of barangays with maximum and minimum prevalence values compared with the observed data. We generated pp*p* values for maximum, minimum and mean prevalence values for the models at five increasing SSA using 75000 simulations. The second method used the area under the curve (AUC) of the receiving operating characteristics (ROC). We applied a threshold of 0.5% (prevalence mean in Mindanao region) since we are interested in knowing the ability of the models to discriminate the mean prevalence level in the study area. We also examined the ability of the model to discriminate the number of positive cases, thus, we used a threshold of 1, which indicates the presence of at least one positive case. We used an AUC value of 70% to indicate acceptable predictive performance [8, 63].

Software

Model implementation was done in the software OpenBUGS 3.2.3 [64, 65] (Medical research Council, Cambridge, UK, and Imperial College London, UK). It was downloaded for free at [66]. We called Open BUGS from R using the package R2OpenBUGS [67]. The spatial models were coded using the GeoBUGS [59] function as an add-on module to OpenBUGS. GeoBUGS provides an interface to work with conditional autoregressive and geo-statistical models. Data pre-processing and Ordinary Kriging was performed in R [68].

Results

Modelling *Schistosoma japonicum* infection under the MAUP

Convolution model

Our findings show that NDVI has a non-significant effect on the prevalence of SCH infection for all SSA, except for SSA = 1 km (Table 2 and Figure 3a). NDVI estimates vary gradually from 0.19 to 0.26 when increasing SSA until 500 m. For SSA = 1 km, the estimate rapidly increases to 0.59. Uncertainties are similar throughout all SSA (Figure 4a and Table 3), slightly increasing when increasing SSA. The highest credible interval value is 0.60 for SSA = 250 m, and the lowest is 0.52 for SSA = 30 m.

NDWI has a significant negative effect on the prevalence of SCH infection throughout all SSAs (Table 2 and Figure 3b). When SSA increases, parameter estimates increase from -1.06 to -0.76, coming somewhat closer to zero. We found similar estimates for SSA = 30 m, 90 m and 250 m (i.e. -1.06 to -1.02), and for SSA = 500 m and 1 km (i.e. -0.8 to -0.76) (Figure 4b). Uncertainty values are similar for all SSAs and show a slight decrease when increasing SSA (Figure 4b and Table 3). The highest uncertainty value equals 0.54 for SSA = 30 m and the lowest value equals 0.44 for SSA = 500 m.

LSTD has a significant negative effect on the prevalence of SCH infection for almost all SSA, except for SSA = 1 km (Table 2 and Figure3c). Similar parameter estimates equal to -0.71 are obtained for SSA = 30 m, 90 m, and 250 m, while the parameter estimate increases slightly to -0.65 for SSA = 500 m. For SSA = 1 km, there is a noticeable increase in the parameter estimate to -0.01 (Figure 3c and Figure 4c). Uncertainty increases from 0.59 to 0.64 when increasing SSA from 30 m to 500 m, but for SSA = 1 km there is a considerable increase in uncertainty to 1.49 (Figure 4c).

LSTN has a significant negative effect on the prevalence of SCH infection for almost all SSA, except for SSA = 1 km (Table 2 and Figure3d). Parameter estimates increase from -0.78 to -0.86 while increasing SSA from 30 m to 500 m. For SSA = 1 km, the parameter estimate rapidly goes up to 0.1 (Figure 3d and Figure 4d). Uncertainty increases slightly from 0.56 to 0.58 when increasing SSA from 30 m to 500 m, but it increases considerably to 1.14 for SSA = 1 km (Table 3 and Figure 4d).

Elevation has a significant negative effect on the prevalence of SCH infection for all SSA, except for SSA = 1 km (Table 2 and Figure 3e). When increasing SSA from 30 m to

500 m, parameter estimates slightly decrease from -0.95 to -1.03. For SSA = 1 km, the parameter estimate considerably increases to -0.04 (Figure 3e and 4e). Uncertainty values vary from 0.59 to 0.64 when increasing SSA from 30 m to 500 m. For SSA = 1 km, uncertainty considerably decreases to 0.35 (Table 3 and Figure 4). The lowest uncertainty value is 0.35 for SSA = 1 km and the highest is 0.66 for SSA = 250 m.

Finally, NDWB has a significant negative effect on the prevalence of SCH infection for all SSA (Table 2 and Figure 3f). We found similar parameter estimates of -0.28, -0.29 and -0.31 for SSA = 30 m, 90 m, and 250 m, respectively, and estimates of -0.38 and -0.4 for SSA = 500 m and 1 km, respectively (Figure 3f). Uncertainties constantly increase from 0.32 to 0.39 (Table 3 and Figure 4f) when increasing SSA.

Intercept values range from -6.02 to -6.17 for almost all SSAs, except for SSA = 1 km, where it is equal to -5.49. The rate of decay of spatial autocorrelation (φ) ranges from $1.65x10^{-5}$ to $2.81x10^{-4}$ for SSAs = 1 km and 500 m, respectively.

Our findings show high and moderate correlation and determination (R^2) coefficient values between the SSAs and all environmental covariates estimates (Table 4) with correlation coefficients ranging from -0.94 to 0.94 and R^2 values from 0.6 to 0.86, respectively. Correlation coefficients between the SSAs and uncertainties are high for LSTD, LSTN and NDWB with the values of 0.91, 0.9 and 0.91, respectively. Determination coefficients R^2 between the SSAs and uncertainties in the covariates estimates are moderate for LSTD, LSTN and NDWB with the values of 0.76, 0.75 and 0.76, respectively (Table 4). Uncertainties in NDVI and NDWI estimates do not show any correlation with SSAs (Table 4).

Table 2. *Regression coefficient estimates for each risk factor at five descending spatial supports of analysis*

Influence on predictions

Differences between observed and predicted prevalence values are similar for the five SSA models (Figure 5). Variation in these differences is highest between the 30 m and 1 km models ($R^2 = 0.94$) and lowest between the 30 m and 90 m models ($R^2 = 0.99$). Figure 5 shows that the maximum and minimum differences are 1.11% and 0.01%, respectively, corresponding to the 1 km SSA model. For fitted prevalence values higher than 2% all models underestimate the prevalence of infection, while, for fitted prevalence values lower than 2%, overestimation and underestimation occur for the five models (Figure 5). A plot of the

residuals against prevalence from Figure 5 serve as a visual inspection of the fit, where we realize that it is based on positive predictive predictions.

