Statistical evaluation of spatial uncertainty in schistosomiasis mapping

Andrea Lucia Araujo Navas

STATISTICAL EVALUATION OF SPATIAL UNCERTAINTY IN SCHISTOSOMIASIS MAPPING

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. T.T.M. Palstra,
on account of the decision of the Doctorate Board,
to be publicly defended
on Thursday September 5, 2019 at 12.45 hrs

by

Andrea Lucia Araujo Navas

born on December 26, 1985

in Quito, Ecuador

This thesis has been approved by
Prof.dr.ir. A. Stein, supervisor
Dr. F.B. Osei, co-supervisor
Prof.dr. R. Soares Magalhães, co-supervisor

ITC dissertation number 360
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISBN 978-90-365-4828-1
DOI 10.3990/1.9789036548281

Data is accessible via DANS

Cover designed by Andrea Araujo.

Printed by ITC Printing Department

ITC
UNIVERSITY OF TWENTE.
FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

Graduation committee:

Chairman / Secretary

Dean of the Faculty	University of Twente

Supervisor(s)

Prof.dr.ir. A. Stein	University of Twente

Co-supervisor(s)

Dr. F.B. Osei	University of Twente
Prof.dr. R. Soares Magalhães	The University of Queensland

Members

Prof.dr.ir. A. Veldkamp	University of Twente
Prof.dr.ir. N.J.J. Verdonschot	University of Twente
Prof.dr. S. Vanwambeke	Catholic University of Louvain
Prof.dr. G.B.M. Heuvelink	Wageningen University

To Leonel and Fernando

Acknowledgements

At this indescribable moment many thoughts come at once. I am certainly not the same person as I was four and a half years ago. In this long way I have learned not only about my own research, but also and most importantly, about life. That life that only asks us one thing, to enjoy it. Despite the difficult moments that blinded my way, the support and guidance of several people made of my Ph.D. life a more enjoyable journey.

I would like to thank Fernando, my beloved life mate. We both came here to accomplish my dream. The decision was not easy. He decided to take the risk knowing that his future was uncertain. Thank you for your courage, patience, dedication, and the peace you brought in the tough moments. Thank you for always be there performing the challenging job of being a father.

Worlds are not enough to express my gratitude to my son Leonel. A two-year-old little one who opens my eyes and brings clearness to my life. By now you do not know what this is all about. You are just there, free of worries, free of the weights we adults carry. I have learned and I still need to learn too much from you. Thank you for showing me that sometimes it could be very hard, but for sure rewarding.

Thanks to my Ecuadorian and Bolivian family. You made our stay here more pleasant despite the distance. To my parents Susana and Miguel, thank you for helping me to become a better person. To my sister and brother, Estefania and Alejandro, thank you for existing dear life buddies. Many thanks to Myri and César, our friends, neighbours and the closest family we had here. Thank you for the support and friendship.

Thanks to those brave and strong mothers I meet in my way. Susana, Jean, Pamela, Lis, Anna, Andrea, Claudia, Myri and many others. We have in our hands perhaps the most important job in the world, in a society that is still too unequal to recognize it. Admiring your everyday dedication, sacrifice and hard work encourages me and gives me hope. Hope to see a better future in the hands of our kids.

Thanks to the colleagues from EOS department, especially to Vera and Caroline the nicest and tallest Dutch girls I know and with whom I shared good and difficult experiences along the Ph.D. way. Thanks to Teresa Brefeld for her valuable job at the department and her help on administrative matters.

My deep thanks to Frank Osei, my daily supervisor. His advice and constructive criticism were always helpful during the last two years of my research. Thank you for the motivation and positivism you transmit. I would also like to thank my second supervisor from the University of Queensland, Ricardo Soares Magalhães. His constructive comments from the epidemiological point of view were always enriching and guided me to clarify the real aim of each objective. I appreciate your interest in my research and the nice interaction we had in our meetings, which was not only professional but also human.

Thank you very much to my promotor Alfred Stein for his constant support along the way. His fast and wise comments were very helpful as they significantly improved the quality of my work. Finally, I would like to thank Lydia Leonardo and Raffy Jay Fornillos from the University of Philippines for providing me with the survey data and clearing the doubts I had about the survey design and data availability.

Summary

The World Health Organization has identified seventeen neglected tropical diseases (NTDs) for targeted control. Schistosomiasis (SCH) is one of the most prevalent NTDs worldwide with high significance in the public health domain. Spatial modelling of SCH using earth observation data informs about the geographic areas where at-risk populations are in need of mass drug anthelminthic treatment. Several sources of uncertainty may decrease the quality and reliability of SCH modelling. This dissertation investigates three methods to reduce the uncertainty derived by the use of earth observation data in SCH modelling studies. It uses spatial statistics for uncertainty quantification and representation, and provides potential consequences when ignoring uncertainties for SCH control.

First, a systematic review and evaluation of uncertainty in SCH and soil-transmitted helminths modelling studies was performed (STH). The definition, quantification, and main sources of uncertainty were investigated as well as implications for SCH and STH control. The literature search was done by grouping three terms referring to uncertainty, geography, and the type of disease (SCH or STH) in the Web of Knowledge and PubMed. Uncertainty was mostly defined as lack of precision. In total, 91% of the studies quantified uncertainty in their predictions, and 23% of the studies mapped uncertainty. Furthermore, uncertainty in the regression coefficients was quantified by 57% of the studies but only 7% incorporated it in the predictions. Uncertainty in the covariates was identified but not quantified in 50% of the studies. Bayesian statistics was used to quantify uncertainty by means of credible intervals. Main sources of uncertainty were related to sample design and spatial aggregation and disaggregation methods.

Second, uncertainty due to positional mismatch between covariate and survey data was addressed using exposure areas as potential locations for SCH transmission. Exposure areas were delineated using a spatial Bayesian network (sBN) with five observable exposure risk factors. Prior and conditional probabilities were obtained from the literature and inserted as weights based upon their relative contribution to exposure. Based on those, joint probabilities of exposure were obtained to be used within sBNs. High probability values of exposure corresponded to areas where snails could be present and where people can easily access water bodies. Extracting covariate values from areas with high probability of exposure, instead of survey locations, is a way to address this mismatch. These results can be used to guide local SCH control teams to exposed communities, and in this way improve the efficiency of mass drug administration campaigns.

Third, uncertainty due to pure specification bias was solved by using a convolutional model. The model used barangay or ecological-level survey data and city-level environmental data. Covariate city–level data were considered as individual-level exposure. Differences between ecological and individual-level estimates and predictions were quantified and compared using Bayesian statistics. The estimated parameter corresponding to the nearest distance to water bodies presented the minimum difference between convolution and ecological models (0.03), whereas the estimated normalized difference water index parameter presented the maximum difference (0.28). Land surface temperature at night and elevation presented high differences with uncertainty vales equal to 0.23 and 0.13, respectively. The convolutional model presented less uncertain parameter estimates showing its good ability to correct for pure specification bias.

Fourth, the effects of the modifiable areal unit problem (MAUP) on environmental drivers of SCH were quantified. Five spatial supports of increasing size were used. All covariates were brought to the same spatial support of analysis (SSA). Differences between individual-level parameter estimates from the models at five increasing SSAs were quantified and compared. Increasing the SSA to 500 m gradually increased the parameter estimates and their associated uncertainties. Abrupt changes in parameter estimates occurred at SSA = 1 km, resulting in loss of significance of almost all covariates on SCH prevalence. These results suggest the use of an adequate spatial data structure to provide more reliable parameter estimates and a realistic relationship between the risk factors and SCH prevalence.

To summarize, the research presented in this dissertation investigates methods to deal with uncertainties derived from the use of earth observation data in SCH modelling. It uses Bayesian statistics for uncertainty quantification and highlights implications of uncertainty interpretation in the public health domain. Such implications aim to enable best practice in survey design and improve the identification of populations at-risk, and quantification of people in need of anthelmintic treatment. This research thus presents a framework for the future development of spatial decision support systems for SCH surveillance and control.

Samenvatting

De Wereldgezondheidsorganisatie heeft zeventien verwaarloosde tropische ziekten (NTD's) geïdentificeerd die in aanmerking komen voor gerichte bestrijding. Schistosomiasis (SCH), één van de meest voorkomende NTD’s wereldwijd, is belangrijk op het gebied van de volksgezondheid. Ruimtelijke modellering van SCH met behulp van aardobservatiegegevens geeft informatie over de geografische gebieden waar populaties die blootstaan aan deze ziekte gebaat zijn bij grootschalige campagnes ter behandeling met anthelminthica. Verschillende bronnen van onzekerheid kunnen de kwaliteit en betrouwbaarheid van SCH-modellering verminderen. Dit proefschrift onderzoekt drie methoden om de onzekerheid te verminderen die ontstaat door het gebruik van aardobservatiegegevens in SCH-modelleringsstudies. Het maakt gebruik van ruimtelijke statistiek voor het kwantificeren en representeren van onzekerheden en laat de mogelijke gevolgen zien bij het negeren van onzekerheden bij een SCH-controle.

Allereerst is een systematische review en evaluatie van studies naar onzekerheid in SCH en naar modellering van helminthen die via de bodem worden overgedragen (STH) uitgevoerd. De definitie, kwantificering en de belangrijkste bronnen van onzekerheid zijn onderzocht, evenals mogelijke implicaties voor controleprogramma's van SCH en STH. Het literatuuronderzoek is gedaan door drie termen te groeperen die verwijzen naar onzekerheid, geografie en het type ziekte (SCH of STH) in het Web of Knowledge en PubMed. Onzekerheid is meestal gedefinieerd als precisie. In totaal heeft 91% van de studies onzekerheid in hun voorspellingen gekwantificeerd en 23% van de studies bracht de onzekerheid in kaart. Daarnaast is de onzekerheid in de regressiecoëfficiënten gekwantificeerd in 57% van de studies, maar slechts 7% nam het op in de voorspellingen. Onzekerheid in de covariabelen is geïdentificeerd maar niet gekwantificeerd in 50% van de studies. Bayesiaanse statistiek is gebruikt om onzekerheid te kwantificeren door middel van geloofwaardigheidsintervallen. De belangrijkste bronnen van onzekerheid zijn gerelateerd aan steekproefontwerp en ruimtelijke aggregatie- en desaggregatiemethoden.

Als tweede is onzekerheid als gevolg van positionele mismatch tussen covariabele en schatting-gegevens aangepakt door blootstellingsgebieden af te bakenen als potentiële locaties voor SCH-transmissie. Dit is gedaan met behulp van een ruimtelijk Bayesiaans netwerk (sBN) met vijf waarneembare factoren die een risico zijn voor blootstelling. *A priori* en voorwaardelijke kansen zijn verkregen uit de literatuur en zijn als gewichten gebruikt naar rato van hun relatieve bijdrage aan blootstelling. Deze zijn gebruikt om simultane blootstellingskansen binnen sBN's te verkrijgen. Hoge blootstellingswaarden komen overeen met gebieden waar een grote kans is op de aanwezigheid van

slakken en waar mensen gemakkelijk toegang hebben tot afgesloten water oppervlaktes. Het verkrijgen van waarden aan de covariabelen uit gebieden met een hoge waarschijnlijkheid van blootstelling, in plaats van op onderzoekslocaties zelf, is een manier om de positionele mismatch aan te pakken. Deze resultaten kunnen worden gebruikt om lokale SCH-controleteams naar blootgestelde gemeenschappen te leiden. Op deze manier wordt de efficiëntie verbeterd van grootschalige campagnes ter behandeling met anthelminthica.

Als derde studie is de onzekerheid als gevolg van pure specificatiebias opgelost met behulp van een convolutie model. Dit model gebruikt onderzoeksgegevens op het niveau van *barangays* of ecologische eenheden en milieugegevens op stadsniveau. Gegevens aan covariabelen op stadsniveau zijn beschouwd als individuele blootstelling. Verschillen tussen ecologische en individuele schattingen en voorspellingen zijn gekwantificeerd en vergeleken met behulp van Bayesiaanse statistiek. De geschatte kortste afstand tot afgesloten water oppervlaktes gaf het kleinste verschil (0,03) tussen convolutie en ecologische modellen, terwijl de geschatte genormaliseerde waterindex parameter het grootste verschil (0,28) gaf. Landoppervlaktemperatuur gedurende de nacht en hoogte lieten grote verschillen zien in onzekerheidswaarden van respectievelijk 0,23 en 0,13. Het convolutiemodel liet minder onzekere parameterschattingen zien, waaruit blijkt dat het goed in staat is te corrigeren voor pure specificatiebias.

De vierde studie kwantificeert de effecten van het Modifiable Area Unit probleem (MAUP) op milieu-factoren van SCH. Vijf ruimtelijke eenheden van een toenemende grootte zijn gebruikt. Alle covariabelen zijn naar dezelfde ruimtelijke analyse eenheid (SSA) gebracht. Verschillen tussen parameter-schattingen op individueel niveau van de modellen bij vijf toenemende SSAs zijn gekwantificeerd en vergeleken. Het verhogen van de SSA tot 500 m verhoogde geleidelijk de parameterschattingen en de bijbehorende onzekerheden. Abrupte veranderingen in parameter schattingen traden op bij een SSA van 1 km, resulterend in verlies van significantie van bijna alle covariabelen op SCH prevalentie. Deze resultaten bieden een adequate ruimtelijke gegevensstructuur aan om betrouwbaardere parameterschattingen en een realistische relatie te verkrijgen tussen de risicofactoren en SCH-prevalentie.

Samenvattend onderzoekt dit proefschrift methoden om met onzekerheden om te gaan die zijn afgeleid van het gebruik van aardobservatiegegevens in SCH-modellen. Het maakt gebruik van Bayesiaanse statistiek voor de kwantificering van onzekerheden en belicht implicaties van de interpretatie van onzekerheden in het domein van de volksgezondheid. Dergelijke implicaties zijn bedoeld om beste praktijken mogelijk te maken bij het verzamelen van gegevens, de

identificatie van risicopopulaties te verbeteren en om het het aantal mensen te schatten die een behandeling met anthelminthica nodig hebben. Dit onderzoek presenteert daarbij een kader voor de toekomstige ontwikkeling van systemen voor ondersteuning van ruimtelijke besluitvorming voor SCH-surveillance en -controle.

Resumen

Diecisiete enfermedades tropicales desatendidas (ETDs) han sido identificadas por la Organización Mundial de la Salud como prioritarias para su control. Schistosomiasis (SCH) es una de las ETDs con mayor prevalencia y de gran importancia para la salúd pública a nivel mundial. El modelamiento de SCH a partir de datos espaciales permite identificar zonas geográficas en donde poblaciones con alto riesgo de infección necesitan un suministro masivo de medicamentos. Varias fuentes de incertidumbre pueden, sin embargo, afectar la calidad y confiabilidad de los modelos espaciales utilizados para este fin. El presente trabajo investiga tres métodos para reducir la incertidumbre en estos modelos mediante el uso de estadísticas espaciales como herramienta de cuantificación y representación de la misma. Esta investigación expone así mismo, las posibles consecuencias en el control de SCH que podrian derivarse de omitir las incertidumbres presentes en su modelamiento.

Primero, se realizó una revisión sistemática y evaluación crítica de la incertidumbre en los estudios de modelamiento de SCH y helmintiasis transmitidas por el suelo (HTS). Se investigaron varios tipos de cuantificación, definición y fuentes de incertidumbre, como también, la connotación que esta tiene en los programas de control de SCH y HTS. La búsqueda bibliográfica se realizó usando una combinación de términos referentes a incertidumbre, geografía o espacio, y el tipo de enfermedad (SCH or HTS). La búsqueda se realizó usando los portales digitales Web of Science y PubMed. En total, 91% de los estudios cuantificaron la incertidumbe en sus predicciones y 23% la mapearon. Adicionalmente, 57% de los estudios cuantificaron la incertidumbre usando coeficientes de regresión, pero sólo un 7% la incorporó en sus predicciones. Finalmente, 50% de los estudios identificaron incertidumbres en las covariables, pero no las cuantificaron. La incertidumbre se define generalmente como precisión, y se cuantifica usando intervalos de confianza por medio de estadística Bayesiana. Las principales fuentes de incertidumbre están relacionadas con el diseño de muestreo, y los métodos de agregación y desagregación espacial.

Segundo, se propone un método para aliviar la incertidumbre causada por la incompatibilidad entre, la posición en la cual las covariables ambientales son extraídas, y la localización donde la encuesta fue abordada. El método propuesto define áreas de exposición como lugares potenciales para la transmisión de SCH. De esta manera las covariables ambientales pueden ser extraídas de las áreas de exposición y relacionadas con la localización de las encuestas. Las áreas de exposición se mapearon usando una red espacial Bayesiana (RsB) en base a cinco factores de riesgo. Las probabilidades condicionales y *a priori* de cada factor de riesgo fueron obtenidas de la literatura e incluidas usando pesos. Los pesos fueron atribuidos en base a la

contribución relativa de cada factor en la exposición a SCH. Estas probabilidades fueron usadas para calcular la probabilidad conjunta de exposición en la RsB. Altos valores de exposición corresponden a áreas de fácil acceso a cuerpos de agua y con potencial presencia de caracoles. Estos resultados pueden guiar a los equipos de control de SCH hacia comunidades con alta exposición a la enfermedad, y de esta forma mejorar la eficiencia en la distribución de medicamentos.

Tercero, se propone un modelo convolucional para reducir la incertidumbre causada por el sesgo en la especificación del modelo. Este sesgo se produce al usar encuestas agregadas a niveles administrativos en la inferencia de parámetros a nivel individual. El modelo convolucional usa encuestas a nivel de barrio, también llamado grupal o ecológico, y datos ambientales a nivel de ciudad como un proxy de exposición individual. Las diferencias entre los parámetros estimados y predicciones, a nivel ecológico e individual, fueron cuantificadas y comparadas usando estadística Bayesiana. Las diferencias mínimas y máximas entre los parámetros estimados en los modelos convolucional y ecológico fueron 0.03 para la distancia mínima a los cuerpos de agua, y 0.28 para la el índice diferencial normalizado de agua (NDWI en inglés), respectivamente. Grandes diferencias en incertidumbre se observaron en la temperatura nocturna terrestre (0.23) y elevación (0.13). El modelo convolucional presentó menor incertidumbre en sus parámetros estimados, demostrando su capacidad para corregir el sesgo en la especificación del modelo.

Cuarto, se cuantificaron los efectos del problema de agregación espacial (MAUP en inglés) en los factores ambientales de SCH. Se usaron cinco tamaños para la resolución espacial (RE). Todas las covariables fueron llevadas a la misma resolución espacial antes de su análisis. Se cuantificaron y compararon las diferencias entre los parámetros estimados a nivel individual para cada uno de los modelos generados en base a los cinco tipos de RE. Un incremento en la RE = 500 m produce un aumento gradual en los parámetros estimados e incertidumbre asociada. Cambios abruptos en los parámetros estimados ocurren a una RE = 1 km, produciendo pérdida de significancia en casi todas las covariables en la prevalencia de SCH. Estos resultados sugieren la importancia de definir una adecuada estructura en los datos espaciales para obtener parámetros estimados más confiables y entender mejor la relación entre los factores de riesgo y la prevalencia de SCH.

En resumen, la presente investigación propone métodos para tratar la incertidumbre derivada del uso de datos espaciales en el modelamiento de SCH. Este trabajo se basa en el uso de estadística Bayesiana como herramienta de cuantificación de la incertidumbre, y resalta las implicaciones de su interpretación en el área de salud pública. Estás implicaciones tienen por fin incentivar mejores prácticas en el diseño de encuestas y, mejorar la identificación de poblaciones en riesgo y la cuantificación de personas con necesidad de un tratamiento antihelmíntico. Esta investigación busca ser una base para el desarrollo futuro de sistemas espaciales de apoyo a la toma de decisiones (SDSS en inglés) para el monitoreo y control de SCH.

Table of Contents

List of figures

List of tables

List of nomenclature

Abbreviation

ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
AUC	Area under the curve
AVHRR	Advanced Very High Resolution Radiometer
BN	Bayesian Network
BUGS	Bayesian inference Using Gibbs Sampling
CI	Confidence Intervals
CrI	Credible intervals
E	Elevation
E	Elevation
EO	Earth observation
GAM	Generalized additive model
GARP	Genetic algorithm for rule-set prediction
GDEM	Global Digital Elevation Map
geoAI	Geospatial Artificial Intelligence
GIS	Geographic Information Systems
GL30	Global Land 30
GLM	Generalized linear model
GLMM	Generalized linear mixed models
GPS	Global Positioning Systems
IRRI	International Rice Research Institute
K	Kappa statistic
LSTD	Land Surface Temperature Day
LSTN	Land Surface Temperature Night
MAE	Mean absolute error
MAUP	Modifiable Areal Unit Problem
MCDA	Multicriteria Decision Analysis
MCMC	Markov Chain Monte Carlo
MDA	Mass Drug Administration
ME	Mean error
MLA	Machine Learning Algorithm
MODIS	Moderate resolution imaging spectro radiometer
NDVI	Normalized Difference Vegetation Index
NDWB	Nearest Distance to Water Bodies
NDWI	Normalized Difference Water Index
NTD	Neglected Tropical Disease
NTP	Node probability tables
OA	Overall Accuracy
OSM	Open Street Map
PA	Producer’s accuracy

PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analysis
RME	Residual mean square
RMSE	Root mean square error
RMSE	Root mean square error
ROC	Receiver operator characteristic
SBD	Spatial Big Data
sBN	spatial Bayesian Network
SCH	Schistosomiasis
SCI	Schistosomiasis control initiative
SD	Standard deviation
SPOT	Satellite Pour l'Observation de la Terre
SSA	Spatial Support of Analysis
STHs	Soil Transmitted Helminths
UA	User’s accuracy
USGS	United States Geological Survey
WASH	Water, Sanitation and Hygiene
WHO	World Health Organization
ZIB	Zero-inflated Binomial

Symbols

$\boldsymbol{\alpha}$	Coordinate regression coefficients
$\boldsymbol{\beta}$	Individual-level covariate coefficients
CC	Community cost
D	Set of discrete random variables
DAG	Directed acyclic graph
DWB	Discretized distance to water bodies
d_{ab}	Distance between barangay centroids a and b
δ^2	Variance of the individual-level regression parameters
E	Discretized elevation values
EX	Exposure
Σ_{ab}	Covariance matrix specified as a function of the distances d_{ab}
F	Findings variables
$\boldsymbol{\gamma}$	Group-level covariate coefficients
H	Information bits
i	Index for a point location in space or individuals
j	Index for a point location in space or cities
k	Index for barangays
κ	Scalar parameter controlling the amount of spatial smoothing
LU	Discretized land use

∇	Degree of entropy reduction
m_k	Number of individual-level exposure locations
n	Total number of environmental covariates
N_k	Number of sampled individuals in barangay k
P	Set of probability distributions
$PA(r)$	Set of parents of the random variable r
PAS	Potential accessible areas
$\hat{p}_k$	Probability of infection in barangay k
$\hat{\hat{p}}_k$	Estimated average probability of infection of the individuals in barangay k
p_{ik}	Probability of infection for the individual i in barangay k
ϕ	Rate of decline of spatial autocorrelation per unit of distance
π	Prior marginal probabilitites
Q	Query variables
R	Set of random variables
$\boldsymbol{s}$	Vector of random variables associated with space locations
s_i	Space location at point i
SI	Discretized snail infection rate
SLP	Discretized slope values
σ^2	Overall spatial autocorrelation variance
t_j	Time location at point j
$\boldsymbol{\theta}$	Set of parameters
μ	Mean of $\boldsymbol{s}$
V	Set of continuous random variables
X	East coordinates
$\boldsymbol{x}$	Set of environmental covariates
$\bar{\boldsymbol{x}}_{\boldsymbol{k}}$	Observed mean exposure within barangay k
$\boldsymbol{x}_{\boldsymbol{ik}}$	Observed exposure at individual level i within barangay k
Y	North coordinates
$\boldsymbol{y}$	Survey infection data
y_k	Number of infected people within barangay k
Z	Interpolated snail infection rate values

Chapter 1. Introduction

1.1 Schistosomiasis (SCH)

This thesis considers Schistosomiasis (SCH), a very common neglected tropical disease (NTD). SCH is a water-borne infection caused by schistosomes. More than 252 million people worldwide are affected by SCH (Hotez et al., 2014), especially regions in sub-Saharan and North Africa, Asia, and the central and Andean regions of Latin America (Hotez et al., 2014). Transmission occurs by skin penetration of the infective stage of schistosomes, a type of helminths, which are parasitic worms transmitted by the ingestion of contaminated food/water, or by skin contact with contaminated soil/water. Helminths have two major phyla: nematodes and platyhelminths. Nematodes encompass soil-transmitted and filial worms. Platyhelminthes include schistosomes and tapeworms (Hotez et al., 2006). Soil-transmitted helminths and schistosomes are parasites in the shape of little worms that spread some of the most prevalent NTDs, affecting human populations living in areas where there is limited access to potable water, and sanitation and hygiene are poor. In human populations, the parasites live and feed on their living hosts, taking all their nutrients and causing malnutrition, stunted growth, and anaemia (Coutinho et al., 2005; Leenstra et al., 2006). Human SCH infections directly influence the nutrition status, individual productivity, and the physical and mental development (Taylor-Robinson et al., 2015). Three schistosome species cause the infection: *Schistosoma mansoni*, *Schistosoma haematobium and Schistosoma japonicum.* Among these, *Schistosoma japonicum* is the hardest one to control due to its zoonotic life cycle (Jia et al., 2007) which includes the infection of an amphibious snail belonging to several subspecies of *Oncomelania hupensis* as the intermediate host, and humans and other mammalians as definite host (Tarafder et al., 2006; Yang et al., 2008).

Control and elimination targets have been established by the World Health Organization (WHO) and the World Bank (Stolk et al., 2016). In order to accomplish these, the 2012 London declaration for Neglected Tropical Diseases and the 2013 World Health Assembly resolution, resolved the implementation of mass drug administration (MDA) campaigns with praziquantel (Keenan et al., 2013; Mccarty, Turkeltaub and Hotez, 2014) to populations at-risk of SCH. MDA campaigns consist on oral therapies of a single dose which are harmless and cheap, but require to be distributed periodically to the at-risk populations. MDA campaigns are commonly school or community-based. Various international initiatives such as Deworm the World (Action, 2018) and the Schistosomiasis Control Initiative (SCI) (Initiative, 2019) have supported mass school-based treatment for SCH. For instance, in 2017, 'Deworm the World' treated over 260 million of children in India, Kenya, Ethiopia, Vietnam and Nigeria, dramatically reducing soil-transmitted and schistosomiasis helminth infections.

Morbidity indicators such as prevalence and intensity of infection can be measured via surveying at-risk populations in order to guide MDA campaigns implementation (Soares Magalhães et al., 2011b). Prevalence (i.e. proportion of infected individuals) is the ration between the number of infected individuals and the total number of sampled individuals in an area. Intensity of infection (i.e. worm load) is derived based upon parasite egg counts or blood smears. Decisions on where to implement SCH control is often based on whether the prevalence exceeds a determined threshold. For SCH, mass treatment is recommended (i) once in a year for all children and adults at risk in communities with prevalence values > 50%. (ii) once every two years for all children and adults at risk in communities with prevalence values ranging from 10 to 50%, and (iii) twice during the primary school age of children in communities with prevalence values < 10%. (Duarte et al., 2014).

1.2 Spatial modelling of SCH

Studying the distribution of at-risk populations enables a better understanding of SCH epidemiology, as well as the impact of the environment on the disease.

Rapid and inexpensive methods that provide reliable estimates of the geographical distribution of SCH infection are needed for an effective control of the disease (Brooker, Clements and Bundy, 2006). Nationwide detailed surveillance data are required for planning optimal control measures, but only few endemic countries provide suitable data for these purposes (Brooker et al., 2000). Ways of maximising the usefulness of available data and reduce the cost of prevalence surveys (Soares Magalhães et al., 2011b), include spatial statistical modelling based on relationships between environmental predictors and the observed risk of infection. Methods for modelling SCH are data-driven, knowledge-driven, and a combination of both. Commonly used data-driven methods are: logistic regression (Brooker et al., 2002) as part of the generalized linear models (GLMs), generalized linear mixed models (GLMM) (Soares Magalhães et al., 2011a), and generalized additive models (GAMs) (Pfukenyi et al., 2006). Commonly used knowledge-driven methods are Maxent (Stensgaard et al., 2013), the genetic algorithm for rule-set prediction (GARP) (Stensgaard et al., 2006), and multi-criteria decision analysis (MCDA). The most commonly used data and knowledge-driven method is Bayesian statistics (Soares Magalhães et al., 2014).

SCH infection risk is determined by various environmental and socio-economic factors (Brooker et al., 2002). Observations on those factors can be derived from earth observation data sources. These sources include both remote sensing and *in situ* observations. During the last decade, SCH infection mapping has used model-based geostatistics for estimation and prediction of spatially continuous prevalence of infection (Brooker et al., 2002). Using model-based geostatistics along with spatial information techniques such as global positioning systems, remote sensing, and geographical information systems (Manyangadze et al., 2015; Walz et al., 2015a), has facilitated the integration of environmental and disease data derived from earth observation (Manyangadze et al., 2015). Thus, they all contribute to more rational resources allocation strategies for optimized cost-effective SCH control efforts.

Unlike other spatial statistical modelling, model-based geostatistics can take into account spatial autocorrelation. The implication of spatial autocorrelation is that nearby locations are likely to have similar risk values than those farther apart (Tobler, 1970). Model-based geostatistics overcomes issues regarding uncertainty quantification for non-Gaussian outcome predictions, such as proportions (i.e. infection prevalence) or counts (i.e. infection intensity) (Diggle, Tawn and Moyeed, 2002). Model-based geostatistics can generally be applied using Bayesian statistics, which allows a direct and intuitive interpretation of uncertainty in the parameter estimates and predictions (Diggle, Tawn and Moyeed, 2002).

1.3 Uncertainty and its sources

Uncertainty, also referred as vagueness (Shi, 2009), ambiguity (Foody, 2003), inaccuracy or imprecision (Dungan, 2002), has been widely discussed in Geographic Information Science (Longley et al., 2015; Tavana et al., 2016; Worboys and Duckham, 2004) as it is present in the acquisition, abstraction and geo-processing of spatial data. In the SCH spatial modelling framework, uncertainties are not only present in spatial data (i.e. disease and covariate data), but also in the selected model and during modelling itself.

Uncertainty sources derived from spatial data – covariate and survey data - arise due to random errors, such as equipment limitations or unfavourable environment conditions during data capture; and systematic errors, such as sampling design (Rothman, 2012) and measurement errors (Zhang et al., 2016). Sampling design errors relate to insufficiently large sample sizes, the type of survey (Chammartin et al., 2014), the selected morbidity indicator (Soares Magalhães et al., 2014), and limited access to geographic areas (Zhang et al., 2009), among others. Examples of measurement errors are positional measurement error (Zhang et al., 2016), geo-coding errors (Atkinson and Graham, 2006), non-calibrated equipment or data (i.e. coordinate inaccuracies) (Curran et al., 2000).

Uncertainty sources from the prediction model have been classified into uncertainties in model structure, model parameters and solution method (Raso et al., 2005). Uncertainties in model structure arise due to the imperfect knowledge about the complex spatial and temporal variability that describe the phenomena of interest (Clements, Moyeed and Brooker, 2006; Schur et al., 2011; Shi, 2009). Uncertainties in model parameters refer to uncertain estimates or measurements derived from empirical quantities applied to specific-case models (Duarte et al., 2014). Uncertainties in the solution method refer to model selection as different mathematical models ranging from simplistic, linear models to rather complex numerical ones could lead to different solutions (Brown and Heuvelink, 2006; Clements et al., 2006).

Finally, uncertainty during modelling could emerge from inconsistencies in the selection of covariates (Clements et al., 2008), their misalignment with survey data and other covariates, i.e. different spatial and temporal scales of analysis, (Soares Magalhães et al., 2014; Sturrock et al., 2013), and inconsistences in the pre-processing of spatial information such as spatial aggregation and disaggregation (Zhang et al., 2009). Also, the lack of radiometric, atmospheric or geometric correction of satellite images is an important source of uncertainty (Curran et al., 2000).

The complex epidemiology of the diseases and the stochastic nature of their environmental risk factors (Briggs, Sabel and Lee, 2009) have made uncertainty quantification more challenging. The complex epidemiology of a disease is characterized by its mobility pattern, global distribution, evolution, and factors influencing its biogeographical and ecological pattern (Hay et al., 2002). These issues are manifest by space and time non-stationarity, and anisotropy. For instance, non-stationarity can be observed in either or both, the space s and time t domains. The disease $\boldsymbol{y}$ is nonstationary if either $\mathrm{E}\{\boldsymbol{y}(\boldsymbol{s},\boldsymbol{t})\}$ varies or the covariance $\mathrm{C}\{y(s_i,t_i),y(s_j,t_j)\}$ depends on locations s_i and s_j and/or time points t_i and t_j, rather than only the space-time lag. The spatio-temporal structure of a disease relies on covariance and variogram modelling which often assume, for simplicity, stationarity. As uncertainty in all these respects is highly complex, I focus in this study on a spatial statistical analysis. Finally, uncertainty is a quantitative data quality indicator. Thus, it should be communicated as a decision support tool for SCH control.

1.4 The Bayesian approach

Uncertainties in SCH mapping often arise from having insufficient information about the underlying causes of the disease infection risk (section 1.3). Uncertainty quantification becomes difficult when these causes cannot be determined directly. Thus, the use of an inductive reasoning, where consequences help to infer possible causes, is needed.

Bayesian statistics allows to quantify uncertainty by using evidence from prior knowledge (Stone, 2013). A Bayesian analysis has two components: prior beliefs and likelihood. Prior beliefs, e.g. prior knowledge, is represented as the probability distribution assigned to the presence of an event. In the context of modelling, prior distributions reflect prior opinions about possible values of the unknown parameters $\boldsymbol{\theta}$ previous to data $\boldsymbol{y}$ collection and inspection (Lawson, 2013). The data likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$, or probability of observing the data $\boldsymbol{y}$ regarded as a function of the parameters $\boldsymbol{\theta}$, in combination with the prior distribution $p(\boldsymbol{\theta})$, provides posterior beliefs about the parameters. This is again expressed as probability distributions $p(\boldsymbol{\theta}|\boldsymbol{y})$. Bayes' theorem is then formulated as: $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})$.

Commonly two types of prior distributions are distinguished. Informative priors are obtained from expert knowledge and historical or experimental data, whereas non-informative priors express the lack of relevant prior information. Non-informative priors have little impact on the posterior distributions, compared to the data likelihood (Lawson, 2013).

Spatial autocorrelation parameter, variance parameter, and covariate regression coefficients can only be estimated numerically. Computational methods like Monte Carlo integration using Markov Chain Monte Carlo (MCMC) are adequate to simulate distributions. MCMC generates random samples from the joint distribution (i.e. prior distributions and likelihood) using the observed data, prior parameters, and the covariates from sampled point locations. MCMC uses a set of chains to calculate approximately the same posterior distribution for each chain, reaching an equilibrium distribution. This approximation to an equilibrium distribution depends upon the number of samples taken from the initial joint distribution and not from the previous state of the chain. Thus, by increasing the number of samples the chain should forget its initial state. The posterior distributions are then estimated after several runs, when the equilibrium is reached.

Bayesian model based geostatistics is a spatial statistical method with useful implications in SCH mapping. In particular, the obtaining of posterior predictive distributions at unsampled locations from both the parameters and the SCH epidemiological data (e.g. prevalence of infection) (Soares Magalhães et al., 2011b). This method is preferred in SCH modelling because it can handle both, several spatial structure of the data, and uncertainty representation on the predictions in a robust way. The uncertainty representation given by model based geostatistics is also practical for control programs as the posterior distributions can be mapped, showing the plausible predicted values for each location.

1.5 Bayesian networks

A Bayesian network (BN) is a probabilistic graphical model that captures the various conditional dependencies of a set of discrete or continuous random variables (Bottcher and Dethlefsen, 2003) into a joint probability distribution using a directed acyclic graph (Fenton and Neil, 2012; Nielsen and Jensen, 2009). A BN consists of the pair (DAG, P), where DAG is the directed acyclic graph and P is the set of probability distributions for a set of random variables R in the network. Each variable r with parents $PA(r)$ has a conditional probability $p(r|PA(r))$. For a set of discrete variables I, the joint probability distribution factorizes into equation 1.1 as the product of all conditional probabilities specified in a BN.

$$P(R) = \prod_{\mathrm{d}=1}^{\mathrm{D}} p\ (r_d|PA(r_d)) \tag{1.1}$$

The spatial Bayesian network (sBN) aims to model the conditional dependence between the a spatial random variable and its parents. In the directed acyclic graph, conditional dependence is represented as edges, whereas spatial random variables are represented as nodes. The direction given by the edges between variable nodes encodes a direct causal dependence of a node on its parent node.

1.6 Pure specification bias

Pure specification bias is a type of ecological fallacy that arises when aggregated survey or covariate data are used for individual-level inferences (Wakefield and Shaddick, 2006; Zhang et al., 2016). Ecological fallacy is an important source of uncertainty in spatial epidemiological studies as any direct link between exposure and health outcomes is imperfectly measured (King, 2013). This means that the real relationship between the affected population and their exposure is incorrectly represented. Pure specification bias arises due to the loss of information when a non-linear model changes its form under aggregation (Gelfand et al., 2010; Wakefield and Lyons, 2010). It is called 'pure' because it specifically addressed model specification bias (Gelfand et al., 2010). Pure specification bias can be reduced as the 'within area' exposure is more homogenous (Wakefield and Lyons, 2010). This could be obtained by dividing the areas of analysis into finer units at which exposure measurements are available (Wakefield and Lyons, 2010; Wakefield and Shaddick, 2006).

Various efforts have been made to address pure specification bias. For instance, Prentice and Sheppard (Prentice and Sheppard, 1995) suggested the creation of models based upon exposure information available for a subset of individuals. Wakefield and Shaddick (Wakefield and Shaddick, 2006) proposed an appropriate likelihood function for aggregated health outcome data and exposure information available at monitoring sites. Wang et al. (Wang et al., 2017) used aggregated disease output counts and point-level exposures to propose a conceptual probability of the incidence surface over the study region as a function of an exposure surface. The probability surface was used to simulate individual disease outcomes to obtain individual-level parameter estimates.

1.7 Modifiable areal unit problem

The modifiable areal unit problem (MAUP) arises when for a specific study area, the use of aggregated survey or covariate data at arbitrary spatial support sizes and shapes, might affect the patterns identified in the data (Schur et al., 2011; Schur et al., 2013) and the relationship between the disease and the environmental risk factors (Dungan et al., 2002). The MAUP is also part of the ecological fallacy issue and represents a source of uncertainty in spatial epidemiological studies. Several studies investigated the consequences of ignoring the MAUP in spatial epidemiology. For instance, Hellsten et al. (Hellsten, 2006) studied the influence of using aggregated covariate data to model ammonia emissions at farm level. They showed that the size and shape of spatial aggregation areas strongly changes the location of the emissions estimated by the model, e.g. too small areas resulting in false emission "hot spots". Schur et al (Schur et al., 2011) and Schur et al (Schur et al., 2013) aggregated SCH prevalence to evaluate endemicity for various administrative units (Schur, Vounatsou and Utzinger, 2012). Such aggregation presented different endemicity patterns and intervention methods. As a consequence, high endemic areas may not be correctly targeted.

1.8 Research gap

Spatial modelling of SCH is now commonplace as it informs decisions regarding the locations to be targeted with mass drug administration (Stensgaard et al., 2005). Mapped outcomes, however, need to be interpreted with care as they could be weakened by several sources of uncertain information (Manyangadze et al., 2015). Although many studies highlight the relevance of uncertainty quantification in the estimates and predictions (Schur et al., 2011), little attention has been put on tackling main sources of uncertainty and investigate their possible consequences in SCH modelling.

The main gaps identified in the study of uncertainty in SCH spatial modelling are the following:

Uncertainty interpretation and communication

The large variety of uncertainty sources can be propagated from the disease and environmental input data through the model definition and structure, reaching the model outputs (Brown and Heuvelink, 2006). Despite the wide use of spatial models for SCH, many studies ignored a proper analysis of the propagated uncertainty sources that influence their results. Moreover, when presenting the level of uncertainty in their findings, its interpretation and communication to the public health scientist, decision makers, and affected community remains weak, despite the large implications it might have for SCH control (Burns et al., 2014; Manyangadze et al., 2015). Such implications are not only related to the target of mass drug administration campaigns, but also to inform about the need of more resources (e.g. improvement of sanitary conditions), or investigative efforts in high risk areas (Clements, Moyeed and Brooker, 2006; Raso et al., 2009).

Spatial locations of disease survey data

Two sources of uncertainty of particular interest arise from the use of disease survey data at non-existing or unreliable locations. One is positional uncertainty, specifically the extraction of disease survey and covariate information from locations where SCH exposure did not occur (Zhang et al., 2016). The affected population is usually concentrated at hospitals, health centres and schools. These places are relatively easy to access and survey, but do not represent the exact locations where exposure could have occurred. This produces a positional mismatch of the surveyed disease values and the actual disease exposure locations (Zhang et al., 2016). The second is a lack of geo-located individual-level survey data. For a detailed inference level, individual survey data are preferred above administrative survey data. Although many studies have individual-level data available, almost none of them have linked them to locations in space. This occurs because disease survey data are usually not intended to be applied for a spatial analysis, thus limiting their usefulness in spatial modelling (Soares Magalhães et al., 2011b).