Uncertainties on the predictions are similar for the five models (Additional file 3). Higher differences in uncertainty were found between the 500 m and 1 km models ($R^2 = 0.96$), and lower differences were found between the 90 m and 250 m models ($R^2 = 0.99$). The highest uncertainty value is 9.23% for all the models, except the 1 km model with 8.9%, and the lowest uncertainty value is 0.006% for the 1 km model.

Model validation

The maximum and minimum observed prevalence values are 8.5% and 0.33%, respectively. The first validation method shows pp*p*-values for all SSA ranging from 0.64 to 0.67 for the first test statistic (Table 5). This means that simulated data slightly deviate from around 0.14 to 0.17 from the maximum observed prevalence value (Figure 6). For all SSA it is likely to see a similar number of predicted maximum prevalence values compared to the observed data. For the second test statistic, pp*p*-values ranged from 0.87 to 0.93 (Table 5). This means that simulated data are biased around 0.36 to 0.43 from the minimum observed prevalence data (Figure 7). For almost all SSA, simulated data predict a higher number of minimum prevalence values compared to the observed data. For the last test statistics, pp*p*-values ranged from 0.59 to 0.67 (Table 5), showing that simulated data deviate from around 0.09 to 0.17 from the mean observed prevalence value (Figure 8).

Results from the second validation method show that all models have a high ability to predict prevalence values, with AUC values of 0.91 for SSA = 30 m, 90 m, 250 m, and 500 m, and 0.93 for SSA = 1 km. All models have a good ability to predict the positive number of SCH cases. Models with SSA = 30 m, 90 m, 250 m, and 500 m models have AUC values of 0.83, while the 1 km SSA model presents a lower AUC value of 0.79, showing a decrease in the ability to predict the positive number of SCH cases.

Discussion

Schistosomiasis modelling studies have commonly used environmental risk factors as drivers for disease exposure and transmission [69, 70]. The studies so far have used spatially misaligned environmental variables at different spatial supports of analysis, ignoring MAUP effects on the parameter estimates, predictions, and the relationship between disease morbidity indicators and risk factors. This study is the first effort to quantify the effects of modifying the areal unit (i.e. spatial support) of NDVI, NDWI, LSTD, LSTN, E, and NDWB,

on model parameter estimates and their uncertainties. Uncertainty may be quantified using measures of accuracy or imprecision [15]. We evaluated uncertainty using measures of imprecision based on the nature of the disease and the survey data available and quantified it using credible intervals in a Bayesian setting. We applied it to *Schistosoma japonicum* infection modelling in the Mindanao region, the Philippines.

Our findings show that the environmental risk factors NDVI, NDWI, LSTD, LSTN, and E behave similarly when increasing the SSA from 30 m to 1 km (Table 4). An increase in SSA from 30 m to 500 m does not represent any significant changes in parameter estimates. Conversely, for SSA = 1km, all show a considerable increase in their estimates. The reasons are explained below.

NDVI has a positive effect on SCH, meaning that higher NDVI values increase the prevalence of infection. This is explained by the positive relationship between vegetation, moisture and snail density [37]. NDVI effects are not significant for SSA < 1 km, because NDVI is an indicator of greenness that is mainly effective for arid areas and Mindanao is not arid. However, the NDVI effect becomes significant on the prevalence of SCH infection for SSA = 1 km. This could be because NDVI effects on SCH prevalence are greater at global scales [8] than at local scales. This might be explained by the fact that prevalence values al local scales can vary significantly at nearby locations, as it depends not only on the nature of the parasite lifecycle, which requires optimal habitat conditions (i.e. environmental conditions), but also on sanitation conditions on the area [71]. The increase in uncertainty values with increasing SSA is due to the coarse areal pixels ≥ 250 m resolution that does not reliably represent rice paddy fields. Those are substantially smaller than 25 ha, i.e. are covered by at most four pixels [72].

For SSA = 30 m, 90 m, 250 m, and 500 m, LSTD, LSTN, and E have a significant negative effect on SCH prevalence. Conversely for SSA = 1 km, their parameter estimates are close to zero. This means that when the areal unit reaches 1 km, the effect of these covariates on the prevalence of SCH infection becomes non-significant. This is also observed from the credible intervals of these covariates for the 1 km SSA model. The reason is that the homogeneity of the covariate values increases when increasing the SSA. LST is a proxy of the ambient temperature of the air, which reflects the thermal conditions of shallow waters [27]. Its negative relationship with the prevalence of infection could be explained by the fact that temperatures above 19 to 20°C do not influence the release of cercariae from the infected host

to the infection foci [73], as well as temperatures below approximately 15°C arrest the development of *S.japonicum* in the snail host [74]. The minimum LST value at night is around 21°C, while the maximum LST value during the day is 31°C. LSTD and LSTN uncertainty values for SSA = 1 km are remarkably high as compared to other SSA. This is explained by the coarse LSTD and LSTN areal pixels of 1 km^2 that cannot reliably represent low and high temperature zones in city areas that range from 0.02 to 3 km^2 [27]. Elevation has a negative effect on SCH. This was expected as in areas with high elevation values (>2300m) the risk of infection is low [75]. Conversely, the risk of infection is high for elevation areas below 900m. Elevation uncertainty values are similar for all SSA, except for SSA = 1 km, where its value considerably decreases to 0.34. Here we see the effect of the gradual changes of elevation in Mindanao region are gradual and without steep slopes [27]. Using data directly at the 1 km SSA could give reliable elevation values, but with a non-significant effect on the disease prevalence.

For NDWB and NDWI, an increase in SSA from 30 m to 250 m represents non-significant changes in parameter estimates, which range from -1.06 to -1.02 for NDWI, and from -0.31 to -0.28 for NDWB. Conversely, when increasing the SSA to 500 m, parameter estimates change to -0.8 and -0.38 for NDWI and NDWB, respectively. For SSA = 500 m and 1 km, NDWI estimates increase, having a less significant effect on SCH prevalence, again due to the increase in the homogeneity of the covariate values when increasing SSA. Higher NDWI values show the presence of potential hidden infection foci. Nevertheless, results show that NDWI presents a negative effect on SCH (Table 2). This could be because NDWI cannot efficiently suppress the signal from build-up land mixing enhanced water features with build-up land noise. Thus, build-up noise could also have high NDWI values [40]. According to Gu et al. [76] NDWI values lower than 0.3 indicate the presence of drought areas. In our study area, we found that around 77% of Mindanao present drought conditions, explaining the negative effect on the disease. NDWB estimates decrease when increasing SSAs (Table 4), specially for SSA = 500 m and 1 km, but their significance on SCH prevalence increases. A possible explanation is that people that move larger distances to water bodies are most likely to get infected. This could be because at spatial supports < 1 km, NDWB values seem to be more homogenous than at smaller spatial supports, showing a weaker relationship with the disease (Table 2). For spatial supports > 1km, neighbouring pixels present more heterogeneous values, which could be because of the aggregation process, but also because of the use of some kind of transportation media that allows apparent reduction of travel distances

503 in a relatively large area (1 km^2). Clearly, transportation (type of road and media of
504 transportation) plays an important role [77].

505 Uncertainty values for NDWI decrease when increasing the SSA, with a minimum of
506 0.44 for SSA = 500 m. Clearly, NDWI data originally available at SSA = 250 m are more
507 reliable than values modified to larger SSAs. Using Ordinary Kriging for interpolation
508 increases the variance in the estimates in a somewhat unrealistic way since it uses a constant
509 mean [58], while in reality, means are different. Uncertainty values of NDWB, for instance,
510 increase with increasing SSA due to the coarse areal pixel units ≥ 0.25 km^2. Such a size is
511 insufficient to reliable define nearest distances to water bodies in city areas of 0.02 to 3 km^2.

512 Our aim was not to compare the performance of the models as we used the same
513 model structure, number and type of covariates in the five models. Thus the model itself is
514 practically the same for all SSA. Although our aim was not focused on model comparison, the
515 resulting DIC values from Table 2 suggest the use of spatial support sizes below or equal to
516 250 m in SCH modelling. This is shown by the low DIC values from 86.67 to 140.5 for SSA
517 $\leq$ 250 m, and high DIC values of 143.7 and 147.5 for the 500 m and 1 km models,
518 respectively (Table 2).

519 When modelling prevalence of *Schistosoma japonicum* infection in Mindanao, the
520 effect of increasing SSA, or modifying the areal unit of analysis, from 30 m to 500 m,
521 produces a gradual and continuous increase on the parameter estimates and their associated
522 uncertainties. For SSA = 1 km, sudden changes occur in the relationship between the risk
523 factors and the prevalence of the disease. This is shown by the non-significant effect of almost
524 all explanatory variables on *Schistosoma japonicum* prevalence. Results suggest that the use
525 of environmental data extracted at SSA = 1 km is not appropriate for the modelling of
526 *Schistosoma japonicum* prevalence.

527 Bayesian statistical methods were used to model the disease, and along with a
528 convolution regression model, they corrected for pure specification bias on our estimates.
529 This is a relevant contribution to the analysis of uncertainties in this type of spatial
530 epidemiological study. For future studies, new trends in geospatial artificial intelligence
531 (geoAI), that could resolve limitations regarding the MAUP for exposure modelling studies,
532 are emerging to model schistosomiasis [78] as well as other diseases [79]. We particularly
533 identified (i) the use of high-performance computing to handle spatiotemporal big data, and
534 (ii) machine and deep learning algorithms implementation to big data infrastructures to extract

relevant disease or environmental information [79, 80]. One example is a data-driven method used to predict particulate matter air pollution ($PM_{2.5}$) in Los Angeles, CA, USA. Here, machine learning was used on spatial big data, i.e. land use and roads, derived from OpenStreeMap, to predict $PM_{2.5}$ concentrations. When generating relative importance measures for the different risk factors, MAUP effects reduced when applying a random forest model that was trained with the distances between the features and the monitoring $PM_{2.5}$ stations, [81]. The rapid development of geoAI methods, their advantage to deal with big data, and their rapid computational time, makes them an attractive and advantageous tool to tackle limitations with modelling schistosomiasis and other diseases. There is still little work done in this field, but we think it is valuable to further explore geoAI solutions to deal with the MAUP, and perhaps other inherent uncertainties produced in disease modelling and mapping.

Finding MAUP effects on the various environmental risk factors used for modelling *Schistosoma japonicum* prevalence, is a step forward to the uncertainty analysis in the schistosomiasis, and possibly other diseases. The present research deals with limitations such as the use of aggregated disease data, due to the lack of geo-located individual-level surveys. It also provides a robust method for the selection of an appropriate spatial data structure, which at the same time, enables the acquisition of more reliable parameter estimates, and defines a clear relationship between the risk factors and the disease. From the public health perspective, this research can support helminth control programs by providing less uncertain models and maps. Epidemiologists and health scientists could use these maps to identify risk areas for the control and prevention of the disease [12, 82], which in the case of schistosomiasis, is generally based on mass drug administration campaigns addressed to the identified at-risk populations. The provision of reliable information is relevant to guide mass drug administration campaigns by enhancing the assessment of the infection risk, understanding its potential impacts on human health [15, 83] and avoiding erroneous conclusions and decisions about the spatial distribution of schistosomiasis [15, 27]. This research is also relevant to evaluate the effectiveness of mass drug administration campaigns, as it could guide the identification of persistent hotspots, or places where prevalence of infection remains despite mass drug administration efforts [71]. It is known that despite the implementation of mass drug administration campaigns, some places do not show a decrease in local SCH transmission. This is because these campaigns do not only depend on the nature of the parasite lifecycle and the poor sanitation conditions, but also on the local environmental

factors, drivers for SCH transmission. Finding relevant environmental factors at local level would allow more intensive efforts at persistent hotspots.