Spatially misaligned data

Environmental covariate data can be obtained from a wide numer of EO data sources: *in situ* measurements and low, moderate and high resolution images. The low cost and rapid availability of data from low resolution (1-8 km) sources, such as the WorldClim global climate data, the moderate resolution imaging spectroradiometer (MODIS), and the advanced very high resolution radiometer (AVHRR) (Araujo Navas et al., 2016); have made them the most used sources for covariate data extraction. Other moderate and high resolution sensors have been used as well, such as SPOT, RapidEye or Landsat (Walz et al., 2015a). Most SCH modelling studies have extracted covariate information directly from these sources and inserted them into the model, using data at various spatial resolutions with misaligned grids. This causes an incompatibility of spatial units, and leads to wrong conclusions about the disease pattern and its relationship with the environment (Dungan et al., 2002; Gotway and Young, 2002).

1.9 Study areas

The Philippines was the selected study area, due to (i) the high endemicity of *Schistosoma japonicum* in the country, with 28 endemic provinces out of 81 (Leonardo et al., 2015), and (ii) the availability of good quality nationwide data. People at-risk surpass five million, and around 1.8 million are infected. For the past 30 years, MDA campaigns have been undertaken to control the infection in the country. However, there are still severe cases of retardation, malnutrition, poor cognitive function and anaemia in endemic areas, especially those with scarce treatment coverage. Control efforts on SCH are complicated due to the zoonotic life cycle of the parasite, as large mammals such as water buffaloes and cattle are known to contribute to the disease transmission SCH keeps on being a major public health problem in the country given the advanced SCH cases reported by the National Department of Health for Mindanao, Leyte, and Oriental Mindoro.

Two endemic areas from the Philippines were selected (Figure 1.1). The first is a local area in the Alangalang municipality in the Leyte province. Its selection was based on a survey collected in 2015 by scientists from the College of Public Health and College of Science from the University of Philippines. The second is the Mindanao region. Selection of this area was due to the good spatial coverage and high response rate (~70%) to the 2008 Nationwide Schistosomiasis Survey in the Philippines.

1.10 Research objectives and questions

The main objective of the proposed research was to analyse uncertainties in spatial modelling of SCH helminth infections. This was done by means of the following sub-objectives:

Objective 1: To identify the gaps in knowledge of the different components of uncertainty associated with mapping and modelling helminth infections.

Research questions:

- How is uncertainty defined and quantified in helminth modelling studies?
- What are the main sources of uncertainty in helminth infections mapping and modelling?
- How is uncertainty informative for decision makers, public health scientist and the affected community?

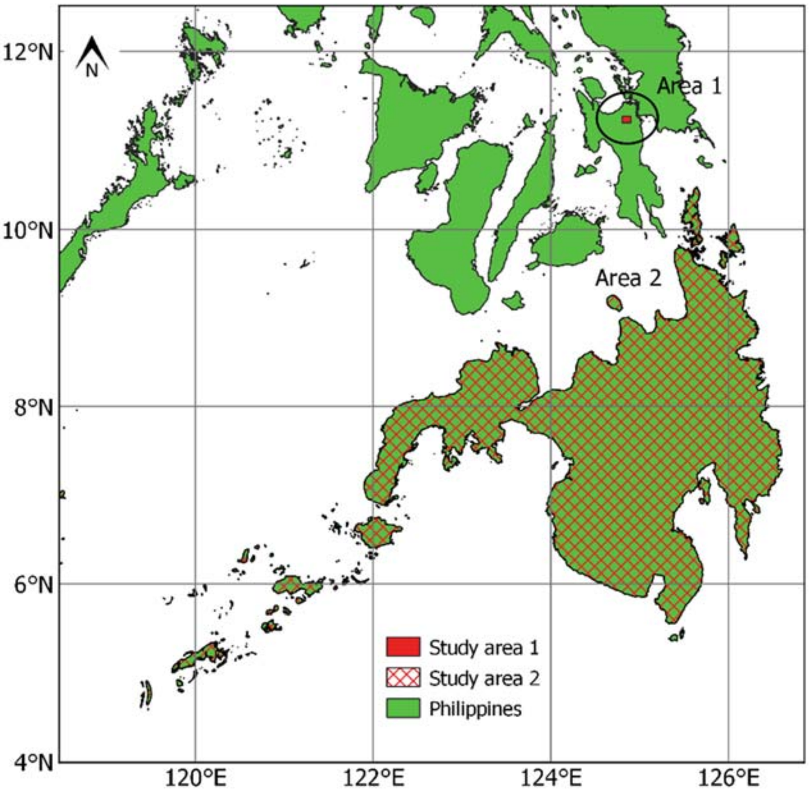


Figure 1.1: *Location of the study areas in The Philippines. Area 1 was used as the study are in Chapter 3. Area 2 was used as the study area in Chapters 4 and 5.*

Objective 2: To map potential areas of exposure to *Schistosoma japonicum* infection using a spatial Bayesian network (sBN) model.

Research question:

Could the positional mismatch between survey and covariate data be corrected?

Objective 3: To quantify the effect of pure specification bias on the parameter estimates of various environmental drivers of *Schistosoma japonicum* infection.

Research questions:

- Can pure specification bias be corrected by using group-level disease data and individual-level covariate data?

- How much does pure specification bias increase or decrease parameter estimates and their uncertainties?

Objective 4: To quantify the modifiable areal unit problem (MAUP) effect on various environmental drivers of *Schistosoma japonicum* infection.

Research questions:

- Does an increase in the spatial support of analysis increase or decrease parameter estimates and their uncertainties?

- What is the suggested spatial support of analysis to model *Schistosoma japonicum* infection?

1.11 Thesis outline

The present thesis contains six chapters. Chapter 1 introduces the research. Chapters 2 to 5 are the key elements of this thesis. They have all resulted in scientific manuscripts that have been published by, or are being reviewed in, ISI journals. Chapter 6 contains the synthesis of the research. The six chapters are organized as follows:

Chapter 1.- gives a general introduction to schistosomiasis and the importance of modelling it for the public health community. It also describes uncertainty and its various sources in SCH modelling; and highlights the use of Bayesian model-based geostatistics for parameter estimation and uncertainty representation. The chapter motivates the work and sharpens the research objectives.

Chapter 2.- presents a systematic review and critical evaluation of the current published literature on the spatial epidemiology of helminth infections. Three analyses are presented: (i) definition and quantification of uncertainty, (ii) identification of the various uncertainty sources, and (iii) implications and communication of uncertainty for soil transmitted helminths control programme managers and scientists.

Chapter 3.- constructs a spatial Bayesian network to delineate exposure areas to *Schistosoma japonicum* infection that could be used to correct for the positional mismatch. It starts by describing the positional mismatch issue in modelling *Schistosoma japonicum* infection. It then presents the defined spatial Bayesian network, using the accessibility cost of people to main sources of infection and the distribution of snail infection. Finally, it describes validation using human positive cases.

Chapter 4.- quantifies the effect of pure *specification* bias on the parameter estimates of six environmental drivers for *Schistosoma japonicum* infection. It proposes a spatial convolution model that uses group-level disease data and individual-level covariate data to correct for pure specification bias. Group and individual level parameter estimates are presented and compared in order to reach conclusions about the implications of pure specification bias on SCH modelling.

Chapter 5.- quantifies the MAUP effects on six environmental drivers for *Schistosoma japonicum* infection. Alignment among all covariates is presented by using aggregation and disaggregation methods. Five increasing spatial support sizes of analysis are used to quantify the differences in parameter estimates, and the MAUP implications on SCH modelling.

Chapter 6.- synthesizes and discusses the results for each chapter and the research as a whole. It also provides reflections on the work and future outlook.

Chapter 2. Mapping Soil Transmitted Helminths and Schistosomiasis under Uncertainty: A Systematic Review and Critical Appraisal of Evidence[1]

[1] This chapter is published as:

Araujo Navas, A. L.; Hamm, N. A. S.; Soares Magalhães, R. J.; Stein, A. Mapping Soil Transmitted Helminths and Schistosomiasis under Uncertainty: A Systematic Review and Critical Appraisal of Evidence. *PLoS Neglected Tropical Diseases* 2016, *10*, e0005208, doi: 10.1371/journal.pntd.0005208.

Abstract

Spatial modelling of STH and schistosomiasis epidemiology is now commonplace. Spatial epidemiological studies help inform decisions regarding the number of people at risk as well as the geographic areas that need to be targeted with mass drug administration; however, limited attention has been given to propagated uncertainties, their interpretation, and consequences for the mapped values. Using currently published literature on the spatial epidemiology of helminth infections we identified: (1) the main uncertainty sources, their definition and quantification and (2) how uncertainty is informative for STH programme managers and scientists working in this domain.

We performed a systematic literature search using the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) protocol. We searched Web of Knowledge and PubMed using a combination of uncertainty, geographic and disease terms. A total of 73 papers fulfilled the inclusion criteria for the systematic review. Only 9% of the studies did not address any element of uncertainty, while 91% of studies quantified uncertainty in the predicted morbidity indicators and 23% of studies mapped it. In addition, 57% of the studies quantified uncertainty in the regression coefficients but only 7% incorporated it in the regression response variable (morbidity indicator). Fifty percent of the studies discussed uncertainty in the covariates but did not quantify it. Uncertainty was mostly defined as precision, and quantified using credible intervals by means of Bayesian approaches.

None of the studies considered adequately all sources of uncertainties. We highlighted the need for uncertainty in the morbidity indicator and predictor variable to be incorporated into the modelling framework. Study design and spatial support require further attention and uncertainty associated with Earth observation data should be quantified. Finally, more attention should be given to mapping and interpreting uncertainty, since they are relevant to inform decisions regarding the number of people at risk as well as the geographic areas that need to be targeted with mass drug administration.

2.1 Introduction

Helminth infections from as soil-transmitted helminths (STHs) and schistosomes are among the most prevalent neglected tropical diseases (NTDs) affecting human populations living in countries where clean water, sanitation, and hygiene (WASH) are limited. STHs and schistosomes, affect more than 1.7 billion and 252 million (Hotez et al., 2014; Pullan et al., 2014) people worldwide respectively. The majority of these infections are concentrated in sub-Saharan (Walz et al., 2015a) and North Africa, Asia, and central and Andean regions of Latin America (Hotez et al., 2014). STH and schistosome infections influence directly the nutrition status, educational development, individual productivity, physical and mental development in human populations (Taylor-Robinson et al., 2015). The World Health Organization (WHO), the World Bank and other agencies defined control and elimination targets in the poorest populations (Stolk et al., 2016). Although the global burden of NTDs declined by 27% from 1990 to 2010 in upper-middle income countries (Stolk et al., 2016), low and lower middle income countries still need attention. Besides, according to the Global Burden of Disease Study 2010 (Hotez et al., 2014), STHs due to intestinal nematode infections, and schistosomiasis, caused the largest number of cases reported in 2010. In order to improve population health and accomplish WHO targets, the 2012 London declaration for Neglected Tropical Diseases and the 2013 World Health Assembly resolution highlighted the importance of mass drug administration (MDA) with benzimidazoles (Keenan et al., 2013; Mccarty, Turkeltaub and Hotez, 2014) to communities at risk.

To identify communities at risk, indirect indicators of morbidity such as prevalence of infection and intensity of infection can be measured via surveying at-risk populations (Soares Magalhães et al., 2011b). Communities at risk can then be categorized into disease prevalence classes (e.g. low, moderate, high) based on WHO guidelines (Duarte et al., 2014). In the absence of empirical data on infection at unsampled communities, one way to identify communities at risk is to study the role of the environment (physical and biological) to characterize potential habitats of parasites and intermediate hosts, as well as to understand the ecology and epidemiology of infections. Statistical modelling of the spatial distribution of helminth infections provides empirical relationships between infections and risk factors, which can then be used to predict the level of infection prevalence at unsampled locations (Cadavid Restrepo et al., 2016; Soares Magalhães et al., 2011b; Weiss et al., 2015). In the statistical model, prevalence or another morbidity indicator, is treated as the response variable.

Although statistical modelling of helminth infections is useful to effectively and efficiently manage surveillance, control and prevention of the infection (Stensgaard et al., 2005), the mapped outputs should be interpreted with care because these can be weakened by several sources of uncertain information

(Manyangadze et al., 2015). Sources of uncertainty that need to be accounted for in the modelling process include differences in variable selection criteria, statistical methods used, selected spatial and temporal scales of analysis (Duarte et al., 2014), sampling design, sensitivity and specificity of diagnostic techniques as well as the quality of the spatial data used.

Uncertainty has been the subject of extensive discussion in Geographic Information Science (GIScience) (Longley et al., 2015; Tavana et al., 2016) and related subjects (Rougier, Sparks and Hall, 2014). Uncertainty may relate to (1) a state of mind and our perception of the world or (2) statements about the world or observations on natural phenomena (Longley et al., 2015; Worboys and Duckham, 2004) and is relevant in terms of specifications and representations, measurement and the transformations, processing and modelling performed on raw data to turn them into usable information (Longley et al., 2015; Worboys and Duckham, 2004). In order to address uncertainty, a more formal approach is required (Duckham et al., 2001; Worboys and Duckham, 2004). Here we conceptualize uncertainty as *imperfection*, which is further categorized as *inaccuracy* or *imprecision*.

Imprecision may arise because the phenomenon is vague (i.e., the phenomenon is not clearly defined), ambiguous (i.e., different definitions can be applied to the phenomenon) (Foody, 2003) or due to the granularity of the observation (Worboys and Duckham, 2004). In the spatial setting granularity relates to the resolution or spatial support (area or volume) of the observation and affects our ability to discern objects (Dungan et al., 2002; Worboys and Duckham, 2004). Imprecision may also arise due to natural variability, measurement error and model variability and may be described statistically, for example by the variance or standard deviation (Atkinson and Graham, 2006; Dungan, 2002). In this context, model variability may arise due to uncertain data, stochastic processes within the model or variability between competing models. The reader may be familiar with the narrow statistical definition of precision as the inverse of the variance, whereas the imprecision that is applied here encompasses a wider set of concepts (Duckham et al., 2001; Worboys and Duckham, 2004). Put another way, in this conceptualization, variance is not the only measure of precision.

Accuracy is a measure of closeness between the observed phenomenon and reference observations, considered representative of the reality (Atkinson and Graham, 2006; Worboys and Duckham, 2004). Accuracy assessment is often referred to as validation (Foody and Atkinson, 2002). Common measures of accuracy include the root mean square error (RMSE) for continuous data (Atkinson and Graham, 2006), the overall accuracy (OA) for categorical data (Congalton, 2010) and the area under the receiver operator characteristic curve (AUC) for binary data (Atkinson and Graham, 2006). Bias relates to accuracy and refers to systematic differences between the observations and reference data.

Accounting for uncertainty in disease mapping is important for the assessment of the applicability and validity of the predicted morbidity indicators (Manyangadze et al., 2015). Furthermore, it will allow a complete risk assessment and the identification of potential sources of bias (Burns et al., 2014). Ignoring uncertainty can lead to incorrect predictions, thus wrong estimates of disease burden, which can result in misleading public health advocacy and decisions regarding disease control. Consideration of information about uncertainty is critical for control programs, health care workers, populations at risk, and other involved users who attempt to reduce prevalence and incidence of helminth infections across the affected areas (Burns et al., 2014). For example, control programs need accurate information to decide about drug distribution strategies and the frequency of treatment of the target populations. Decision makers can use information about uncertainty to target more resources (e.g., data acquisition) or to focus investigative efforts on low or highly uncertain risk areas (Clements, Moyeed and Brooker, 2006; Raso et al., 2009).

This paper is a systematic review that aims at the identification of the gaps in knowledge of the different components of uncertainty associated with mapping and modelling helminth infections. It also aims at providing a basis for a complete uncertainty communication, by evaluating the impact of uncertainty on the predicted morbidity indicators. This paper starts by investigating how uncertainty is informative for decision makers, public health scientists and the affected community. It then identifies main sources of uncertainty in helminth infection mapping studies, and how uncertainties have been defined and quantified. Regarding the sources of uncertainty, their definition and quantification, the focus will be put on sources relating to Earth Observation. The significance of this paper is that it contributes to inform control programs and health workers about the importance of uncertainty in mapping and modelling helminth infections, by putting special attention on relevant sources of uncertainty, and analyzing their real influence on the predicted morbidity indicator values used to guide mass drug administration strategies and their cost effectiveness.

2.2 Methods for the search strategy and data collection process

2.2.1 Search Strategy

An online search was performed using two search engines, the Web of Knowledge (Core collection and MEDLINE) and PubMed. Only articles published in English were considered. The date range was 1 January 1980 to 24 October 2016. The search strategy aimed at the identification of primary research studies that have looked into establishing the geographical limits of STH and schistosomiasis present only in humans; therefore the search strategy combined variations of three terms: spatial, helminth infection, and uncertainty terms. The full list of terms used in the systematic review is shown in Table 2.1 Six searches were performed by combining the three terms in each search engine, using the keywords described in Table 2.2.

After removing duplicates, the abstracts of 139 papers were read. Papers written in languages other than English (11 papers) were automatically excluded. Review papers (14 papers) were also excluded. Further criteria were then applied to select the final papers to read, but also to make the reading process more efficient. The inclusion criteria considered were (i) the presence of the three spatial, uncertainty and helminth infection search terms in the abstracts and (ii) also articles related to only STH and schistosomiasis helminth infections. The papers were classified into schistosomiasis and soil transmitted helminth studies. The selection of the papers, data acquisition and analysis was undertaken by the first author. The PRISMA flow diagram is given in Figure 2.1.

2.2.2 Data collection process

Data collection from each paper focused on addressing three main research questions. (1) How is uncertainty informative for decision making in the public health context? (2) What are the different uncertainty sources reported in the reviewed studies? (3) How were uncertainty and its sources defined and quantified in the studies? Papers addressing these questions were enumerated.

Figure 2.2 illustrates the relevant three uncertainty stages that drive the final mapping and modelling of STH and schistosomiasis infections. The first stage (pink box) describes the origin of uncertainty coming from data sources, including uncertainties in the response variable and covariates. The second stage (orange box) shows how uncertainty from the pink box propagates through the predictive model (green box). The green box incorporates uncertainties derived from the selection of the predictive model, considering that there could be different ways to model the same helminth infection. It also includes uncertainties in model structure, which refers to all possible limitations and assumptions in the selected model, such as: the lack of understanding about the interaction between the environment, helminth infections and human populations, as well as the assumptions of stationarity and spatial isotropy (Soares Magalhães et al., 2011b). Finally, the green box includes uncertainties in the methods used to estimate the model parameters.

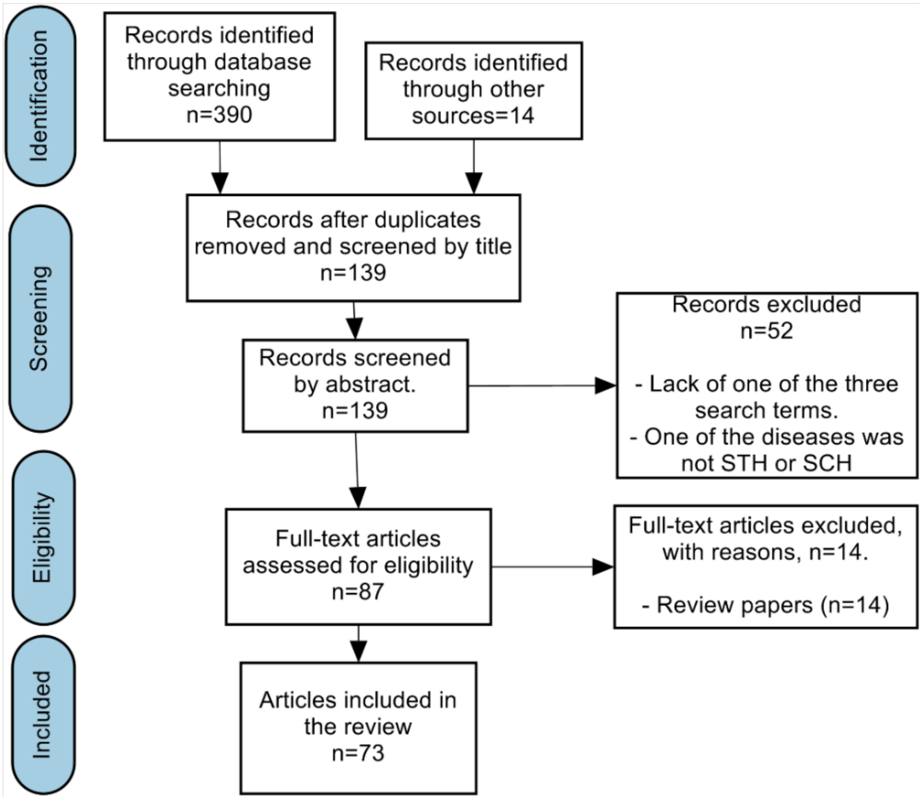

Identification
Screening
Eligibility
Included
Records identified through database searching n=390
Records identified through other sources=14
Records after duplicates removed and screened by title n=139
Records screened by abstract. n=139
Records excluded n=52
- Lack of one of the three search terms.
- One of the diseases was not STH or SCH
Full-text articles assessed for eligibility n=87
Full-text articles excluded, with reasons, n=14.
- Review papers (n=14)
Articles included in the review n=73

Figure 2.1: *PRISMA flow diagram*

Table 2.1: *Classification of search terms*

Uncertainty term		**Spatial term**		**Disease term**	
1	TS=uncertain*	3	TS=geogra* OR TS=spatial OR TS=geo$spatial OR TS= "remote* sens*"	4	TI= schistosom*
2	TS= vague* OR TS=*precision OR TS=*precise OR TS=*accura* OR TS=fuzz* OR TS=error* OR TS=bias			5	TI= hookworm* OR TI="trichuris trichiura" OR TI="ascaris lumbricoides"
				6	TI= helminth* OR TI="soil$transmitted helminth*" OR TS= “neglected tropical disease*”

The third stage (yellow box), shows how uncertainty in the predicted morbidity indicator is addressed, firstly in policy and decision making settings and secondly in a scientific setting. This stage aims to understand how information on uncertainty is used practically and how is it defined and quantified. The blue box represents different elements of data quality that relate to the sources of information (pink box), and the predicted morbidity indicators (yellow box), which due to its wide field of study and importance was separated into a different box.

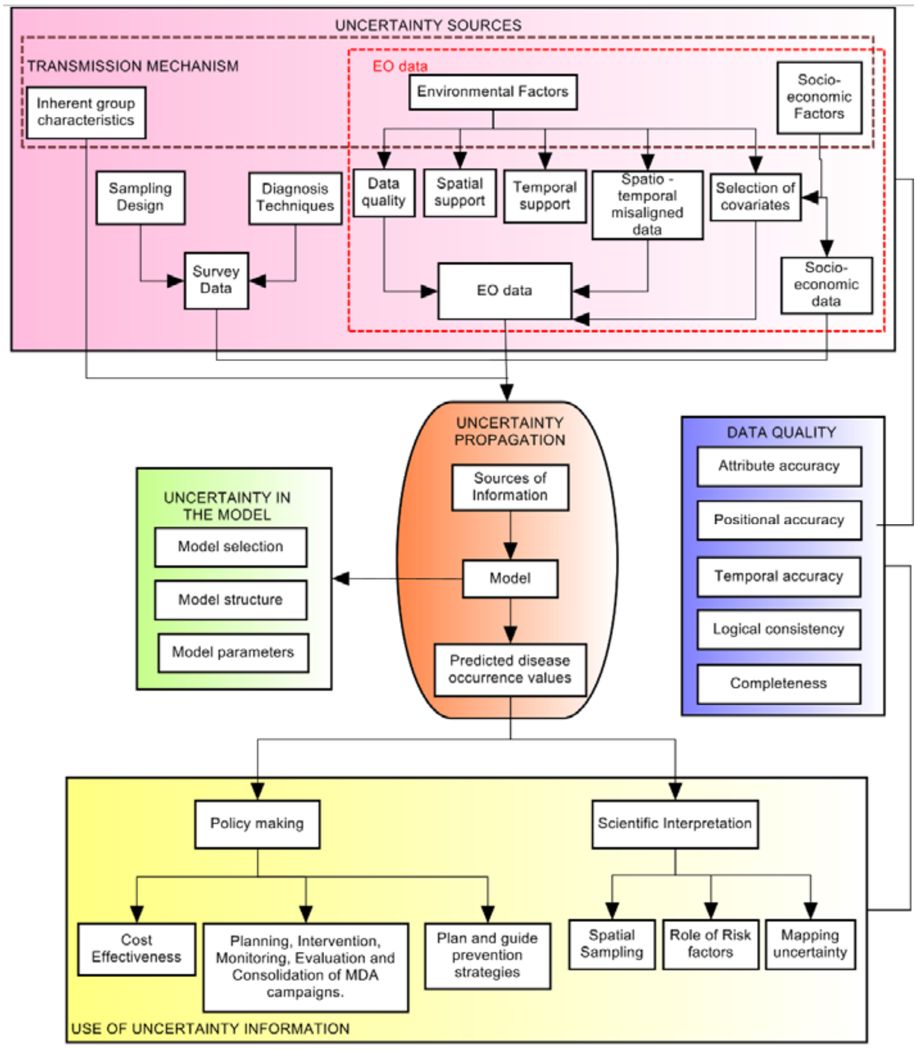


Figure 2.2: *Uncertainty propagation through the process chain of mapping and modelling helminth infections. Pink box: uncertainty from information data sources. Orange box: uncertainty from the predictive model. Yellow box: uncertainty in the predictions.*

Uncertainty use in helminth infection mapping for morbidity control (Uncertainty Interpretation)

Two approaches were considered to describe the possible usage of uncertainty in helminth infections modelling. The first approach indicates that uncertainty could be used in policy making in order to support public health institutions, governments and national or international organizations involved in the control and prevention of STH and schistosome infections. Three foci of attention for policy making were considered: (1) plan and guide prevention strategies, (2) plan the intervention, monitoring, evaluation and consolidation of MDA campaigns, (3) evaluate cost-effectiveness of control programmes. The second approach proposes to use uncertainty to support scientific interpretation by looking at the influence of different information sources on the modelling process, and decide about new improvements or conclusions that need to be considered. Three foci of attention for scientific research were considered: (1) spatial sampling, (2) the role of risk factors (covariates in the statistical model), (3) the mapping of uncertainty. An overview of the different foci of attention of uncertainty information is explained in Table 2.3.

Uncertainty sources in modelling and mapping helminth infections (Uncertainty in the data)

Sources of uncertainty shown in the red box in Figure 2.2 were classified into four: (1) survey, (2) Earth observation, and (3) socio-economic data, (4) inherent group characteristics. Survey data encompassed uncertainties in the response variable, while Earth observation and socio-economic data were uncertainty sources coming from the covariates. Survey data contained uncertainty from the sampling design and diagnostic technique.

Sampling design refers to the type of survey used, sample manipulation, sample size selection, incomplete sample coverage, logistic limitations, survey registration method, adjustment for confounding and the measured morbidity indicator. Uncertainty in the diagnostic technique arises due to the lack of sensitivity and specificity in the methods used to detect helminth parasites eggs in the stool or urine of affected individuals. Uncertainties derived from Earth observation data arise due to spatio-temporal misaligned data, incorrect selection of significant environmental and socio-economic variables, as well as selection of spatial and temporal support of analysis which do not fit the study purpose. The term *misaligned data* refers to the combination of multiple datasets that may be defined on different or non-aligned spatial units, whereas the support refers to size, shape and orientation of the spatial units (Gotway and Young, 2002).

The term *scale* can have multiple meanings in geographical information science (GIScience) (Dungan et al., 2002); here we consider scale in terms of the *support* of the data and the *extent* of the study domain (Atkinson and Graham, 2006).

Table 2.2: Keywords used in the literature search,* indicates wildcard

Uncertainty term (UT)	**Spatial term (ST)**	**Disease term (UT)**
Uncertainty, uncertain, uncertainties	Geographic, geographical, geography	helminth(s), helminthiasis, soil-transmitted helminths, soil-transmitted helminthiasis, neglected tropical diseases
Vagueness, vague	Spatial, geospatial	Schistosome, Schistosoma, schistosomiasis
Imprecision, precision, precise, imprecise	Remote sensing, remotely sensed	Hookworm(s)
Accuracy, inaccuracy, accurate, inaccurate		Trichuris trichiura
Fuzzy, fuzziness Error(s) Bias		Ascaris lumbricoides

Data quality refers to the evaluation in terms of fitness-for-use for a given application . This evaluation addresses the completeness, logical consistency, time, attribute and positional accuracy of spatial data (Iso, 2013; Shi, 2009). Different measurements of the same variable may even have different qualities according to the sensitivity, specificity and accuracy of the instrument or measurement technique.

Scale is a major concern in spatial epidemiology (Atkinson and Graham, 2006; Walz et al., 2015c). Different environmental and socio-economic risk factors may be relevant according to the scale of the analysis (Simoonga et al., 2009a). For a given extent the choice of support may affect the patterns identified in the data (Schur et al., 2011; Schur et al., 2013) as well as the relationship between the response variable and covariates. This is known as the modifiable areal unit problem (MAUP) in GIScience (Dungan et al., 2002). Different datasets may be misaligned and need to be brought to a common grid prior to analysis (Schur et al., 2013). Hence it may be necessary to aggregate, disaggregate or interpolate data prior to analysis (Raj, Hamm and Kant, 2013). All of these operations may be applied in time and space and all have an associated uncertainty. Issues about the selection of significant environmental and socio-economic variables referred to: (1) the exclusion of some socio-economic and climatic factors, which due to logistics or lack of reliable information have not been included in the modelling process; (2) the uncertain choice of covariates produced by the lack of knowledge about the influence of risk factors depending on the spatial support of analysis, the spatial support of the data and other aspects of data quality. Sources of uncertainty derived from inherent group characteristics refer to the heterogeneous distribution of parasites in the population, and the influence of polyparasitism (infection due to multiple parasites also termed coinfections) on the risk of infection.

Table 2.3: *Description of communication of uncertainty*

Uncertainty informs about		Description
Policy Making	Planning, Intervention, Monitoring, Evaluation and Consolidation of MDA campaigns.	• *Plan* spatial targeting and the frequency of deworming campaigns to estimate required drug supplies. • Guide *interventions* towards high risk populations. • *Monitoring:* Maintain success and long term sustainability of control programs. • *Evaluation:* compare and choice more efficient strategies to control the disease. • *Consolidate* control and move towards disease elimination.
	Cost effectiveness	• Inform about the cost associated with the health benefit acquired by implementing a specific control strategy. • Ensure the resources are distributed efficiently by channel funds to high risk populations.
	Plan and guide prevention Strategies	• Plan and guide hygiene education and infrastructure programs in water sanitation and hygiene, as well as implement environmental educational health awareness programs. • Control intermediate host or parasite sources to prevent transmission to definitive hosts.
Scientific Interpretation	Sampling	• Define uncertain risk areas where further data collection is required. • Guarantee the safety of local citizens from future infection resurgence by determining appropriate surveys and monitoring strategies.
	Role of risk factors	• Investigate the effect of environmental risk factors on transmission of parasites. • Guide control efforts in the absence of epidemiological information.
	Mapping Uncertainties	• Spatial representation of uncertainty as a necessary resource for decision making.

Table 2.4: *Measures of uncertainty corresponding to different types of data*

Categories of Imperfection	**Element of uncertainty**	**Measure of Uncertainty**	**Abbreviation**
Imprecision	Continuous data	Standard deviation Credible intervals Confidence Intervals	SD CrI CI
	Categorical data (Vagueness)	Fuzzy sets Rough sets	
Inaccuracy	Continuous data	Root mean square error Mean absolute error Residual mean square Mean error (bias)	RMSE MAE RME ME
	Categorical data	Overall accuracy User's accuracy Producer's accuracy Kappa statistic	OA UA PA K
	Binary data	Area under the receiver operator characteristic curve	AUC

Uncertainty definition and quantification in helminth infections mapping

As mentioned in the introduction, uncertainty was conceptualized as imperfection and further categorized as accuracy and imprecision (Duckham et al., 2001; Worboys and Duckham, 2004). Accuracy may be evaluated by comparison with a reference dataset (Atkinson and Graham, 2006; Congalton, 2010) and different quantitative measures may be used depending on the type of data. Continuous data may be evaluated using the root mean square error (RMSE) or mean absolute error (MAE), which are both measures of the average error. Bias can be evaluated using the mean error. Categorical data are typically evaluated using a confusion matrix with summary measures including the overall accuracy, user's and producer's accuracy and kappa statistic. Binary data may be evaluated using the area under the receiver operator curve (ROC) (AUC). Measures of accuracy are summarized in Table 2.4.

Evaluation of imprecision depends on the nature of the phenomena and data being studied. Where these are well defined, imprecision may be defined statistically (Tavana et al., 2016) and applied in both Bayesian and frequentist settings. The error variance is the usual measure here, although this is commonly expressed as the standard deviation or standard error or as an interval – such as the 95% confidence interval (frequentist) or credible/credibility interval (Bayesian). Vagueness may be evaluated using fuzzy set or rough set theory (Tavana et al., 2016). Table 2.4 shows the elements and measures of uncertainty conceptualized as imperfection.

2.3 Results on uncertainty definition and quantification in helminth infections mapping

2.3.1 Search Strategy

The total number of papers found in each search is shown in Table 2.5. Table 2.6 shows the resulting number of read and discarded papers presented per infection. In total 73 papers were selected, from which 14 were review papers. While the identified review papers were not included in this review we examined their reference lists; this yielded another 14 valuable references that had not been identified by our original search. Finally 73 primary research papers were included in our systematic review. Our results demonstrate that the annual number of publications on mapping and modelling STH and schistosome infections was constant until the year 2007 and steadily increased since then; since 2008 a total of 49 (67% of the total) papers were published (Figure 2.3).

2.3.1 Data collection process

Uncertainty use in helminth infection mapping for morbidity control

For policy making 47 (64%) studies used uncertainty information, in planning, intervention, monitoring, evaluation and consolidation of MDA campaigns (Table 2.7). This was followed by 15 (21%) studies that focused on increasing cost effectiveness of these programmes. Five studies (7%) used uncertainty in disease maps to inform about prevention strategies such as to plan and guide hygiene education and infrastructure WASH programmes. For scientific interpretation, only seven studies (10%) used uncertainty to improve spatial sampling, eight studies (11%) used it to investigate the role of environmental and socio-economic risk factors on the infections, and 17 (23%) papers mapped uncertainty.

Uncertainty sources in modelling and mapping helminth infections

Table 2.8 shows that, from the total number of reviewed papers, sampling design was the most highlighted source of uncertainty, with a total of 42 (58%) papers acknowledging it. The second and third most highlighted sources of uncertainty were diagnostic techniques, with a total of 29 (40%) papers acknowledging it, and selection of significant environmental and socio-economic variables, acknowledged by 22 (30%) papers. The last highlighted uncertainty source was related to spatial support, with 19 (26%) papers acknowledging it.

Table 2.5: *Results of the search performed in the Web of Knowledge and PubMed, using the search terms and the corresponding keywords given in Table 2.1 and Table 2.2, respectively.*

UT	ST	DT	Results Web of Science	Results PubMed
1	3	4	24	23
2	3	4	72	65
1	3	5	0	5
2	3	5	7	18
1	3	6	19	13
2	3	6	52	90

Table 2.6: *Total number of read and discarded papers presented per infection*

	Read papers	Discarded papers
Schistosomes	47	26
STH	26	26

Table 2.7: *Use of information on uncertainty in the public health context*

Uncertainty informs about		Papers SCH	Papers STH	Total
Policy Making	**Cost effectiveness**	[1-10]	[11-15]	15
	Planning, intervention, monitoring, evaluation and consolidation of MDA campaigns.	[1-8, 16-44]	[11-13, 45-52]	47
	Plan and guide prevention strategies	[4, 28, 38]	[13, 46]	5
Scientific Interpretation	**Sampling**	[5, 17-18, 31, 33, 53]	[54]	7
	Role of risk factors	[39, 55-59]	[60-61]	8
	Mapping uncertainty	[1-2, 4-6, 17-18, 24-25, 27-28, 33, 36, 62-63]	[13, 45]	17

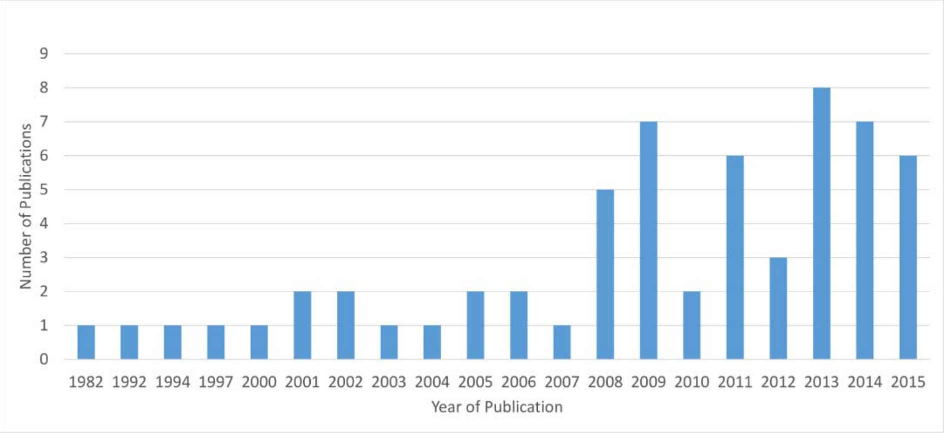


Figure 2.3: *Year of publication of studies included in this review*

The least highlighted uncertainty sources were: inherent group characteristics, use of data with insufficient quality, temporal support, and spatio-temporal misalignment, with 15 (20%), 15, 7 (10%) and 5 (7%) papers acknowledging them respectively. From the category sampling design, the most highlighted sources of uncertainty were: incomplete sample coverage and sample size, with respectively 16 (37%) and 22 (51%) papers acknowledging them respectively (table 2.9). Heterogeneity and polyparasitism were acknowledged by nine (12%) and six papers (8%) respectively.

Regarding uncertainty relating to the model, model structure was the most highlighted source of uncertainty, with 19 (26%) papers acknowledging it, followed by, uncertainty in model selection and uncertainty in model parameters with 3 (4%) papers each (Table 2.10).

Uncertainty definition and quantification in helminth infections mapping

Four ways to define uncertainty were found: *accuracy*, *imprecision*, *bias* and *vagueness*. Sixty-one (83%) papers expressed uncertainty in the modelled results using measures of imprecision and credible intervals were the most frequently used measure of imprecision (Table 2.11). Thirty-nine (53%) papers defined uncertainty by means of accuracy, using mostly the area under the curve of the receiver operating characteristic and the percentage of correctly predicted morbidity indicators.