Conclusions

The present study shows a clear MAUP effect on *Schistosoma japonicum* modelling. An increase in parameter estimates and their associated uncertainties occurs when increasing the spatial support of analysis (SSA). It also showed that using environmental data extracted at SSA = 1 km is not relevant for *Schistosoma japonicum* prevalence of infection at this specific extent of analysis, as this leads to wrong conclusions about the distribution of the disease and its relationship with the potential risk factors. Thus, the use of maps based upon this information is to be avoided as these may guide health scientists in the control or prevention of the disease astray. The results from this study could guide other disease modelling studies as they suggest a spatial support sizes at which environmental information has no longer a significant effect on the disease, and which data structure is recommended for the modelling. Epidemiologists, decision makers, and health scientists could thus benefit from those, e.g. to better understand and quantify MAUP effects on the relationship between the disease and its risk factors, and to provide reliable maps that are useful for disease control and prevention.

Additional files

Additional file 1: Table S1. Survey data aggregated at the barangay level. These data show the number of positive cases (y) and the total number of sampled people (n) in a barangay (k).

Additional file 2: Text S1. Code for the convolution model used in OpenBUGS. It includes the prior distributions used for the covariate and spatial parameters, and the model itself. The model uses two indexes, *k* for the barangay and *j* for the number of city pixels within barangays. The betas are the individual-level regression coefficients, *m* is the number of city pixels in a barangay. Covariates are *ndvi*: normalized difference vegetation index, *ndwi*: normalized difference water index, *lstd*: land surface temperature day, *lstn*: land surface temperature night, *e*: elevation, *ndwb*: nearest distance to water bodies. The spatial parameter is represented as *s*.

Additional file 3: Figure S1. Residual plot for the five increasing spatial supports of analysis. a) SSA = 30m, b) SSA = 90m, c) SSA= 250m, d) SSA= 500m, e) SSA= 1km. The x axis represents the fitted prevalence values for the five spatial supports of analysis. The y axis

represents the residuals calculated by the difference between the observed and predicted prevalence values.

Abbreviations

ASTER: Advanced spaceborne thermal emission and reflection radiometer; AUC: Area under the curve; BUGS: Bayesian inference using Gibbs sampling; E: Elevation; EO: Earth observation; geoAI: Geospatial artificial intelligence; GDEM: global digital elevation map; GIS: Geographic information systems; LSTD: Land surface temperature day; LSTN: Land surface temperature night; MAUP: Modifiable areal unit problem; MCMC: Markov chain Monte Carlo; MODIS: Moderate resolution imaging spectroradiometer; NDVI: Normalized difference vegetation index; NDWB: Nearest distance to water bodies; NDWI: Normalized difference water index; ROC: Receiver operator characteristic; SCH: Schistosomiasis; SSA: Spatial support of analysis; USGS: United states geological survey.

Declarations

Ethics Approval

The data used in this study were collected in 2005 when there was no requirement for ethical review and clearance. This study used aggregated survey data at the barangay level, which enabled the full de-identification of individuals involved in the survey.

The study results represent part of the PhD thesis entitled “Statistical evaluation of spatial uncertainty in schistosomiasis mapping”, which was published as: Araujo Navas AL. Statistical evaluation of spatial uncertainty in schistosomiasis mapping. Enschede: University of Twente, Faculty of Geo-Information Science and Earth Observation (ITC), 2019. 157 p. (ITC Dissertation). https://doi.org/10.3990/1.9789036548281.

Consent for publication

Not applicable

Availability of data and material

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests

Funding

The research team acknowledges the World Health Organization funds provided to conduct the surveys. This research received no external funding. ALAN's doctoral research is funded by the University of Twente. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

Conceptualization, ALAN, FO and RJSM; Formal analysis, ALAN; Investigation, ALAN; Methodology, ALAN and FO; Resources, LRL and RJSM; Software, ALAN; Supervision, FO, RJSM and AS; Validation, ALAN; Visualization, ALAN; Writing – original draft, ALAN; Writing – review & editing, ALAN, FO, RJSM and AS. All authors read and approved the final

Acknowledgments

Not applicable

References

1. Walz Y, Wegmann M, Dech S, Vounatsou P, Poda J-N, N'Goran EK, et al. Modeling and validation of environmental suitability for schistosomiasis transmission using remote sensing. PLoS Negl Trop Dis. 2015;9:e0004217.
2. Leenstra T, Acosta LP, Langdon GC, Manalo DL, Su L, Olveda RM, et al. *Schistosoma japonicum*, anemia, and iron status in children, adolescents, and young adults in Leyte, Philippines. Am J Clin Nutr. 2006;83:371–79.
3. Coutinho HM, McGarvey ST, Acosta LP, Manalo DL, Langdon GC, Leenstra T, et al. Nutritional status and serum cytokine profiles in children, adolescents, and young adults with *Schistosoma japonicum*-associated hepatic fibrosis, in Leyte, Philippines. J Infect Dis. 2005;192:528–36.
4. Jia TW, Zhou XN, Wang XH, Utzinger J, Steinmann P, Wu XH. Assessment of the age-specific disability weight of chronic *Schistosomia japonicum*. Bull World Health Organ. 2007;85:458–65.
5. Tarafder MR, Balolong E, Carabin H, Belisle P, Tallo V, Joseph L, et al. A cross-sectional study of the prevalence of intensity of infection with *Schistosoma japonicum* in 50 irrigated and rain-fed villages in Samar province, the Philippines. Bmc Public Health. 2006;6:10.
6. Yang K, Wang XH, Yang GJ, Wu XH, Qi YL, Li HJ, et al. An integrated approach to identify distribution of *Oncomelania hupensis,* the intermediate host of *Schistosoma japonicum*, in a mountainous region in China. Int J Parasitol. 2008;38:1007–16.
7. Hotez PJ, Alvarado M, Basanez MG, Bolliger I, Bourne R, Boussinesq M, et al. The global burden of disease study 2010: interpretation and implications for the neglected tropical diseases. PLoS Negl Trop Dis. 2014;8:e2865.
8. Soares Magalhães RJ, Salamat MS, Leonardo L, Gray DJ, Carabin H, Halton K, et al. Geographical distribution of human *Schistosoma japonicum* infection in The Philippines: tools to support disease control and further elimination. Int J Parasitol. 2014;44:977–84.
9. Herbreteau V, Salem G, Souris M, Hugot J-P, Gonzalez J-P. Thirty years of use and improvement of remote sensing, applied to epidemiology: from early promises to lasting frustration. Health Place. 2007;13:400–3.