Table 2.8: *Uncertainty sources in modelling and mapping helminth infections*

Uncertainty sources			Papers using different measures of uncertainty	Papers highlighting the importance of uncertainty sources		Total
				Papers SCH	Papers STH	
ID	**Survey Data**	Sampling design	ROC (AUC) [5] Credible intervals [26]	[3-6, 8-9, 17, 20, 22-28, 30, 32, 34, 36-37, 40, 42, 44, 52, 55, 62, 64-65]	[11-12, 14, 45-48, 50-51, 66-70]	42
		Diagnostic Techniques	Credible intervals [11, 44]	[4, 6-8, 10, 20, 22-28, 37-40, 43-44, 53, 55-56]	[11-13, 15, 46, 49, 66]	29
	EO data	Spatial support		[1, 4-5, 8-9, 22, 25, 27, 34, 38, 44, 53, 56]	[13, 45, 47, 60-61, 68]	19
		Temporal support		[53, 64]	[14, 45, 60-61, 66]	7
		Data quality		[1, 3, 16, 18, 28, 36, 42, 52-53, 56-57]	[14, 46, 48, 50]	15
		Spatio-temporal misaligned data		[9, 38]	[18, 46, 48]	5
	Socio-economic data	Selection of significant environmental and socio-economic risk factors	*Credible Intervals:* SCH: [2-7, 9-10, 18, 22, 24-26, 29-30, 34, 39-44, 53, 56-59, 64, 71] STH: [11-14, 45-47, 49-50, 60-61, 70] *Confidence Intervals:* SCH: [23, 53, 64] STH: [60-61, 66, 68]	[4-5, 8, 19-21, 25-28, 32, 34, 38, 44, 59, 65]	[11-13, 46-47, 66]	22
IGC	Heterogeneity		ROC (AUC) [3]	[3, 6-7, 20, 44, 59]	[12, 46, 69]	9
	Polyparasitism			[6-7, 10, 25]	[48, 70]	6

ID: Input Data; *IGC:* Inherent group characteristics

Table 2.9: *Categories of sources of uncertainty and papers included in this review grouped into categories*

Categories	Uncertainty sources	Papers focusing on SCH	Papers focusing on STH	Total
Sampling Design	Type of survey	[22, 26, 30, 32, 69]	[47]	6
	Samples manipulation	[23]		1
	Sample size	[4, 6, 9, 17, 20, 24-25, 27, 30, 32, 36-37, 62, 64]	[11-12, 45-47, 66-67, 70]	22
	Sample coverage	[3-4, 17, 24-25, 34, 42, 44]	[11-12, 14, 46, 48, 50-51, 68]	16
	Logistics	[3, 8, 27, 55, 65]	[51, 66]	6
	Survey registration method	[5, 9, 52]		3
	Adjust for confounders	[26]		1
	Selection of the measure of risk	[32]	[46, 69]	3
Diagnostic Techniques	Sensitivity and specificity of diagnostic methods	[4, 6-8, 10, 20, 22-28, 37-40, 43-44, 53, 55-56]	[11-13, 15, 46, 49, 66]	29
Spatial support	Spatial aggregation and disaggregation	[1, 4-5, 8-9, 22, 25, 27, 34, 38, 44, 53, 56]	[13, 45, 47, 60-61, 68]	19
Temporal support	Temporal aggregation and disaggregation	[53, 64]	[14, 45, 60-61, 66]	7
Data quality	Position accuracy, logical consistency, time accuracy, completeness, attribute accuracy (pre-processing)	[1, 3, 16, 18, 28, 36, 42, 52-53, 56-57]	[14, 46, 48, 50]	15
Spatio-temporal misaligned EO data	Spatial and temporal misaligned EO data.	[9, 38]	[18, 46, 48]	5
Selection of environmental and socio-economic variables	*Environmental:* Distance to water bodies, land surface temperature, soil moisture, vegetation cover, Rainfall.	[4-5, 8, 19-21, 25-28, 32, 34, 38, 44, 59, 65]	[11-13, 46-47, 66]	22
	Socio-Economic: poverty, clean water, sanitation and hygiene, urbanization, land use.			
Inherent group characteristics	Heterogeneity	[3, 6-7, 20, 44, 59]	[12, 46, 69]	9
	Polyparasitism	[6-7, 10, 25]	[48, 70]	6

Table 2.10: *Model sources of uncertainty*

Model uncertainty sources	Papers SCH	Papers STH	Total
Model parameters	[3, 16, 55]		3
Model selection	[18]	[46, 69]	3
Model structure	[1, 3-4, 6-8, 18, 20, 26, 30, 33-34, 41, 44, 53, 64]	[12-13, 46, 61]	20

Bias and vagueness were the least used measure of uncertainty with only five (7%) and one (1%) papers quantifying uncertainty in their results by means of mean error and fuzzy sets respectively.

A total of 57 (78%) studies evaluated regression coefficient parameters by means of precision, and quantified them using Bayesian approaches (57%), and frequentist approaches (52%). This overlap arose because several authors first used frequentist non-spatial approaches to identify the significant covariates (Soares Magalhães et al., 2015) and then applied these covariates in a Bayesian geostatistical model (Chammartin et al., 2014; Pullan et al., 2014; Walz et al., 2015a). Two papers (3%) quantified the uncertainty arising due to questionnaires data, as well as the uncertainty arising due to combining age-groups in the predictions (Clements et al., 2008; Schur, Utzinger and Vounatsou, 2011). Regarding diagnostic techniques, two studies (3%) addressed diagnostic uncertainty by modelling sensitivity and specificity as random variables, specified as beta distributions, and quantified as posterior credible intervals (Soares Magalhães et al., 2014; Soares Magalhães et al., 2015).

2.4 Discussion

2.4.1 Uncertainty use in helminth infections mapping for morbidity control

Most of the studies used information on uncertainty to guide MDA campaigns and evaluate their cost effectiveness. Information on uncertainty was also used to evaluate the role of risk factors in mapping helminth infections. Nevertheless, prevention strategies, improvements in sampling design, and mapping of uncertainty have not yet been addressed (Walz et al., 2015b). We advise to use information on uncertainty not only to inform about MDA campaigns, but also to inform about prevention strategies such as improving sanitation and hygiene education or delineating potential transmission sites (Walz et al., 2015b).

Table 2.11: *Uncertainty definition and quantification*

Uncertainty definition	Uncertainty quantification	Model + parameters		Total	Parameters	
		Papers SCH	Papers STH		Papers SCH	Papers STH
Accuracy	Residual mean square.	[1]		1		
	Mean absolute error.	[6, 19, 22, 26]	[13, 43, 45, 70]	8		
	Percentage of locations that were predicted within a 95% confidence/credible interval.	[2, 4, 6-7, 10, 24, 26, 57, 72]	[12, 45, 70]	12		
	Receiving operating characteristics (AUC).	[2-3, 5, 8, 18, 25, 30, 32, 34, 42, 44, 73]	[11, 13, 46, 49-50]	18	[3, 5]	
	Point-wise standard error.	[17]		1		
	Log likelihood ratio.	[21]		1		
	Root mean square error.	[62-63, 73]		3		
	Kappa statistic.	[36]	[54]	2		
Precision	Bayesian approaches (Credible Intervals).	[2-7, 9-10, 18, 20, 22, 24-26, 29-30, 34, 39-44, 53, 56-59, 64]	[11-14, 45-47, 49-50, 60-61, 70]	42	[2-7, 9-10, 18, 20, 22, 24-26, 29-30, 34, 39-44, 53, 56-59, 64]	[11-14, 45-47, 49-50, 60-61, 70]
	Standard deviation.	[27, 33, 35, 63]		4		
	Standard deviational ellipse.	[28]		1		
	Frequentist approaches (Confidence intervals, R squared).	[1, 4-6, 8, 16-17, 23, 28, 33, 36, 38, 40-42, 44, 52-53, 55-58, 62-64, 70, 72]	[11, 14, 50-51, 54, 59-61, 66, 68]	38	[1, 4-6, 8, 17, 23, 28, 33, 36, 40-42, 44, 52-53, 55-58, 62-64, 70]	[11, 14, 50-51, 54, 59-61, 66, 68]
	Ranking statistic based on maximum likelihood.	[16]		1		
Bias	Residual, mean error	[6, 9, 43, 63]	[13]	5		
Vagueness	Fuzzy theory	[74]		1		

Transmission control is important for its public health relevance, since potential disease transmission sites could guide direct intervention measures at the place of infection (Walz et al., 2015b; Walz et al., 2015c). Likewise, mapping of uncertainty is also recommended, since it is known to be an important tool for public health decision making, especially to determine the geographical distribution of areas for which information is lacking [112]. Mapping could be used as a tool to improve the sampling strategy and modelling efforts. Maps of uncertainty could also support communication of uncertainty to the affected communities. A complete exploration and judgement of uncertainty information would enhance the assessment of the risk of getting these infections, and would allow to understand potential impacts on human health (Burns et al., 2014).

While most studies identified and discussed different sources of uncertainty, this was mainly limited to a qualitative discussion, rather than a quantitative one (Jurek et al., 2006) (Table 2.11). For instance, 38 (52%) papers highlighted qualitatively the importance of sampling design in mapping helminth infections, but only two studies (3%) have quantified their possible effects on the accuracy of the predicted morbidity indicator. An example is given by Clements et al (Clements et al., 2006), where uncertainties in the predictions were used to identify areas requiring further data collection before programme implementation. The lack of a quantitative assessment limits the utility of the findings in both policy/decision making setting and a scientific setting (Burns et al., 2014; Stürmer et al., 2007). Communication of uncertainty will never be complete without an extensive quantification of uncertainties in all possible information sources (Burns et al., 2014), where model assumptions, selection of covariates and acquisition of survey data are clearly explained, either within the publication or as supplementary information.

2.4.2 Uncertainty sources in modelling and mapping helminth infections

Figure 2.4 shows the three uncertainty stages previously described in Figure 2.2 where these stages encompass specific uncertainty components, which need to be considered for a complete uncertainty communication. Each of these components is analyzed in the next sections.

Uncertainty in the response variable (morbidity indicator)

This uncertainty belongs to the first uncertainty stage (uncertainty coming from different data sources) and is described in Box A from Figure 2.4. This type of uncertainty exists as a function of the measurement (Dungan, 2002) or data collection. Uncertainty in the response variable depends on the survey data quality, generated based on the sampling design, and the used diagnostic approach (Figure 2.2). A total of 68% of the papers mentioned the importance of sampling design as the main source of uncertainty, supporting the idea that significantly biased results may be produced due to an inappropriate sampling design (Rothman, 2012). When mapping helminth infections, it is suggested to document the sample size calculation method, together with the analysis of a certain target group selection. Other sources of uncertainty in sampling design are related to the type of survey, type of morbidity indicator and the use of misaligned survey data. For instance, Chammartin et al. (Chammartin et al., 2014) argued that cross sectional studies might not capture well the focal pattern of schistosomiasis, since their information is based on an specific point in time. Likewise, prevalence as the most frequently used morbidity indicator, underestimates morbidity values (Mccreesh, Nikulin and Booth, 2015; Soares Magalhães et al., 2014) and was considered a biased and poor indicator of risk (Mccreesh, Nikulin and Booth, 2015; Rothman, 2012).

Also, combining data from different sources of information, with different survey times and diagnosis methods may result in inaccurate estimates (Clements et al., 2008; Schur et al., 2013; Schur, Utzinger and Vounatsou, 2011). This is why it is suggested to document all possible drawbacks in the selected type of survey and measure of risk, and document all problems when using misaligned survey data.

Data collection also influenced the results when there was a lack of spatial and laboratory sampled data in areas where the presence of infection was suspected to be high (Hu et al., 2015; Mccreesh, Nikulin and Booth, 2015; Scholte et al., 2014). This could be due to inaccurate and missing reports (Hu et al., 2015), lack of people's participation and limited access to geographical areas (Zhang et al., 2009).

All these potential causes should be reported as well as issues regarding high costs of the survey, diagnosis, delivery of drugs, type of registration resource and limited training and expertise of field personnel, which might also influence the quality of the results (Hu et al., 2015; Zhang et al., 2009). For instance, the use of questionnaires might underestimate prevalence data, since their discriminatory performance differs among regions, and these are not always completely returned by surveyed people (Clements et al., 2008; Sturrock et al., 2013).

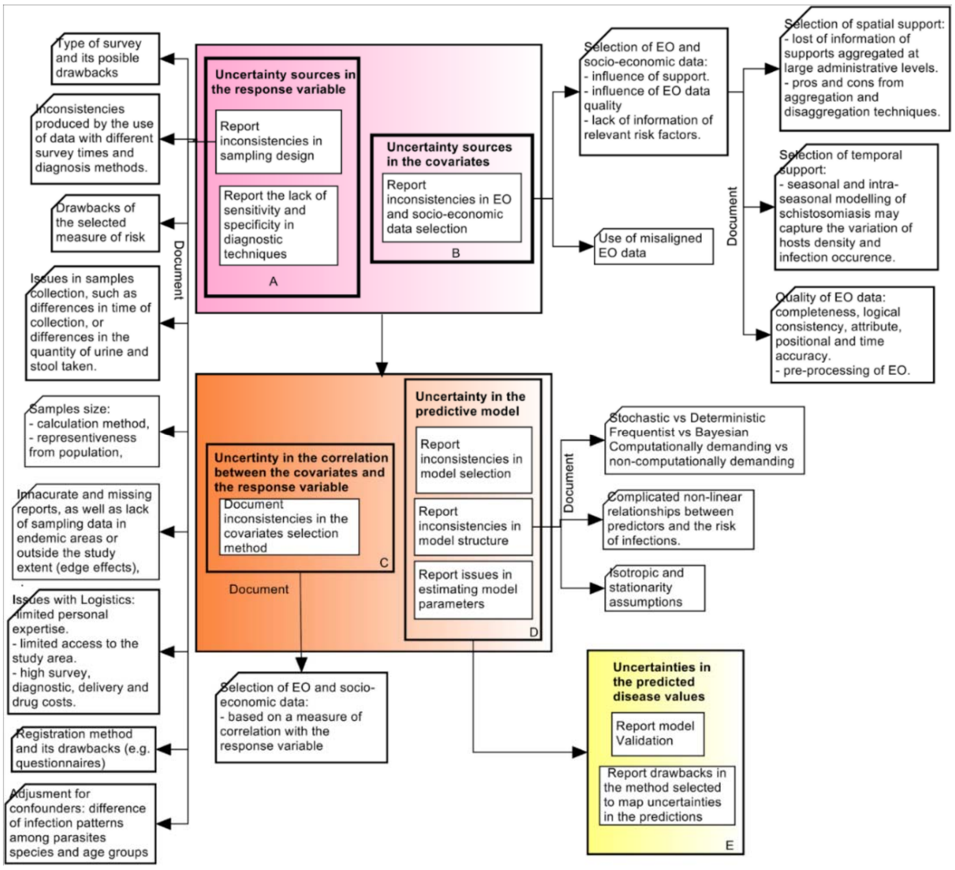

Type of survey and its posible drawbacks
Inconsistencies produced by the use of data with different survey times and diagnosis methods.
Drawbacks of the selected measure of risk
Issues in samples collection, such as differences in time of collection, or differences in the quantity of urine and stool taken.
Samples size:
- calculation method,
- representiveness from population,
Innacurate and missing reports, as well as lack of sampling data in endemic areas or outside the study extent (edge effects),
Issues with Logistics:
- limited personal expertise.
- limited access to the study area.
- high survey, diagnostic, delivery and drug costs.
Registration method and its drawbacks (e.g. questionnaires)
Adjusment for confounders: difference of infection patterns among parasites species and age groups
Document
Uncertainty sources in the response variable
Report inconsistencies in sampling design
Report the lack of sensitivity and specificity in diagnostic techniques
A
Uncertainty sources in the covariates
Report inconsistencies in EO and socio-economic data selection
B
Selection of EO and socio-economic data:
- influence of support.
- influence of EO data quality
- lack of information of relevant risk factors.
Use of misaligned EO data
Document
Selection of spatial support:
- lost of information of supports aggregated at large administrative levels.
- pros and cons from aggregation and disaggregation techniques.
Selection of temporal support:
- seasonal and intra-seasonal modelling of schistosomiasis may capture the variation of hosts density and infection occurence.
Quality of EO data: completeness, logical consistency, attribute, positional and time accuracy.
- pre-processing of EO.
Uncertinty in the correlation between the covariates and the response variable
Document inconsistencies in the covariates selection method
C
Document
Uncertainty in the predictive model
Report inconsistencies in model selection
Report inconsistencies in model structure
Report issues in estimating model parameters
D
Document
Stochastic vs Deterministic
Frequentist vs Bayesian
Computationally demanding vs non-computationally demanding
Complicated non-linear relationships between predictors and the risk of infections.
Isotropic and stationarity assumptions
Selection of EO and socio-economic data:
- based on a measure of correlation with the response variable
Uncertainties in the predicted disease values
Report model Validation
Report drawbacks in the method selected to map uncertainties in the predictions
E

Figure 2.4: *Stages of uncertainty analysis when mapping STH and schistosome helminth infections. Colour coding as for Figure 2.2*

Finally, issues related to diagnostic technique, sample manipulation (Krauth et al., 2012), and lack of stratification due to confounders (Schur, Utzinger and Vounatsou, 2011) are also important to be considered and should also be reported and analyzed.

Uncertainty in the covariates (EO data)

This uncertainty is also part of the first uncertainty stage and is represented in Box B of Figure 2.4. Main sources of uncertainty in the covariates were related to the selection of significant environmental and socio-economic risk factors, the type of environmental data, and also to the selection of the spatial support of analysis. The importance of including risk factors such as sewage system, water supply and other climatic, demographic and socio-economic variables were the most highlighted issues (Table 2.8). Soares Magalhães et al (Soares Magalhães, Barnett and Clements, 2011) found that including WASH indicators as random variables in the model contributed to improved definition of the areas to target for integrated helminth control and improvement of WASH risk factors. The selection of EO data depends on the selected spatial support, defined based on the research objective and analysis method used (Nijland et al., 2009), but also on the quality of EO data itself. In addition Walz et al. (Walz et al., 2015a) argued that the relevance of environmental variables are expected to vary between different landscapes and ecological regions, having an impact on the predicted morbidity indicators. Likewise, socio-economic and ecological processes that govern schistosomiasis transmission operate and vary across different scales of observation (Schur, Vounatsou and Utzinger, 2012). Since statistical correlation can vary according to the extent of the studied area and the scale of aggregation (Walz et al., 2015b), quantitative methods to select the optimal support of analysis, such as aggregation and disaggregation process should be documented. Clear guidance on the selection of the optimal support of EO data does not exist (Hamm, Soares Magalhães and Clements, 2015), and this remains an open topic of research. Nevertheless the choices made as well as an applied aggregation or disaggregation should be documented. Although few studies highlighted the relevance of data quality, temporal support and extent, and spatio-temporal misaligned data (Table 2.9), these sources of uncertainty cannot be ignored. Data quality elements (i.e completeness, logical consistency, temporal accuracy, spatial accuracy, and attribute accuracy (Iso, 2013) relate to the identification of uncertainty sources, and have been shown to influence the predicted disease risk (Hamm, Soares Magalhães and Clements, 2015). EO quality elements should also be addressed and analyzed, as well as possible inconsistencies in their pre-processing. Attention should also be put to the selection of the temporal support of analysis (Jong and Bruin, 2012), which need to be defined depending on the study objective and the host and vectors epidemiology and ecology. Finally, both temporal and spatial supports need to be adjusted into a common temporal and spatial grid since different spatial and temporal supports, could lead to erroneous conclusions in the predictions (Gotway and Young, 2002).

According to our analysis, although uncertainty in the covariates has been highlighted by most studies, almost none of them have quantified their impact on the disease risk predictions, and just a few have incorporated uncertainty in the response variable. Uncertainty quantification and documentation is suggested in order to completely inform about uncertainty and help decision makers and public health scientists to undertake independent uncertainty assessments (Jurek et al., 2007) and better communicate uncertainty (Burns et al., 2014; Stürmer et al., 2007).

Uncertainty in the EO data selection, predictive model and predicted disease values

Spatial prediction of parasitic disease risk patterns are explained by the statistical relationships between environmental and socio-economic covariates, individuals, and observed risk of infection (Soares Magalhães et al., 2011b). Setting initial candidate environmental and socio-economic covariates and their inclusion in the predictive model is one of the first steps for geostatistical modelling of helminth infections. Thus the methods used for this selection should be explained and documented explicitly such that the statistical method itself and the measure used for covariates inclusion are clearly interpreted in the mapping process (Box C from Figure 2.4). The selection of the predictive model, its possible limitations (when estimating model parameters, predicting morbidity indicators, or handling non-linear relations between response variables and covariates) and assumptions made, should also be reported and justified, explaining step by step the reasoning behind the use of the specific model (Box D from Figure 2.4). Boxes C and D in Figure 2.4 relate to the green box (uncertainty in the predictive model) in Figure 2.2, whereas Box E relates to the model output (yellow Box from Figure 2.2).

The mean predicted values are often aggregated to different administrative supports, without considering the uncertainty in the predictions (Kabore et al., 2013). This could lead to a biased estimate of treatment needs (Kabore et al., 2013; Schur, Vounatsou and Utzinger, 2012). Uncertainty can and should be incorporated into the aggregation process, yielding measures of precision (e.g., credible intervals) in the aggregated predictions. Where feasible, we advise validation of the predicted aggregated morbidity indicators (Box E in Figure 2.4) against empirical observations (Kabore et al., 2013). This will facilitate a more appropriate spatial target of intervention and prevention strategies.

2.5 Conclusions

Acknowledging and incorporating uncertainty in mapping and modelling helminth infections is a step-by-step process, which should be considered formally when developing geographical models of helminth infection. Geographical models aim at informing, not only about MDA campaigns and their cost-effectiveness, but also prevention strategies, where it is necessary to define transmission areas and plan and guide hygiene education and infrastructure programs in water sanitation and hygiene. A quantitative and qualitative analysis of uncertainty is necessary for a complete assessment of risk, to understand potential impacts on human health, and to allow a complete uncertainty communication to public health managers. Five components of uncertainty analysis were recognized: (1) uncertainty in the response variable,

(2) uncertainty in the covariates, (3) uncertainty in the relationship between them, (4) uncertainty in the predictive model, and (5) the propagated uncertainty on the results. Our conclusions are shown diagrammatically in Figure2.5, which aims at providing a framework for a full uncertainty evaluation when undertaking spatial modelling of helminth infections for policy formulation. Uncertainty analysis should start by identifying possible sources of uncertainty in the studies and categorize them such that at least the most important ones can be incorporated into the predictive model. Sampling design and EO data have been acknowledged as the major sources of uncertainty and should be given primary attention in the modelling process. In particular, sampling design, diagnosis, selection of significant risk factors, and selection of an adequate spatial support of analysis. Next, uncertainties in the response variable and covariates should be quantified and incorporated into the model. Methods used to define the relationship between covariates and response variables should also be documented, as well as the selection of the predictive model and its limitations. Finally, uncertainties in the parameters and response variables should be quantified, and uncertainty mapping should be performed as a valuable element for uncertainty communication and policy formulation.

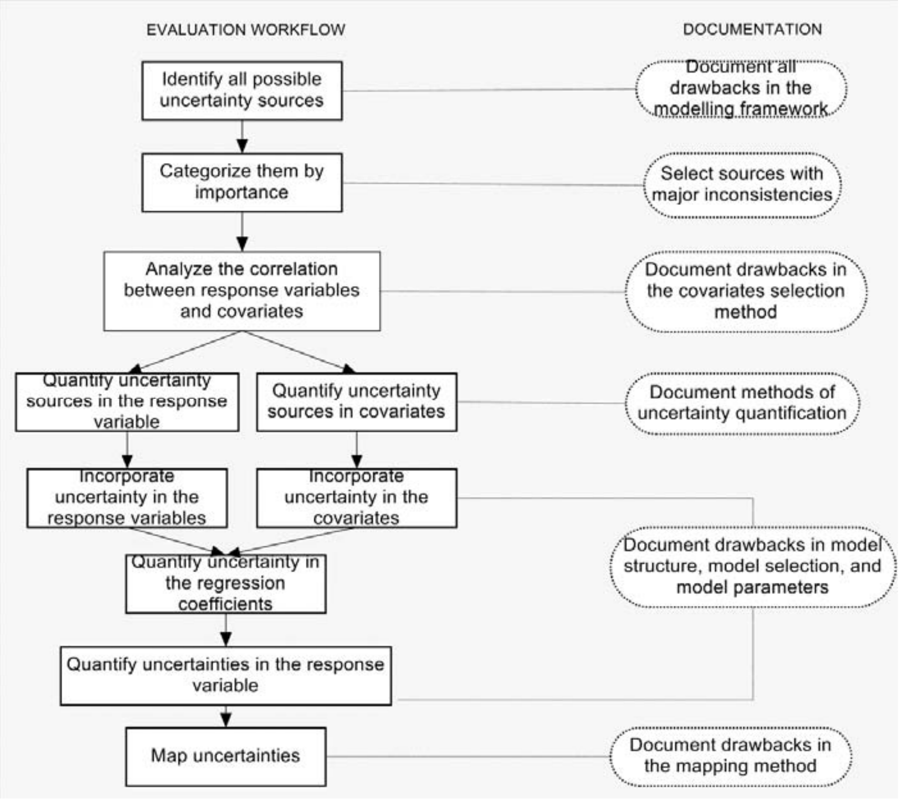

EVALUATION WORKFLOW
DOCUMENTATION
Identify all possible uncertainty sources
Document all drawbacks in the modelling framework
Categorize them by importance
Select sources with major inconsistencies
Analyze the correlation between response variables and covariates
Document drawbacks in the covariates selection method
Quantify uncertainty sources in the response variable
Quantify uncertainty sources in covariates
Document methods of uncertainty quantification
Incorporate uncertainty in the response variables
Incorporate uncertainty in the covariates
Document drawbacks in model structure, model selection, and model parameters
Quantify uncertainty in the regression coefficients
Quantify uncertainties in the response variable
Map uncertainties
Document drawbacks in the mapping method

Figure 2.5: *Framework for the evaluation and utilization of uncertainty in mapping soil transmitted helminth infections and schistosomiasis.*

Chapter 3. Modelling local areas of exposure to *Schistosoma japonicum* in a limited survey data environment [2]

[2] This chapter is published as:

Araujo Navas, A. L.; Soares Magalhaes, R. J.; Osei, F.; Fornillos, R. J. C.; Leonardo, L. R.; Stein, A. Modelling local areas of exposure to *Schistosoma japonicum* in a limited survey data environment. *Parasites & Vectors* 2018, *11*,doi: 10.1186/s13071-018-3039-6.

Abstract

Spatial modelling studies of schistosomiasis (SCH) are now commonplace. Covariate values are commonly extracted at survey locations, where infection does not always take place, resulting in an unknown positional exposure mismatch. The present research aims to: (i) describe the nature of the positional exposure mismatch in modelling SCH helminth infections; (ii) delineate exposure areas to correct for such positional mismatch; and (iii) validate exposure areas using human positive cases.

To delineate exposure areas to *Schistosoma japonicum*, a spatial Bayesian network (sBN) was constructed. It uses data on exposure risk factors such as: potential sites for snails' accessibility, geographical distribution of snail infection rate, and cost of the community to access nearby water bodies. Prior and conditional probabilities were obtained from the literature and inserted as weights based on their relative contribution to exposure; these probabilities were then used to calculate joint probabilities of exposure within the sBN.

High values of probability of *S. japonicum* exposure correspond to polygons where snails could potentially be present, for instance in wet soils and areas with low slopes, but also where people can easily access water bodies. Low correlation ($R^2 = 0.3$) was found between the percentage of human cases and the delineated probabilities of exposure when validation buffers are generated over the human cases.

The utility of a probabilistic method for the identification of exposure areas for *S. japonicum*, with wider application for other water-borne infections, was demonstrated. From a public health perspective, the schistosomiasis exposure sBN developed in this study could be used to guide local schistosomiasis control teams to specific potential areas of exposure, and improve efficiency of mass drug administration campaigns in places where people are likely to be exposed to the infection.

3.1 Introduction

Schistosomiasis (SCH) is a water-borne neglected tropical disease of global public health significance (King, Dickman and Tisch, 2005; Walz et al., 2015b). It affects more than 252 million people worldwide (Hotez et al., 2014), especially human populations living in places where clean water and sanitation are limited (Araujo Navas et al., 2016). Schistosomiasis is known to lead to anaemia, stunted growth and other organ pathologies in school-aged children (Coutinho et al., 2005; Leenstra et al., 2006). Three schistosome species cause the infection*: Schistosoma mansoni*, *S. japonicum* and *S. haematobium. Schistosoma japonicum* is presently endemic in China, Indonesia and the Philippines, and is hard to control due to its zoonotic life-cycle (Jia et al., 2007). The life-cycle of *S. japonicum* includes infection of an amphibious snail belonging to several subspecies of *Oncomelania hupensis* as the intermediate host, and humans and other mammalians as definitive hosts (Tarafder et al., 2006; Yang et al., 2008).

Traditionally, schistosomiasis risk mapping has enabled the identification of at risk populations for targeting mass drug administration campaigns, thus increasing the efficiency of schistosomiasis disease control (Soares Magalhães et al., 2014). Schistosomiasis mapping has been supported by the use of spatial information techniques, such as geographical information systems (GIS), remote sensing and global positioning systems (GPS). Spatial information techniques allow the manipulation of spatially referenced infection data and data on the physical and biological environmental variables (Hamm, Soares Magalhães and Clements, 2015; Herbreteau et al., 2007). Modelling those data in combination allows studying the distribution of communities most at risk schistosomiasis and the role of the geographical variation of environmental exposure factors on schistosomiasis risk (Zhang et al., 2016).

There are a number of errors inherent to spatial information used in geographical epidemiological studies (Araujo Navas et al., 2016). Most of these errors involve positional measurement errors, where observation and prediction locations are affected by various factors such as GPS inaccuracies, the presence of multiple addresses, geocoding errors, outcome or covariate aggregations, and misalignment between covariates of exposure and disease outcome estimates (Zhang et al., 2016). The last one is of our current interest and may occur when covariates of exposure are extracted from locations where exposure has not occurred.

Statistical modelling of the spatial distribution of schistosome infections estimates empirical relationships between morbidity indicators (e.g. prevalence or intensity of infection) and risk factors. Risk factors for schistosome exposure include various environmental and socio-economic covariates that help to interpolate the level of infection at unsampled locations (Cadavid Restrepo et al., 2016; Hamm, Soares Magalhães and Clements, 2015; Weiss et al., 2015). Covariates and morbidity indicators are commonly extracted from survey locations such as health centres, hospitals and schools. In most cases, exposure to infection did not occur at survey data locations but at locations where environmental and geographical conditions, together with the level of accessibility to contaminated sites, are optimally exposed. Such exposure locations are usually unknown, resulting in positional mismatch of the surveyed disease values, and the covariates in the model.

To date, methods to account for this type of positional misalignment are scarce. Several studies have used remote sensing data to determine biophysical features of habitats in relation to snail prevalence (Stensgaard et al., 2006; Stensgaard et al., 2013; Zhang et al., 2013), acknowledging that *S. japonicum* transmission is closely related to the distribution of its intermediate host in the environment (Yang et al., 2008). Only one study (Walz et al., 2015b) has used these habitats to correct for the positional mismatch when modelling disease infection risk in human populations. Walz et al. (Walz et al., 2015b) used high-resolution remote sensing data, environmental field measurements, and ecological data, to model environmental suitability for schistosomiasis-related parasites and snail species. They represented environmental suitability as potential transmission areas that could guide public health interventions to places where people could potentially be infected. Although potential transmission areas were delineated, interactions between humans, hosts, and suitable environments were not taken into account.

These studies suggest that ignoring positional mismatch and its impact on spatial prediction remains largely unquantified in schistosomiasis modelling. Furthermore, the extraction of covariate values in the presence of positional mismatch is a significant source of uncertainty that may influence the efficacy of schistosomiasis control strategies (Araujo Navas et al., 2016). Therefore, methods to correct for this positional mismatch need to be further investigated (Araujo Navas et al., 2016; King, Dickman and Tisch, 2005).

The objective of this study is to develop a schistosomiasis exposure sBN model that maps potential areas of exposure to *S. japonicum*, taking into account human interactions with main sources of infection (i.e. water bodies). To accomplish this objective, we aimed to (i) describe the positional mismatch problem in modelling *S. japonicum* infection; (ii) delineate exposure areas that take into consideration the accessibility cost of people to main sources of infection, and that could be used to correct for this positional mismatch; and to (iii) validate the delineated exposure areas.

3.2 Methods to construct the spatial Bayesian network

3.2.1 Data on human and snail *S. japonicum* infection

In the Philippines *S. japonicum* is endemic in 28 of its 81 provinces (Leonardo et al., 2015), with approximately 1.8 million estimated infected people (Leonardo et al., 2002). The disease affects children, adolescents and individuals with high-risk occupations, such as farmers and fishermen (Leonardo et al., 2002; Zhou et al., 2010). In the Philippines, the smallest administrative division is the barangay, numbering about 22–50 in a municipality.

We used data on human schistosomiasis and snail prevalence of infection, collected in six barangays from Alangalang municipality in Leyte province in 2015 and 2016. Data were collected by researchers from the College of Public Health and College of Science from the University of the Philippines. Surveyors selected Alangalang municipality because it has the highest prevalence of schistosomiasis (7.5%) from all the 43 municipalities of Leyte Province; within this municipality, they visited the barangays with the highest prevalence of infection from the 54 barangays in Alangalang municipality.

Human positive cases (12 records) were georeferenced at household locations and snails surveys (8 records) were taken from water bodies in close proximity to surveyed households. The recording of all the human case locations (also including negative cases) was not possible due to a lack of manpower and material resources, such as the availability of only one GPS device in the field.

Diagnosis of schistosomiasis in humans was performed using stool examination. Single stool sample was requested per participant with informed consent, coded and prepared following the Kato-Katz method. Each slide prepared was read in the field using a microscope and the presence of *S. japonicum* eggs indicated active infection.

Infection among *O. h. quadrasi* snails was determined by manually crushing the snails in aliquots on a glass slide. Each snail was placed in an aliquot droplet of distilled water, usually three aliquots per glass slide. Snails were gently crushed in between slides and were examined under a conventional stereomicroscope (40×) using forceps for separating snail tissues to detect the presence of sporocysts or furcocercous cercariae characteristic of *S. japonicum*.

3.2.2 Study Area

For the purpose of this study, it was decided to work at a local spatial scale in the Province of Leyte, due to the localized nature of the surveys and the high endemicity of the disease (Olveda et al., 2016). For the analysis, we identified a small area surrounding surveyed points (Figure 3.1). This was done in order to select only surveyed barangays and to include information of all risk factors, avoiding areas without survey information (Figure 3.1).

3.2.3 Environmental and geographical data

Exposure risk factors of SCH transmission are associated with the environment (i.e. moisture, temperature, rainfall and water characteristics), the topography (i.e. elevation, slope) of the area (Soares Magalhães et al., 2014; Stensgaard et al., 2013; Walz et al., 2015b) and snail infection status (Gao et al., 2014; Zhang et al., 2013). In the endemic provinces of the Philippines, exposure to snails is mostly driven by the local topography, land use and the physical and chemical components of the water and soil (Pesigan et al., 1958). We included elevation, slope, land use, nearest distance to water bodies and snail infection rates as exposure risk factors. Elevation was obtained as a raster file from ASTER GDEM version 2 from USGS (Geological Survey, 2017). Vector layers for land use, river and road network were obtained from the OpenStreetMap (OSM) project (Project, 2017). OSM land use and land cover products use information from GlobeLand30 (GL30), which is a new generation of 30 m land cover maps (Schultz et al., 2017). The OSM road and river networks are incomplete and contain errors in their connectivity. To account for this, we edited roads and rivers, and digitalized footpaths using Google Earth images. The vector layer for snail infection rate was obtained from the recorded surveys (Table 1). Slope was derived from elevation by using the Terrain Analysis tool from Quantum GIS version 2.6 (Project, 2018).

Distance to water bodies was calculated using the closest facility network analysis tool from ArcGIS version 10 (Esri, 2011). Firstly, we corrected for topology errors such as duplicate lines, presence of dangles and multipart geometries in the river and road network.

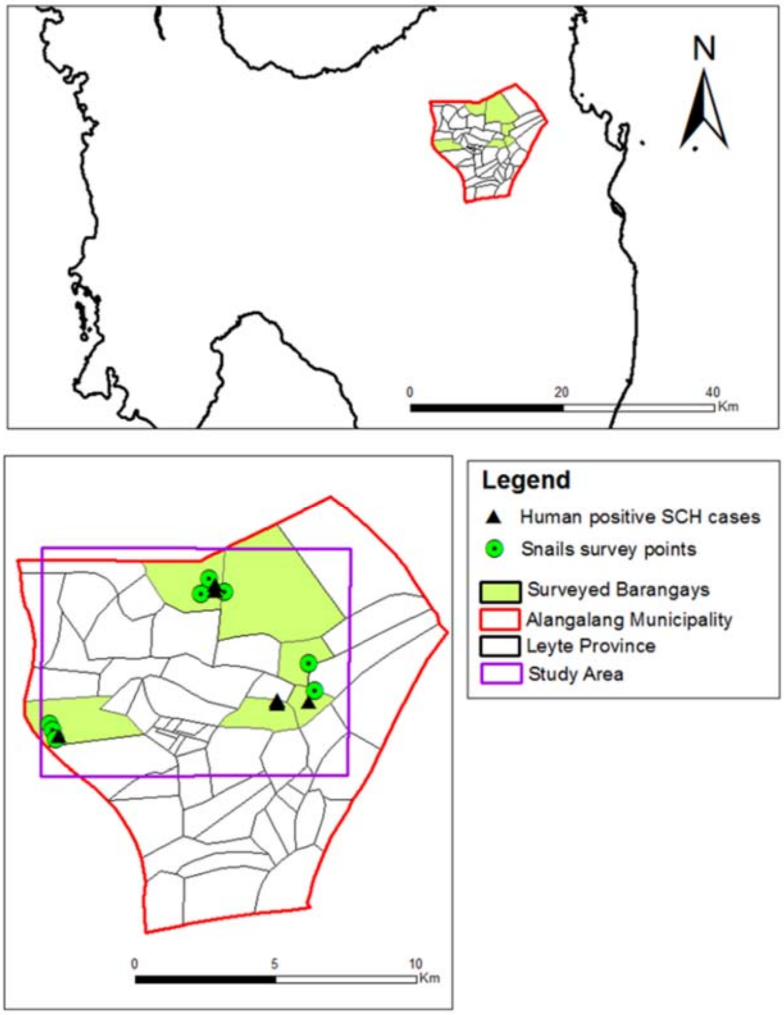


Figure 3.1: *Selected study area represented by the purple polygon. Rounded green and black triangles represent the snails and human positive cases, respectively.*

Secondly, communities were loaded as incidents (261 points), and contact river points as facilities (42 points). Thirdly, we used the closest facility tool to find the nearest river from an urban area following a road. Finally, we interpolated the distance to the nearest water source using ordinary kriging from the *gstat* package in R (Pebesma and Graeler, 2017) and saved the map as a raster file.

3.2.4 Snail infection rate map

We constructed a trend surface that represents snail infection rate for the whole study area, thus using data of all the points to predict at unknown locations (i.e. global interpolation). It fits a mathematically defined surface through the data points (i.e. deterministic interpolation) to discover smoother (i.e. inexact interpolation) regional and local trends. It is similar to a three dimensional regression surface obtained with linear regression, where coordinates $s_i = (\mathrm{x}_i, \mathrm{y}_i)$ are used as predictors.

The interpolated value $z(s_i)$ for a first and second order polynomial is represented in equations 3.1 and 3.2, respectively. $z(s_i)$ represents infection rate values (number of positive cases/number of sampled snails) at location i.

$$z^*(s_i) = \alpha_0 + \alpha_1 \mathrm{x}_i + \alpha_2 \mathrm{y}_i \qquad (3.1)$$

$$z^*(s_i) = \alpha_0 + \alpha_1 \mathrm{x}_i + \alpha_2 \mathrm{y}_i + \alpha_3 \mathrm{x}_i^2 + \alpha_4 \mathrm{y}_i^2 + \alpha_5 \mathrm{x}_i \mathrm{y}_i \qquad (3.2)$$

Figure 3.2a, b shows the resulting surfaces for the first and second order polynomials, respectively. Figure 3.2a shows low risk probability values (Table 3.1), from -0.003 to 0.008. These values do not match the original surveyed values. Figure 3.2b shows low and medium risk probability values from -0.01 to 0.035. These values show a better fit to the original surveyed values showed in red.

To remove the occurring negative values, we fitted a multiple linear regression by applying a generalized linear regression model using equation 3.2. In this case $z(s_i)$ was the infection status for each location i, 1 indicates an infected case and 0 a non-infected case. The resulting prediction from Figure 3.3 shows only positive predicted values but very large standard errors (28.7 to 3e+13). Besides, none of the predictions approximate the original surveyed values. Finally, the second order trend surface (Figure 3.2b) map was used for the analysis since it better fitted the original surveyed values.

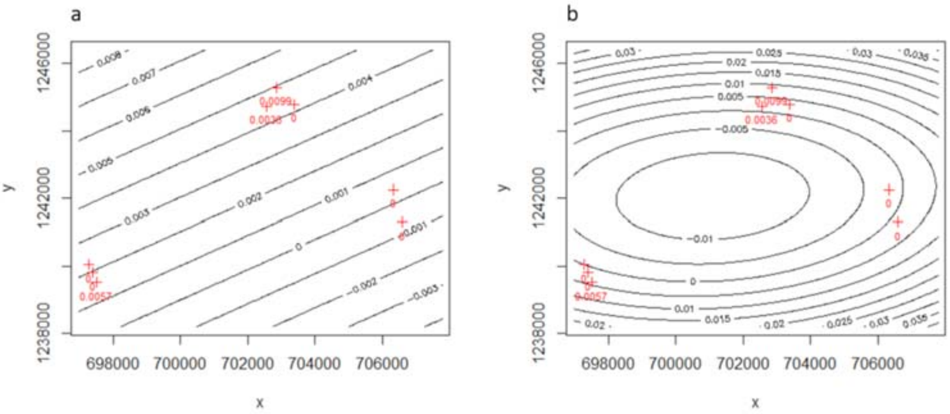


Figure 3.2: First order (a) and second order (b) polynomial trend surface. Red crosses represent the original surveyed snail infection locations

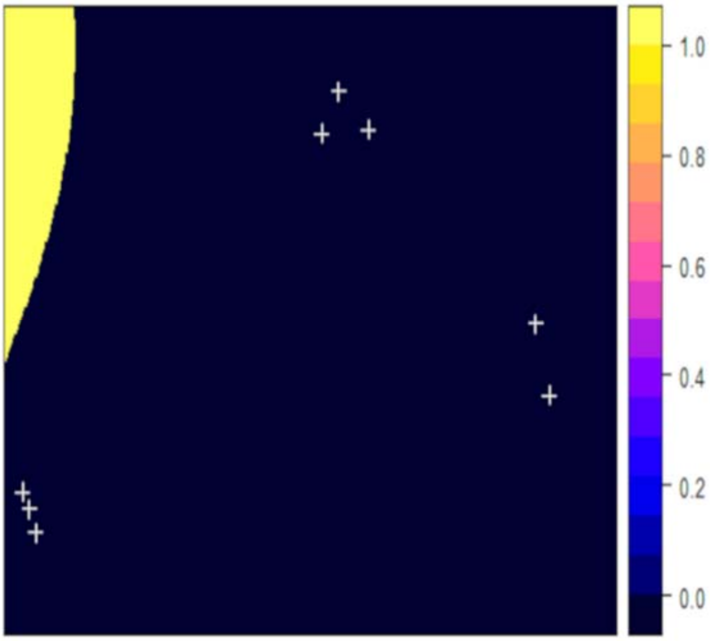


***Figure 3.3**: Predicted probability of snail infection values using generalized linear regression model. Colour scale represent probability values from 0 to 1. Snail survey locations are represented by white crosses:*

3.2.5 Spatial Bayesian network of Schistosoma japonicum exposure

We have conceptually designed a model that represents the positional mismatch between survey locations and exposure sites (Figure 3.4). Locations s_1 and s_2 represent the schools, households, or other survey locations from which morbidity indicators are extracted, while ex_{mn} represents the various exposure points where infections could have taken place, m is the corresponding number of exposure points and n is the corresponding survey locations related to the exposure.

Exposure areas were delineated by using spatial Bayesian networks (sBN) (Corporation, 1998). A Bayesian network (BN) is a probabilistic graphical model that captures the various conditional dependencies of a set of random variables (discrete or continuous) (Bishop, 2006; Bottcher and Dethlefsen, 2003), into a joint probability distribution by means of a directed acyclic graph (DAG) (Fenton and Neil, 2012; Nielsen and Jensen, 2009).

Table 3.1: *Categorization of exposure risk factors*

Risk factor (weight)	Spatial resolution	Temporal resolution	Data type	Coordinate system	Data source	Hypothetical link	Classification	π weights	Based upon
Elevation (0.03)	~ 30 m at equator	na	Raster	EPSG:4326	ASTER GDEM V2 from USGS	While elevation decreases, the risk of infection increases	High risk: < 900 m	0.70	(Pesigan et al., 1958; Zhu et al., 2015)
							Medium risk: 900–2300 m	0.25	
							Low risk: > 2300 m	0.05	
Land use (0.26)	~ 30 m	2-3-2017	Vector	EPSG:4326	OpenStreetMap project	Wet surfaces are more suitable to a higher risk of infection	Very high risk: wet soils	0.42	(Pesigan et al., 1958)
							High risk: water bodies	0.29	
							High and medium risk: Agriculture land and grass	0.16	
							Medium and low risk: forest and natural areas	0.08	
							Low risk: barren land	0.02	
							Very low risk: built land	0.03	
Slope (0.13)	~ 30 m at equator	na	Raster	EPSG:4326	Derived from elevation	At more flat surfaces the risk of infection increases	High risk: < 11 degrees	0.70	(Ajakaye, Adedeji and Ajayi, 2017; Zhu et al., 2015)
							Medium risk: 11–30 degrees	0.23	
							Low risk: > 30 degrees	0.07	
Distance to water bodies (0.50)	30 m	2-3-2017	Raster	EPSG:32651	Derived from roads, urban areas, river network and water bodies from the OpenStreetMap project	While distance to water bodies decreases, the risk of infection increases	High risk: < 1000 m	0.74	(Clements et al., 2006; Zhu et al., 2015)
							Medium risk: 1000–5000 m	0.21	
							Low risk: > 5000 m	0.05	
Snail infection rate (0.06)	na	2015–2016	Vector	EPSG:4326	Derived from recorded surveys	While snail infection rate increases, the risk of infection increases	High risk: > 3.6%	0.65	(Gao et al., 2014; Zhang et al., 2013)

Abbreviation: na, not applicable

A BN for a set of random variables R is defined by the pair (DAG, P). Here, P is the set of probability distributions for all variables in the network. Each variable r with parents $PA(r)$ has a conditional probability $p(r|PA(r))$. For a BN with a set of discrete (D) variables, the joint probability distribution factorizes into equation 3.3 (Bottcher and Dethlefsen, 2003). This is the joint probability distribution as the product of all conditional probabilities specified in a BN:

$$P(R) = \prod_{\mathrm{d}=1}^{\mathrm{D}} p\ (r_d | PA(r_d)) \tag{3.3}$$

The schistosomiasis exposure sBN defines exposure areas in a probabilistic way, by allowing the combination of various probability distributions from a set of random spatial variables (Nielsen and Jensen, 2009). We have constructed a DAG for exposure areas (Figure 3.5), where each random variable is represented as a node. Nodes are connected by directed links or edges that express probabilistic relationships between the variables (Bishop, 2006). Three types of random variables can be found including (i) an observable discrete variable [land use (LU)]; (ii) observable continuous variables [elevation (E), slope (SLP), distance to water bodies (DWB) and snail infection rates (SI)]; and (iii) latent discrete variables [potential accessible sites for snails (PAS), community cost (CC) and exposure (EX)]. The direction given in the link between variables, for instance from LU to PAS, encodes a direct causal dependence of PAS on LU; the node LU is known then as the parent of PAS (Fenton and Neil, 2012).

All continuous variables (E, SLP, DWB and SI) were discretized into different categories, given that high or low levels of exposure could occur at various ranges of risk factor values. We established hypothetical relationships between the risk factors and the disease, and categorized the risk factors based on literature (Table 3.1).

Exposure is a discrete child node, which has three discrete parent nodes: PAS, CC and SI; its conditional probability is expressed as $p(EX|PAS, CC, SI)$. PAS and CC are at the same time child nodes conditional on discrete parents. Their conditional probabilities are derived by $p(PAS|LU, E, SLP)$ and $p(CC|DWB)$, respectively. The joint probability distribution for our Bayesian network is given as:

$$P(R) = p(EX|PAS, CC, SI)\ .\ p(PAS|LU, E, SLP)\ .\ p(LU)\ .\ p(E)\ .\ p(SLP)\ .$$
$$p(CC|DWB)\ .\ p(DWB)\ .\ p(SI) \tag{3.4}$$

Equation 3.4 encodes assumptions of this research about direct dependencies between variables and indicates which node probability tables need to be defined (Fenton and Neil, 2012).

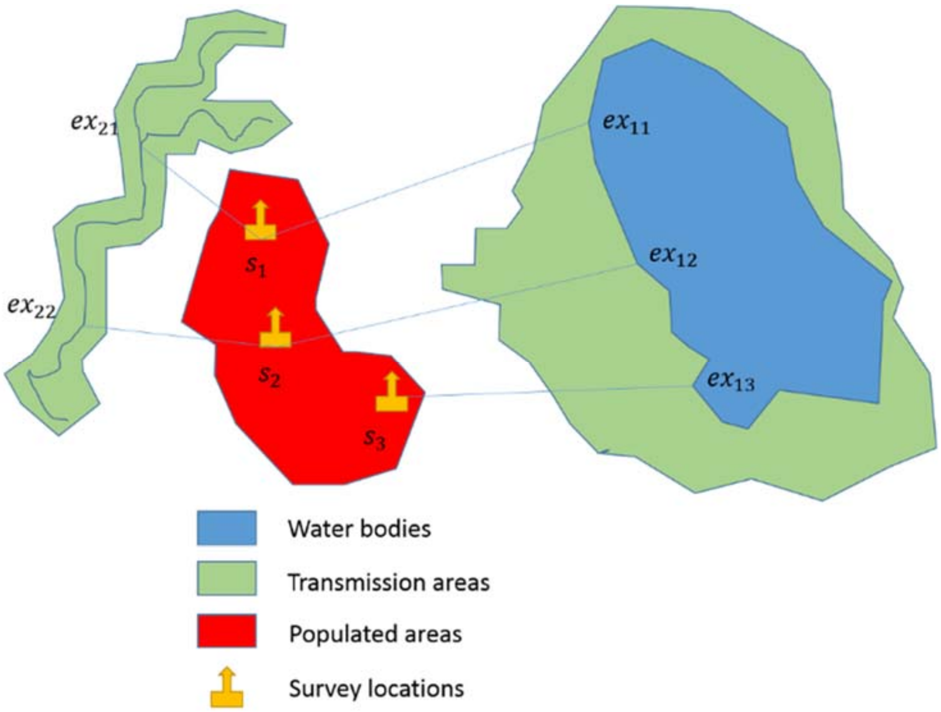


Figure 3.4: *Positional mismatch in SCH modelling. Blue lines represent the difference between survey and transmission locations*

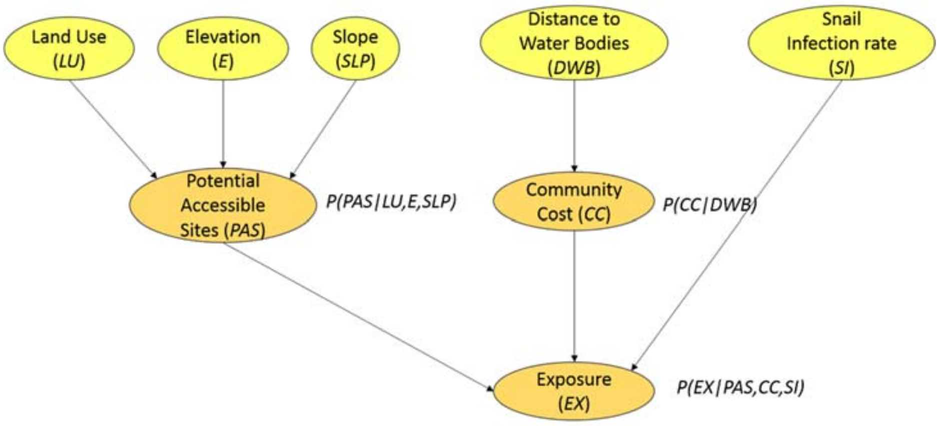


Figure 3.5: *Spatial Bayesian network for SCH exposure. Yellow and orange nodes are observable and latent risk factors, respectively.*

3.2.6 Construction of node probability tables

After defining the structure of our sBN, a main challenge is to construct the node probability tables (NPT). NPT are probability tables associated to each child node r given every possible state of the set of parents of r. NPT are intended to capture the strength of the relationship between the node and its parents (Fenton and Neil, 2012). The practicality of doing this depends on the number of states of the parent and child nodes. In our sBN eight NPTs were constructed, five NPTs as prior marginal probabilities (π) were inserted for the set of parent nodes (LU, E, SLP, DWB and SI) and three NPTs as conditional probabilities linking parent and child nodes (PAS, CC and EX).

We inserted prior marginal probabilities for the set of discrete parent nodes as weights. Weights were calculated using the eigen vector derived from a pairwise comparison matrix using Saaty's comparison table (Saaty, 2008). Saaty (Saaty, 2008) uses a scale of numbers (i.e. scale of judgement) to indicate how many times a factor is more dominant than another with respect to a criterion used for their comparison. In this case, the criterion is the risk of infection assigned to each parent node category given by literature (Table 3.1). Consistency indexes and ratios were calculated in order to measure the consistency of the judgements. Consistency ratios lower than 10%, indicate that our judgements are acceptable, while consistency ratios higher than 10% indicate untrustworthy judgements or random decisions. Saaty's pairwise matrices as well as consistency indexes and ratios are included in the Appendices in Tables 3A.1-A.7. Prior marginal probabilities for the parent nodes are shown in Table 3.1.

Latent variables PAS, CC and EX were divided into three probability categories: high, medium and low risk. Conditional probabilities for these child nodes are associated with the edges that link them to the parent nodes, and were also assigned using a pair-wise comparison matrix. The criterion used to assign the scale of judgement is the strength of the hypothetical link between the risk factors and exposure. The strength of the hypothetical link was evaluated based upon three studies that evaluated the risk factors associated with schistosomiasis infection (Ajakaye, Adedeji and Ajayi, 2017; Hu et al., 2017; Zhang et al., 2009).

Hu et al. (Hu et al., 2017) ranked the potential importance of the schistosomiasis risk factors by means of a power detector. According to this detector, distance to water bodies is the most significant factor for disease risk, and elevation the least significant. Zhang et al. (Zhang et al., 2009) used environmental, topographical and human behavioural factors to locate schistosomiasis active transmission sites. Their predictor capacity was compared by means of deviance analysis, used to determine the important variables to be included in a generalized additive model. As in the previous study, distance to water bodies was the most significant factor because of the smallest deviance, and elevation the least significant. Finally, Ajakaye et al. (Ajakaye, Adedeji and Ajayi, 2017) evaluated physical and environmental risk factors to identify areas with suitable conditions for schistosomiasis transmission. They used Saaty's comparison matrix to assign weights to each risk factor. Distance to water bodies and land use were the most significant factors, followed by elevation and slope as the least significant.

Weights obtained for each risk factor are shown in Table 3.1 and the conditional probabilities linking parent and child nodes are shown in the Appendices in Tables 3A.8-A.10.

3.2.7 Deriving join probabilities

To compute the probabilities for each category of the child nodes, PAS, CC and EX, conditional and marginal probabilities were used by applying equations 3.5, 3.6 and 3.7, respectively. Joint probability values of exposure were calculated for each polygon of analysis. In order to update the prior marginal probabilities, evidence is inserted for each spatial polygon into the observable variables ($\boldsymbol{SI}, \boldsymbol{LU}, \boldsymbol{E}, \boldsymbol{SLP}, \boldsymbol{DWB}$). Bold facing indicates the insertion of evidence. Variables notation can be found in the Appendices in Table 3A.11.

$$p(PAS, \boldsymbol{LU}, \boldsymbol{E}, \boldsymbol{SLP}) = \sum_{LU,E,SLP} p(\boldsymbol{LU}) * p(\boldsymbol{E}) * p(\boldsymbol{SLP}) * p(PAS|\boldsymbol{LU}, \boldsymbol{E}, \boldsymbol{SLP}) \quad (3.5)$$

$$p(CC, \boldsymbol{DWB}) = \sum_{DWB} p(\boldsymbol{DWB}) * p(CC|\boldsymbol{DWB}) \quad (3.6)$$

$$p(EX, PAS, CC, \boldsymbol{SI}) = \left(\sum_{PAS,CC,SI} p(EX|PAS, CC, \boldsymbol{SI}) * p(SI)\left(\sum_{LU,E,SLP} p(\boldsymbol{LU}) * p(\boldsymbol{E}) * p(\boldsymbol{SLP}) * p(PAS|\boldsymbol{LU}, \boldsymbol{E}, \boldsymbol{SLP})\right)\left(\sum_{DWB} p(\boldsymbol{DWB}) * p(CC|\boldsymbol{DWB})\right)\right) \quad (3.7)$$

For the implementation, polygons of analysis were constructed based on the overlaying of each risk factor (i.e. parent node). To overlay all risk factors, they were first transformed into vectors and then corrected for topology errors. Topology errors included duplicated polygons, multipart geometries and overlapping polygons.

Sensitivity analysis was used to see the relative influence of the risk factors on PAS and CC, and the relative influence of PAS, CC and SI on exposure. We used the sensitivity function, calculated as the degree of entropy reduction. Degree of entropy reduction ∇ is the degree of change or expected difference in information bits H between a query variable Q (exposure) with q states and findings variable F (risk factors) with *f* states (Marcot et al., 2006) (equation

3.8). A degree of entropy reduction of 0 means a query variable is independent of the varying variable.

$$\nabla = H(Q) - H(F) = \sum_q \sum_f \frac{P(q,f)\log_2[P(q,f)]}{P(q)P(f)} \tag{3.8}$$

3.2.8 Software

To work within the spatial domain we used the software Netica^{TM} 6.03 (Corporation, 1998), which works with Bayesian networks, decision nets and influence diagrams. Evidence is inserted as cases for each polygon of analysis, and prior and conditional probabilities are inserted as tables.

3.2.9 Validation

Validation was first performed by counting all surveyed positive SCH human cases falling inside the various categories of exposure in the map. However, this introduces a positional mismatch as the surveyed positive cases were not necessarily acquired at those specific exposure points.

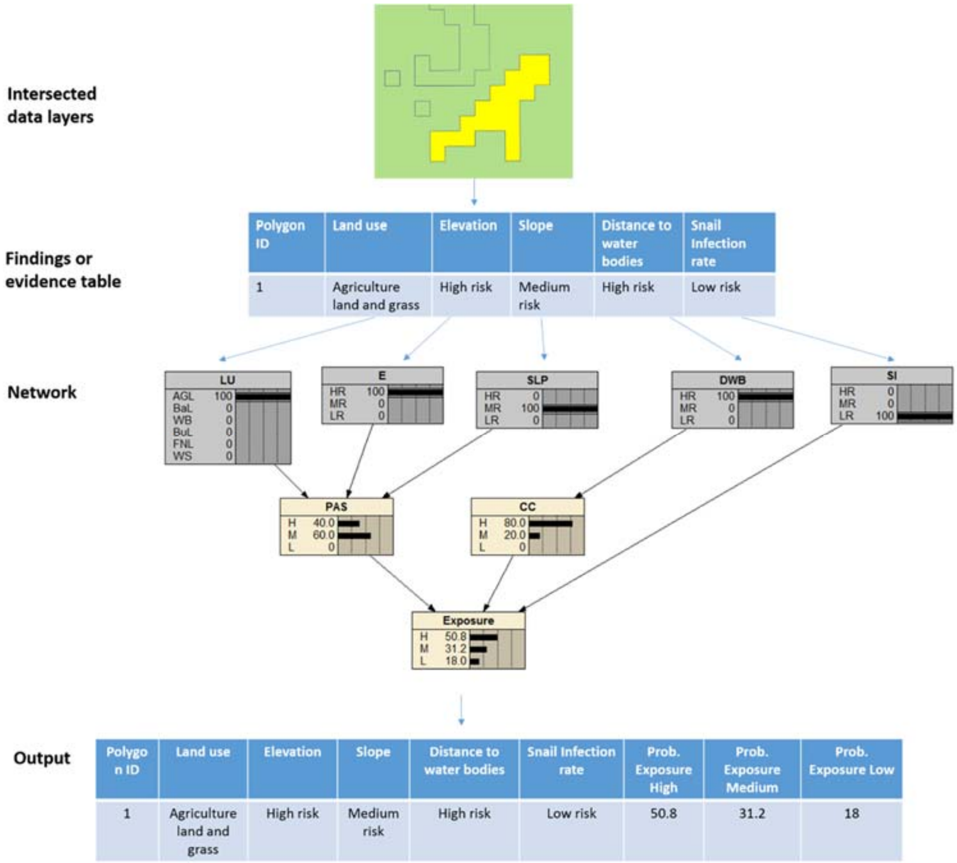
As a second approach for validation, we defined potential validation areas by constructing buffers around each of the positive cases. We extracted the distance to the nearest water body for each surveyed point using the distance map previously generated. Extracted distance values were used as distance buffers generated around positive cases. Buffers completely containing other buffers were grouped. We counted the number of positive cases falling inside each group and calculated the mean probability of exposure within the grouped buffers.

3.3 Resulting spatial Bayesian network

3.3.1 Exposure network

High (> 50%), medium (35–50%), low (20–35%) and very low (< 20%) probabilities of exposure were derived from the proposed exposure network. This is exemplified in Figure 3.6 for only one polygon. For this particular polygon, the probability is predominantly high (50.8%) for a high-risk elevation (< 900 m), *DWB* (< 1 km), and *LU* (agriculture land and grass), a medium risk slope (11–30°), and a low risk *SI* (< 0.5%).

Very low probability values of exposure (< 20%) were found in built-up areas, medium risk *DWB* (1–5 km), slopes < 30° and low and medium (0.5–3.6%) risk of snail infection, but also in agriculture and grass land with *DWB* > 5 km and slopes > 30° (Figure 3.7).



Polygon ID	Land use	Elevation	Slope	Distance to water bodies	Snail Infection rate
1	Agriculture land and grass	High risk	Medium risk	High risk	Low risk

Polygon ID	Land use	Elevation	Slope	Distance to water bodies	Snail Infection rate	Prob. Exposure High	Prob. Exposure Medium	Prob. Exposure Low
1	Agriculture land and grass	High risk	Medium risk	High risk	Low risk	50.8	31.2	18

Figure 3.6: *Probabilities of exposure in the Bayesian Network*

Low probabilities of exposure (20–35%) were found in built-up areas with slopes < 30°, low risk of snail infection, and within *DWB* < 1 km, but also in agriculture and grass land in *DWB* > 5 km. Medium probability values (35–50%) were found in agriculture and grass land and forest areas, in slopes > 11°, low risk of snail infection, and *DWB* < 1 km, but also in slopes < 30°, medium risk of snail infection and *DWB* from 1 to 5 km. High probability of exposure values (> 50%) were found in wet soils with slopes < 30°, with *DWB* from 1 to 5 km and medium risk of snail infection, but also in agriculture and grass land with *DWB* < 1 km and low risk of snail infection.

Based on the degree of entropy reduction, our sensitivity results show that the risk factor with the highest degree of change is *PAS* followed by *SI* and *CC.* Within PAS, land use has the highest degree of change and elevation has the lowest, showing that the most influential risk factors on exposure are land use, snail infection rate and distance to water bodies in that order, and the least influential factors are slope and elevation (Table 3.2).

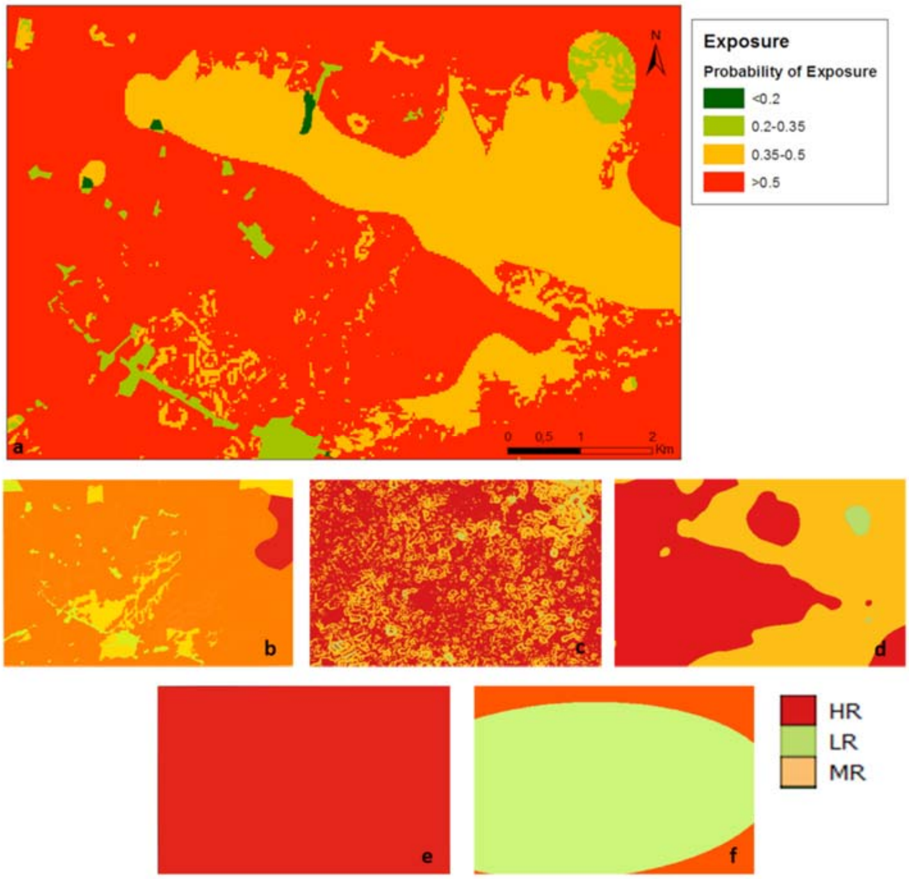


Figure 3.7: (a) *Probability of exposure map.* **(b-f)** *Risk factors of exposure: land use* **(b)***; slope* **(c)***; distance to water bodies* **(d)***; elevation* **(e)***; snail infection rates* **(f)**

Our findings show that approximately 63% of the study area has high probability of exposure values (> 50%). This is mainly explained by the predominance of agricultural fields in the area (Figure 3.7b) and the distance to water bodies results, which indicate that approximately 80% of the urban areas can access water bodies following routes < 500 m. Lowest and highest distance values between urban areas and water bodies are 7.6 m and 5.7 km, respectively, with a mean of 1.4 km (Figure 3.8).

Table 3.2: *Sensitivity of exposure to risk factors using entropy reduction (variables are listened in order of influence on exposure)*

Node	Degree of entropy reduction	% of influence to the network
PAS	0.07149	28.0
SI	0.06524	25.3
CC	0.04708	18.3
LU	0.04138	16.0
DWB	0.02868	11.1
SLP	0.00291	1.1
E	0.00066	0.2

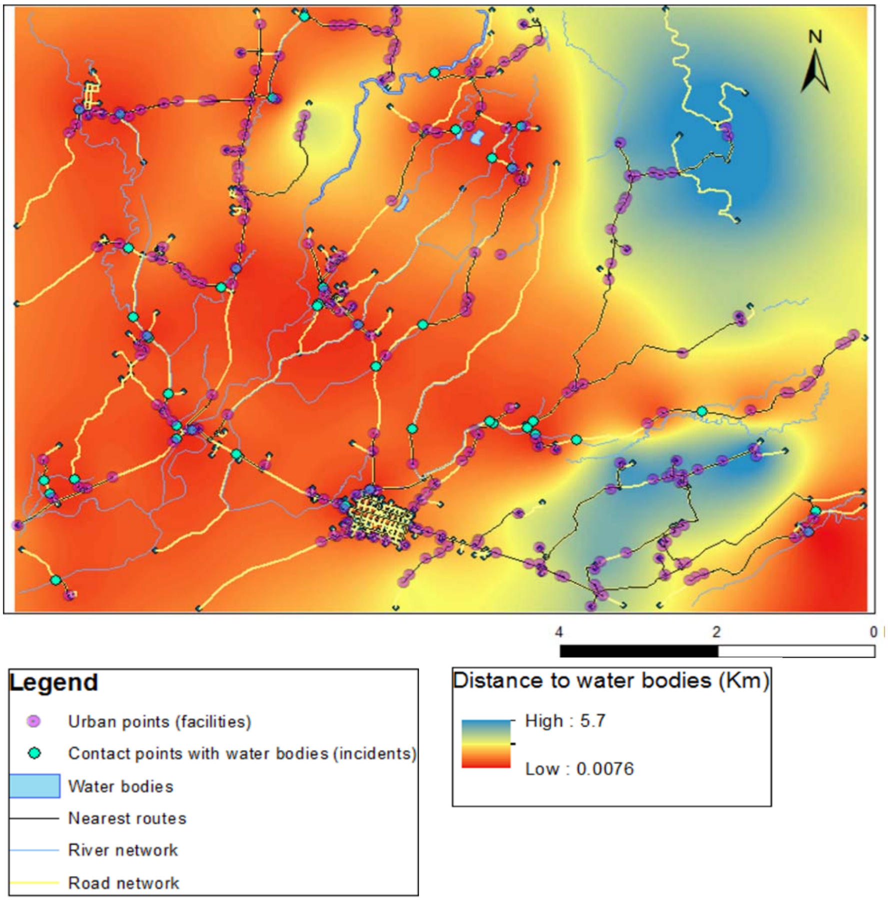

N
4
2
0
Legend
Urban points (facilities)
Contact points with water bodies (incidents)
Water bodies
Nearest routes
River network
Road network
Distance to water bodies (Km)
High : 5.7
Low : 0.0076

Figure 3.8: *Nearest route calculation from urban points to water bodies and DWB ordinary kriging interpolation.*

3.3.2 Validation

For the first validation, the results show an increase in the probability of exposure as the proportion of human cases also increases, except for 17% of human cases where a reduction in the probability of exposure of 35.8% can be observed (Table 3.3). For the second validation, four groups of buffers were observed: Group A with one positive case, Group B with two positive cases and Groups C and D with four and five positive cases, respectively (Figure 3.9).

A low correlation was found between probability of exposure and percentage of human cases within the groups (linear correlation, $R^2 = 0.3$). For the first three groups (A, B and C) the probability of exposure increases while the percentage of human cases also increases. For Group D, the group with more positive cases, a minor decrease in the probability of exposure can be observed (Figure 3.10). This could be explained by the distance to water bodies that has a slightly positive correlation with the probability of exposure values (0.47–0.55) calculated from our sBN for groups C ($R^2 = 0.98$) and D ($\mathrm{R}^2 = 0.96$) (Figure 3.11).

For instance, for Groups A and B with one and two positive cases respectively, the distance to water bodies is higher for Group A (~980 m) than for Group B (~177 m), with an average exposure value of approximately 0.47 and 0.48, respectively (Figure 3.10). Likewise, for Groups C and D, the distance to water bodies is higher for Group D (~1100 m) than for Group C (~490 m), with an average probability of exposure values equal to 0.55 and 0.49, respectively (Figure 3.10).

3.4 Discussion

Several studies have modelled snail distribution as input information for risk prediction of schistosomiasis (Hu et al., 2017; Walz et al., 2015b; Zhu et al., 2015), in order to guide prevention (sanitary and hygiene conditions of the population) and control (mass drug administration campaigns in the community) strategies for schistosomiasis infection. These approaches are inadequate spatial decision support tools since they have not accounted for snails' infection status or people's exposure to infection (i.e. contact of people with snails' sites). In this study we demonstrate a novel approach to delineate spatial areas of exposure to *S. japonicum* infection by accounting for the distribution of infected and non-infected snails, and considering the human interaction with active transmission sites. This was done by accounting for the cost of the community to access water bodies and potential sites where snails may be present.

Table 3.3*: Percentage of human cases falling within probabilities of high exposure values*

No. of human cases	% of human cases	Probabilities of exposure
1	8.3	41.2
2	16.7	35.8
3	25.0	50.8
6	50.0	55.6

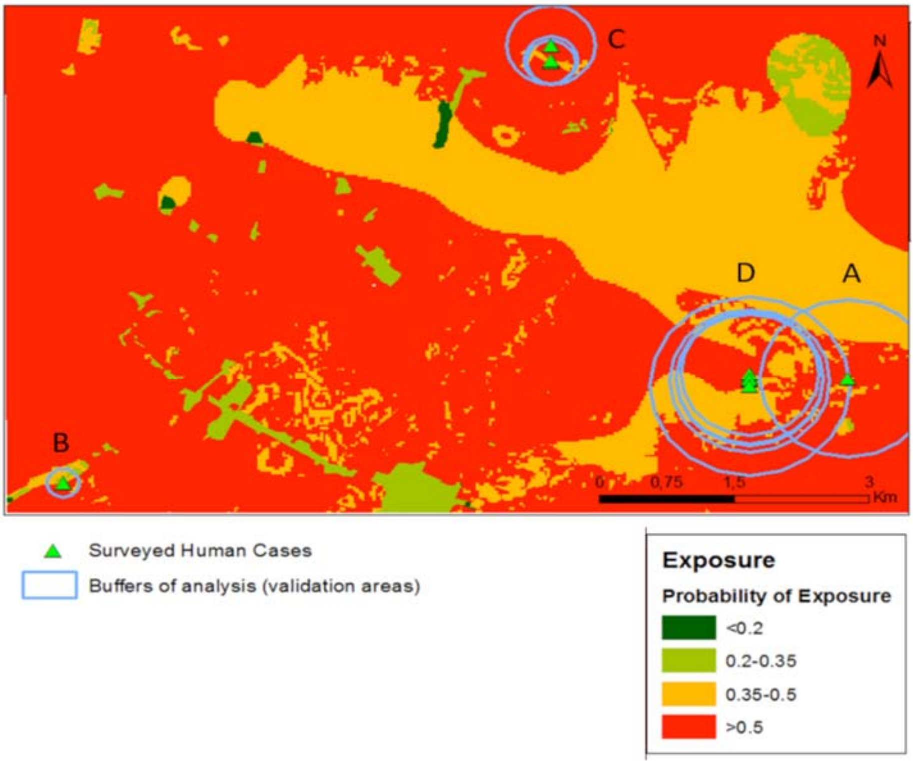


Figure 3.9: *Buffers around surveyed human cases points. Letters show the grouped buffers based on points location*

The results suggest that the predominance of high probabilities of exposure values (> 50%) in the study area are explained by the presence of wet soils and agriculture land in the zone, but also by the distance from urban areas to nearby water bodies (< 5 km). This was expected given that land use is a highly influencing risk factor on exposure after potential accessible sites (Table 3.2), and also because of the initial high weights given to *LU* and *DWB* (Table 3.1).

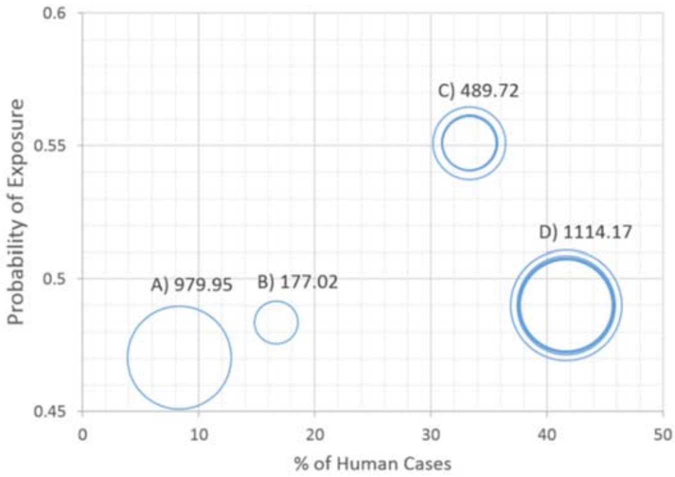


Figure 3.10: *Probability of exposure vs percentage of human cases. Labels correspond to the grouped buffers visualized in Figure 9. The values of the buffers are shown in meters.*

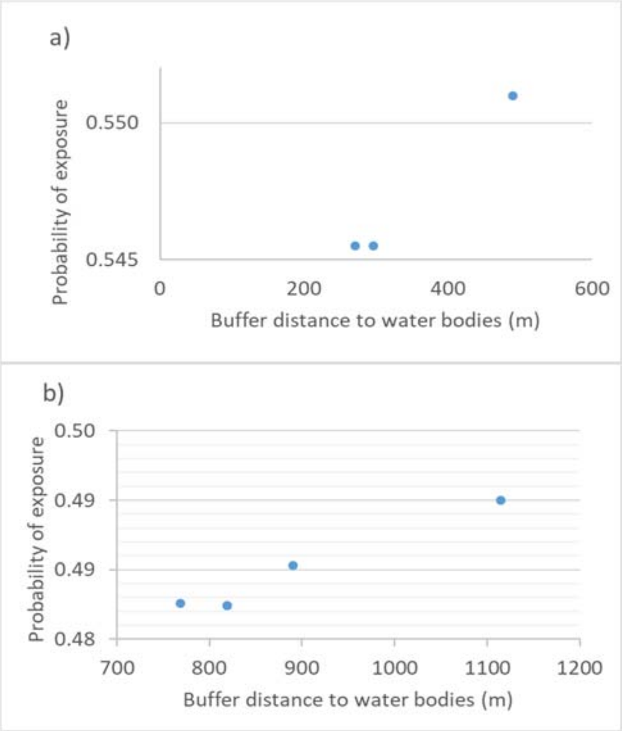


Figure 3.11: *Distance to water bodies versus probability of exposure. Plotted values for **(a)** Group C and **(b)** Group D*

The results demonstrate that for short distances to water bodies, the probability of a community to be exposed to *S. japonicum* is high (Figure 3.8). This was explained by the probability of exposure map and the relative influence of *DWB* on exposure. Although *DWB* is the fifth influencing factor on exposure (Table 3.2), it is the only influencing factor on community cost, which is the third most important variable of the network (Table 3.2). Based on our results we propose that future studies utilise the nearest distance to water bodies following a road instead of the commonly used Euclidean distance (Clements et al., 2006; Zhu et al., 2015), since the former provides a more accurate representation of community access to water bodies, as it accounts for the nearest path from human dwellings to potential infection foci.

We postulated that the proportion of human *S. japonicum* cases was higher in areas predicted to have a higher probability of exposure. Our validation procedure using overlaying proportions in the four groups of buffers surrounding nearby *S. japonicum* cases, demonstrated a positive correlation for three groups. Although the number of validation points is somewhat low for a total validation, overlying proportions of exposure to schistosomiasis infection suggest a correlation between potential areas of exposure and the disease in the presence of limited survey data.

3.4.1 Utility of modelling the geographical probability of S. japonicum exposure

Modelled schistosomiasis exposure areas account for the transmission processes occurring between the environment containing infective stages of *S. japonicum* or intermediary hosts (snails), and the susceptible hosts (humans and livestock). From a public health perspective, the provision of maps that define the geographical limits of probability of exposure to *S. japonicum* infected areas could help target local schistosomiasis control strategies to communities more likely to contact contaminated environments and thereby improve the efficiency of mass drug administration campaigns. From a spatial modelling perspective, the availability of a predictive exposure map could serve as an important base map to obtain covariate values. By relating them to indicators of disease, we could possibly account for the positional mismatch between epidemiological survey data and environmental covariates, and improve the statistical modelling of *S. japonicum* infection.

3.4.2 Limitations of the study

A number of limitations should be accounted for in the interpretation of our results. Firstly, estimates of the probability of exposure are highly influenced by the availability of snail infection estimates (Table 3.2). Due to the localized nature of the study, it was difficult to generate an adequate surface map that could properly explain snail infection distribution, constraining this map into a binary output with low and medium risk values (Figure 3.7f). This might have an impact on the results and could be further improved by an increase of the study extent, and the number of survey points. In addition, whenever these data or new knowledge becomes available, the sBN developed in this study will enable a "rapid delineation" of potential exposure areas of *S. japonicum* by facilitating a flexible integration of exposure data as risk factors, and prior information derived from literature or expert knowledge (Smith et al., 2007).

Secondly, model validation procedures could be improved by including positive and negative human cases. Collecting data on livestock infection (Gao et al., 2014; Zhang et al., 2013) could also serve for validation as livestock infection, particularly carabao, has been suggested to play an important role in the transmission of *S. japonicum* in the Philippines (Gordon et al., 2012).

3.5 Conclusions

In conclusion, the present study describes the nature of the positional exposure mismatch in the modelling of *S. japonicum* infection. Results of the present study suggest that the best way to address this mismatch should include the extraction of covariate values from potential exposure areas. A probabilistic method to delineate exposure areas in the absence of sufficient empirical survey data is proposed. Unlike other studies, the present sBN is adequate to delineate exposure areas based upon the contact of communities to water bodies and other potential sites of infection. We conclude that even with limited disease survey data, it is possible to define potential exposure areas for schistosomiasis. Modelled exposure areas might be used to correct for positional mismatches and significantly improve disease predictions to better guide control programs to prevent and control schistosomiasis and other water-borne infections.

Chapter 4. Modelling *Schistosoma japonicum* Infection under Pure Specification Bias: Impact of Environmental Drivers of Infection [3]

[3] This chapter is published as:

Araujo Navas, A. L.; Osei, F.; Leonardo, L. R.; Soares Magalhães, R. J.; Stein, A. Modeling *Schistosoma japonicum* Infection under Pure Specification Bias: Impact of Environmental Drivers of Infection. *International Journal of Environmental Research and Public Health* 2019, *16*, 176, doi: 10.3390/ijerph16020176

Abstract

Uncertainties in spatial modelling studies of schistosomiasis (SCH) are relevant for the reliable identification of at-risk populations. Ecological fallacy occurs when ecological or group-level analyses, such as spatial aggregations at a specific administrative level, are carried out for an individual-level inference. This could lead to the unreliable identification of at-risk populations, and consequently to fallacies in the drugs' allocation strategies and their cost-effectiveness. A specific form of ecological fallacy is pure specification bias. The present research aims to quantify its effect on the parameter estimates of various environmental covariates used as drivers for SCH infection. This is done by (i) using a spatial convolution model that removes pure specification bias, (ii) estimating group and individual-level covariate regression parameters, and (iii) quantifying the difference between the parameter estimates and the predicted disease outcomes from the convolution and ecological models. We modeled the prevalence of *Schistosoma japonicum* using group-level health outcome data, and city-level environmental data as a proxy for individual-level exposure. We included environmental data such as water and vegetation indexes, distance to water bodies, day and night land surface temperature, and elevation. We estimated and compared the convolution and ecological model parameter estimates using Bayesian statistics. Covariate parameter estimates from the convolution and ecological models differed between 0.03 for the nearest distance to water bodies (NDWB), and 0.28 for the normalized difference water index (NDWI). The convolution model presented lower uncertainties in most of the parameter estimates, except for NDWB. High differences in uncertainty were found in night land surface temperature (0.23) and elevation (0.13). No significant differences were found between the predicted values and their uncertainties from both models. The proposed convolution model is able to correct for a pure specification bias by presenting less uncertain parameter estimates. It shows a good predictive performance for the mean prevalence values and for a positive number of infected people. Further research is needed to better understand the spatial extent and support of analysis to reliably explore the role of environmental variables.