10. Kalluri S, Gilruth P, Rogers D, Szczur M. Surveillance of arthropod vector-borne infectious diseases using remote sensing techniques: a review. PLoS Pathogens. 2007;3:e116.
11. Zhang ZJ, Manjourides J, Cohen T, Hu Y, Jiang QW. Spatial measurement errors in the field of spatial epidemiology. Int J Health Geogr. 2016;15:12.
12. Soares Magalhães RJ, Clements ACA, Patil AP, Gething PW, Brooker S. The applications of model-based geostatistics in helminth epidemiology and control. Adv Parasitol. 2011;74:267–96.
13. Cadavid Restrepo AM, Yang YR, McManus DP, Gray DJ, Giraudoux P, Barnes TS, et al. The landscape epidemiology of echinococcoses. Infec Dis Poverty. 2016;5.
14. Weiss DJ, Mappin B, Dalrymple U, Bhatt S, Cameron E, Hay SI, et al. Re-examining environmental correlates of *Plasmodium falciparum* malaria endemicity: a data-intensive variable selection approach. Malar J. 2015;14.
15. Araujo Navas AL, Hamm NAS, Soares Magalhães RJ, Stein A. Mapping soil transmitted helminths and schistosomiasis under uncertainty: a systematic review and critical appraisal of evidence. PLoS Negl Trop Dis. 2016;10:e0005208.
16. Wang XH, Zhou XN, Vounatsou P, Chen Z, Utzinger J, Yang K, et al. Bayesian spatio-temporal modeling of *Schistosoma japonicum* prevalence data in the absence of a diagnostic 'gold' standard. PLoS Negl Trop Dis. 2008;2:e250.
17. Walz Y, Wegmann M, Leutner B, Dech S, Vounatsou P, N'Goran EK, et al. Use of an ecologically relevant modelling approach to improve remote sensing-based schistosomiasis risk profiling. Geospat Health. 2015;10:271–79.
18. Young LJ, Gotway CA, Yang J, Kearney G, DuClos C. Assessing the association between environmental impacts and health outcomes: a case study from Florida. Stat Med. 2008;27:3998–4015.
19. Simoonga C, Utzinger J, Brooker S, Vounatsou P, Appleton CC, Stensgaard AS, et al. Remote sensing, geographical information system and spatial analysis for schistosomiasis epidemiology and ecology in Africa. Parasitology. 2009;136:1683–93.
20. Atkinson PM, Graham AJ. Issues of scale and uncertainty in the global remote sensing of disease. In: Hay SI, Graham A, Rogers DJ, editors. Global mapping of infectious diseases: methods, examples and emerging applications. San Diego: Elsevier Academic Press Inc; 2006. p. 79–118
21. Schur N, Hurlimann E, Garba A, Traore MS, Ndir O, Ratard RC, et al. Geostatistical model-based estimates of schistosomiasis prevalence among individuals aged <= 20 Years in west Africa. PLoS Negl Trop Dis. 2011;5:e1194.
22. Schur N, Hurlimann E, Stensgaard AS, Chimfwembe K, Mushinge G, Simoonga C, et al. Spatially explicit schistosoma infection risk in eastern Africa using bayesian geostatistical modelling. Acta Trop. 2013;128:365–77.
23. Dungan JL, Perry JN, Dale MRT, Legendre P, Citron-Pousty S, Fortin MJ, et al. A balanced view of scale in spatial statistical analysis. Ecography. 2002;25:626–40.
24. Openshaw S. The modifiable areal unit problem. : Norwick: GeoBooks; 1984.
25. Hellsten AS. A spatio-temporal ammonia emissions inventory for the UK. https://www.era.lib.ed.ac.uk/handle/1842/24693. Accessed April 2018.
26. Schur N, Vounatsou P, Utzinger J. Determining treatment needs at different spatial scales using geostatistical model-based risk estimates of schistosomiasis. PLoS Negl Trop Dis. 2012;6:e1773.
27. Araujo Navas AL, Osei F, Leonardo LR, Soares Magalhães RJ, Stein A. Modeling *Schistosoma japonicum* Infection under pure specification bias: impact of environmental drivers of infection. Int J Env Res Pub He. 2019;16:176.
28. Leonardo L, Rivera P, Saniel O, Solon JA, Chigusa Y, Villacorte E, et al. New endemic foci of schistosomiasis infections in the Philippines. Acta Trop. 2015;141:354–60.
29. Leonardo L, Acosta LP, Olveda RM, Aligui GDL. Difficulties and strategies in the control of schistosomiasis in the Philippines. Acta Trop. 2002;82:295–99.
30. Zhou XN, Bergquist R, Leonardo L, Yang GJ, Yang K, Sudomo M, et al. *Schistosoma japonicum*: control and research needs. In: Zhou XN, Bergquist R, Olveda R, Utzinger J, editors. Important