4.1 Introduction

Schistosomiasis (SCH) is a water-borne infection caused by parasitic worms known as schistosomes. People get infected by skin penetration of the infective stage of the parasite. Three schistosomes species cause the infection: *Schistosoma mansoni*, *Schistosoma japonicum*, and *Schistosoma haematobium*. Among these, *S. japonicum* is the hardest one to control due to its zoonotic life cycle (Jia et al., 2007), which includes the infection of an amphibious snail from the species *Oncomelania hupensis quadrasi* as the intermediate host, and humans and other mammals as definitive hosts (Tarafder et al., 2006; Yang et al., 2008). Schistosomiasis is a disease of public health significance (King, Dickman and Tisch, 2005; Walz et al., 2015b) since it affects more than 252 million people worldwide (Hotez et al., 2014). This especially concerns communities in tropical and subtropical areas, where access to clean water and sanitation is limited. Schistosomiasis leads to malnutrition, which causes anemia and stunted growth in school-aged children (Coutinho et al., 2005; Leenstra et al., 2006).

Schistosomiasis risk mapping has enabled the identification of at-risk populations to target mass drug administration campaigns for disease control (Soares Magalhães et al., 2014). Mapping SCH involves the use of geographic information systems, remote sensing, and global positioning systems (GPS). These help to allocate data about infection and the physical and biological environmental variables in space. Environmental variables together with various statistical methods have been combined to model the distribution of the disease (Herbreteau et al., 2007) and to quantify the role of the environmental exposure factors on SCH risk (Zhang et al., 2016).

Spatial epidemiological studies are susceptible to uncertainties, which are inherent in spatial information (Araujo Navas et al., 2016; Zhang et al., 2016). Most of these uncertainties are caused by positional measurement errors due to GPS inaccuracies, multiple addresses, geocoding errors, misalignments between covariates and disease outcomes, and disease outcome or covariate aggregations (Zhang et al., 2016). Particularly, disease outcome and covariate aggregations at a specific administrative level could be incorrectly obtained for individual-level inference (Richardson and Monfort, 2000). Health data are often available at a specific administrative level (i.e., group-level) while environmental data consists of a set of recorded values aggregated at monitor sites or gridded data derived from remote sensing images (Wakefield and Shaddick, 2006). Spatial aggregation occurs due to the lack of geolocated information at the individual level, caused by the scarcity of sampling resources, availability of associated data or the need to protect confidentiality (Zhang et al., 2016).

Disease outcome and covariate aggregations cause an ecological bias or ecological fallacy. This represents an important source of uncertainty (Araujo Navas et al., 2016; King, 2013; Zhang et al., 2016) because any direct link between exposure and health outcomes is imperfectly measured. For instance, Wakefield and Lyons (Wakefield and Lyons, 2010) mention that the fundamental problem of this kind of spatial aggregations is the loss of information. Thus, the function used in the regression modelling does not represent the real relationship between the affected population and their exposure. This type of ecological fallacy is also called pure specification bias and arises due to the loss of information when a non-linear model changes its form under aggregation (Gelfand et al., 2010; Wakefield and Lyons, 2010). It is called 'pure' because it specifically addresses a model specification bias (Gelfand et al., 2010).

Several efforts have been made to address pure specification bias resulting in disaggregation methods. For instance, Prentice and Sheppard (Prentice and Sheppard, 1995) suggest an 'aggregated data' method to create models based on exposure information available for a subset of individuals. Richardson et al. (Richardson, Stucker and Hemon, 1987) assume parametric distributions for within-area exposures to derive accurate risk functions. Wakefield and Shaddick (Wakefield and Shaddick, 2006) propose a convolution model and derive an appropriate likelihood function for a scenario where health outcome data are aggregated at the district level, and exposure information is known at monitoring sites. Wang et al. (Wang et al., 2017) address pure specification bias in the least informative data scenario for aggregated disease counts with associated counts of the population at-risk, and a separate set of point level exposures from monitoring stations. They propose a conceptual probability of the incidence surface over the entire study region as a function of an exposure surface. This probability surface was then used to simulate individual disease outcomes and to obtain individual-level parameter estimates.

For a tropical disease such as SCH, the availability of individual-level infection data obtained from schools or health care centers is common (Chammartin et al., 2014; Hu et al., 2015; Sturrock et al., 2013). Nevertheless, there are several SCH epidemiological studies (Scholte et al., 2014; Soares Magalhães et al., 2014; Soares Magalhães et al., 2015) that limit their modelling to the outcome and covariate data aggregated at a specific administrative level (i.e., ward, municipal, province, county, district, barangays, among others) without taking pure specification bias into consideration. Moreover, there are no up-to-date studies that have quantified the effect of the covariate information on the parameter estimates when accounting for pure specification bias in SCH modelling.

The objective of this study is to quantify the effect of pure specification bias on the parameter estimates of various environmental covariates used as drivers for SCH infection. To achieve this objective we aim to: (i) use a spatial convolution model that removes pure specification bias by using group-level health outcome data and individual-level environmental covariate data, (ii) estimate group and individual-level covariate regression parameters, and (iii) quantify the difference between the parameter estimates and predicted disease outcomes from the convolution and ecological models.

4.2 Methods to reduce pure specification bias

4.2.1 Data on Human Schistosoma japonicum Infection, Study Area, and Sampling Design

We use *S. japonicum* infection data collected as part of the 2008 Nationwide Schistosomiasis Survey in The Philippines (Leonardo et al., 2012). In this case, *S. japonicum* is endemic in 28 of its 81 provinces (Leonardo et al., 2015), with approximately 1.8 million estimated infected people (Leonardo et al., 2002). The disease affects children, adolescents, and individuals with high-risk occupations, such as farmers and fishermen (Leonardo et al., 2002; Zhou et al., 2010).

A two-stage systematic cluster sampling was used in the sampling design. Stratification was done by a prevalence level (high, medium, and low), obtained from the 1994 World Bank-assisted Philippine Health Development Program. Provinces and barangays were the primary and secondary sampling units respectively. A barangay is the smallest administrative division in the Philippines, numbering about 22–50 in a single municipality. Provinces with high and moderate prevalence rates were included, while the random selection was done among the low-prevalence provinces and non-endemic provinces. Within the selected provinces, high prevalence barangays were also selected. We decided to work in the Mindanao region due to the good spatial coverage of the sampling over the whole region (Figure 4.1), and the high response rate of 70 percent (Leonardo et al., 2012; Leonardo et al., 2008; Soares Magalhães et al., 2014). In total, 22 provinces were surveyed, and between 2 and 10 barangays were surveyed per province. In total, 108 out of 10,021 barangays were surveyed in Mindanao.

Data from 19,763 individuals were recorded in the survey but not georeferenced. Information regarding the corresponding barangay and province were recorded for each individual. For this reason, individual-level survey data were aggregated and geo-located to the centroids of 108 barangays in Mindanao. We used a probability of infection p in barangay k as our disease outcome variable.

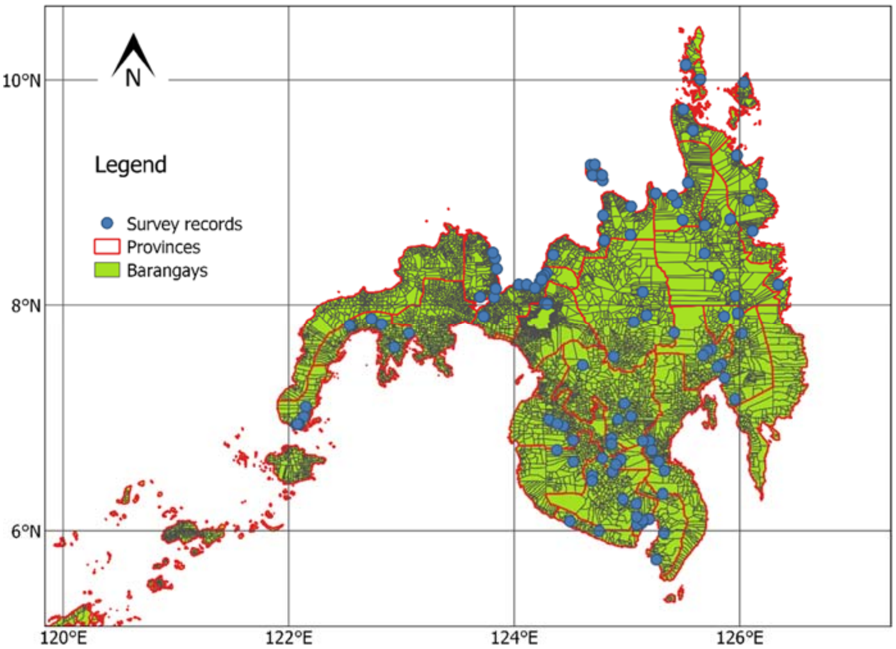


Figure 4.1: *The Mindanao region in the Philippines is the study area. Blue dots show the surveyed data aggregated at the barangay centroids.*

Kato-Katz thick smear examination (Santos, Cerqueira and Soares, 2005) was used to diagnose *S. japonicum* infection based on two-sample stool collection. Each sample was read using a microscope and the presence of *S. japonicum* eggs indicated active infection. Due to inconsistencies in the submission of the second stool sample, however, only the results of the first stool sample were available (Soares Magalhães et al., 2014) from people aged two years and above. Information such as gender and age were recorded for each individual. More details about the sampling design can be found in Leonardo et al. (Leonardo et al., 2012; Leonardo et al., 2008).

4.2.2. Environmental and Geographical Data

Six environmental variables were included in the analysis: the nearest distance to water bodies (NDWB), normalized difference vegetation index (NDVI), normalized difference water index (NDWI), day (LSTD) and night (LSTN) land surface temperature, and elevation (E). The nearest distance to water bodies was calculated using the closest facility network analysis tool in ArcGIS version 10 (Esri, 2011). We used rivers and lakes as water bodies. As input for the network, we used the river and road networks, and the location of cities and hamlets. We first calculated the NDWB for each city point and then interpolated the distance values for all the surveyed barangays. Interpolation was performed using Ordinary Kriging. The nearest distance to water bodies shows the accessibility of people to water bodies since they represent the main infection foci. We used NDVI obtained from the MODIS MOD13Q1 product.The normalized difference vegetation index served as an indicator of vegetation presence and greenness, particularly the presence of flooded agricultural land such as paddy fields, which is an important factor for Asian schistosomiasis (Zhou, Liang and Jiang, 2012). We included NDWI from the Google Earth Engine as an annual Landsat 7 composite for the year 2008. The normalized difference water index was used as a proxy indicator of flooding (Walz et al., 2015b; Xu, 2006). The day land surface temperature was also included and obtained from the MODIS MOD11A2_LST product. The day land surface temperature is determinant for the survival of larval stages of snails (Pietrock and Marcogliese, 2003; Prah and James, 1977; Woolhouse and Chandiwana, 1990) and is used as a proxy for water temperature given that the thermal condition of shallow waters usually reflects the ambient temperature of the air (Soares Magalhães et al., 2014). Elevation was also included and obtained from ASTER GDEM version 2 from United States Geological Survey (USGS) (Geological Survey, 2017). In the Philippines, the presence of snails is also driven by the local topography (Pesigan et al., 1958; Stensgaard et al., 2006; Stensgaard et al., 2013). At lower altitudes, the risk of finding snails increases. Table 4.1 summarizes the information about environmental information and their sources.

4.2.3 Convolution Model (Individual-Level Model)

An individual is considered infected if at least one parasite egg is found. *S. japonicum* infection data y are available at individual-level i recorded within a barangay k. Various n environmental variables $\boldsymbol{x}$ are available for each image pixel. Because the exact response locations of the y_{ik} are unknown, individual-level data are aggregated to their corresponding barangay centroid and are, thus, denoted by y_k. A naive group-level model is given by equation 4.1.

$$y_k|\bar{\boldsymbol{x}}_{\boldsymbol{k}}, \boldsymbol{\gamma} \sim Binomial(N_k, \hat{p}_k) \qquad (4.1)$$

where N_k and $\hat{p}_k$ are the number of sampled individuals and the probability of infection in barangay k, respectively, and $\bar{x}_k$ is the observed mean exposure within barangay k. The probability $\hat{p}_k$ is modelled based on equation 4.2 using environmental variables as predictors, where $\boldsymbol{\gamma}$ are the group-level covariate coefficients.

$$logit(\hat{p}_k) = \gamma_0 + \gamma_1 \cdot \bar{x}_{1_k} + \gamma_2 \cdot \bar{x}_{2_k} + \cdots + \gamma_n \cdot \bar{x}_{n_k} \qquad (4.2)$$

Table 4.1*: 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
NDWI	500 m	2008	Raster	EPSG:32651	Landsat 7, one-year composite
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: day land surface temperature; NDWB: nearest distance to water bodies; USGS: United States Geological Survey

We suppose that, for an individual i in area k, y_{ik} follows a Bernoulli distribution (Equation 4.3). Then an individual level model is presented in equation 4.4, where the parameters $\boldsymbol{\beta}$ are the individual-level coefficients. However, this model assumes that we know the individual level locations.

$$y_{ik}|\boldsymbol{x_{ik}},\boldsymbol{\beta}\sim Bernoulli(p_{ik}) \tag{4.3}$$

$$logit(p_{ik}) = \beta_0 + \beta_1 \cdot x_{1_{ik}} + \beta_2 \cdot x_{2_{ik}} + \cdots + \beta_n \cdot x_{n_{ik}} \tag{4.4}$$

Pure specification bias will result in $\boldsymbol{\gamma} \neq \boldsymbol{\beta}$, where the relationship between aggregated disease risk and exposure on areal units differs from the relationship between the individual disease risk and the associated exposure. In a non-linear model, as in our case, this difference is produced by a loss of information due to aggregation known as pure specification bias. Pure specification bias is reduced in size since the 'within area' exposure is more homogenous (Wakefield and Lyons, 2010). This could be obtained by having a finer partition of space in which exposure measurements are available (Wakefield and Lyons, 2010; Wakefield and Shaddick, 2006).

As we know the individual-level responses y_{ik} but not their locations, we minimized the pure specification bias by extracting covariate information from cities within the barangays. We selected cities since they were the finest units we found available by taking them as a proxy for exposure locations at an individual level. Cities were extracted from the 2010 build-up data base from the National Mapping and Resource Information Authority from The Philippines ("National Mapping and Resource Information Authority," 2018). Cities were not available for all the surveyed barangays. Therefore, we digitalized them using Google Earth images.

We used the aggregate data method proposed by Prentice and Sheppard, 2001 (Prentice and Sheppard, 1995). Let exposure or covariate data $\boldsymbol{x_{jk}}$ be measured at locations s_{jk} $j = 1, \dots, m_k \leq N_k$, for a subset of individuals. Then, we estimated the average risk of the individuals in area k, and the individual level coefficients $\boldsymbol{\beta}$. This was done by calculating the mean of the risk function (Equation 4.5) instead of evaluating the risk function at the mean exposure (Equation 4.2). In this way, the average $\hat{\hat{p}}_k$ of the function over the exposures corrects for the pure specification bias and differs from the function evaluated at the average exposure ($\hat{p}_k$). Thus, $\hat{\hat{p}}_k$ is the estimated average probability of infection of the individuals in area k.

$$\hat{\hat{p}}_k = \frac{1}{m_k} \cdot \sum_{j=1}^{m_k} \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \cdot \bar{x}_{1_{jk}} + \beta_2 \cdot \bar{x}_{2_{jk}} + \cdots + \beta_n \cdot \bar{x}_{n_{jk}}))} \tag{4.5}$$

In our study, covariate data were obtained at a finer level of analysis than the barangay. Therefore, we assumed that averaged covariate values at city level j represent individual covariate values (i.e., $x_{jk} = x_{ik}$). We then know x_{jk} but not the geographical linkage with individuals. One way to account for this is to allocate N_{jk} individuals to measurement x_{jk} by equally dividing the population. Therefore, $N_{jk} = N_k/m_k$, is a simple version of the convolution model.

The spatial convolution model is represented in equation 4.6, where the risk function also accounts for the spatial variability by including spatial (s_k) structured random effects.

$$\hat{\hat{p}}_k = \frac{1}{m_k} \cdot \sum_{j=1}^{m_k} \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \cdot \bar{x}_{1_{jk}} + \beta_2 \cdot \bar{x}_{2_{jk}} + \cdots + \beta_n \cdot \bar{x}_{n_{jk}} + s_k))} \tag{4.6}$$

The implemented convolution model that corrects for the ecological fallacy is of the form.

$$y_k|\overline{\boldsymbol{x}}_{\boldsymbol{jk}}, \boldsymbol{\beta} \sim Binomial(N_k, \hat{\hat{p}}_k) \tag{4.7}$$

The model includes an intercept (β_0), averaged city-level environmental variables ($\overline{\boldsymbol{x}}_{\boldsymbol{jk}} = NDVI, NDWI, LSTD, LSTN, E, NDWB$), and their corresponding individual-level coefficients $\boldsymbol{\beta}$, and a spatial random effect (s_k) as described in equation 4.6. All covariates were standardized to have mean = 0 and standard deviation = 1. Collinearity between covariates was assessed with the Pearson correlation coefficient.

Prior information for the intercept β_0 was given as a proper uniform distribution with wide bounds (−100, 100). The other $\boldsymbol{\beta}$ parameters were given non-informative normal distributions $\boldsymbol{\beta} \sim N[0, \frac{1}{\delta^2}]$, with δ uniformly distributed on a wide range $\delta \sim U[0, 100]$. These distributions are recommended in order to avoid overestimations on the parameters (Gelman, 2006). These parametrizations allow a good mixing of the sequences used for Markov Chain Monte Carlo simulations and contribute to their faster convergence (Gelman et al., 1995).

Spatial dependence was modelled using a spatially structured random effects distribution based upon a geo-statistical model. This model can be used as a sampling distribution for continuous spatial data (Diggle, Tawn and Moyeed, 2002). The vector of random variables $\boldsymbol{s}$ associated with space locations $(\mathrm{x}_k, \mathrm{y}_k)$, $k = 1, \dots, K$, was modelled with a multivariate normal distribution $s \sim MVN_K[\mu, \Sigma_{ab}]$ with mean μ = 0 and a covariance matrix Σ_{ab} defined by a powered exponential spatial correlation function from equation 4.8.

$$\Sigma_{ab} = \sigma^2 \cdot \exp[-(\phi \cdot d_{ab})^\kappa] \tag{4.8}$$

The covariance matrix is specified as a function of the distances d_{ab} between barangay centroids a and b, 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. Since it is often difficult to learn much about the κ parameter, and large values of κ could lead to smoothing, we used $\kappa = 1$. The prior distribution for ϕ was set uniform: $\phi \sim U[2\mathrm{E}10^{-7}, 3\mathrm{E}10^{-3}]$. 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), which assists identification (Thomas et al., 2004). The variance parameter σ^2 was given a half-normal distribution $\sigma^2 \sim HN[0,1]$. Half-normal was selected in order to restrict our prior to positive values and avoid problems with convergence (Gelman, 2006; Lunn et al., 2013).

The model was run using three sequences or chains with 50,000 iterations that ensured simulations representative of target distributions (Gelman et al., 1995) and a good stability for convergence (Gelman et al., 1995). A burn-in of 25,000 iterations was used, discarding the first half of each sequence that is used to diminish the influence of starting values (Gelman et al., 1995). We monitored convergence visually and statistically, first by inspecting at the trace plots and then by checking the $\hat{R}$ statistic (Brooks and Gelman, 1998; Gelman and Rubin, 1992), which are also called a potential scale reduction factor. This assesses sequences mixing by comparing the between and within variation. $\hat{R}$ values < 1.1 indicate evidence that sequences had converged (Brooks and Gelman, 1998), while high values suggest that an increase in the number of simulations may improve our inferences (Gelman et al., 1995). Data were structured in a rectangular format, where the columns are headed by the array name. The survey data and the codes in bugs for the convolution and ecological models are provided in the Appendices 4A.1 and 4A.2, respectively.

4.2.4. Ecological Model (Group-Level Model)

As a comparison, we estimated the group-level covariate regression parameters $\boldsymbol{\gamma}$ by using equations 4.1 and 4.2. We used $\hat{p}_k$ as a function of the same covariate information, $\bar{\boldsymbol{x}}_{\boldsymbol{k}} = NDVI, NDWI, LSTD, LSTN, E, NDWB$, but averaged for each surveyed barangay, and added the spatial random effects term s_k (Equation 4.9). Prior distributions for $\boldsymbol{\gamma}$ and s_k were the same as the ones given for the $\boldsymbol{\beta}$ parameters and the geo-statistical spatial term.

$$\hat{p}_k = \sum_{k=1}^{K} \frac{1}{1 + \exp(-(\gamma_0 + \gamma_1 \cdot \bar{x}_{1_k} + \gamma_2 \cdot \bar{x}_{2_k} + \cdots + \gamma_n \cdot \bar{x}_{n_k} + s_k))} \qquad (4.9)$$

We compared estimated covariate regression coefficients and their credible intervals from both models, and also the data generated from the posterior predictive distribution to the observed data. Posterior predictive distributions were generated using simulations. Residuals were calculated by subtracting the simulated values from the observed values. We created correlation and residual plots for convolution and ecological models for three different simulations. Lastly, we compared the predicted prevalence values for both models and their corresponding credible intervals for each barangay.

4.2.5 Model Validation

In order to assess the model fit, we compared the deviance information criterion (DIC) values between simple and spatial models from the convolution and ecological models. Convolution and ecological models were validated using two methods. First, we used the posterior predictive distribution to check our model assumptions by comparing the data generated from the simulations of the predictive distribution to the observed data using a test statistic. The test statistic generates a posterior predictive *p*-value (pp*p*-value) by calculating the proportion of the predicted values that are more extreme for the test statistic than the observed value for that statistic. If the data violated our model assumptions, the observed test statistic should differ from most of the replicated test statistics from our model (i.e., pp*p*-value close to 0 or 1). If the model fits the data, a pp*p*-value around 0.5 is expected. Test statistics and pp*p*-values were created for maximum, minimum, and mean values for both, convolution, and ecological models.

Second, we used the area under the curve (AUC) of the receiving operating characteristics (ROC) for applying a threshold of 0.5%, which is the prevalence mean in the Mindanao region. Thus, we would like to know the ability of the model to discriminate the mean prevalence level in the study area. We also investigated the ability of the models to discriminate the number of positive cases. Thus, a threshold of 1 was used, which indicates the presence of at least one positive case. An AUC value of 70% was taken to indicate acceptable predictive performance (Brooker, Hay and Bundy, 2002; Soares Magalhães et al., 2014).

4.2.5 Software and Data sources

Barangay centroids were obtained from an up-to-date barangay shape file from a DIVA geographic information system (Hijmans R., 2018). River and road networks were obtained from the Open Street Map Project in the Philippines (Project, 2017). Locations of cities and hamlets were extracted from the National Mapping and Resource Information Authority from The Philippines (Ocha, 2018) from 2010.

The model was implemented in the software OpenBUGS 3.2.3 (Spiegelhalter et al., 2007, 2003) (Medical Research Council, Cambridge, UK and Imperial College London, UK). The software is available for free at (Lunn D., 2018). We used the package R2OpenBUGS (Sturtz, Ligges and Gelman, 2010) to call OpenBUGS from R. The spatial models were coded using functions from GeoBUGS (Thomas et al., 2004), which is an add-on module to OpenBUGS that provides an interface to work with conditional autoregressive and geo-statistical models. Data pre-processing and Ordinary Kriging was performed in R (Team, 2013).

4.3 Resulting convolutional and ecological models

4.3.1 Convolution model

Posterior means and credible intervals resulting from the simple version of the convolution model (Equation 4.5) are given in Table 4.2. The credible intervals did not include zero values, which shows that all covariates have a strong effect in the observed outcomes except NDVI. We decided to discard NDVI from the spatial convolution and ecological model (Equation 4.6). Deviance information criterion values for the simple and spatial models were 419.8 and 64.13, respectively. This shows that the spatial model performs better than the simple model. This is also supported by the residual spatial variation of the survey data presented in Mindanao (Soares Magalhães et al., 2014). Table 4.2 shows the resulting parameter estimates and credible intervals for the spatial convolution model.

4.3.2 Ecological model

As in the convolution model, credible intervals resulting from the simple version of the ecological model (Equation 4.2) showed that all covariates have a strong effect in the observed outcomes except NDVI. The normalized difference vegetation index was also discarded from the spatial ecological model (Equation 4.9). Deviance information criterion values for the simple and spatial models were 388.1 and 126.4, respectively. This shows that the spatial model is more adequate. Table 4.2 shows the resulting parameter estimates and credible intervals for the spatial ecological model.

4.3.3 Convolution Versus Ecological model

Figure 4.2 shows the resulting density plots from the regression parameter estimates derived from the ecological and convolution models. For the intercept parameter (Figure 4.2a), the convolution model estimated a lower mean value (difference = 0.11) (Table 4.2) with a higher uncertainty, or credible interval width, than the ecological model (Table 4.2). The regression parameter estimate for NDWI from the convolution model shows a higher mean value than the one from the ecological model (Table 4.2 and Figure 4.2b). The difference between these estimates is around 0.28 with the same uncertainty for both models (Table 4.2). In the case of LSTD (Figure 4.2c), the estimated mean values from the convolution model are higher but with a slightly lower uncertainty than the ecological model. Difference between estimates are around 0.16 (Table 4.2). For the LSTN parameter (Table 4.2) and Figure 2d, the convolution model estimated a lower mean value with lower uncertainty than the ecological model (Table 4.2). The difference between these estimates is approximately 0.2. In the case of the elevation parameter (Figure 4.2e), estimated mean values from the convolution model are slightly higher than estimates from the ecological model (difference = 0.08) (Table 4.2) and have lower uncertainty than the ecological model estimates (Table 4.2). The estimated parameter value for NDWB (Figure 4.2f) is lower in the convolution than in the ecological model (Table 4.2). The difference in this value is relatively small at around 0.03. Nevertheless, uncertainty is higher in the convolution than in the ecological model (Table 4.2).

Differences between observed and predicted prevalence values are similar for both models (R^2 = 0.9). For the convolution and ecological model, the maximum difference between the predicted and the observed prevalence values is around 1% (Figure 4.3a and Figure 4.3b). Figure 4.3a and Figure 4.3b show that, for fitted prevalence values higher than 2%, both models underestimate the prevalence of infection, while, for fitted prevalence values lower than 2%, both overestimation and underestimation occur. Both models show similar predicted values with a maximum difference of 0.3%. The uncertainty in the predictions from both models is the same in 88 barangays. The ecological model presents a higher uncertainty than the convolution model in 9 barangays, while the convolution model shows higher uncertainty than the ecological model in 11 barangays.

4.3.4 Model Validation

The maximum and minimum observed prevalence values are 0.085 and 0, respectively. For the convolution model, the pp$p$-values for the maximum and minimum observed values are 0.65 and 1, respectively. This shows that it is likely to see the maximum and minimum prevalence values from the observed data in the predicted data. The highest pp$p$-value of 1 assures that 100% of our predicted data contain the minimum observed value. This could be due to an over fit to the data for small prevalence values. For the ecological model, the maximum and minimum pp$p$-values are 0.67 and 1, respectively. Like the convolution model, it is likely to see the maximum and minimum observed prevalence values in our predicted data and there might be an over fit to the data for small prevalence values.

Table 4.2: *Estimated regression coefficients (mean and 95% credible intervals).*

Estimated Parameters	**Posterior Mean (95 % CrI)**		**Standard Deviation**		**Credible Intervals Width (Uncertainty)**	
	Convolution Model	**Ecological Model**	**Convolution Model**	**Ecological Model**	**Convolution Model**	**Ecological Model**
Intercept	−5.79 (−6.11,−5.5)	−5.67 (−5.93,−5.41)	0.16	0.13	0.62	0.53
NDWI	−0.74 (−0.96,−0.55)	−1.02 (−1.24,−0.80)	0.11	0.11	0.44	0.44
LSTD	−0.63 (−0.92,−0.38)	−0.79 (−1.08,−0.49)	0.14	0.15	0.56	0.59
LSTN	−0.84 (−1.13,−0.55)	−0.65 (−1.05,−0.24)	0.15	0.21	0.59	0.82
Elevation	−1.05 (−1.4,−0.71)	−1.13 (−1.53,−0.70)	0.18	0.22	0.71	0.84
NDWB	−0.28 (−0.51,−0.05)	−0.24 (−0.43,−0.05)	0.13	0.09	0.48	0.38
φ	4×10^{-5} (−0.004,0.004)	2×10^{-5} (−0.0004,0.0004)	2.00×10^{-4}	1.00×10^{-5}	6.80×10^{-5}	$3.50 \times 10^{-}$
Variance of spatial random effect	2.58 (1.7,3.6)	2.6 (1.8,3.61)	0.48	0.47	1.9	1.82

CrI: Credible interval

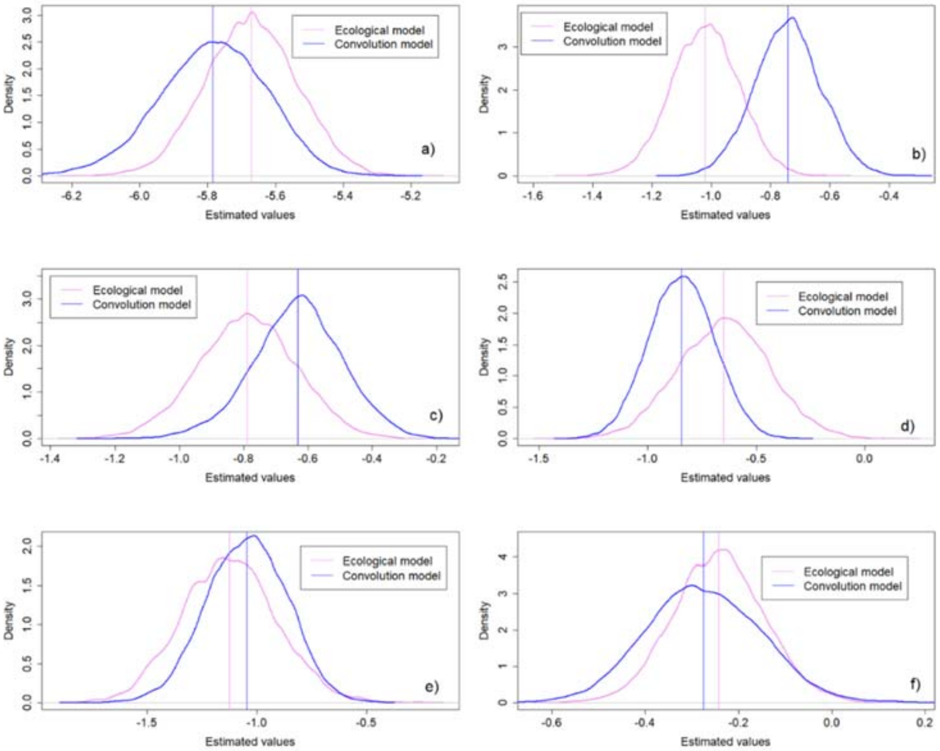


***Figure 4.2:** Covariate regression coefficients density plots for the convolution and ecological models. (**a**) Intercept, (**b**) Normalized Difference Water Index, (**c**) Day Land Surface Temperature, (**d**) Night Land Surface Temperature, (**e**) Elevation, and (**f**) Nearest Distance to Water Bodies.*

Results from the second validation method show the high ability of both models to predict prevalence values with an AUC equal to 93% and 94% for the convolution and ecological models, respectively. The AUC values with respect to predictive ability of the number of cases are 81% and 94% for the convolution and ecological models, respectively. Both models can discriminate the number of positive cases of schistosomiasis and mean prevalence values. Nevertheless, as we can see, the ecological model might over fit the data as compared to the convolution model with respect to the number of positive cases.

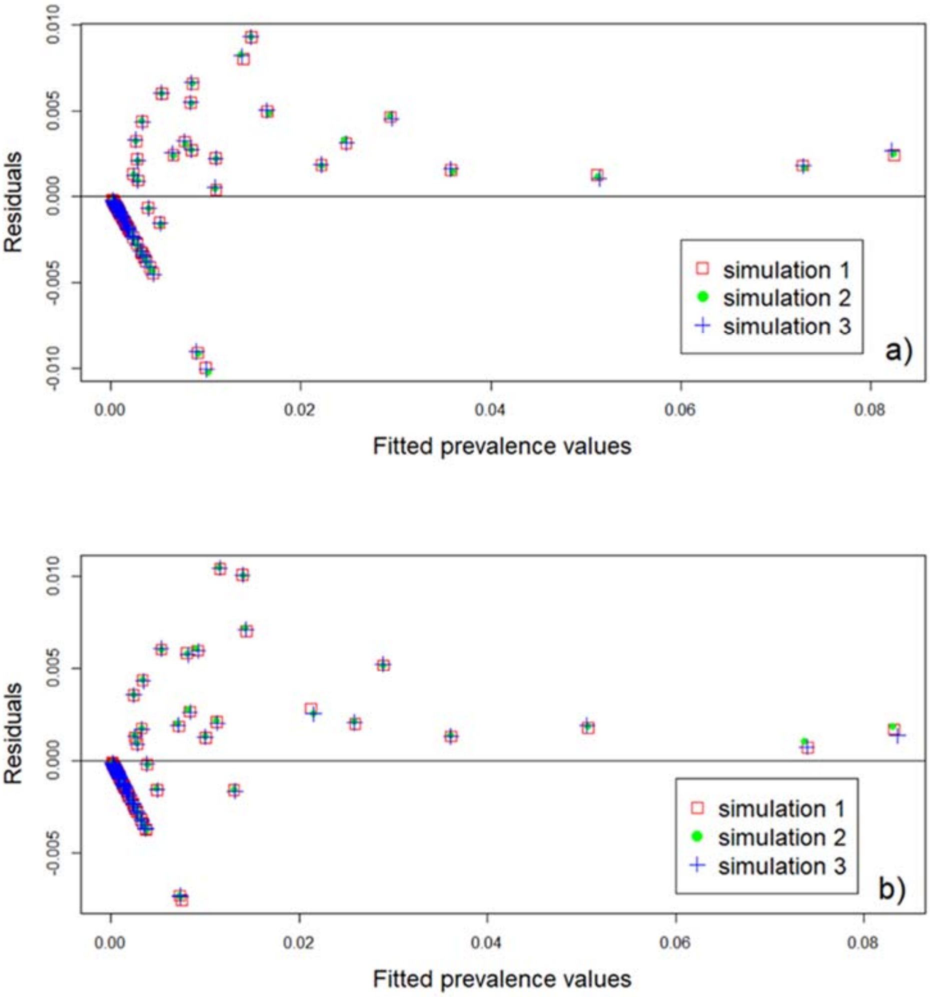


Figure 4.3: *Residual plots for the (**a**) convolution and (**b**) ecological models.*

4.4 Discussion

Several studies have modelled SCH disease risk using surveyed and environmental aggregated information at an administrative-level (Clements et al., 2008; Scholte et al., 2014; Soares Magalhães et al., 2014). These studies so far ignored the pure specification bias caused by the use of ecological or group-level estimates as individual-level estimates. Only a few studies (Wang et al., 2017) have quantified the influence of pure specification bias on the regression parameter estimates and all studies on SCH ignored the influence of pure specification bias on disease predictions. In our paper, we quantified the effect of pure specification bias on assessing the parameters for environmental covariates that are used for the mapping of *S. japonicum* infection risk. Our contribution is both methodological and practical.

Our starting point was that NDVI, NDWI, LSTD, LSTN, elevation, and NDWB are relevant for SCH transmission (Brooker et al., 2002; Kristensen, Malone and Mccarroll, 2001). For instance, NDVI is an indicator of flooded vegetation (Soares Magalhães et al., 2014), specifically rice paddy fields, and environmental moisture (Malone et al., 2001; Walz et al., 2015a). In both models, all variables have a strong effect in the observed range of outcomes, except NDVI. An explanation could be the effect of spatial support of this variable at 250 m. The International Rice Research Institute (IRRI) estimated an area substantially smaller than 25 ha for rice paddy fields ("Rice science for a better world," 2018). The area covered by an NDVI pixel equals 6.25 ha. This shows that a spatial support of 250 m is still too coarse to reliably represent paddy fields. It could be relevant to use a higher spatial support to reliably assess the role of NDVI. For instance, Walz et al. (Walz et al., 2015a) have successfully delineated paddy fields by using NDVI from RapidEye imagery at a higher spatial support of 5 m.

In the case of NDWI, the convolution model estimates a higher NDWI parameter value than the ecological model, but closer to zero (Figure 4.2b). Hence, it may not have a strong effect on the observed range of outcomes. The difference between convolution and ecological models when estimating the NDWI parameter is high (0.28) as compared to the differences for other variables. This could be due to (i) the decrease of the spatial extent of analysis from barangay to a city-level for covariate extraction, and (ii) the coarse spatial support of the variable at 500 m. A correspondence between a decrease in the extent of analysis and the within cities variability of NDWI values would change the average value used for parameter estimation. This could yield the high differences in the estimates and, together with the coarse support of the variable, could lead to a weaker effect of NDWI in SCH prevalence. Hence, NDWI pixels of 0.25 km^2 are too coarse to reliably define flooded zones in city areas that range from 0.02 to 3 km^2. The uncertainty in NDWI parameter estimate is similar for both models, as shown by the credible interval with of 0.44 (Figure 4.2b and Table 4.2). A possible explanation is that NDWI values are similar between and within cities, and between and within barangays. For instance, NDWI values between cities and between barangays range from 0.095 to 0.51 and from 0.099–0.51, respectively. The average NDWI value within a specific barangay is 0.3, while NDVI values range from 0.29–0.31 for the cities it contains.

For LSTD, a higher estimated value in the convolution model was observed than in the ecological model (Figure 4.2c). As for NDWI, this could be (i) due to the decrease in the within cities variability corresponding to a decrease in spatial extent for covariate values extraction, and (ii) the coarse LSTD spatial support of 1 km. The day land surface temperature, area pixels of 1 km^2 are too coarse to reliably define low and high-temperature zones in city areas ranging from 0.02 to 3 km^2. The uncertainty is slightly lower in the convolution model than in the ecological model possibly due to the similarity of the LSTD variability between and within cities and between and within barangays. For instance, similar LSTD values ranging from 25 to 35.7 °C and from 21.7 to 36 °C were found for the convolution and ecological models, respectively.

A different pattern is observed for the estimate of the LSTN parameter. In contrast to LSTD, the estimate and uncertainty from the convolution model are lower than for the ecological one (Figure 4.2d). This could be explained by the difference between the LSTN variability within cities and within barangays. For instance, we selected a barangay bigger in size than the cities it contains. We compared LSTN values from the cities and from the barangay and found differences from 4 to 7 °C, although the spatial support of LSTN is coarse at 1 km. These differences are small, but could be determinant for the parasite presence, as the distribution of SCH is driven by water temperatures from 15 to 20 °C (Stensgaard et al., 2005). This means that we could find the parasite in the barangay with an average LSTN value of 20 °C, but we could not find it in the cities inside the barangay, with LSTN values ranging from 22–24 °C. Thus, there is a clear loss of information produced by a pure specification bias.

Elevation presents an estimate closer to zero in the convolution model as compared to the ecological model (Figure 4.2e). This is possibly related to the decrease in the within cities elevation variability as a result of the decrease in the spatial extent of covariate extraction, from barangays to cities. Although its spatial support is relatively high, i.e., 30 m, within cities variability decreases because the Mindanao region does not contain steep slopes or sharp changes in elevation. Changes in elevation are gradual, which means that a city can share a single elevation value. The uncertainty in the convolution model is lower than in the ecological model. The reason could be the large difference in the between and within cities and barangays variability. For instance, elevation values between the cities ranged from 0 to 0.9 km, while elevation values between barangays ranged from 0 to 1.3 km. We also compared the elevation values averaged in a specific barangay and the range of values from the cities within the barangay. The averaged barangay value was around 1.3 km, while the city values within the barangay ranged from 0.87 to 0.89 km. These differences are high and give an idea of the large amount of information that could be lost when estimating individual parameters at ecological levels of analysis.

Lastly, for NDWB, the convolution model estimates a lower parameter value but with a higher uncertainty (Figure 4.2f). This could be explained by the discrepancy between the within barangay and city variability. For instance, for a specific barangay, the averaged NDWB value was approximately 3.75 km, while NDWB values for the cities within this barangay ranged from 0.4 to 6.4 km. In addition, the NDWB values between the cities ranged from 0.17 to 26.2 km, while the NDWB values between the barangays ranged from 0.26 to 15.5 km. The higher uncertainty in the convolution model could be explained by the use of ordinary Kriging in the NDWB calculation. The use of interpolation increases the variance in the estimates in a somewhat unrealistic way since it uses a constant mean (Diggle, Tawn and Moyeed, 2002) while, in reality, the different cities and barangays have different means.

The present research shows that the ecological and convolution models present similar prediction results. However, the proposed convolution model is preferred based on the lower uncertainties found in most of its parameter estimates, which shows that it corrects for pure specification bias. Moreover, according to our validation results, the convolution model has a high predictive ability to detect a positive number of cases (81%) and mean prevalence values (93%). From a public health perspective, the provision of regression coefficients that are less uncertain and better approach the individual-level estimates is a step forward to the desired uncertainty analysis in a schistosomiasis-modelling framework. This could be used as a decision support tool for helminth control programs. Moreover, less uncertain models and maps would avoid erroneous conclusions and decisions about the spatial distribution of schistosomiasis. Lastly, information on uncertainty regarding pure specification bias could guide mass drug administration campaigns by enhancing the assessment of the infection risk and understand potential impacts on human health (Araujo Navas et al., 2016).

Spatial extent and support of analysis are relevant drivers for the model parameter estimates and their associated uncertainties. The choice of support may affect the pattern identified from the data and the relationship between environmental variables and SCH prevalence (Schur et al., 2011; Schur et al., 2013). We recommend bringing all covariates to a common spatial support prior to analysis. A suggestion would be to start at e.g., 30 m and examine larger supports to more precisely quantify the role of an environmental variable in the disease modelling process (Hamm, Soares Magalhães and Clements, 2015; Simoonga et al., 2009b).

The trend in the residual plots from both models (Figure 4.3) points to a dependence between the residuals and the fitted prevalence values. This dependence could be due to heteroscedasticity (White, 1980), which means that similar interactions between the variables could lead to different prevalence values. This does not represent a problem in the model parameter estimates but is an indicator that the model can be improved (Fox, 1997).

Although our aim is not about model fit, we could possibly improve our predictions by exploring other disease driven factors and include them in the model or explore the model specification. Perhaps fitting a zero-inflated binomial (ZIB) model could help improve the predictions given that our data show a high number of barangays with zero prevalence (~77%).

4.5 Conclusions

The present study proposes a convolution model that removes pure specification bias by using ecological, group-level, health outcome data and city-level, individual-level, environmental data. For most covariates, the uncertainties in the convolution model are lower than those in the ecological model.

The spatial extent of the covariates values and the spatial support or resolution of these covariates are relevant for the parameter estimates and their uncertainties. The spatial extent and support also influence the role of the covariates in SCH modelling. Why this happens should be further explored. Additionally, between-covariate and within-covariate variability resulted in similar uncertainties in both models. Conversely, differences in between and within covariate variability explain the loss of information produced by a pure specification bias, which leads to lower uncertainties in the convolution model.

Lastly, this study shows no significant differences in the predicted values from both models. Predicted values from the convolution model are as uncertain as the predicted values from the ecological model in the majority of surveyed barangays (81.5%). The convolution model, however, shows a good predictive performance for the mean prevalence values and a positive number of infected people.

Chapter 5. Modelling the impact of MAUP on environmental drivers for *Schistosoma japonicum* prevalence [4]

[4] This chapter is submitted as:

Araujo Navas, A. L.; Osei, F.; Leonardo, L. R.; Soares Magalhães, R. J.; Stein, A. Modelling the impact of MAUP on environmental drivers for *Schistosoma japonicum* prevalence, 2019.

Abstract

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.

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. 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.

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.

5.1 Introduction

Schistosomiasis (SCH) is a water-borne neglected tropical disease of public health significance associated with important morbidity outcomes in school-aged children such as malnutrition, anaemia and stunted growth in school-aged children (Coutinho et al., 2005; Leenstra et al., 2006). 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 (Jia et al., 2007), *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 (Tarafder et al., 2006; Yang et al., 2008). SCH affects more than 252 million people worldwide (Hotez et al., 2014) especially populations living at 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 (Soares Magalhães et al., 2014). 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 (Herbreteau et al., 2007; Kalluri et al., 2007). Modelling those infections using various statistical methods have enabled the study of the distribution of populations at-risk (Herbreteau et al., 2007; Kalluri et al., 2007), and the role of the environmental variation on the geographical heterogeneity of infection burden (i.e. prevalence or intensity of infection) . 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 (Cadavid Restrepo et al., 2016; Soares Magalhães et al., 2011b).

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 (Araujo Navas et al., 2016). EO data are generally constrained by their spatial and temporal scale of sampling (Wang et al., 2008). In this study we focus on spatial scale. Scale is a major concern in spatial epidemiology (Walz et al., 2015c) since it determines the significance of the various environmental risk factors on the disease distribution (Simoonga et al., 2009b). Spatial scale encompasses the spatial support and the spatial extent of analysis (Atkinson and Graham, 2006). 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 (Atkinson and Graham, 2006). For a given extent the support size may affect the patterns identified in the survey and environmental data (Schur et al., 2013) and the relationship between the disease morbidity indicators and the environmental risk factors. This is known as the modifiable areal unit problem (MAUP) (Dungan et al., 2002).

Various studies investigated the consequences of ignoring MAUP effects in spatial epidemiological modelling. For instance, Hellsten et al. (Hellsten, 2006) studied the influence of using aggregated covariate data to model ammonia emissions at 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 (Schur et al., 2011) and Schur et al (Schur et al., 2013) aggregated SCH prevalence maps to estimate endemicity for various administrative units (Schur, Vounatsou and Utzinger, 2012). Such aggregation showed different methods of intervention and endemicity patterns. As a consequence, localized areas of high endemicity may not be addressed properly. In a recent study (Araujo Navas et al., 2019) we concluded that the changes in spatial support and their effects on the model parameter estimates, and their associated uncertainties, should be further investigated as it might be a significant source of uncertainty in SCH modelling (Araujo Navas et al., 2016).

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 (Schur et al., 2011; Schur et al., 2013). 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 (Hotez et al., 2014; Simoonga et al., 2009b). 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.

5.2 Methods for modelling Schistosoma japonicum under the MAUP

5.2.1 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 (Leonardo et al., 2015), with approximately 1.8 million estimated infected people (Leonardo et al., 2002). The disease affects children, adolescents, and individuals with high-risk occupations, like fishermen and farmers (Leonardo et al., 2002; Zhou et al., 2010). The area of study is the region of Mindanao in the Philippines (Figure 4.1). This area was selected due to the high response rate of 70.9 percent of the individuals to the 2008 survey (Leonardo et al., 2012; Leonardo et al., 2008; Soares Magalhães et al., 2014), 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 about 58-1158 within a single province. In total 11 Provinces with high (≥ 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, 20 provinces were surveyed, and 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 (Leonardo et al., 2008) was used based on two-sample stool collection. Due to inconsistencies in the second stool sample submission, however, only the results of the first sample were available (Soares Magalhães et al., 2014). 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 updated barangay centroids shape file from DIVA geographic information system (Hijmans R., 2018). More details about the sampling design and surveyed information can be found in Leonardo et al (Leonardo et al., 2012; Santos, Cerqueira and Soares, 2005).

5.2.2 Environmental risk factors

We included in the analysis six relevant environmental risk factors for SCH transmission (Brooker et al., 2002; Kristensen, Malone and Mccarroll, 2001). 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 (Soares Magalhães et al., 2014). NDVI is and indicator of flooded vegetation (Soares Magalhães et al., 2014), particularly rice paddy fields, and environmental moisture (Malone et al., 2001; Walz et al., 2015a). Both are an important risk factor for Asian SCH (Zhou, Liang and Jiang, 2012). NDWI was used as a proxy indicator of flooding (Walz et al., 2015a; Xu, 2006) showing potentially hidden water bodies. LSTD and LSTN are determinant for the survival of larval stages of snails (Pietrock and Marcogliese, 2003; Prah and James, 1977; Woolhouse and Chandiwana, 1990) and are used as proxies for water temperature given that the thermal condition of shallow waters usually reflects the temperature of the air (Soares Magalhães et al., 2014). Elevation is relevant for SCH transmission as the local topography of the area determines the presence of snails (Pesigan et al., 1958; Stensgaard et al., 2006; Stensgaard et al., 2013). For instance, at lower altitudes the risk of finding snails increases.

NDWB values were calculated using the closest facility network analysis tool from ArcGIS (Esri, 2011). 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 (Project, 2017). Cities and hamlets locations were obtained from the National Mapping and Resource Information Authority from The Philippines (Ocha, 2018) 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 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 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 and LSTN values were derived from MODIS MOD11A2_LST product with a spatial support of 1 km. Finally, Elevation was obtained from ASTER GDEM version 2 from USGS (Geological Survey, 2017) with a spatial support of 30 m. All covariates were set to a common coordinate system UTM zone 51N before being used. Table 5.1 summarizes all sources of information.

Modifying the areal units of analysis

From now onwards we will refer to an aerial unit as the spatial support of analysis (SSA). We used five SSAs, equal to 30 m, 90 m, 250 m, 500 m, and 1 km, respectively. These spatial supports increase when going from less to more 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 interpolation. 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.

5.2.3 Modelling Schistosoma japonicum infection under the MAUP

Convolution model

Data on the human *Schistosoma japonicum* infection variable y 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. We assigned to y_k a binomial distribution with parameters N_k and $\hat{p}_k$ 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 $\boldsymbol{\bar{x}_k}$, where $\boldsymbol{\gamma}$ are the barangay-level coefficients (Equation 5.1).

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

$$logit(\hat{p}_k) = \gamma_0 + \gamma_1 \cdot \bar{x}_{1_k} + \gamma_2 \cdot \bar{x}_{2_k} + \cdots + \gamma_n \cdot \bar{x}_{n_k}. \quad (5.1)$$

Table 5.1: *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.

We modelled human S*chistosoma japonicum* infection at five increasing SSAs using a convolution model that accounts for pure specification bias (Araujo Navas et al., 2019). This bias is a source of uncertainty (Araujo Navas et al., 2016; King, 2013; Richardson and Monfort, 2000) and is produced by the loss of information when using aggregated survey data in a non-linear model, for individual-level inferences (Wakefield and Lyons, 2010). It is called 'pure' because it specifically addresses model specification bias (Gelfand et al., 2010), and it biases the estimates because any direct link between exposure and health outcomes is imperfectly measured (Richardson and Monfort, 2000). This is because the regression function does not approximate the real relationship between the affected population and their exposure (Araujo Navas et al., 2019). Pure specification bias can be reduced as the within area exposure is more homogenous (Wakefield and Lyons, 2010). This could be done by having a finer partition of space at which environmental risk factors are available (Wakefield and Lyons, 2010).

We propose to minimize pure specification bias by extracting covariate information from cities within barangays (Figure 5.2). The city level is the finest available one. Cities are thus considered as a proxy for individual-level exposure locations. They were extracted from the 2010 build-up data base from the National Mapping and Resource Information Authority from the Philippines ("National Mapping and Resource Information Authority," 2018). We completed unavailable cities using Google Earth Images.

For the convolution model, we used the aggregate data method proposed by Prentice and Sheppard (Prentice and Sheppard, 1995). For each SSA, we obtained covariate information $\boldsymbol{x}$ for image pixel i belonging to a city j within a specific barangay k (Figure 5.1). Let n 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. 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 5.2). We accounted for the spatial variability at barangay-level by adding a spatial structure random effects term s_k. Thus, the convolution model that we used for each SSA is of the form:

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

$$\hat{\hat{p}}_k = \frac{1}{m_k} \cdot \sum_{i=1}^{m_k} \frac{1}{1+\exp(-(\beta_0+\beta_1 \cdot x_{1_{ijk}}+\beta_2 \cdot x_{2_{ijk}}+\cdots+\beta_n \cdot x_{n_{ijk}}+s_k))}. \qquad (5.2)$$

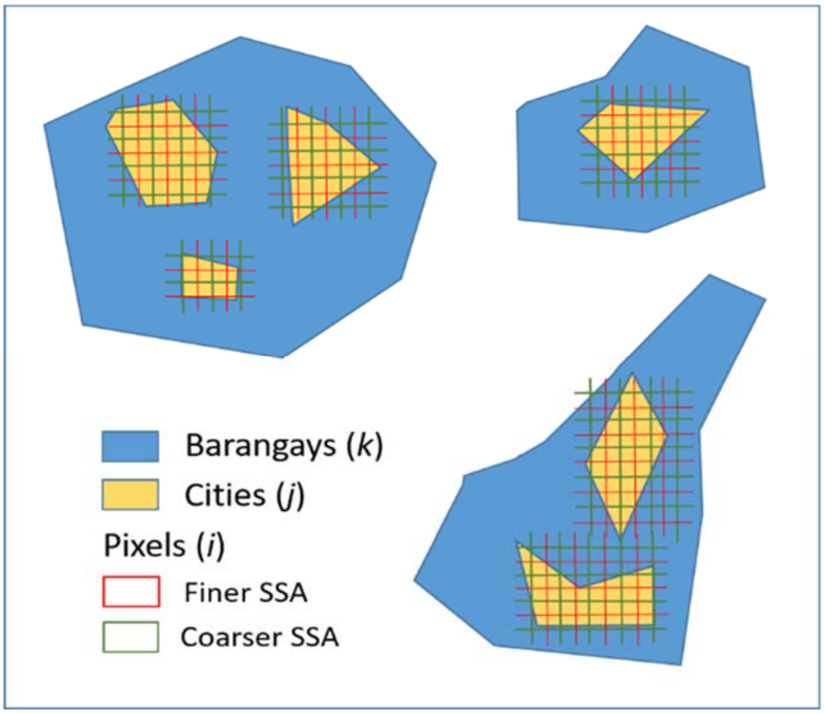


Figure 5.1: *Pure specification bias minimization: Environmental risk factors extraction at pixel-level from cities within barangays*

Model implementation

Five models were implemented, all include 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 an uninformative uniform prior distribution with wide bounds $\beta_0 \sim U[-100{,}100]$. The other $\boldsymbol{\beta}$ parameters were given uninformative normal distributions $\boldsymbol{\beta} \sim N[0, \frac{1}{\delta^2}]$, with δ uniformly distributed on a wide range $\delta \sim U[0, 100]$. These distributions avoid overestimating the parameters (Gelman, 2006) and allow a good sequences mixing used for Markov Chain Monte Carlo (MCMC) simulations, contributing to a fast convergence (Gelman et al., 1995).

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 (Diggle, Tawn and Moyeed, 2002). The vector of random variables $\boldsymbol{s}$ associated with point locations $(\mathrm{x}_k, \mathrm{y}_k)$, $k = 1, \ldots, K$, was modelled with a multivariate normal distribution $s{\sim}MVN_K[\mu,\Sigma_{ab}]$ with mean $\mu = 0$ 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, 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 uniform: $\phi \sim U[2\mathrm{E}10^{-7}, 3\mathrm{E}10^{-3}]$. 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 (Thomas et al., 2004). 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 (Gelman, 2006; Lunn et al., 2013).

To run the model we used three sequences or chains with 50000 iterations. This number of iterations ensured simulations representativeness of target distributions and good stability for convergence (Gelman et al., 1995). In order to diminish the influence of starting values, we discarded the first half of each sequence (Gelman et al., 1995) using a burn-in of 25000 iterations. Convergence was monitored visually and statistically. First by inspecting the trace plots, and second by checking the $\hat{R}$ statistic (Brooks and Gelman, 1998; Gelman and Rubin, 1992) also called the potential scale reduction factor. The scale reduction factor assesses sequences mixing by comparing the between and within variation. $\hat{R}$ values < 1.1 indicate evidence that sequences had converged (Brooks and Gelman, 1998), while high values suggest that an increase in the number of simulations may improve our inferences (Gelman et al., 1995).

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 Appendices 4A.1 and 4A.2, respectively.

5.2.4 Model validation

The five models were validated using two methods. First, we compared the data generated from the simulations of the predictive distribution to the observed data using a test statistic. This test statistic uses a posterior predictive p-value (pp$p$-value) generated by calculating the proportion of the predicted values which are more extreme for the test statistic than the observed value for that statistic. If the model fits the data, we expect a pp$p$-value of around 0.5. Otherwise, the pp$p$-value will be close to 0 or 1. We generated pp*p* values for maximum, and minimum values for the models at five increasing SSA. Second, we 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 (Brooker, Hay and Bundy, 2002; Soares Magalhães et al., 2014).

5.2.5 Software

Model implementation was done in the software OpenBUGS 3.2.3 (Spiegelhalter et al., 2003, 2007) (Medical research Council, Cambridge, UK, and Imperial College London, UK). It was downloaded for free at (Lunn D., 2018). We called Open BUGS from R using the package R2OpenBUGS (Sturtz, Ligges and Gelman, 2010). The spatial models were coded using the GeoBUGS (Thomas et al., 2004) 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 (Team, 2013).

5.3 Resulting MAUP effects on modelling Schistosoma japonicum infection

5.3.1 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 5.2 and Figure 5.2a). 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 5.3a and Table 5.3), slightly increasing when increasing SSA. The highest uncertainty 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 5.2 and Figure 5.2b). When SSA increases, parameter estimates increase from -1.06 to -0.76, approaching 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 5.3b and Table 5.3). The highest uncertainty value equals 0.44 for SSA = 500 m and the lowest value equals 0.5 for SSA = 30 m.

LSTD has a significant negative effect on the prevalence of SCH infection for almost all SSA, except for SSA = 1 km (Table 5.2 and Figure 5.2c). Similar parameter estimates equal to -0.71 are obtained for SSA = 30 m, 90 m and 250 m, while it increases slightly to -0.65 for SSA = 500 m. For SSA = 1 km, there is a noticeable increase in the estimate to -0.01 (Figure 5.2c and Figure 5.3c). 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 5.3c).

LSTN has a significant negative effect on the prevalence of SCH infection for almost all SSA, except for SSA = 1 km (Table 5.2 and Figure 5.2d). 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 5.2d and Figure 5.3d). Uncertainty varies from 0.56 to 0.58 when increasing SSA from 30 m to 500 m, but it considerably increases to 1.14 for SSA = 1 km (Table 5.3 and Figure 5.3d).

Elevation has a significant negative effect on the prevalence of SCH infection for all SSA, except for SSA = 1 km (Table 5.2 and Figure 5.2e). 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 5.2e and 5.3e). 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 5.3 and Figure 5.3). 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 5.2 and Figure 5.2f). 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 5.2f). Uncertainties constantly increase from 0.32 to 0.39 (Table 5.3 and Figure 5.3f) when increasing SSA.

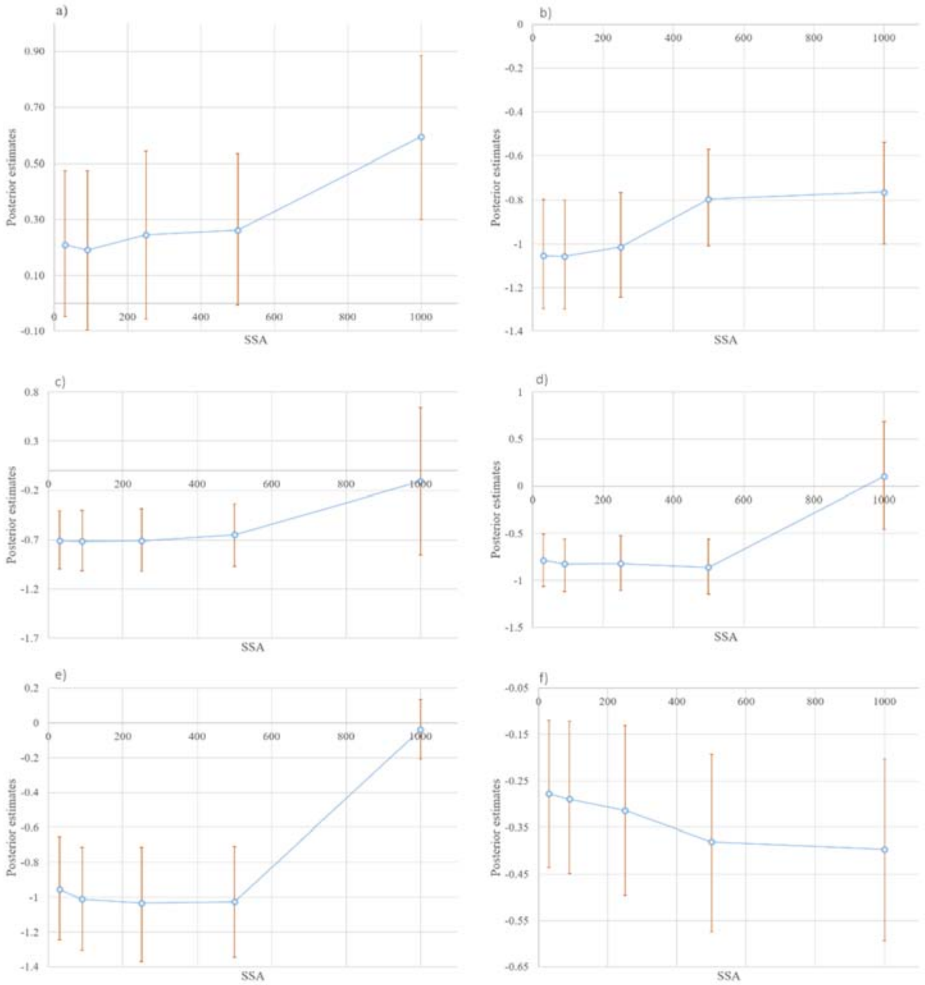
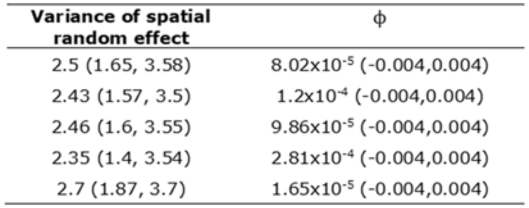


Figure 5.2: *Posterior estimates and their credible intervals.* ***(a)*** *Normalized difference vegetation index;* ***(b)*** *Normalized difference water index;* ***(c)*** *Land surface temperature day;* ***(d)*** *Land surface temperature night;* ***(e)*** *Elevation;* ***(f)*** *Nearest distance to water bodies.*
SSA: Spatial support of analysis.

Table 5.2: *Regression coefficient estimates for each risk factor at five descending spatial supports of analysis*

Spatial Supports of analysis	Posterior Mean (95% CrI)					
	NDVI	**NDWI**	**LSTD**	**LSTN**	**E**	**NDWB**
30 m	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	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	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	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	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
2.5 (1.65, 3.58)	8.02×10^{-5} (-0.004,0.004)
2.43 (1.57, 3.5)	1.2×10^{-4} (-0.004,0.004)
2.46 (1.6, 3.55)	9.86×10^{-5} (-0.004,0.004)
2.35 (1.4, 3.54)	2.81×10^{-4} (-0.004,0.004)
2.7 (1.87, 3.7)	1.65×10^{-5} (-0.004,0.004)

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

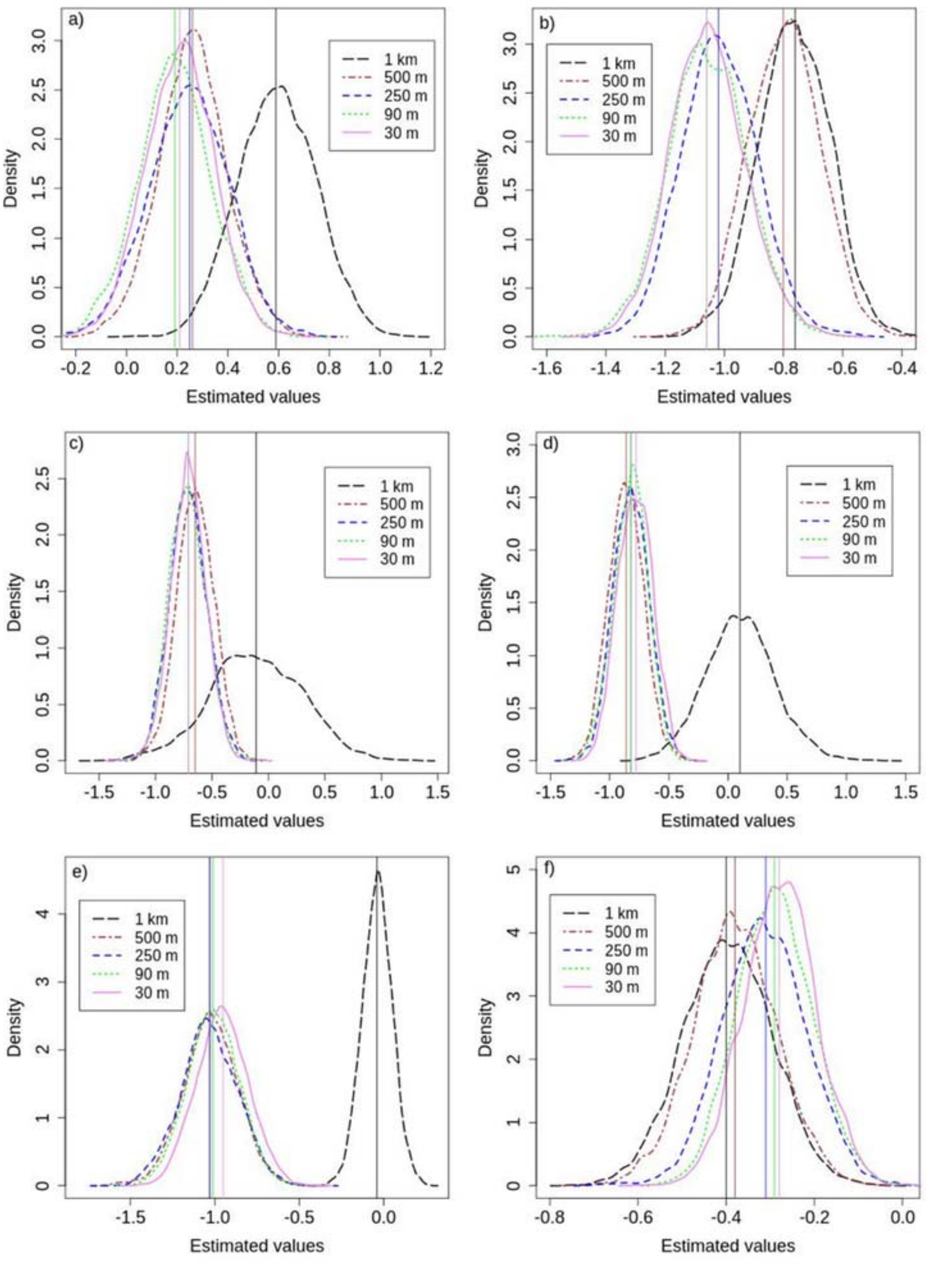


Figure 5.3: *Density plots for the risk factors regression coefficients.* ***(a)*** *Normalized difference vegetation index;* ***(b)*** *Normalized difference water index;* ***(c)*** *Land surface temperature day;* ***(d)*** *Land surface temperature night;* ***(e)*** *Elevation;* ***(f)*** *Nearest distance to water bodies.*

Table 5.3: *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.

Influence on predictions

Differences between observed and predicted prevalence values are similar for the five SSA models (Figure 5.4). Variation in these differences is the 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 5.4 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 5.4).

Uncertainties on the predictions are similar for the five models. 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.

5.3.2 Model validation

The maximum and minimum observed prevalence values are 8.5% and 0%, respectively. The first validation method shows pp$p$-values for all SSA ranging from 0.65 to 0.67 for the maximum observed values (Table 5.4). This means that for all SSA it is likely to see the maximum prevalence values from the observed data in the predicted data.

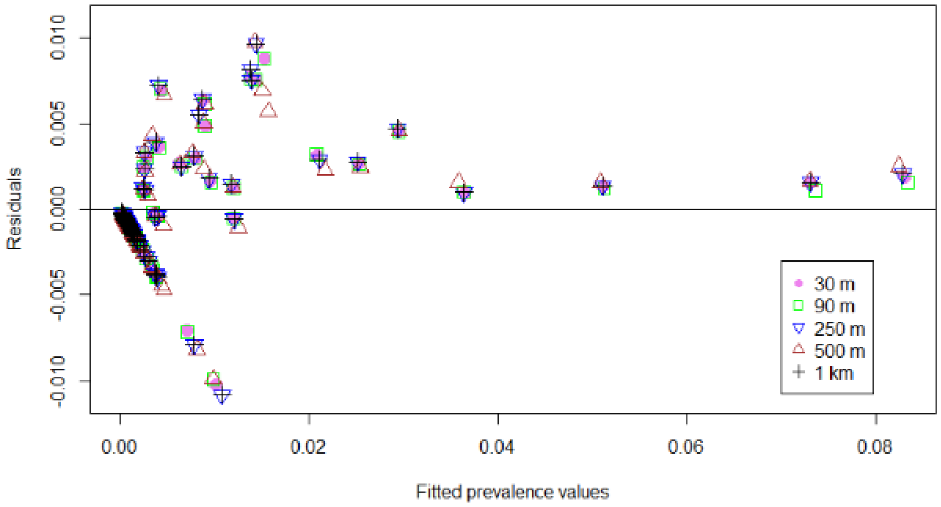


Figure 5.4: *Residual plots for the five increasing spatial supports of analysis.*

All models present pp*p*-values of 1 for the minimum observed values (Table 5.4). This means that 100% of the predicted data in all models contain the minimum observed value, showing an over fit to the data for small prevalence values.

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.

5.4 Discussion

Schistosomiasis modelling studies have commonly used environmental risk factors as drivers for disease exposure and transmission (Hu et al., 2015; Scholte et al., 2014). 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.

Table 5.4: *Resulting ppp-values for the first validation method for the five increasing SSA*

Spatial Supports of analysis	ppp-value (maximum)	ppp-value (minimum)
m	0.66	1
90 m	0.67	1
250 m	0.66	1
500 m	0.66	1
1 km	0.65	1

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

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. We applied it on *Schistosoma japonicum* infection modelling in the Mindanao region, the Philippines.

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

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, because NDVI effects on SCH prevalence are global in the sense of Soares Magalhães et al (Soares Magalhães et al., 2014), where significant effects of NDVI were found for almost the whole Philippines. The increase in uncertainty values with increasing SSA is due to the coarse areal pixels ≥ 250 m resolution that do not reliably represent rice paddy fields. Those are substantially smaller than 25 ha, i.e. are covered by at most four pixels ("Rice science for a better world," 2018).

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. 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 (Araujo Navas et al., 2019). 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 (Araujo Navas et al., 2019). 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 does not represent 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. NDWB estimates decrease 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. Clearly, transportation plays an important role (Kummu et al., 2011) in SCH transmission.

Uncertainty values for NDWI decrease when increasing the SSA, with a minimum of 0.44 for SSA = 250 m. Clearly, NDWI data originally available at SSA = 250 m are more reliable than values modified to larger SSAs. Using Ordinary Kriging for interpolation increases the variance in the estimates in a somewhat unrealistic way since it uses a constant mean (Diggle, Tawn and Moyeed, 2002), while in reality, means are different. Uncertainty values of NDWB, for instance, increase with increasing SSA due to the coarse areal pixel units ≥ 0.25 km^2. Such a size is insufficient to reliable define nearest distances to water bodies in city areas of 0.02 to 3 km^2.

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

Bayesian statistical methods were used to model the disease, and along with a convolution regression model, they corrected for pure specification bias on our estimates. This is a relevant contribution in the analysis of uncertainties on this type of spatial epidemiological studies. For future studies, new trends in geospatial artificial intelligence (geoAI), that could resolve limitations regarding the MAUP for exposure modelling studies, are emerging to model schistosomiasis (Mari et al., 2017) as well as other diseases (Vopham et al., 2018). We particularly identified (i) the use of high-performance computing to handle spatiotemporal big data, and (ii) machine and deep learning algorithms implementation to big data infrastructures to extract relevant disease or environmental information (Baker and Nieuwenhuijsen, 2008; Vopham et al., 2018). 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, (Lin et al., 2017). 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. Epidemiologist and health scientist could use these maps to identify risk areas for the control and prevention of the disease (Soares Magalhães et al., 2011b), 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 (Araujo Navas et al., 2016) and avoiding erroneous conclusions and decisions about the spatial distribution of schistosomiasis (Araujo Navas et al., 2016; Araujo Navas et al., 2019).

5.5 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 scientist in the control or prevention of the disease astray.

The results from this study could guide other disease modelling studies as it suggests the spatial supports 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 benefit, as it could help to better understand and quantify MAUP effects on the relationship between the disease and its risk factors, as well as to provide reliable maps useful for disease control and prevention.

Chapter 6. Synthesis

6.1 Research findings and conclusions

The main objective of this thesis was to analyse uncertainty in modelling SCH helminth infections by: (i) identifying the gaps in knowledge regarding uncertainty, its sources, definition, quantification, and public health implications; (ii) mapping potential areas of exposure to correct for positional mismatch; (iii) quantifying the effect of pure specification bias on parameter estimates; and (iv) quantifying MAUP effects on parameter estimates. The following section describes the main conclusions per objective. The section also reflects upon the implications of this research, and discusses and recommends further research.

Objective 1. To identify the gaps in knowledge of the different components of uncertainty associated with mapping and modelling helminth infections.

Research question 1: How is uncertainty defined and quantified in helminth modelling studies?

Results show four definitions of uncertainty, the most commonly used are imprecision and accuracy, and the least used are bias and vagueness. Credible intervals and the area under the curve of the receiving operating characteristics were the most commonly used measures of imprecision and accuracy, respectively. Bias and vagueness were quantified by using mean error and fuzzy sets, respectively. Uncertainty in parameter estimates was most quantified by Bayesian and frequentist methods. Identification and incorporation of uncertainty regarding the use of questionnaires, diagnostic uncertainty, and the combination of age-groups in the predictions was limited.

Research question 2: What are the main sources of uncertainty in helminth infections mapping and modelling?

Sampling design, diagnostic techniques and the selection of environmental and socio-economic risk factors are sources of uncertainty that were commonly addressed. These should be given primary attention when modelling helminth infections. Less treated uncertainty sources were input data quality, spatio-temporal misaligned data and inherent group characteristics. Although these being largely ignored, they are relevant for helminth modelling and should be further investigated.

Research question 3: How is uncertainty informative for decision makers, public health scientist and the affected community?

Uncertainty is informative in two ways: policy making and scientific interpretation. Regarding policy making, uncertainty information is mostly used for planning, intervention, monitoring, evaluation and consolidation of mass drug administration campaigns, but also for the increase in the cost-effectiveness of these programmes. Uncertainty is used, at a minor scale, for prevention strategies such as to plan and guide hygiene education and infrastructure water sanitation and hygiene programmes. Regarding scientific interpretation, uncertainty was used to improve spatial sampling, and explore the role of environmental and socio-economic risk factors on helminth infections.

In Chapter 2 I provided a framework for an uncertainty evaluation in the spatial modelling of helminth infections. This framework was based on the identification of various uncertainty sources, definitions and quantification measures used in the current literature. I proposed a quantitative and qualitative analysis of uncertainty for a complete assessment of risk, its impacts and implications in the public health domain.

Objective 2: To map potential areas of exposure to *Schistosoma japonicum* infection using a spatial Bayesian network (sBN) model.

Research question 1: Could the positional mismatch between survey and covariate data be corrected?

The answer is Yes. Spatial modelling studies of SCH commonly use covariate values extracted at survey locations (e.g. schools, hospitals), where SCH infection does not occur. This produces a positional mismatch between survey and exposure locations. By using a model that represents the positional mismatch between survey locations and exposure sites, potential exposures areas could be delineated, and be used to extract covariate information

In order to delineate exposure areas, I used an sBN. The structure of the sBN was defined by five observable random variables: land use, elevation, slope, nearest distance to water bodies and snail infection rate; and three latent random variables: potential accessible sites for snails, community cost, and exposure. In order to compute the probabilities for each latent random variable, node probability tables and marginal probabilities were used. Probabilities were computed for each polygon of analysis, which were constructed based on the overlaying of each risk factor, i.e. the parent node. High, medium, low and very low probabilities of exposure were derived from the proposed exposure network.

In Chapter 3 I show that the positional mismatch problem can be addressed by extracting covariate values that are not at the survey locations, but from potential exposure areas. These areas are represented as a probability of exposure surface, driven mainly by the presence of specific land use types and distances to water bodies. Although the availability of data was limited to construct and validate the sBN, whenever new knowledge is available the sBN will enable a rapid delineation of potential exposure areas by facilitating a flexible integration of exposure data and their prior information.

Objective 3: To quantify the effect of pure specification bias on the parameter estimates of various environmental drivers of *Schistosoma japonicum* infection.

Research question 1: Can pure specification bias be corrected by using group-level disease data and individual-level covariate data?

Again the answer is Yes. It was achieved by the implementation of a convolution model that calculates the mean of the risk function over the total number of cities, used as proxies for exposure locations. By dividing the area of analysis into finer units that contain exposure measurements, pure specification bias is diminished as the within-area exposure variability decreases. Since individual level responses are known, but not their locations, pure specification bias was reduced by extracting covariate information from cities within barangays.

Research question 2: How much does pure specification bias increase or decrease parameter estimates and their uncertainties?

The convolution model increased the parameter estimates of NDWI[1*], LSTD[2] and Elevation to 27%, 20% and 7%, respectively; and decreased the parameter estimated of LSTN[3] and NDWB[4] to 29% and 17%, respectively. The convolution model decreased the uncertainty for LSTD, LSTN and elevation estimates to 5%, 28% and 15%, respectively. These results show that there is a clear loss of information produced by pure specification bias. For instance, we could find the parasite in a barangay with an average LSTN value of 20℃, but we could not find it in cities inside this barangay with LSTN values ranging from 22-24℃. The convolution model also increased the uncertainty for NDWB to 26% and maintained the same uncertainty (0.44) for NDWI as the ecological model.

Chapter 4 proposes the use of a convolution model that reduces pure specification bias by providing lower uncertainty values for most of the parameter estimates. Results from validation show that the convolution model has a higher predictive ability for the number of positive cases (81%) and the mean prevalence values (93%). The provision of reliable individual-level

[*] 1: Normalized difference water index; 2: Land surface temperature day; 3: Land surface temperature night; 4: Nearest distance to water bodies.

estimates would enable a less uncertain mapping process. Thus, correct conclusions and decision about the spatial distribution of SCH can be taken to support mass drug administration campaigns.

Objective 4: To quantify the modifiable areal unit problem (MAUP) effect on various environmental drivers of *Schistosoma japonicum* prevalence of infection.

Research question 1: Does an increase in the spatial support of analysis increase or decrease parameter estimates and their uncertainties?

By increasing the SSA[1*], the parameter estimates for almost all covariates increase, except for NDWB. For NDVI[2], LSTD, LSTN and E[3], increasing the SSA from 30 m to 500 m does not indicate significant changes in their parameter estimates. Conversely, for SSA = 1 km, parameter estimates considerably increase. LSTD, LSTN and E have a negative correlation with SCH prevalence. However, when SSA = 1 km, their estimates approach zero, losing their significance on SCH prevalence. For NDWI and NDWB, an increase in SSA from 30 m to 250 m does not indicate significant changes in parameter estimates. However if SSA ≥ 500 m noticeable changes occur in their estimates.

By increasing the SSA, the uncertainty in the covariate parameter estimates increases, except for NDWI. LSTD and LSTN have noticeable high uncertain estimates at SSA = 1 km. For NDVI and E, uncertainty increases until the SSA = 250 m, and then decreases. This is explained by their inherent characteristics in the study area and their global effects on SCH. A decrease in uncertainty for NDWI is attributed to the various aggregation and disaggregation processes. Originally NDWI data were obtained at 500 m. Thus, values at this SSA are more reliable than at other SSAs.

Research question 2: What is the suggested spatial support of analysis to model *Schistosoma japonicum* infection?

Results suggest to use all covariates for SSA ≤ 500 m as most of them have a significant negative effect on SCH prevalence and present more reliable parameter estimates. Only NDVI has a significant positive effect on SCH prevalence. Nevertheless, at SSA = 1 km, its estimate is unreliable.

In conclusion, chapter 5 shows changes in covariate estimates and their uncertainties when increasing the SSA. It also evaluates the relationship between SCH prevalence and the environmental risk factors. This chapter

[*] 1: Spatial support of analysis; 2: Normalized difference vegetation index;
3: Elevation.

recommends modelling *Schistosma japonicum* infection in Mindanao region at a SSA ≤ 500 m. It advises not to use covariate data at a SSA = 1 km as it could lead to wrong conclusions about the distribution of the disease and unreliable decisions about targeting mass drug administration campaigns.

6.2 Reflections

The motivation of the present research was (i) to perform a systematic appraisal of uncertainty in modelling SCH; and (ii) to propose methods to reduce various sources of uncertainty, specifically those coming from the use of earth observation data. The first motivation aimed to contribute to the optimal interpretation and communication of uncertainty in SCH modelling. The second motivation aimed to design effective methods to reduce uncertainties derived by using incorrectly geo-located and aggregated survey data, and spatially misaligned environmental risk factors. This research is thus the basis for the analysis of various sources of uncertainty intrinsict to SCH modelling studies. The reflection on this work is detailed below.

Practical implications from the public health perspective of SCH in The Philippines.

Uncertainty identification and quantification in SCH modelling is important to assess applicability and validity of the parameter estimates and the predicted outcomes. This, at the same time, is important for an adequate survey design and quantification of future investments in treatment needs. Ignoring uncertainties can lead to a weak assessment of SCH risk, incorrect geo-location of at-risk populations, and inaccurate quantification of the number of people in need of treatment. As mass drug administration campaigns depend upon SCH modelling outputs, uncertainty communication is critical for control programs, health care workers, public health scientists, decision makers and other involved stakeholders who attempt to reduce prevalence or incidence of SCH infection across the affected areas. Control programs and health care workers could use uncertainty outcomes to decide about drug distribution strategies and the frequency of treatment to target populations. Decision makers could use it to target more resources in terms of data acquisition or the required amount of anthelminthic drugs. Finally, public health scientist could use uncertainty information to focus investigative efforts on high risk areas such as improving methods for uncertainty quantification and mapping. Details about the practical implications of this research for uncertainty stakeholders in SCH modelling are given in figure 6.1

All research outcomes provide a framework for the future development of a spatial decision support system for SCH surveillance and control. For instance, five components of uncertainty were recognized in research outcome 1 shown in Figure 6.1. This facilitates the identification and priorization of main sources of uncertainty such that at least the most important ones can be incorporated into the predictive model.