719 helminth infections in southeast Asia: diversity and potential for control and elimination, Part A. San
720 Diego Elsevier Academic Press Inc; 2010. p. 145–78.
721 31. Leonardo L, Rivera P, Saniel O, Villacorte E, Lebanan MA, Crisostomo B, et al. A national
722 baseline prevalence survey of schistosomiasis in the Philippines using stratified two-step systematic
723 cluster sampling design. J Trop Med. 2012;2012:8.
724 32. Leonardo LR, Rivera P, Saniel O, Villacorte E, Crisostomo B, Hernandez L, et al. Prevalence
725 survey of schistosomiasis in Mindanao and the Visayas, The Philippines. Parasitol Int. 2008;57:246–
726 51.
727 33. DIVA-GIS free, simple and effective. Hijmans R, Rojas E, Cruz M, O'Brien R, Barrantes I.,
728 University of California and International Potato Center in Peru 2018. http://www.diva-gis.org/Data.
729 Accessed 8 April 2018.
730 34. Santos FLN, Cerqueira EJL, Soares NM. Comparison of the thick smear and Kato-Katz
731 techniques for diagnosis of intestinal helminth infections. Rev Soc Bras Med Trop. 2005;38:196–98.
732 35. Brooker S, Hay S, Issae W, Hall A, Kihamia C, Lawambo N, et al. Predicting the distribution of
733 urinary schistosomiasis in Tanzania using satellite sensor data. Trop Med Int Health. 2001;6:998–
734 1007.
735 36. Brooker S, Hay SI, Tchuente L-AT, Ratard R. Using NOAA-AVHRR data to model human
736 helminth distributions in planning disease control in Cameroon, West Africa. Photogramm Eng Rem
737 S. 2002;68:175–79.
738 37. Walz Y, Wegmann M, Dech S, Raso G, Utzinger J. Risk profiling of schistosomiasis using
739 remote sensing: approaches, challenges and outlook. Parasit Vectors. 2015;8:16.
740 38. Malone JB, Yilma JM, McCarroll JC, Erko B, Mukaratirwa S, Zhou XY. Satellite climatology and
741 the environmental risk of *Schistosoma mansoni* in Ethiopia and east Africa. Acta Trop. 2001;79:59–
742 72.
743 39. Zhou YB, Liang S, Jiang QW. Factors impacting on progress towards elimination of
744 transmission of *Schistosoma japonicum* in China. Parasit Vectors. 2012;5:7.
745 40. Xu H. Modification of normalised difference water index (NDWI) to enhance open water
746 features in remotely sensed imagery. Int J Remote Sens. 2006;27:3025–33.
747 41. Woolhouse M, Chandiwana S. Population dynamics model for Bulinus globosus, intermediate
748 host for *Schistosoma haematobium*, in river habitats. Acta Trop. 1990;47:151–60.
749 42. Pietrock M, Marcogliese DJ. Free-living endohelminth stages: at the mercy of environmental
750 conditions. Trends Parasitol. 2003;19:293–99.
751 43. Stensgaard AS, Jorgensen A, Kabatereine NB, Rahbek C, Kristensen TK. Modeling freshwater
752 snail habitat suitability and areas of potential snail-borne disease transmission in Uganda. Geospat
753 Health. 2006;1:93-104.
754 44. Stensgaard AS, Utzinger J, Vounatsou P, Hurlimann E, Schur N, Saarnak CFL, et al. Large-scale
755 determinants of intestinal schistosomiasis and intermediate host snail distribution across Africa: does
756 climate matter? Acta Trop. 2013;128:378–90.
757 45. ESRI. ArcGIS desktop. New release simplifies your work, provides new ways to share
758 information, supplies GIS in the cloud. 10 ed: Environmental systems research institute; 2011.
759 46. Planet OSM. Open Street Map Project. 2017. https://planet.osm.org. Accessed 21 Nov 2017.
760 47. Humanitarian data exchange v1.25.3. United nations office for the coordination of
761 humanitarian affairs, New York and Geneva. 2018.
762 https://data.humdata.org/search?groups=phl&q=&ext_page_size=25. Accessed 10 April 2018.
763 48. Global Data Explorer. U.S. Department of interior 2017. https://gdex.cr.usgs.gov/gdex/.
764 Accessed 7 August 2017.
765 49. King G. A solution to the ecological inference problem: reconstructing individual behavior
766 from aggregate data. 1st ed. Princeton: Princeton University Press; 2013.
767 50. Wakefield J, Lyons H. Spatial aggregation and the ecological fallacy. In: G. F, editor. Handbook
768 of modern statistical methods. Boca Raton: Chapman & Hall/CRC Press; 2010. p. 541–58.

769 51. Gelfand AE, Diggle P, Guttorp P, Fuentes M. Handbook of spatial statistics. 1st ed. Boca
770 Raton: Taylor & Francis Group; 2010.
771 52. Richardson S, Monfort C. Ecological correlation studies. In: Elliot P, Wakefield JC, Best NG,
772 Briggs DJ, editors. Spatial epidemiology: methods and applications. Oxford: Oxford University Press;
773 2000. p. 205–20.
774 53. Wakefield J, Shaddick G. Health-exposure modeling and the ecological fallacy. Biostatistics.
775 2006;7:438–55.
776 54. National Mapping and Resource Information Authority (NAMRIA). Department of
777 Environment and Natural Resources. 2018. http://www.namria.gov.ph/. Accessed 3 February 2018.
778 55. Prentice RL, Sheppard L. Aggregate data studies of disease risk factors. Biometrika.
779 1995;82:113–25.
780 56. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on an
781 article by Browne and Draper). Bayesian Anal. 2006;1:515–33.
782 57. Gelman A, Carlin JB, Stern HS, Dunson D, Rubin DB. Bayesian data analysis. 3rd ed. New York:
783 Taylor & Francis Group; 2013.
784 58. Diggle PJ, Tawn J, Moyeed R. Model-based geostatistics. J R Stat Soc Ser C Appl Stat.
785 2002;47:299–350.
786 59. Thomas A, Best N, Lunn D, Arnold R, Spiegelhalter D. GeoBugs user manual. In: MRC
787 Biostatistics Unit. 2004. https://www.mrc-bsu.cam.ac.uk/software/bugs/thebugs-project-geobugs/.
788 Accessed February 2018.
789 60. Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. The BUGS book: a practical
790 introduction to bayesian analysis. 1st ed. Boca raton: Taylor & Francis Group; 2012.
791 61. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. J
792 Comput Graph Stat. 1998;7:434–55.
793 62. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci.
794 1992;7:457–72.
795 63. Brooker S, Hay SI, Bundy DAP. Tools from ecology: useful for evaluating infection risk
796 models? Trends Parasitol. 2002;18:70–4.
797 64. Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS user manual. In: MRC Biostatistics Unit.
798 2003. http://www.mrc-bsu.cam.ac.uk/bugs. Accessed February 2018.
799 65. Spiegelhalter D, Thomas A, Best N, Lunn D. OpenBUGS user manual, version 3.0. 2. In: MRC
800 Biostatistics Unit. 2007. http://www.openbugs.net/w/Manuals. Accessed February 2018.
801 66. Lunn D. SD, Thomas A. and Best N. OpenBUGS version 3.0.2. In: Downloads. 2018.
802 http://www.openbugs.net/w/Downloads. Accessed 25 June 2018.
803 67. Sturtz S, Ligges U, Gelman A. R2OpenBUGS: a package for running OpenBUGS from R. Journal
804 of Statistical Software. 2005;12:1–16.
805 68. Team RDC. R: A language and environment for statistical computing. Vienna, Austria: The R
806 foundation for statistical computing; 2013.
807 69. Hu Y, Bergquist R, Lynn H, Gao F, Wang Q, Zhang S, et al. Sandwich mapping of
808 schistosomiasis risk in Anhui province, China. Geospat Health. 2015;10:111–16.
809 70. Stensgaard AS, Vounatsou P, Sengupta ME, Utzinger J. Schistosomes, snails and climate
810 change: current trends and future expectations. Acta Trop. 2019;190:257–68.
811 71. Kittur N, Binder S, Campbell Jr CH, King CH, Kinung'hi S, Olsen A, et al. Defining persistent
812 hotspots: areas that fail to decrease meaningfully in prevalence after multiple years of mass drug
813 administration with praziquantel for control of schistosomiasis. Am J Trop Med Hyg. 2017;97:1810–
814 17.
815 72. Rice science for a better world. In: International Rice Research Institute. 2018.
816 http://irri.org/our-work/research/policy-and-markets/mapping-rice-in-the-philippines-where.
817 Accessed 22 October 2018.
818 73. Bauman PM, Bennett HJ, Ingalls Jr JW. The Molluscan intermediate host and *Schistosoma*
819 *japonicum*. Am J Trop Med Hyg. 1948;1:567–75.

74. Yang G-J, Utzinger J, Sun L-P, Hong Q-B, Vounatsou P, Tanner M, et al. Effect of temperature on the development of *Schistosoma japonicum* within *Oncomelania hupensis*, and hibernation of O. hupensis. Parasitology Research. 2007;100:695–700.
75. Pesigan TP, Hairston NG, Jauregui JJ, Garcia EG, Santos AT, Santos BC, et al. Studies on *Schistosoma japonicum* infection in The Philippines 2. The molluscan host. Bull World Health Organ. 1958;18:481–578.
76. Gu Y, Brown JF, Verdin JP, Wardlow B. A five-year analysis of MODIS NDVI and NDWI for grassland drought assessment over the central great plains of the United States. Geophys Res Lett. 2007;34.
77. Kummu M, de Moel H, Ward PJ, Varis O. How close do we live to water? A global analysis of population distance to freshwater bodies. PLoS One. 2011;6:e20578.
78. Mari L, Gatto M, Ciddio M, Dia ED, Sokolow SH, De Leo GA, et al. Big-data-driven modeling unveils country-wide drivers of endemic schistosomiasis. Sci Rep. 2017;7:11.
79. VoPham T, Hart JE, Laden F, Chiang YY. Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. Environ Health. 2018;17.
80. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. In: Fielding JE, Brownson RC, Green LW, editors. Annual review of public health 2018. p. 95–112.
81. Lin Y, Chiang Y-Y, Pan F, Stripelis D, Ambite JL, Eckel SP, et al. Mining public datasets for modeling intra-city PM2.5 concentrations at a fine spatial resolution. Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems; Redondo Beach, CA, USA. 3140013: ACM; 2017. p. 1-10.
82. Montresor A, Crompton DW, Hall A, Bundy D, Savioli L, Organization WH. Guidelines for the evaluation of soil-transmitted helminthiasis and schistosomiasis at community level: a guide for managers of control programmes. Geneva: World Health Organization; 1998.
83. Burns CJ, Wright M, Pierson JB, Bateson TF, Burstyn I, Goldstein DA, et al. Evaluating uncertainty to strengthen epidemiologic data for use in human health risk assessments. Environ Health Perspect. 2014;122:1160–65.

Figure Legends

Fig 1. Study Area: The Mindanao region in The Philippines. Blue dots are the aggregated survey data at barangay-level.

Fig 2. Environmental risk factors extraction at pixel-level from cities within barangays

Fig 3. Posterior estimates and their credible intervals: a) Normalized difference vegetation index; b) Normalized difference water index; c) Land surface temperature day d) Land surface temperature night; e) Elevation; f) Nearest distance to water bodies. *Abbreviations:* SSA, Spatial support of analysis

Fig 4. Density plots for the risk factors regression coefficients: a) Normalized difference vegetation index; b) Normalized difference water index; c) Land surface temperature day d) Land surface temperature night; e) Elevation; f) Nearest distance to water bodies

Fig 5. Residual plot for the five increasing spatial supports of analysis. The x axis represents the fitted prevalence values for the five spatial supports of analysis. The y axis represents the residuals calculated by the difference between the observed and predicted prevalence values.

Fig 6. Proportion of simulated prevalence data that fit the observed maximum prevalence value. a) SSA=30m, b) SSA=90m, c) SSA=250m, d) SSA=500m, e) SSA=1 km. *Abbreviations:* SSA, Spatial support of analysis.

Fig 7. Proportion of simulated prevalence data that fit the observed minimum prevalence value. a) SSA=30m, b) SSA=90m, c) SSA=250m, d) SSA=500m, e) SSA=1 km. *Abbreviations:* SSA, Spatial support of analysis.

Fig 8. Proportion of simulated prevalence data that fit the observed mean prevalence value. a) SSA=30m, b) SSA=90m, c) SSA=250m, d) SSA=500m, e) SSA=1 km. *Abbreviations:* SSA, Spatial support of analysis.

Tables

Table 1. Environmental variables description

Environmental variable	Spatial resolution	Temporal resolution	Data Type	Original coordinate system	Data Source
Elevation	30 m	NA	Raster	EPSG:4326	ASTER GDEM V2 from USGS
NDVI	250 m	2008	Raster	EPSG:4326	MOD13Q1
	30 m	2008	Raster	EPSG:4326	Landsat 5
NDWI	500 m	2008	Raster	EPSG:32651	Landsat 7, 1-year composite
	30 m	2008	Raster	EPSG:4326	Landsat 5
LST	1 km	2008	Raster	EPSG:4326	MOD11A2
NDWB	250 m	2010	Raster	EPSG:32651	Derived from closest facility network using roads, urban areas, river network, and water bodies

NDVI: Normalized difference vegetation index; NDWI: Normalized difference water index; LST: Land surface temperature day and night; NDWB: Nearest distance to water bodies. USGS: United States Geological Survey.

882

883

884

885

Table 2. Regression coefficient estimates for each risk factor at five descending spatial supports of analysis

Spatial Supports of analysis	Posterior Mean (95% CrI)						
	Intercept	**NDVI**	**NDWI**	**LSTD**	**LSTN**	**E**	**NDWB**
30 m	-6.17 (-6.6, -5.71)	0.21 (-0.05, 0.48)*	-1.06 (-1.3, -0.8)	-0.71 (-1, -0.41)	-0.78 (-1.06, -0.51)	-0.95 (-1.25, -0.65)	-0.28 (-0.43, -0.12)
90 m	-6.16 (-6.61, -5.71)	0.19 (-0.1, 0.48)*	-1.06 (-1.3, -0.8)	-0.71 (-1.01, -0.39)	-0.82 (-1.12, -0.55)	-1.01 (-1.3, -0.71)	-0.29 (-0.45, -0.12)
250 m	-6.08 (-6.48, -5.69)	0.25 (-0.06, 0.54)*	-1.02 (-1.25, -0.76)	-0.71 (-1.02, -0.38)	-0.82 (-1.11, -0.52)	-1.03 (-1.37, -0.71)	-0.31 (-0.5, -0.13)
500 m	-6.02 (-6.43, -5.61)	0.26 (-0.01, 0.53)*	-0.8 (-1.01, -0.57)	-0.65 (-0.97, -0.33)	-0.86 (-1.14, -0.56)	-1.03 (-1.34, -0.71)	-0.38 (-0.57, -0.19)
1 km	-5.49 (-5.76, -5.21)	0.59 (0.3, 0.88)	-0.76 (-1, -0.54)	-0.11 (-0.85, 0.64)*	0.1 (-0.46, 0.68)*	-0.04 (-0.21, 0.14)*	-0.4 (-0.59, -0.2)

Variance of spatial random effect	φ	**DIC**
2.5 (1.65, 3.58)	8.02×10^{-5} (-0.004,0.004)	127.7
2.43 (1.57, 3.5)	1.2×10^{-4} (-0.004,0.004)	140.5
2.46 (1.6, 3.55)	9.86×10^{-5} (-0.004,0.004)	86.67
2.35 (1.4, 3.54)	2.81×10^{-4} (-0.004,0.004)	147.5
2.7 (1.87, 3.7)	1.65×10^{-5} (-0.004,0.004)	143.7

NDVI: Normalized difference vegetation index; NDWI: Normalized difference water index; LSTD: Land surface temperature day; LSTN: Land surface temperature night; NDWB: Nearest distance to water bodies; CrI: Credible Interval.
*Non-significant variable in the model

Table 3. Credible interval widths (Uncertainty) at five increasing spatial supports of analysis

Spatial Supports of analysis	Credible intervals width (Uncertainty)					
	NDVI	**NDWI**	**LSTD**	**LSTN**	**E**	**NDWB**
30 m	0.52	0.50	0.59	0.56	0.59	0.32
90 m	0.57	0.50	0.62	0.56	0.59	0.33
250 m	0.60	0.48	0.64	0.59	0.66	0.36
500 m	0.54	0.44	0.64	0.58	0.64	0.38
1 km	0.58	0.46	**1.49**	**1.14**	0.34	0.39

NDVI: Normalized difference vegetation index; NDWI: Normalized difference water index; LSTD: Land surface temperature day; LSTN: Land surface temperature night; NDWB: Nearest distance to water bodies.

High uncertainty values are present in bold

Table 4. Correlation and determination coefficients between the spatial supports of analysis (SSAs) and environmental covariates estimates and uncertainties.

	Estimates			**Uncertainties**		
	Correlation coefficient	**Determination coefficient ($R^{2)}$**	**p-value**	**Correlation coefficient**	**Determination coefficient ($R^{2)}$**	**p-value**
NDVI	0.94	0.85	0.02	0.32	-0.2	0.6
NDWI	0.93	0.81	0.02	-0.3	-0.2	0.6
LSTD	0.92	0.8	0.03	0.91	0.76	0.03
LSTN	0.86	0.6	0.06	0.9	0.75	0.04
E	0.86	0.65	0.06	-0.78	0.48	0.12
NDWB	-0.94	0.86	0.02	0.91	0.76	0.03
Variance	0.64	0.21	0.25	-0.54	0.06	0.35

Table 5. Resulting ppp-values for the test statistics: maximum (8.5%), minimum (0.33%) and mean (0.5%) prevalence values at five increasing SSA.

Spatial Supports of analysis	**pp*p*-value (maximum)**	**pp*p*-value (minimum)**	**pp*p*-value (mean)**
30 m	0.66	0.87	0.66
90 m	0.67	0.86	0.66
250 m	0.66	0.88	0.63
500 m	0.66	0.87	0.67
1 km	0.64	0.93	0.59

pp*p*-value: posterior predictive *p*-value