Research outcome 2, also shown in Figure 6.1, would enable best practices in survey design by providing maps that show high probability of exposure areas. These areas allow the geolocation of populations at-risk in need of being surveyed. Likewise, Figure 6.1 shows research outcome 3 which suggests the use of geo-located individual-level surveyed data such that pure specification bias can be avoided.

Research outcomes 3 and 4 (see Figure 6.1) provide more accurate parameter estimates by reducing uncertainties due to pure specification bias and MAUP. Using reliable posterior predictive distributions would trigger the prediction of reliable prevalence of infection values at unsampled locations, providing an accurate identification of at-risk populations and quantification of the number of people in need of treatment. Consequently, results also provide a better account of future investments for treatment needs of at-risk populations.

Epidemiological implications

SCH intermediate snail hosts are found in stagnant water from ditches, rice paddy fields, and ponds, or moisture areas with aquatic vegetation or inundated grass (De Roeck et al., 2014), which assure snails survival (Brown, 1994). Because the distribution of anthelminthic medication is inexpensive, understanding the dynamics of snail populations and SCH transmission would help to increase its cost effectiveness and thus, improve SCH control.

Results from this research show that, working with aggregated survey data for individual level inferences, and using environmental risk factors at various spatial supports of analysis, strongly influence the spatial correlation patterns of SCH. For instance, as environmental factors determine the habitat requirements of intermediate snail hosts (Clennon et al., 2007), changes in their values, based upon the spatial support of analysis, will reflect different patterns on the disease spatial distribution. Results show that for a spatial correlation close to zero, the range or maximum correlation distance varies at different spatial supports. For spatial supports from 30 m to 250 m, the range has a low variability from 19.2 km to 28.7 km. Conversely, for spatial supports of 500 m and 1 km, range values are equal to 8.2 km and 139.6 km, respectively. The local dynamics of SCH seem to be properly reflected when modelling the disease at finer spatial supports (< 500 m), while for coarser supports (≤ 500 m), SCH is estimated to be spatially correlated either at remarkably short (8.2 km) or extremely long distances (139.6 km).

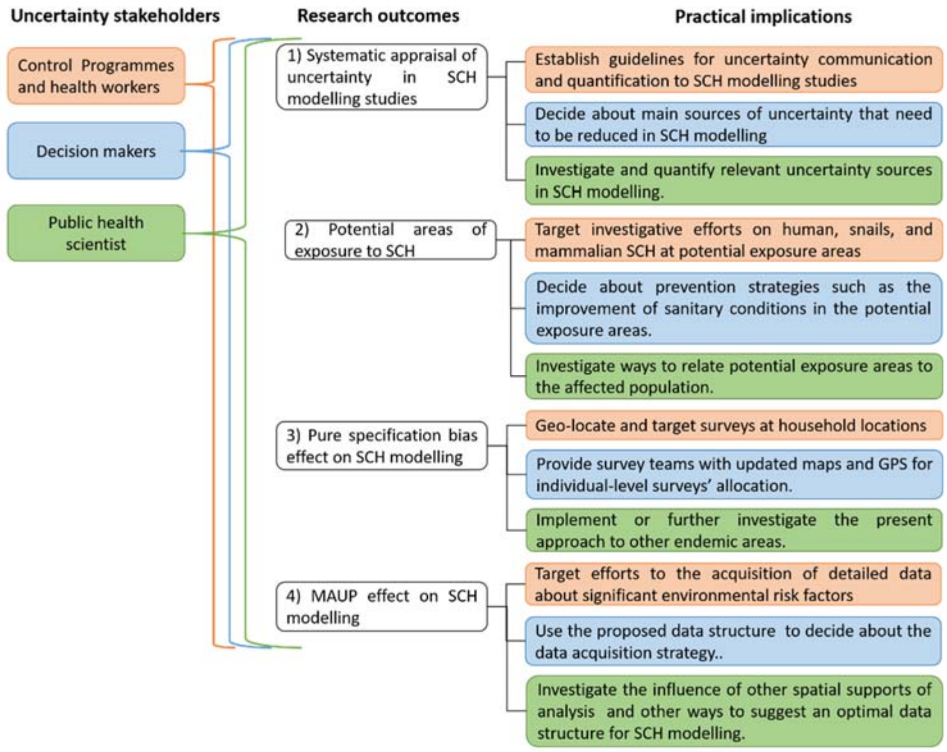

Uncertainty stakeholders
Research outcomes
Practical implications
Control Programmes and health workers
Decision makers
Public health scientist
1) Systematic appraisal of uncertainty in SCH modelling studies
2) Potential areas of exposure to SCH
3) Pure specification bias effect on SCH modelling
4) MAUP effect on SCH modelling
Establish guidelines for uncertainty communication and quantification to SCH modelling studies
Decide about main sources of uncertainty that need to be reduced in SCH modelling
Investigate and quantify relevant uncertainty sources in SCH modelling.
Target investigative efforts on human, snails, and mammalian SCH at potential exposure areas
Decide about prevention strategies such as the improvement of sanitary conditions in the potential exposure areas.
Investigate ways to relate potential exposure areas to the affected population.
Geo-locate and target surveys at household locations
Provide survey teams with updated maps and GPS for individual-level surveys' allocation.
Implement or further investigate the present approach to other endemic areas.
Target efforts to the acquisition of detailed data about significant environmental risk factors
Use the proposed data structure to decide about the data acquisition strategy..
Investigate the influence of other spatial supports of analysis and other ways to suggest an optimal data structure for SCH modelling.

Figure 6.1: *Practical implications of the research output for the main uncertainty stakeholders of SCH modelling and mapping.*

Local dynamics of SCH are driven by the spatial connectivity between the mechanisms of human mobility, snails dispersal, and hydrological transport of schistosome larvae (Ciddio et al., 2017). Human mobility is important for the spread of SCH at a local scale. Its patterns at various scale levels, from villages to potential infection foci can be captured from anonymized mobile phone records, where information such as the transport pathway, date of the year, activity type or weather conditions can be recorded and used for further big data analysis. This may lead to a larger and more realistic data set with potential consequences in epidemiology.

Research applicability and transferability

Three different methods have been proposed to address relevant sources of uncertainty coming from the use of earth observation data: positional mismatch, pure specification bias, and the MAUP. The findings from these analyses could be linked to other SCH modelling studies. For instance, methods from Chapter 3 related to the positional mismatch issue, could be used to improve survey data collection. Souza Gomez et al. (Gomes et al., 2018) collected data on snails at demarcated collection points to identify sites with infected (foci sites) and non-infected (breeding sites) snails. The nature of these points is not clearly determined but by providing high probability areas of exposure could increase efficiently in data collection and help to rapidly identifying breeding or foci sites by taking environmental drivers of SCH into account.

Methods aimed to reduce pure specification bias and results from Chapter 4 could be applied in studies such as the one from Kulinkina et al (Kulinkina et al., 2018). They aggregated SCH prevalence at community-level and used random forest models to evaluate the relevance of fifteen environmental risk factors on SCH. Either the extraction of variables at a finer spatial extent than the community level, or the use of uncertainty information to weight the variables in the random forest models, could guide the study to find more reliable results and improve the quality of their predictions.

MAUP effects from Chapter 5 could guide other SCH studies to find a suitable spatial support for their analysis, as well as provide information about uncertainty derived from the use of specific spatial supports. For instance, the study by Manyangadze et al (Manyangadze et al., 2016) used a jackknifing procedure to determine the significance of the environmental variables in the spatio-temporal distribution of SCH snail hosts. In this research, the variable selection procedure could be affected by the use of various spatial resolutions varying from 250 m to 1 km. Uncertainty values could inform these selection procedures in order to get more reliable results.

These three sources of uncertainty affect not only SCH modelling, but also other infectious disease modelling efforts. The findings from chapter 3, 4 and 5, found for one research study area and a single disease, could transfer the proposed methods to other infectious diseases, in different areas of study. I will give some considerations below.

A positional mismatch is present in the spatial modelling of leptospiriosis, echinococcosis, soil-transmitted helminths, among others (Cadavid Restrepo et al., 2016; Dhewantara et al., 2019). For instance, for leptospiriosis modelling, different wild or domestic animals carry the bacterium that can get into the soil/water and survive there even several months (Dhewantara et al., 2019). Cases, however, are not reported at the exact places of exposure. Therefore, my research could guide the delineation of potential exposure areas, which can then be related to the disease. Pure specification bias and MAUP are also typical sources of uncertainty in the modelling of soil-transmitted helminths (Karagiannis-Voules et al., 2015). In these studies, pure specification bias has been treated as a positional uncertainty issue (Cressie and Kornak, 2003): the attributes of a variable are first recorded and then assigned a location (e.g. barangay centroid). This leads to the disease outcome variable being linked to the erroneous covariate value. The MAUP emerges in many studies that base interventions on misaligned covariate data (Clements et al., 2006; Soares Magalhães, Barnett and Clements, 2011) or maps aggregated at different administrative units (Schur, Vounatsou and Utzinger, 2012). This latter form of the MAUP leads to different patterns of endemicity. It shows the importance of both the size of support, as well as its shape (Schur et al., 2011; Schur et al., 2013). As long as the uncertainty sources are the same, or at some extent similar, as in the case of positional uncertainty as a source of pure specification bias, it would be possible to reproduce and transfer the methods to various infectious disease modelling approaches for a significant range of infectious such as the ones outlined above.

On the use of Bayesian statistics for uncertainty quantification

Thanks to the onset of powerful computers and algorithms, scientists started using Bayesian statistics by the end of the 20th century, and is currently being broadly studied and applied. Bayesian statistics has several advantages over frequentist statistics as it proposes practical ways to solve everyday problems, irrespective of the data size, in a direct, intuitive, and informative way (Kruschke and Liddell, 2018). Benefits of using Bayesian statistics for data analysis encompass the integration of parameter estimation and hypothesis testing in a coherent predictive framework (Wagenmakers, Morey and Lee, 2016), which provides measures of uncertainty in a one-go analysis.

In this research, I proposed a convolutional model to make individual-level inferences of SCH prevalence based on group or ecological-level survey data. Here, Bayesian parameter estimation captured the prior knowledge about the parameters through the use of probability distributions. These prior knowledge, represented as prior distributions, was updated to posterior distributions by using the SCH data evidence. Posterior distributions represent the parameter estimates with a measure of uncertainty lower than the one of the prior distributions. Bayesian hypothesis testing uses the parameter posterior distributions to (i) update the prior knowledge about the plausibility of SCH prevalence, and (ii) quantify the parameters ability to predict the observed prevalence values (Wagenmakers, Morey and Lee, 2016). Bayesian statistics allow one to monitor changes in our prior beliefs indefinitely as more data are added providing an inherent measure of evidence (i.e. uncertainty).

Emerging trends in mapping SCH and uncertainty quantification

Spatial modelling and mapping of SCH aims to describe the spatial distribution of populations at risk by using statistical models that link SCH infection data to environmental or socio-economic variables as drives of infection. Spatial modelling of SCH has been supported by the use of geographic information systems (GIS), and EO data, to explore the significance of various environmental factors on SCH infection (Manyangadze et al., 2015).

Geospatial artificial intelligence (geoAI) has emerged thanks to the development of innovative approaches in artificial intelligence (AI) such as machine learning, data mining (i.e. big data, data science), and fast computing, combined with spatial sciences (Vopham et al., 2018). Innovations in this field aim to address real-world problems like the ones related to human health, in particular, spatial epidemiology (Baker and Nieuwenhuijsen, 2008). geoAI in spatial epidemiology looks to target issues related to inefficient computational processing and data constraints regarding coarse spatial and temporal supports, for exposure assessment (Lin et al., 2017; Vopham et al., 2018).

Three main advantages of applying geoAI in SCH mapping were identified. First, the use of spatial big data (SBD) coming from diverse sources, formats and structures (Dietrich, 2015). Data such as direct measures, surrogate variables or electronic health records might be generated at high velocity and derived from large geographic study areas. These data in combination with applied machine and deep learning algorithms could extract relevant information regarding SCH. The use of SBD could improve the spatio-temporal resolutions of the output predictions, as well as identify significant risk factors of SCH prevalence (Vopham et al., 2018). For instance, Lin et al. (Lin et al., 2017) eliminated the need of previous selection of predictors of air pollution, by letting the data decide which risk factors were significant for exposure. This was done using a data mining-based algorithm. Second, geoAI is flexible when to addressing properties of spatial processes such as spatial non-stationarity and anisotropy (Vopham et al., 2018). For instance, Lin et al. (Brown and Heuvelink, 2006) addressed spatial non-stationarity by grouping air monitoring stations into similar geographic features based on unique temporal geo-contexts. Anisotropy could be addressed by modifying algorithms to include more data to predict different environmental exposures at various geographic areas (Vopham et al., 2018). Third, geoAI involves the development of algorithms to identify and classify objects with spatio-temporal limitations using EO data (Vopham et al., 2018). This decreases computational time by getting a faster and accurate use of environmental drivers of SCH infection.

Despite the potential advantages of geoAI in spatial SCH mapping, the presence and use of SBD is a key challenge that needs to be carefully considered (Vopham et al., 2018). SBD offers more data from diverse sources. However, it also could bring uncertainties in three ways. First, multisource data have been generated by different methods, units, and are of a varying quality (Shi et al., 2018). Second, results may be affected by aggregation or disaggregation of multisource data into different areas (Openshaw, 1984) or time periods (Cheng and Adepeju, 2014). Third, the chances of getting noise, erroneous data correlations, and unreliable or biased results increase from multiple data sets (Fan, Han and Liu, 2014). SBD quality by means of variety of sources, veracity of data, computing velocity and data volume (Shi et al., 2018) needs to be evaluated before SBD can be used for mapping SCH. Higher accuracy of geoAI algorithms depends on higher sets of training data. Disease training data require continuous classification. Re-labelling these data over time could reduce cost-effectiveness of geoAI applications, limiting its feasibility (Shi et al., 2018). Some solutions to this problem have been proposed such as the use of weakly supervised learning that works on incomplete low-quality labels (Han et al., 2015), or crowdsourcing of training data (Chen and Zipf, 2017).

Applications of geoAI in the mapping of SCH and other diseases require the inclusion of uncertainty as a quality measure of the results. Uncertainty quantification in geoAI could be challenging due to the unknown complexity of the proposed algorithms, which limits the generation of an analytical solution for uncertainty quantification. Some work has been done on this aspect. For example, Gal and Ghahrami (Gal and Ghahramani, 2016) evaluated the uncertainty derived from the use of deep neural networks using Bayesian methods. They used a dropout training approach to avoid overfitting caused by the deep neural network. The reliability of geoAI applications needs to be further explored. For this, uncertainty quantification of these applications applied to SCH mapping is essential not only for epidemiologists, but also scientists and decision makers in this field.

6.3 Recommendations

In Chapter 3 I used a sBN with purely discrete variables. This was done to represent the various risk factors ranges at which exposure could occur. Nevertheless, the construction of node probability tables might become challenging for a larger number of discrete parent nodes. Therefore, for the definition of exposure areas with more complex network configurations, it is suggested to explore the use of a mixed Bayesian network. These networks mix a set of D discrete and V continuous random variables, factorizing the joint probability distribution from equation 1.1 as:

$$P(R) = \prod_{d=1}^{D} p\left(r_d \middle| PA(r_d)\right) \cdot f\left(r_v \middle| PA(r_v)\right) \tag{6.1}$$

In Chapters 3, 4, and 5, issues of data quality were identified as a limitation to the network and model input data as well as for validation. In Chapter 3 I used a limited number of survey data, and in Chapters 4 and 5, I used barangay-level aggregated survey data. Further research in this regard could perform a simulation back of survey data using the estimated model parameters from Chapters 4 and 5. This would be done in order to generate a more reliable snail infection rate map and get enough human positive cases points for validation.

Other issues regarding the selection of environmental risk factors, and different spatial supports of analyses need further investigation. In Chapters 3, 4, and 5, environmental risk factors were pre-selected from literature and assumed to be potential drivers for SCH prevalence. In Chapter 5, I presented five increasing spatial supports. It would be interesting to investigate other spatial supports, especially between 500 m and 1 km, to more precisely explore the shift of a variable from significant to non-significant or vice versa. To further explore these issues I encourage to investigate geoAI applications for spatial epidemiology. For instance, Lin et al (Lin et al., 2017) presented a data mining approach to automatically select important geographic features for $PM_{2.5}$ concentration predictions. These features were selected using a random forest classifier that quantified their influence on SCH prevalence. This was done in order to avoid the use of expert-selected predictors. Another advantage of geoAI is that it allows the inclusion of data at fine and coarse spatial supports to perform predictions at fine spatial supports. Lin et al. (Lin et al., 2017) avoids the selection of a specific spatial support size by creating geo-contexts that use various feature types extracted at different buffer sizes. The creation of geo-contexts could be an alternative to deal with the MAUP in SCH predictions, as different geographic characteristics for each surveyed location could be inserted as layers in a learning classifier, such a classifier then decides which risk factors at which spatial supports are relevant for SCH prevalence.

Current trends in artificial intelligence are for instance machine learning algorithms (MLAs). These algorithms parse the data and learn from them to make informed decisions based upon both inputs and the desired outputs (supervised learning) or only inputs (unsupervised learning). In Chapters 3, 4, and 5, I have addressed three sources of uncertainty coming from the use of EO data: positional mismatch, pure specification bias, and MAUP, respectively. In order to incorporate their uncertainties into a MLA for SCH prediction I suggest (i) to perform a supervised learning using uncertainty probability distributions, coming from pure specification bias and MAUP (Chapters 4 and 5), as input parameters for the MLA; and (ii) to include distance proximity scenarios between survey locations and potential exposure areas (Chapter 3), also as input parameters for the supervised MLA. This could allow the incorporation of uncertainties as layers in a deep learning method, or as nodes and weights in a random forest classifier for SCH prediction.

Last but not least, there are many more uncertainty sources that need attention, in particular the ones related to the use of EO data such as uncertainties derived from temporal aggregation and temporal misaligned EO data. By using Bayesian statistics or exploring the applications of geoAI as a tool for uncertainty quantification, we could tackle uncertainties necessary for the control and spread of SCH.

Bibliography

Action, E. (2018). The Problem: Parasitic Worms. Retrieved from https://www.evidenceaction.org/

Ajakaye, O. G., Adedeji, O. I., and Ajayi, P. O. (2017). Modeling the risk of transmission of schistosomiasis in Akure North Local Government Area of Ondo State, Nigeria using satellite derived environmental data. *PLoS Neglected Tropical Diseases*, *11*(7). doi:10.1371/journal.pntd.0005733

Araujo Navas, A. L., Hamm, N. A. S., Soares Magalhães, R. J., and Stein, A. (2016). Mapping Soil Transmitted Helminths and Schistosomiasis under Uncertainty: A Systematic Review and Critical Appraisal of Evidence. *PLoS Neglected Tropical Diseases*, *10*(12), e0005208. doi:10.1371/journal.pntd.0005208

Araujo Navas, A. L., Osei, F., Leonardo, L. R., Soares Magalhães, R. J., and Stein, A. (2019). Modeling Schistosoma japonicum Infection under Pure Specification Bias: Impact of Environmental Drivers of Infection. *International Journal of Environmental Research and Public Health*, *16*(2), 176. doi:10.3390/ijerph16020176

Atkinson, P. M., and Graham, A. J. (2006). Issues of Scale and Uncertainty in the Global Remote Sensing of Disease. *Advances in Parasitology*, *62*, 79-118. doi:10.1016/S0065-308X(05)62003-9

Baker, D. B., and Nieuwenhuijsen, M. J. (2008). *Environmental epidemiology: study methods and application*. Oxford: Oxford University Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer Science and Business Media.

Bottcher, S. G., and Dethlefsen, C. (2003). deal: A package for learning Bayesian networks. *Journal of Statistical Software*. doi:10.18637/jss.v008.i20

Briggs, D. J., Sabel, C. E., and Lee, K. (2009). Uncertainty in epidemiology and health risk and impact assessment. *Environmental Geochemistry and Health*, *31*(2), 189-203. doi:10.1007/s10653-008-9214-5

Brooker, S., Clements, A. C. A., and Bundy, D. A. P. (2006). Global Epidemiology, Ecology and Control of Soil-Transmitted Helminth Infections. *Advances in Parasitology*, *62*, 221-261. doi:10.1016/s0065-308x(05)62007-6

Brooker, S., Hay, S. I., and Bundy, D. A. P. (2002). Tools from ecology: useful for evaluating infection risk models? *Trends in Parasitology*, *18*(2), 70-74. doi:10.1016/s1471-4922(01)02223-1

Brooker, S., Hay, S. I., Tchuente, L. A. T., and Ratard, R. (2002). Using NOAA-AVHRR data to model human helminth distributions in planning disease control in Cameroon, West Africa. *Photogrammetric Engineering and Remote Sensing*, *68*(2), 175-179.

Brooker, S., Rowlands, M., Haller, L., Savioli, L., and Bundy, D. A. P. (2000). Towards an atlas of human helminth infection in sub-Saharan Africa: The use of geographical information systems (GIS). *Parasitology Today, 16*(7), 303-307. doi:10.1016/s0169-4758(00)01687-2

Brooks, S. P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434-455. doi:10.2307/1390675

Brown, D. S. (1994). *Freshwater snails of Africa and their medical importance*: CRC press.

Brown, J. D., and Heuvelink, G. B. (2006). Assessing Uncertainty Propagation through Physically Based Models of Soil Water Flow and Solute Transport. In M. Anderson & J. McDonnell (Eds.), *Encyclopedia of hydrological sciences*.

Burns, C. J., Wright, J. M., Pierson, J. B., Bateson, T. F., Burstyn, I., Goldstein, D. A., Klaunig, J. E., Luben, T. J., Mihlan, G., and Ritter, L. (2014). Evaluating Uncertainty to Strengthen Epidemiologic Data for Use in Human Health Risk Assessments. *Environmental Health Perspectives, 122*(11), 1160-1165. doi:10.1289/ehp.1308062

Cadavid Restrepo, A. M., Yang, Y. R., McManus, D. P., Gray, D. J., Giraudoux, P., Barnes, T. S., Williams, G. M., Soares Magalhães, R. J., Hamm, N. A. S., and Clements, A. C. A. (2016). The landscape epidemiology of echinococcoses. *Infectious Diseases of Poverty, 5*(13). doi:10.1186/s40249-016-0109-x

Chammartin, F., Houngbedji, C. A., Huerlimann, E., Yapi, R. B., Silue, K. D., Soro, G., Kouame, F. N., N'Goran, E. K., Utzinger, J., Raso, G., and Vounatsou, P. (2014). Bayesian Risk Mapping and Model-Based Estimation of Schistosoma haematobium-Schistosoma mansoni Co-distribution in Cote d'Ivoire. *PLoS Neglected Tropical Diseases, 8*(12), e3407. doi:10.1371/journal.pntd.0003407

Chen, J., and Zipf, A. (2017). *DeepVGI: Deep learning with volunteered geographic information.* Paper presented at the Proceedings of the 26th International Conference on World Wide Web Companion.

Cheng, T., and Adepeju, M. (2014). Modifiable Temporal Unit Problem (MTUP) and Its Effect on Space-Time Cluster Detection. *PLoS One, 9*(6), 10. doi:10.1371/journal.pone.0100465

Ciddio, M., Mari, L., Sokolow, S. H., De Leo, G. A., Casagrandi, R., and Gatto, M. (2017). The spatial spread of schistosomiasis: A multidimensional network model applied to Saint-Louis region, Senegal. *Advances in Water Resources, 108*, 406-415. doi:10.1016/j.advwatres.2016.10.012

Clements, A. C. A., Brooker, S., Nyandindi, U., Fenwick, A., and Blair, L. (2008). Bayesian spatial analysis of a national urinary schistosomiasis questionnaire to assist geographic targeting of schistosomiasis control in Tanzania, East Africa. *International Journal for Parasitology, 38*(3-4), 401-415. doi:10.1016/j.ijpara.2007.08.001

Clements, A. C. A., Lwambo, N. J. S., Blair, L., Nyandindi, U., Kaatano, G., Kinung'hi, S., Webster, J. P., Fenwick, A., and Brooker, S. (2006). Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania. *Tropical Medicine and International Health, 11*(4), 490-503. doi:10.1111/j.1365-3156.2006.01594.x

Clements, A. C. A., Moyeed, R., and Brooker, S. (2006). Bayesian geostatistical prediction of the intensity of infection with Schistosoma mansoni in East Africa. *Parasitology, 133*(6), 711-719. doi:10.1017/S0031182006001181

Clennon, J. A., King, C. H., Muchiri, E. M., and Kitron, U. (2007). Hydrological modelling of snail dispersal patterns in Msambweni, Kenya and potential resurgence of Schistosoma haematobium transmission. *Parasitology, 134*, 683-693. doi:10.1017/s0031182006001594

Congalton, R. G. (2010). How to Assess the Accuracy of Maps Generated from Remotely Sensed Data. In J. D. Bossler, J. B. Campbell, R. B. McMaster, & C. Rizos (Eds.), *Manual of Geospatial Science and Technology* (Second ed., pp. 403-421). London: CRC Press.

Corporation, N. S. (1998). NeticaTM Application for Belief Networks and Influence Diagrams: User's Guide. Retrieved from www.norsys.com/downloads/

Coutinho, H. M., McGarvey, S. T., Acosta, L. P., Manalo, D. L., Langdon, G. C., Leenstra, T., Kanzaria, H. K., Solomon, J., Wu, H. W., Olveda, R. M., Kurtis, J. D., and Friedman, J. F. (2005). Nutritional status and serum cytokine profiles in children, adolescents, and young adults with *Schistosoma japonicum*-associated hepatic fibrosis, in Leyte, Philippines. *Journal of Infectious Diseases, 192*(3), 528-536. doi:10.1086/430929

Cressie, N., and Kornak, J. (2003). Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science, 18*(4), 436-456. doi:10.1214/ss/1081443228

Curran, P. J., Atkinson, P. M., Foody, G. M., and Milton, E. J. (2000). Linking remote sensing, land cover and disease. In S. I. Hay, S. E. Randolph, & D. J. Rogers (Eds.), *Remote Sensing and Geographical Information Systems in Epidemiology* (Vol. 47, pp. 37-80). London: Academic Press

De Roeck, E., Van Coillie, F., De Wulf, R., Soenen, K., Charlier, J., Vercruysse, J., Hantson, W., Ducheyne, E., and Hendrickx, G. (2014). Fine-scale mapping of vector habitats using very high resolution satellite imagery: a liver fluke case-study. *Geospatial Health, 8*(3), 671-683. doi:10.4081/gh.2014.296.

Dhewantara, P. W., Lau, C. L., Allan, K. J., Hu, W. B., Zhang, W. Y., Mamun, A. A., and Soares Magalhaes, R. J. (2019). Spatial epidemiological approaches to inform leptospirosis surveillance and control: A systematic review and critical appraisal of methods. *Zoonoses and Public Health*, *66*(2), 185-206. doi:10.1111/zph.12549

Dietrich, D. (2015). *Data science & big data analytics: discovering, analyzing, visualizing and presenting data*: Wiley.

Diggle, P. J., Tawn, J., and Moyeed, R. (2002). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *47*(3), 299-350. doi:10.1111/1467-9876.00113

Duarte, H. d. O., Droguett, E. L., Moura, M. d. C., De Souza Gomes, E. C., Barbosa, C., Barbosa, V., and Araujo, M. (2014). An Ecological Model for Quantitative Risk Assessment for Schistosomiasis: The Case of a Patchy Environment in the Coastal Tropical Area of Northeastern Brazil. *Risk Analysis*, *34*(5), 831-846. doi:10.1111/risa.12139

Duckham, M., Mason, K., Stell, J., and Worboys, M. (2001). A formal approach to imperfection in geographic information. *Computers, Environment and Urban Systems*, *25*, 89-103. doi:10.1016/S0198-9715(00)00040-5

Dungan, J. L. (2002). Toward a Comprehensive View of Uncertainty in Remote Sensing Analysis. In P. M. Atkinson & G. M. Foody (Eds.), *Uncertainty in Remote Sensing and GIS* (Vol. 3, pp. 25-35). Chichester: John Wiley & Sons Ltd.

Dungan, J. L., Perry, J. N., Dale, M. R. T., Legendre, P., Citron-Pousty, S., Fortin, M. J., Jakomulska, A., Miriti, M., and Rosenberg, M. S. (2002). A balanced view of scale in spatial statistical analysis. *Ecography*, *25*(5), 626-640. doi:10.1034/j.1600-0587.2002.250510.x

ESRI. (2011). ArcGIS Desktop (Version 10): Environmental Systems Research Institute. Retrieved from http://www.esri.com/news/releases/10_2qtr/arcgis10-download.html

Fan, J. Q., Han, F., and Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*(2), 293-314. doi:10.1093/nsr/nwt032

Fenton, N., and Neil, M. (2012). *Risk assessment and decision analysis with Bayesian networks*. Boca Raton, FL, USA: CRC Press.

Foody, G. M. (2003). Uncertainty, knowledge discovery and data mining in GIS. *Progress in Physical Geography*, *27*(1), 113-121. doi:10.1191/0309133303pp345pr

Foody, G. M., and Atkinson, P. M. (Eds.). (2002). *Uncertainty in Remote Sensing and GIS*. Chichester: John Wiley & Sons Ltd.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, United States: Sage Publications, Inc.

Gal, Y., and Ghahramani, Z. (2016). *Dropout as a bayesian approximation: Representing model uncertainty in deep learning.* Paper presented at the international conference on machine learning.

Gao, F. H., Abe, E. M., Li, S. Z., Zhang, L. J., He, J. C., Zhang, S. Q., Wang, T. P., Zhou, X. N., and Gao, J. (2014). Fine scale Spatial-temporal cluster analysis for the infection risk of Schistosomiasis japonica using space-time scan statistics. *Parasites & Vectors, 7*. doi:10.1186/s13071-014-0578-3

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of spatial statistics*. Boca Raton, United States: Taylor & Francis Group.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on an Article by Browne and Draper). *Bayesian Analysis, 1*(3), 515-533. doi:10.1214/06-ba117a

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457-472.

Geological Survey, U. S. (2017). Global Data Explorer. Retrieved 7 August 2017, from U.S. Department of Interior https://gdex.cr.usgs.gov/gdex/

Gomes, E. C. D., Mesquitta, M. C. D., Wanderley, L. B., de Melo, F. L., Guimaraes, R., and Barbosa, C. S. (2018). Spatial risk analysis on occurrences and dispersal of Biomphalaria straminea in and endemic area for schistosomiasis. *Journal of Vector Borne Diseases, 55*(3), 208-214. doi:10.4103/0972-9062.249142

Gordon, C. A., Acosta, L. P., Gray, D. J., Olveda, R. M., Jarilla, B., Gobert, G. N., Ross, A. G., and McManus, D. P. (2012). High prevalence of Schistosoma japonicum infection in carabao from Samar province, the Philippines: implications for transmission and control. *PLoS Neglected Tropical Diseases, 6*(9), e1778. doi:10.1371/journal.pntd.0001778

Gotway, C. A., and Young, L. J. (2002). Combining Incompatible Spatial Data. *Journal of the American Statistical Association, 97*(458), 632-648. doi:10.1198/016214502760047140

Hamm, N. A. S., Soares Magalhães, R. J., and Clements, A. C. A. (2015). Earth Observation, Spatial Data Quality, and Neglected Tropical Diseases. *PLoS Neglected Tropical Diseases, 9*(12), e0004164. doi:10.1371/journal.pntd.0004164

Han, J., Zhang, D., Cheng, G., Guo, L., and Ren, J. (2015). Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *IEEE Transactions on Geoscience and Remote Sensing, 53*(6), 3325-3337. doi:10.1109/TGRS.2014.2374218

Hay, S. I., Rogers, D. J., Randolph, S. E., Stern, D. I., Cox, J., Shanks, G. D., and Snow, R. W. (2002). Hot topic or hot air? Climate change and malaria resurgence in East African highlands. *Trends in Parasitology, 18*(12), 530-534. doi:10.1016/s1471-4922(02)02374-7

Hellsten, A. S. (2006). *A spatio-temporal ammonia emissions inventory for the UK.* University of Edinburgh.

Herbreteau, V., Salem, G., Souris, M., Hugot, J.-P., and Gonzalez, J.-P. (2007). Thirty years of use and improvement of remote sensing, applied to epidemiology: from early promises to lasting frustration. *Health & Place*, *13*(2), 400-403. doi:10.1016/j.healthplace.2006.03.003

Hijmans R., R. E., Cruz M., O'Brien R., Barrantes I. (2018). DIVA-GIS free, simple and effective. *Free Spatial Data.* Retrieved from http://www.diva-gis.org/Data

Hotez, P. J., Alvarado, M., Basanez, M. G., Bolliger, I., Bourne, R., Boussinesq, M., Brooker, S. J., Brown, A. S., Buckle, G., Budke, C. M., Carabin, H., Coffeng, L. E., Fevre, E. M., Furst, T., Halasa, Y. A., Jasrasaria, R., Johns, N. E., Keiser, J., King, C. H., Lozano, R., Murdoch, M. E., O'Hanlon, S., Pion, S. D. S., Pullan, R. L., Ramaiah, K. D., Roberts, T., Shepard, D. S., Smith, J. L., Stolk, W. A., Undurraga, E. A., Utzinger, J., Wang, M. R., Murray, C. J. L., and Naghavi, M. (2014). The Global Burden of Disease Study 2010: Interpretation and Implications for the Neglected Tropical Diseases. *PLoS Neglected Tropical Diseases*, *8*(7), e2865. doi:10.1371/journal.pntd.0002865

Hotez, P. J., Molyneux, D. H., Fenwick, A., Ottesen, E., Sachs, S. E., and Sachs, J. D. (2006). Incorporating a rapid-impact package for neglected tropical diseases with programs for HIV/AIDS, tuberculosis, and malaria - A comprehensive pro-poor health policy and strategy for the developing world. *Plos Medicine*, *3*(5), 576-584. doi:10.1371/journal.pmed.0030102

Hu, Y., Bergquist, R., Lynn, H., Gao, F., Wang, Q., Zhang, S., Li, R., Sun, L., Xia, C., Xiong, C., Zhang, Z., and Jiang, Q. (2015). Sandwich mapping of schistosomiasis risk in Anhui Province, China. *Geospatial Health*, *10*(324), 111-116. doi:10.4081/gh.2015.324

Hu, Y., Xia, C. C., Li, S. Z., Ward, M. P., Luo, C., Gao, F. H., Wang, Q. Z., Zhang, S. Q., and Zhang, Z. J. (2017). Assessing environmental factors associated with regional schistosomiasis prevalence in Anhui Province, Peoples' Republic of China using a geographical detector method. *Infectious Diseases of Poverty*, *6*, 8. doi:10.1186/s40249-017-0299-x

Initiative, S. C. (2019). Our vision is a world free of parasitic worm infections. Retrieved from https://www.schistosomiasiscontrolinitiative.org/

ISO. (2013). *ISO 19157: Geographic Information - Data Quality* Retrieved from https://www.iso.org/standard/32575.html

Jia, T. W., Zhou, X. N., Wang, X. H., Utzinger, J., Steinmann, P., and Wu, X. H. (2007). Assessment of the age-specific disability weight of chronic schistosomiasis japonica. *Bulletin of the World Health Organization*, *85*(6), 458-465. doi:10.2471/blt.06.033035

Jong, R. d., and Bruin, S. d. (2012). Linear trends in seasonal vegetation time series and the modifiable temporal unit problem. *Biogeosciences*, *9*(1), 71-77. doi:10.5194/bg-9-71-2012

Jurek, A. M., Maldonado, G., Greenland, S., and Church, T. R. (2006). Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *European Journal of Epidemiology, 21*(12), 871-876. doi:10.1007/s10654-006-9083-0

Jurek, A. M., Maldonado, G., Greenland, S., and Church, T. R. (2007). Uncertainty analysis: an example of its application to estimating a survey proportion. *Journal of Epidemiology and Community Health, 61*(7), 650-654. doi:10.1136/jech.2006.053660

Kabore, A., Biritwum, N.-K., Downs, P. W., Magalhaes, R. J. S., Zhang, Y., and Ottesen, E. A. (2013). Predictive vs. Empiric Assessment of Schistosomiasis: Implications for Treatment Projections in Ghana. *PLoS Neglected Tropical Diseases, 7*(3), e2051. doi:10.1371/journal.pntd.0002051

Kalluri, S., Gilruth, P., Rogers, D., and Szczur, M. (2007). Surveillance of arthropod vector-borne infectious diseases using remote sensing techniques: a review. *PLoS Pathogens, 3*(10), e116. doi:10.1371/journal.ppat.0030116

Karagiannis-Voules, D.-A., Biedermann, P., Ekpo, U. F., Garba, A., Langer, E., Mathieu, E., Midzi, N., Mwinzi, P., Polderman, A. M., and Raso, G. (2015). Spatial and temporal distribution of soil-transmitted helminth infection in sub-Saharan Africa: a systematic review and geostatistical meta-analysis. *Lancet Infectious Diseases, 15*(1), 74-84. doi:10.1016/S1473-3099(14)71004-7

Keenan, J. D., Hotez, P. J., Amza, A., Stoller, N. E., Gaynor, B. D., Porco, T. C., and Lietman, T. M. (2013). Elimination and Eradication of Neglected Tropical Diseases with Mass Drug Administrations: A Survey of Experts. *PLoS Neglected Tropical Diseases, 7*(12). doi:10.1371/journal.pntd.0002562

King, C. H., Dickman, K., and Tisch, D. J. (2005). Reassessment of the cost of chronic helmintic infection: a meta-analysis of disability-related outcomes in endemic schistosomiasis. *Lancet, 365*(9470), 1561-1569. doi:10.1016/S0140-6736(05)66457-4

King, G. (2013). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton, United States: Princeton University Press.

Krauth, S. J., Coulibaly, J. T., Knopp, S., Traore, M., N'Goran, E. K., and Utzinger, J. (2012). An In-Depth Analysis of a Piece of Shit: Distribution of Schistosoma mansoni and Hookworm Eggs in Human Stool. *PLoS Neglected Tropical Diseases, 6*(12), e1969. doi:10.1371/journal.pntd.0001969

Kristensen, T., Malone, J., and McCarroll, J. (2001). Use of satellite remote sensing and geographic information systems to model the distribution and abundance of snail intermediate hosts in Africa: a preliminary model for Biomphalaria pfeifferi in Ethiopia. *Acta Tropica, 79*(1), 73-78. doi:10.1016/S0001-706X(01)00104-8

Kruschke, J. K., and Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178-206. doi:10.3758/s13423-016-1221-4

Kulinkina, A. V., Walz, Y., Koch, M., Biritwum, N. K., Utzinger, J., and Naumova, E. N. (2018). Improving spatial prediction of Schistosoma haematobium prevalence in southern Ghana through new remote sensors and local water access profiles. *PLoS Neglected Tropical Diseases, 12*(6), 22. doi:10.1371/journal.pntd.0006517

Kummu, M., de Moel, H., Ward, P. J., and Varis, O. (2011). How Close Do We Live to Water? A Global Analysis of Population Distance to Freshwater Bodies. *PLoS One, 6*(6), 13. doi:10.1371/journal.pone.0020578

Lawson, A. B. (2013). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Boca Raton, FL, USA: Chapman and Hall/CRC.

Leenstra, T., Acosta, L. P., Langdon, G. C., Manalo, D. L., Su, L., Olveda, R. M., McGarvey, S. T., Kurtis, J. D., and Friedman, J. F. (2006). Schistosomiasis japonica, anemia, and iron status in children, adolescents, and young adults in Leyte, Philippines. *American Journal of Clinical Nutrition, 83*(2), 371-379. doi:10.1093/ajcn/83.2.371

Leonardo, L., Acosta, L. P., Olveda, R. M., and Aligui, G. D. L. (2002). Difficulties and strategies in the control of schistosomiasis in the Philippines. *Acta Tropica, 82*(2), 295-299. doi:10.1016/s0001-706x(02)00022-0

Leonardo, L., Rivera, P., Saniel, O., Solon, J. A., Chigusa, Y., Villacorte, E., Chua, J. C., Moendeg, K., Manalo, D., Crisostomo, B., Sunico, L., Boldero, N., Payne, L., Hernandez, L., and Velayudhan, R. (2015). New endemic foci of schistosomiasis infections in the Philippines. *Acta Tropica, 141*, 354-360. doi:10.1016/j.actatropica.2013.03.015

Leonardo, L., Rivera, P., Saniel, O., Villacorte, E., Lebanan, M. A., Crisostomo, B., Hernandez, L., Baquilod, M., Erce, E., and Martinez, R. (2012). A national baseline prevalence survey of schistosomiasis in the Philippines using stratified two-step systematic cluster sampling design. *Journal of Tropical Medicine, 2012*, 8. doi:10.1155/2012/936128

Leonardo, L. R., Rivera, P., Saniel, O., Villacorte, E., Crisostomo, B., Hernandez, L., Baquilod, M., Erce, E., Martinez, R., and Velayudhan, R. (2008). Prevalence survey of schistosomiasis in Mindanao and the Visayas, The Philippines. *Parasitology International, 57*(3), 246-251. doi:10.1016/j.parint.2008.04.006

Lin, Y., Chiang, Y.-Y., Pan, F., Stripelis, D., Ambite, J. L., Eckel, S. P., and Habre, R. (2017). *Mining Public Datasets for Modeling Intra-City PM2.5 Concentrations at a Fine Spatial Resolution*. Paper presented at the

Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Redondo Beach, CA, USA.

Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W. (2015). *Geographic Information Science and Systems* Chichester: John Wiley & Sons.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013). *The BUGS book: A practical introduction to Bayesian Analysis*. Boca Raton, FL, USA: CRC press.

Lunn D., S. D., Thomas A. and Best N. (2018). OpenBUGS license. *Downloads.* Retrieved from http://www.openbugs.net/w/Downloads

Malone, J. B., Yilma, J. M., McCarroll, J. C., Erko, B., Mukaratirwa, S., and Zhou, X. Y. (2001). Satellite climatology and the environmental risk of Schistosoma mansoni in Ethiopia and east Africa. *Acta Tropica*, *79*(1), 59-72. doi:10.1016/s0001-706x(01)00103-6

Manyangadze, T., Chimbari, M. J., Gebreslasie, M., Ceccato, P., and Mukaratirwa, S. (2016). Modelling the spatial and seasonal distribution of suitable habitats of schistosomiasis intermediate host snails using Maxent in Ndumo area, KwaZulu-Natal Province, South Africa. *Parasites & Vectors*, *9*, 10. doi:10.1186/s13071-016-1834-5

Manyangadze, T., Chimbari, M. J., Gebreslasie, M., and Mukaratirwa, S. (2015). Application of geo-spatial technology in schistosomiasis modelling in Africa: a review. *Geospatial Health*, *10*(2), 99-110. doi:10.4081/gh.2015.326

Marcot, B. G., Steventon, J. D., Sutherland, G. D., and McCann, R. K. (2006). Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere*, *36*(12), 3063-3074. doi:10.1139/x06-135

Mari, L., Gatto, M., Ciddio, M., Dia, E. D., Sokolow, S. H., De Leo, G. A., and Casagrandi, R. (2017). Big-data-driven modeling unveils country-wide drivers of endemic schistosomiasis. *Scientific Reports*, *7*, 11. doi:10.1038/s41598-017-00493-1

McCarty, T. R., Turkeltaub, J. A., and Hotez, P. J. (2014). Global progress towards eliminating gastrointestinal helminth infections. *Current Opinion in Gastroenterology*, *30*(1), 18-24. doi:10.1097/mog.0000000000000025

McCreesh, N., Nikulin, G., and Booth, M. (2015). Predicting the effects of climate change on Schistosoma mansoni transmission in eastern Africa. *Parasites & Vectors*, *8*(4), 1-9. doi:10.1186/s13071-014-0617-0

National Mapping and Resource Information Authority. (2018). Retrieved 3 February 2018 http://www.namria.gov.ph/

Nielsen, T. D., and Jensen, F. V. (2009). *Bayesian Networks and Decision Graphs*. New York, NY, USA: Springer Science & Business Media.

Nijland, W., Addink, E., De Jong, S., and Van der Meer, F. (2009). Optimizing spatial image support for quantitative mapping of natural vegetation. *Remote Sensing of Environment*, *113*(4), 771-780. doi:10.1016/j.rse.2008.12.002

OCHA. (2018). Humanitarian Data Exchange v1.25.3. *OCHA Services.* Retrieved from https://data.humdata.org/search?groups=phl&q=&ext_page_size=25

Olveda, R. M., Tallo, V., Olveda, D. U., Inobaya, M. T., Chau, T. N., and Ross, A. G. (2016). National survey data for zoonotic schistosomiasis in the Philippines grossly underestimates the true burden of disease within endemic zones: implications for future control. *International Journal of Infectious Diseases*, *45*, 13-17. doi:10.1016/j.ijid.2016.01.011

Openshaw, S. (1984). The modifiable areal unit problem. *Concepts and techniques in modern geography*.

Pebesma, E., and Graeler, B. (2017). Spatial and Spatio-Temporal Geostatistical Modelling, Prediction, Package 'gstat'. *The Comprehensive R Archive Network*. Retrieved from https://cran.r-project.org/web/packages/gstat

Pesigan, T. P., Hairston, N. G., Jauregui, J. J., Garcia, E. G., Santos, A. T., Santos, B. C., and Besa, A. A. (1958). Studies on Schistosoma japonicum infection in the Philippines 2. The molluscan host. *Bulletin of the World Health Organization*, *18*(4), 481-578.

Pfukenyi, D. M., Mukaratirwa, S., Willingham, A. L., and Monrad, J. (2006). Epidemiological studies of Schistosoma mattheei infections in cattle in the highveld and lowveld communal grazing areas of Zimbabwe. *Onderstepoort Journal of Veterinary Research*, *73*(3), 179-191. doi:10.4102/ojvr.v73i3.144

Pietrock, M., and Marcogliese, D. J. (2003). Free-living endohelminth stages: at the mercy of environmental conditions. *Trends in Parasitology*, *19*(7), 293-299. doi:10.1016/S1471-4922(03)00117-X

Prah, S., and James, C. (1977). The influence of physical factors on the survival and infectivity of miracidia of *Schistosoma mansoni* and *S. haematobium* I. Effect of temperature and ultra-violet light. *Journal of Helminthology*, *51*(1), 73-85. doi:10.1017/S0022149X00007288

Prentice, R. L., and Sheppard, L. (1995). Aggregate data studies of disease risk factors. *Biometrika*, *82*(1), 113-125. doi:10.1093/biomet/82.1.113

Project, O. S. G. F. (2018, Accessed 29 November 2017). QGIS, A Free and Open Source Geographic Information System. Retrieved from https://www.qgis.org/en/site/

Project, O. S. M. (2017, Accessed 21 Nov 2017). Planet OSM. Retrieved from https://planet.osm.org

Pullan, R. L., Smith, J. L., Jasrasaria, R., and Brooker, S. J. (2014). Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. *Parasites & Vectors, 7*(1), 1-19. doi:10.1186/1756-3305-7-37

Raj, R., Hamm, N. A. S., and Kant, Y. (2013). Analysing the effect of different aggregation approaches on remotely sensed data. *International Journal of Remote Sensing, 34*(14), 4900-4916. doi:10.1080/01431161.2013.781289

Raso, G., Li, Y., Zhao, Z., Balen, J., Williams, G. M., and McManus, D. P. (2009). Spatial Distribution of Human Schistosoma japonicum Infections in the Dongting Lake Region, China. *PLoS One, 4*(9), e6947. doi:10.1371/journal.pone.0006947.

Raso, G., Matthys, B., N'goran, E., Tanner, M., Vounatsou, P., and Utzinger, J. (2005). Spatial risk prediction and mapping of Schistosoma mansoni infections among schoolchildren living in western Côte d'Ivoire. *Parasitology, 131*(01), 97-108. doi:10.1017/S0031182005007432

Rice science for a better world. (2018). Retrieved from http://irri.org/our-work/research/policy-and-markets/mapping-rice-in-the-philippines-where

Richardson, S., and Monfort, C. (2000). Ecological correlation studies. In P. Elliot, J. C. Wakefield, N. G. Best, & D. J. Briggs (Eds.), *Spatial Epidemiology: methods and applications* (pp. 205-220). Oxford, United Kingdom: Oxford University Press.

Richardson, S., Stucker, I., and Hemon, D. (1987). Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology, 16*(1), 111-120. doi:10.1093/ije/16.1.111

Rothman, K. J. (2012). *Epidemiology: An Introduction* (Vol. 2). New York: Oxford University Press.

Rougier, J., Sparks, S., and Hall, L. (2014). *Risk and Uncertainty Assessment for Natural Hazards*. Cambridge: Cambridge University Press.

Saaty, T. L. (2008). Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors The analytic hierarchy/network process *Revista De La Real Academia De Ciencias Exactas Fisicas Y Naturales Serie a-Matematicas, 102*(2), 251-318. doi:10.1007/BF03191825

Santos, F. L. N., Cerqueira, E. J. L., and Soares, N. M. (2005). Comparison of the thick smear and Kato-Katz techniques for diagnosis of intestinal helminth infections. *Revista da Sociedade Brasileira de Medicina Tropical, 38*(2), 196-198. doi:10.1590/S0037-86822005000200016

Scholte, R. G. C., Gosoniu, L., Malone, J. B., Chammartin, F., Utzinger, J., and Vounatsou, P. (2014). Predictive risk mapping of schistosomiasis in

Brazil using Bayesian geostatistical models. *Acta Tropica*, *132*, 57-63. doi:10.1016/j.actatropica.2013.12.007

Schultz, M., Voss, J., Auer, M., Carter, S., and Zipf, A. (2017). Open land cover from OpenStreetMap and remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, *63*, 206-213. doi:10.1016/j.jag.2017.07.014

Schur, N., Huerlimann, E., Garba, A., Traore, M. S., Ndir, O., Ratard, R. C., Tchuente, L.-A. T., Kristensen, T. K., Utzinger, J., and Vounatsou, P. (2011). Geostatistical Model-Based Estimates of Schistosomiasis Prevalence among Individuals Aged <= 20 Years in West Africa. *PLoS Neglected Tropical Diseases*, *5*(6), e1194. doi:10.1371/journal.pntd.0001194

Schur, N., Huerlimann, E., Stensgaard, A.-S., Chimfwembe, K., Mushinge, G., Simoonga, C., Kabatereine, N. B., Kristensen, T. K., Utzinger, J., and Vounatsou, P. (2013). Spatially explicit Schistosoma infection risk in eastern Africa using Bayesian geostatistical modelling. *Acta Tropica*, *128*(2), 365-377. doi:10.1016/j.actatropica.2011.10.006

Schur, N., Utzinger, J., and Vounatsou, P. (2011). Modelling age-heterogeneous Schistosoma haematobium and S.mansoni survey data via alignment factors. *Parasites & Vectors*, *4*(142), 1-10. doi:10.1186/1756-3305-4-142

Schur, N., Vounatsou, P., and Utzinger, J. (2012). Determining treatment needs at different spatial scales using geostatistical model-based risk estimates of schistosomiasis. *PLoS Neglected Tropical Diseases*, *6*(9), e1773. doi:10.1371/journal.pntd.0001773

Shi, W. (2009). *Principles of Modeling Uncertainties in Spatial Data and Spatial Analyses*. Boca Raton: Taylor & Francis.

Shi, W. Z., Zhang, A. S., Zhou, X. L., and Zhang, M. (2018). Challenges and Prospects of Uncertainties in Spatial Big Data Analytics. *Annals of the American Association of Geographers*, *108*(6), 1513-1520. doi:10.1080/24694452.2017.1421898

Simoonga, C., Utzinger, J., Brooker, S., Vounatsou, P., Appleton, C., Stensgaard, A.-S., Olsen, A., and Kristensen, T. K. (2009a). Remote sensing, geographical information system and spatial analysis for schistosomiasis epidemiology and ecology in Africa. *Parasitology*, *136*(13), 1683-1693. doi:10.1017/S0031182009006222

Simoonga, C., Utzinger, J., Brooker, S., Vounatsou, P., Appleton, C. C., Stensgaard, A. S., Olsen, A., and Kristensen, T. K. (2009b). Remote sensing, geographical information system and spatial analysis for schistosomiasis epidemiology and ecology in Africa. *Parasitology*, *136*(13), 1683-1693. doi:10.1017/s0031182009006222

Smith, C. S., Howes, A. L., Price, B., and McAlpine, C. A. (2007). Using a Bayesian belief network to predict suitable habitat of an endangered mammal - The Julia Creek dunnart (Sminthopsis douglasi). *Biological Conservation*, *139*(3-4), 333-347. doi:10.1016/j.biocon.2007.06.025

Soares Magalhães, R. J., Barnett, A. G., and Clements, A. C. A. (2011). Geographical analysis of the role of water supply and sanitation in the risk of helminth infections of children in West Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(50), 20084-20089. doi:10.1073/pnas.1106784108

Soares Magalhães, R. J., Biritwum, N.-K., Gyapong, J. O., Brooker, S., Zhang, Y., Blair, L., Fenwick, A., and Clements, A. (2011a). Mapping Helminth Co-Infection and Co-Intensity: Geostatistical Prediction in Ghana. *PLoS Neglected Tropical Diseases*, *5*(6), e1200. doi:10.1371/journal.pntd.0001200

Soares Magalhães, R. J., Clements, A. C. A., Patil, A. P., Gething, P. W., and Brooker, S. (2011b). The Applications of Model-Based Geostatistics in Helminth Epidemiology and Control. *Advances in Parasitology*, *74*, 267-296. doi:10.1016/b978-0-12-385897-9.00005-7

Soares Magalhães, R. J., Salamat, M. S., Leonardo, L., Gray, D. J., Carabin, H., Halton, K., McManus, D. P., Williams, G. M., Rivera, P., Saniel, O., Hernandez, L., Yakob, L., McGarvey, S., and Clements, A. (2014). Geographical distribution of human Schistosoma japonicum infection in The Philippines: tools to support disease control and further elimination. *International Journal for Parasitology*, *44*(13), 977-984. doi:10.1016/j.ijpara.2014.06.010

Soares Magalhães, R. J., Salamat, M. S., Leonardo, L., Gray, D. J., Carabin, H., Halton, K., McManus, D. P., Williams, G. M., Rivera, P., Saniel, O., Hernandez, L., Yakob, L., McGarvey, S. T., and Clements, A. C. A. (2015). Mapping the Risk of Soil-Transmitted Helminthic Infections in the Philippines. *PLoS Neglected Tropical Diseases*, *9*(9), e0003915. doi:10.1371/journal.pntd.00039115

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2007, February 2018). OpenBUGS user manual, version 3.0. 2. Retrieved from http://www.openbugs.net/w/Manuals

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003, February 2018). WinBUGS user manual. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs

Stensgaard, A., Jorgensen, A., Kabatareine, N., Malone, J., and Kristensen, T. (2005). Modeling the distribution of Schistosoma mansoni and host snails in Uganda using satellite sensor data and Geographical Information Systems. *Parassitologia*, *47*(1), 115.

Stensgaard, A. S., Jorgensen, A., Kabatereine, N. B., Rahbek, C., and Kristensen, T. K. (2006). Modeling freshwater snail habitat suitability and areas of potential snail-borne disease transmission in Uganda. *Geospatial Health*, *1*(1), 93-104. doi:10.4081/gh.2006.284

Stensgaard, A. S., Utzinger, J., Vounatsou, P., Hurlimann, E., Schur, N., Saarnak, C. F. L., Simoonga, C., Mubita, P., Kabatereine, N. B., Tchuente, L. A. T., Rahbek, C., and Kristensen, T. K. (2013). Large-scale determinants of intestinal schistosomiasis and intermediate host snail distribution across Africa: does climate matter? *Acta Tropica*, *128*(2), 378-390. doi:10.1016/j.actatropica.2011.11.010

Stolk, W. A., Kulik, M. C., le Rutte, E. A., Jacobson, J., Richardus, J. H., de Vlas, S. J., and Houweling, T. A. J. (2016). Between-Country Inequalities in the Neglected Tropical Disease Burden in 1990 and 2010, with Projections for 2020. *PLoS Neglected Tropical Diseases*, *10*(5), e0004560. doi:10.1371/journal.pntd.0004560

Stone, J. V. (2013). *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*: Sebtel Press.

Stürmer, T., Glynn, R. J., Rothman, K. J., Avorn, J., and Schneeweiss, S. (2007). Adjustments for Unmeasured Confounders in Pharmacoepidemiologic Database Studies Using External Information. *Medical Care*, *45*(10), 158-165. doi:10.1097/MLR.0b013e318070c045

Sturrock, H. J. W., Pullan, R. L., Kihara, J. H., Mwandawiro, C., and Brooker, S. J. (2013). The Use of Bivariate Spatial Modeling of Questionnaire and Parasitology Data to Predict the Distribution of Schistosoma haematobium in Coastal Kenya. *PLoS Neglected Tropical Diseases*, *7*(1), e2016. doi:10.1371/journal.pntd.0002016

Sturtz, S., Ligges, U., and Gelman, A. (2010, February 2018). R2OpenBUGS: a package for running OpenBUGS from R. Retrieved from https://cran.r-project.org/web/packages/R2OpenBUGS/index.html

Tarafder, M. R., Balolong, E., Carabin, H., Belisle, P., Tallo, V., Joseph, L., Alday, P., Gonzales, R. O., Riley, S., Olveda, R., and McGarvey, S. T. (2006). 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*, *6*(61), 10. doi:10.1186/1471-2458-6-61

Tavana, M., Liu, W., Elmore, P., Petry, F. E., and Bourgeois, B. S. (2016). A practical taxonomy of methods and literature for managing uncertain spatial data in geographic information systems. *Measurement*, *81*, 123-162. doi:10.1016/j.measurement.2015.12.007

Taylor-Robinson, D. C., Maayan, N., Soares-Weiser, K., Donegan, S., and Garner, P. (2015). Deworming drugs for soil-transmitted intestinal worms in children: effects on nutritional indicators, haemoglobin, and school performance. *The Cochrane Library*(7), CD000371. doi:10.1002/14651858.CD000371

Team, R. D. C. (2013). R: A language and environment for statistical computing. Vienna, Austria: The R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Thomas, A., Best, N., Lunn, D., Arnold, R., and Spiegelhalter, D. (2004, February 2018). GeoBugs user manual. Retrieved from https://www.mrc-bsu.cam.ac.uk/software/bugs/thebugs-project-geobugs/

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, *46*(suppl), 234-240. doi:10.2307/143141

VoPham, T., Hart, J. E., Laden, F., and Chiang, Y. Y. (2018). Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology. *Environmental Health*, *17*(1), 40. doi:10.1186/s12940-018-0386-x

Wagenmakers, E. J., Morey, R. D., and Lee, M. D. (2016). Bayesian Benefits for the Pragmatic Researcher. *Current Directions in Psychological Science*, *25*(3), 169-176. doi:10.1177/0963721416643289

Wakefield, J., and Lyons, H. (2010). Spatial Aggregation and the Ecological Fallacy. *Handbook of Modern Statistical Methods*, 541-558. doi:10.1201/9781420072884-c30

Wakefield, J., and Shaddick, G. (2006). Health-exposure modeling and the ecological fallacy. *Biostatistics*, *7*(3), 438-455. doi:10.1093/biostatistics/kxj017

Walz, Y., Wegmann, M., Dech, S., Raso, G., and Utzinger, J. (2015a). Risk profiling of schistosomiasis using remote sensing: approaches, challenges and outlook. *Parasites & Vectors*, *8*(1), 163. doi:10.1186/s13071-015-0732-6

Walz, Y., Wegmann, M., Dech, S., Vounatsou, P., Poda, J.-N., N'Goran, E. K., Utzinger, J., and Raso, G. (2015b). Modeling and Validation of Environmental Suitability for Schistosomiasis Transmission Using Remote Sensing. *PLoS Neglected Tropical Diseases*, *9*(11), e0004217. doi:10.1371/journal.pntd.0004217

Walz, Y., Wegmann, M., Leutner, B., Dech, S., Vounatsou, P., N'Goran, E. K., Raso, G., and Utzinger, J. (2015c). Use of an ecologically relevant modelling approach to improve remote sensing-based schistosomiasis risk profiling. *Geospatial Health*, *10*(2), 271-279. doi:10.4081/gh.2015.398

Wang, F. F., Wang, J., Gelfand, A., and Li, F. (2017). Accommodating the ecological fallacy in disease mapping in the absence of individual exposures. *Statistics in Medicine*, *36*(30), 4930-4942. doi:10.1002/sim.7494

Wang, X. H., Zhou, X. N., Vounatsou, P., Chen, Z., Utzinger, J., Yang, K., Steinmann, P., and Wu, X. H. (2008). Bayesian Spatio-Temporal Modeling of Schistosoma japonicum Prevalence Data in the Absence of a Diagnostic 'Gold' Standard. *PLoS Neglected Tropical Diseases*, *2*(6), e250. doi:10.1371/journal.pntd.0000250

Weiss, D. J., Mappin, B., Dalrymple, U., Bhatt, S., Cameron, E., Hay, S. I., and Gething, P. W. (2015). Re-examining environmental correlates of Plasmodium falciparum malaria endemicity: a data-intensive variable selection approach. *Malaria Journal*, *14*(1), 68. doi:10.1186/s12936-015-0574-x

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, *48*(4), 817-838. doi:10.2307/1912934

Woolhouse, M., and Chandiwana, S. (1990). Population dynamics model for Bulinus globosus, intermediate host for *Schistosoma haematobium*, in river habitats. *Acta Tropica*, *47*(3), 151-160. doi:10.1016/0001-706X(90)90021-Q s

Worboys, M., and Duckham, M. (2004). *GIS: A Computing Perspective* (Vol. 2). Boca Raton: CRC press.

Xu, H. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, *27*(14), 3025-3033. doi:10.1080/01431160600589179

Yang, K., Wang, X. H., Yang, G. J., Wu, X. H., Qi, Y. L., Li, H. J., and Zhou, X. N. (2008). An integrated approach to identify distribution of Oncomelania hupensis, the intermediate host of *Schistosoma japonicum*, in a mountainous region in China. *International Journal for Parasitology*, *38*(8-9), 1007-1016. doi:10.1016/j.ijpara.2007.12.007

Zhang, Z. J., Bergquist, R., Chen, D. M., Yao, B. D., Wang, Z. L., Gao, J., and Jiang, Q. W. (2013). Identification of Parasite-Host Habitats in Anxiang County, Hunan Province, China Based on Multi-Temporal China-Brazil Earth Resources Satellite (CBERS) Images. *PLoS One*, *8*(7), 9. doi:10.1371/journal.pone.0069447

Zhang, Z. J., Carpenter, T. E., Lynn, H. S., Chen, Y., Bivand, R., Clark, A. B., Hui, F. M., Peng, W. X., Zhou, Y. B., Zhao, G. M., and Jiang, Q. W. (2009). Location of active transmission sites of Schistosoma japonicum in lake and marshland regions in China. *Parasitology*, *136*(7), 737-746. doi:10.1017/s0031182009005885

Zhang, Z. J., Manjourides, J., Cohen, T., Hu, Y., and Jiang, Q. W. (2016). Spatial measurement errors in the field of spatial epidemiology. *International Journal of Health Geographics*, *15*(1), 21. doi:10.1186/s12942-016-0049-5

Zhou, X. N., Bergquist, R., Leonardo, L., Yang, G. J., Yang, K., Sudomo, M., and Olveda, R. (2010). Schistosomiasis Japonica: Control and Research Needs. In X. N. Zhou, R. Bergquist, R. Olveda, & J. Utzinger (Eds.), *Advances in Parasitology*, *Vol 72: Important Helminth Infections in Southeast Asia: Diversity and Potential for Control and Elimination*, *Pt A* (Vol. 72, pp. 145-178). San Diego: Elsevier Academic Press Inc.

Zhou, Y. B., Liang, S., and Jiang, Q. W. (2012). Factors impacting on progress towards elimination of transmission of schistosomiasis japonica in China. *Parasites & Vectors*, *5*(275), 7. doi:doi: 10.1186/1756-3305-5-275

Zhu, H. R., Liu, L., Zhou, X. N., and Yang, G. J. (2015). Ecological Model to Predict Potential Habitats of Oncomelania hupensis, the Intermediate Host of Schistosoma japonicum in the Mountainous Regions, China. *PLoS Neglected Tropical Diseases*, *9*(8). doi:10.1371/journal.pntd.0004028

Appendices

List 2A.1: *List of papers that fulfilled the inclusion criteria and were included in the review.*

1. Souza Guimaraes, R. J.,Freitas, C. C.,Dutra, L. V.,Carvalho Scholte, R. G.,Martins-Bede, F. T.,Fonseca, F. R.,Amaral, R. S.,Drummonds, S. C.,Felgueiras, C. A.,Oliveira, G. C.,Carvalho, O. S. (2010). A geoprocessing approach for studying and controlling schistosomiasis in the state of Minas Gerais, Brazil. Memorias Do Instituto Oswaldo Cruz, 105(4), 524-531. doi:10.1590/s0074-02762010000400030
2. Clements, A. C. A.,Garba, A.,Sacko, M.,Toure, S.,Dembele, R.,Landoure, A.,Bosque-Oliva, E.,Gabrielli, A. F.,Fenwick, A. (2008). Mapping the Probability of Schistosomiasis and Associated Uncertainty, West Africa. Emerging Infectious Diseases, 14(10), 1629-1632. doi:10.3201/eid1410.080366
3. Clements, A. C. A.,Firth, S.,Dembele, R.,Garba, A.,Toure, S.,Sacko, M.,Landoure, A.,Bosque-Oliva, E.,Barnett, A. G.,Brooker, S.,Fenwick, A. (2009). Use of Bayesian geostatistical prediction to estimate local variations in Schistosoma haematobium infection in western Africa. Bulletin of the World Health Organization, 87(12), 921-929. doi:10.2471/blt.08.058933
4. Scholte, R. G. C.,Gosoniu, L.,Malone, J. B.,Chammartin, F.,Utzinger, J.,Vounatsou, P. (2014). Predictive risk mapping of schistosomiasis in Brazil using Bayesian geostatistical models. Acta Tropica, 132, 57-63. doi:10.1016/j.actatropica.2013.12.007
5. Clements, A. C. A.,Brooker, S.,Nyandindi, U.,Fenwick, A.,Blair, L. (2008). Bayesian spatial analysis of a national urinary schistosomiasis questionnaire to assist geographic targeting of schistosomiasis control in Tanzania, East Africa. International Journal for Parasitology, 38(3-4), 401-415. doi:10.1016/j.ijpara.2007.08.001
6. Schur, N.,Huerlimann, E.,Stensgaard, A.-S.,Chimfwembe, K.,Mushinge, G.,Simoonga, C.,Kabatereine, N. B.,Kristensen, T. K.,Utzinger, J.,Vounatsou, P. (2013). Spatially explicit Schistosoma infection risk in eastern Africa using Bayesian geostatistical modelling. Acta Tropica, 128(2), 365-377. doi:10.1016/j.actatropica.2011.10.006
7. Vounatsou, P.,Raso, G.,Tanner, M.,N'Goran, E. K.,Utzinger, J. (2009). Bayesian geostatistical modelling for mapping schistosomiasis transmission. Parasitology, 136(13), 1695-1705. doi:10.1017/s003118200900599x
8. Zhang, Z. J.,Carpenter, T. E.,Lynn, H. S.,Chen, Y.,Bivand, R.,Clark, A. B.,Hui, F. M.,Peng, W. X.,Zhou, Y. B.,Zhao, G. M.,Jiang, Q. W. (2009). Location of active transmission sites of Schistosoma japonicum in lake and marshland regions in China. Parasitology, 136(7), 737-746. doi:10.1017/s0031182009005885

9. Sturrock, H. J. W.,Pullan, R. L.,Kihara, J. H.,Mwandawiro, C.,Brooker, S. J. (2013). The Use of Bivariate Spatial Modeling of Questionnaire and Parasitology Data to Predict the Distribution of Schistosoma haematobium in Coastal Kenya. PLoS Neglected Tropical Diseases, 7(1), e2016. doi:10.1371/journal.pntd.0002016

10. Raso, G.,Vounatsou, P.,Singer, B. H.,Eliézer, K.,Tanner, M.,Utzinger, J. (2006). An integrated approach for risk profiling and spatial prediction of Schistosoma mansoni–hookworm coinfection. Proceedings of the National Academy of Sciences, USA, 103(18), 6934-6939. doi:10.1073/pnas.0601559103

11. Soares Magalhães, R. J.,Salamat, M. S.,Leonardo, L.,Gray, D. J.,Carabin, H.,Halton, K.,McManus, D. P.,Williams, G. M.,Rivera, P.,Saniel, O.,Hernandez, L.,Yakob, L.,McGarvey, S. T.,Clements, A. C. A. (2015). Mapping the Risk of Soil-Transmitted Helminthic Infections in the Philippines. PLoS Neglected Tropical Diseases, 9(9), e0003915. doi:10.1371/journal.pntd.00039115

12. Chammartin, F.,Scholte, R. G. C.,Malone, J. B.,Bavia, M. E.,Nieto, P.,Utzinger, J.,Vounatsou, P. (2013). Modelling the geographical distribution of soil-transmitted helminth infections in Bolivia. Parasites & Vectors, 6(152), 1-14. doi:10.1186/1756-3305-6-152

13. Pullan, R. L.,Gething, P. W.,Smith, J. L.,Mwandawiro, C. S.,Sturrock, H. J. W.,Gitonga, C. W.,Hay, S. I.,Brooker, S. (2011). Spatial Modelling of Soil-Transmitted Helminth Infections in Kenya: A Disease Control Planning Tool. PLoS Neglected Tropical Diseases, 5(2), e958. doi:10.1371/journal.pntd.0000958

14. Sturrock, H. J. W.,Picon, D.,Sabasio, A.,Oguttu, D.,Robinson, E.,Lado, M.,Rumunu, J.,Brooker, S.,Kolaczinski, J. H. (2009). Integrated Mapping of Neglected Tropical Diseases: Epidemiological Findings and Control Implications for Northern Bahr-el-Ghazal State, Southern Sudan. PLoS Neglected Tropical Diseases, 3(10), e537. doi:10.1371/journal.pntd.0000537

15. Clasen, T.,Boisson, S.,Routray, P.,Cumming, O.,Jenkins, M.,Ensink, J. H. J.,Bell, M.,Freeman, M. C.,Peppin, S.,Schmidt, W.-P. (2012). The effect of improved rural sanitation on diarrhoea and helminth infection: design of a cluster-randomized trial in Orissa, India. Emerging Themes in Epidemiology, 9(7), 1-10. doi:10.1186/1742-7622-9-7

16. Duarte, H. d. O.,Droguett, E. L.,Moura, M. d. C.,De Souza Gomes, E. C.,Barbosa, C.,Barbosa, V.,Araujo, M. (2014). An Ecological Model for Quantitative Risk Assessment for Schistosomiasis: The Case of a Patchy Environment in the Coastal Tropical Area of Northeastern Brazil. Risk Analysis, 34(5), 831-846. doi:10.1111/risa.12139

17. Zhang, Z. J.,Davies, T. M.,Gao, J.,Wang, Z.,Jiang, Q.-W. (2013). Identification of high-risk regions for schistosomiasis in the Guichi region of China: an adaptive kernel density estimation-based approach. Parasitology, 140(7), 868-875. doi:10.1017/s0031182013000048

18. Clements, A. C. A.,Lwambo, N. J. S.,Blair, L.,Nyandindi, U.,Kaatano, G.,Kinung'hi, S.,Webster, J. P.,Fenwick, A.,Brooker, S. (2006). Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania. Tropical Medicine and International Health, 11(4), 490-503. doi:10.1111/j.1365-3156.2006.01594.x

19. Medina, D. C.,Findley, S. E.,Doumbia, S. (2008). State-Space Forecasting of Schistosoma haematobium Time-Series in Niono, Mali. PLoS Neglected Tropical Diseases, 2(8), e276. doi:10.1371/journal.pntd.0000276

20. Tarafder, M. R.,Balolong, E.,Carabin, H.,Belisle, P.,Tallo, V.,Joseph, L.,Alday, P.,Gonzales, R. O.,Riley, S.,Olveda, R.,McGarvey, S. T. (2006). 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, 6(61), 10. doi:10.1186/1471-2458-6-61

21. Gao, F.-h.,Abe, E. M.,Li, S.-z.,Zhang, L.-j.,He, J.-c.,Zhang, S.-q.,Wang, T.-p.,Zhou, X.-n.,Gao, J. (2014). Fine scale Spatial-temporal cluster analysis for the infection risk of Schistosomiasis japonica using space-time scan statistics. Parasites & Vectors, 7(578), 1-11. doi:10.1186/s13071-014-0578-3

22. Chammartin, F.,Houngbedji, C. A.,Huerlimann, E.,Yapi, R. B.,Silue, K. D.,Soro, G.,Kouame, F. N.,N'Goran, E. K.,Utzinger, J.,Raso, G.,Vounatsou, P. (2014). Bayesian Risk Mapping and Model-Based Estimation of Schistosoma haematobium-Schistosoma mansoni Co-distribution in Cote d'Ivoire. PLoS Neglected Tropical Diseases, 8(12), e3407. doi:10.1371/journal.pntd.0003407

23. Krauth, S. J.,Coulibaly, J. T.,Knopp, S.,Traore, M.,N'Goran, E. K.,Utzinger, J. (2012). An In-Depth Analysis of a Piece of Shit: Distribution of Schistosoma mansoni and Hookworm Eggs in Human Stool. PLoS Neglected Tropical Diseases, 6(12), e1969. doi:10.1371/journal.pntd.0001969

24. Wang, X.-H.,Zhou, X.-N.,Vounatsou, P.,Chen, Z.,Utzinger, J.,Yang, K.,Steinmann, P.,Wu, X.-H. (2008). Bayesian Spatio-Temporal Modeling of Schistosoma japonicum Prevalence Data in the Absence of a Diagnostic 'Gold' Standard. PLoS Neglected Tropical Diseases, 2(6), e250. doi:10.1371/journal.pntd.0000250

25. Soares Magalhães, R. J.,Biritwum, N.-K.,Gyapong, J. O.,Brooker, S.,Zhang, Y.,Blair, L.,Fenwick, A.,Clements, A. (2011). Mapping Helminth Co-Infection and Co-Intensity: Geostatistical Prediction in Ghana. PLoS Neglected Tropical Diseases, 5(6), e1200. doi:10.1371/journal.pntd.0001200

26. Schur, N.,Utzinger, J.,Vounatsou, P. (2011). Modelling age-heterogeneous Schistosoma haematobium and S.mansoni survey data via alignment factors. Parasites & Vectors, 4(142), 1-10. doi:10.1186/1756-3305-4-142

27. Hu, Y.,Bergquist, R.,Lynn, H.,Gao, F.,Wang, Q.,Zhang, S.,Li, R.,Sun, L.,Xia, C.,Xiong, C.,Zhang, Z.,Jiang, Q. (2015). Sandwich mapping of schistosomiasis risk in Anhui Province, China. Geospatial Health, 10(324), 111-116. doi:10.4081/gh.2015.324

28. Yang, K.,Li, W.,Sun, L.-P.,Huang, Y.-X.,Zhang, J.-F.,Wu, F.,Hang, D.-R.,Steinmann, P.,Liang, Y.-S. (2013). Spatio-temporal analysis to identify determinants of Oncomelania hupensis infection with Schistosoma japonicum in Jiangsu province, China. Parasites & Vectors, 6(138), 1-8. doi:10.1186/1756-3305-6-138

29. Spear, R. C.,Hubbard, A.,Liang, S.,Seto, E. (2002). Disease Transmission Models for Public Health Decision Making: Toward an Approach for Designing Intervention Strategies for Schistosomiasis japonica. Environmental Health Perspectives, 110(9), 907-915. doi:10.1007/978-1-4419-6064-1_12

30. Clements, A. C. A.,Bosque-Oliva, E.,Sacko, M.,Landoure, A.,Dembele, R.,Traore, M.,Coulibaly, G.,Gabrielli, A. F.,Fenwick, A.,Brooker, S. (2009). A Comparative Study of the Spatial Distribution of Schistosomiasis in Mali in 1984-1989 and 2004-2006. PLoS Neglected Tropical Diseases, 3(5), e431. doi:10.1371/journal.pntd.0000431

31. Nihei, N.,Komagata, O.,Kobayashi, M.,Saitoh, Y.,Mochizuki, K.-i.,Nakamura, S. (2009). Spatial Analysis and Remote Sensing for Monitoring Systems of Oncomelania nosophora Following the Eradication of Schistosomiasis Japonica in Yamanashi Prefecture, Japan. Japanese Journal of Infectious Diseases, 62(2), 125-132.

32. McCreesh, N.,Nikulin, G.,Booth, M. (2015). Predicting the effects of climate change on Schistosoma mansoni transmission in eastern Africa. Parasites & Vectors, 8(4), 1-9. doi:10.1186/s13071-014-0617-0

33. Martins-Bede, F. T.,Freitas, C. C.,Dutra, L. V.,Sandri, S. A.,Fonseca, F. R.,Drummond, I. N.,Souza Guimaraes, R. J. d. P.,Amaral, R. S.,Carvalho, O. S. (2009). Risk Mapping of Schistosomiasis in Minas Gerais, Brazil, Using MODIS and Socioeconomic Spatial Data. IEEE Transactions on Geoscience and Remote Sensing, 47(11), 3899-3908. doi:10.1109/tgrs.2009.2028332

34. Kabore, A.,Biritwum, N.-K.,Downs, P. W.,Magalhaes, R. J. S.,Zhang, Y.,Ottesen, E. A. (2013). Predictive vs. Empiric Assessment of Schistosomiasis: Implications for Treatment Projections in Ghana. PLoS Neglected Tropical Diseases, 7(3), e2051. doi:10.1371/journal.pntd.0002051

35. Seto, E.,Liang, S.,Qiu, D.,Gu, X.,Spear, R. C. (2001). A Protocol for Geographically Randomized Snail Surveys in Schistosomiasis Fieldwork Using the Global Positioning System. American Journal of Tropical Medicine and Hygiene, 64(1-2), 98-9.

36. Martins-Bede, F. T.,Dutra, L. V.,Freitas, C. C.,Guimardes, R. J. P. S.,Amaral, R. S.,Drummond, S. C.,Carvalho, O. S. (2010). Schistosomiasis risk mapping in the state of Minas Gerais, Brazil, using a decision tree approach, remote sensing data and sociological indicators. Memorias Do Instituto Oswaldo Cruz, 105(4), 541-548. doi:10.1590/s0074-02762010000400033

37. Scott, D.,Senker, K.,England, E. C. (1982). Epidemiology of human Schistosoma-haematobium infection around Volta Lake, Ghana, 1973-75 Bulletin of the World Health Organization, 60(1), 89-100.

38. Seto, E.,Xu, B.,Liang, S.,Gong, P.,Wu, W. P.,Davis, G.,Qiu, D. C.,Gu, X. G.,Spear, R. (2002). The Use of Remote Sensing for Predictive Modeling of Schistosomiasis in China. Photogrammetric Engineering and Remote Sensing, 68(2), 167-174.

39. Raso, G.,Vounatsou, P.,Gosoniu, L.,Tanner, M.,N'Goran, E. K.,Utzinger, J. (2006). Risk factors and spatial patterns of hookworm infection among schoolchildren in a rural area of western Côte d'Ivoire. International Journal for Parasitology, 36(2), 201-210. doi:10.1016/j.ijpara.2005.09.003

40. Raso, G.,Vounatsou, P.,McManus, D. P.,Utzinger, J. (2007). Bayesian risk maps for Schistosoma mansoni and hookworm mono-infections in a setting where both parasites co-exist. Geospatial Health, 2(1), 85-96. doi:10.4081/gh.2007.257

41. Clements, A. C. A.,Moyeed, R.,Brooker, S. (2006). Bayesian geostatistical prediction of the intensity of infection with Schistosoma mansoni in East Africa. Parasitology, 133(6), 711-719. doi:10.1017/S0031182006001181

42. Hodges, M. H.,Magalhaes, R. J. S.,Paye, J.,Koroma, J. B.,Sonnie, M.,Clements, A.,Zhang, Y. (2012). Combined Spatial Prediction of Schistosomiasis and Soil-Transmitted Helminthiasis in Sierra Leone: A Tool for Integrated Disease Control. PLoS Neglected Tropical Diseases, 6(6), e1694. doi:10.1371/journal.pntd.0001694

43. Schur, N.,Huerlimann, E.,Garba, A.,Traore, M. S.,Ndir, O.,Ratard, R. C.,Tchuente, L.-A. T.,Kristensen, T. K.,Utzinger, J.,Vounatsou, P. (2011). Geostatistical Model-Based Estimates of Schistosomiasis Prevalence among Individuals Aged <= 20 Years in West Africa. PLoS Neglected Tropical Diseases, 5(6), e1194. doi:10.1371/journal.pntd.0001194

44. Soares Magalhães, R. J.,Salamat, M. S.,Leonardo, L.,Gray, D. J.,Carabin, H.,Halton, K.,McManus, D. P.,Williams, G. M.,Rivera, P.,Saniel, O. (2014). Geographical distribution of human Schistosoma japonicum infection in the Philippines: tools to support disease control and further elimination. International Journal for Parasitology, 44(13), 977-984. doi:10.1016/j.ijpara.2014.06.010.

45. Scholte, R. G. C.,Schur, N.,Bavia, M. E.,Carvalho, E. M.,Chammartin, F.,Utzinger, J.,Vounatsou, P. (2013). Spatial analysis and risk mapping of soil-transmitted helminth infections in Brazil, using Bayesian geostatistical models. Geospatial Health, 8(1), 97-110. doi:10.4081/gh.2013.58

46. Soares Magalhães, R. J.,Barnett, A. G.,Clements, A. C. A. (2011). Geographical analysis of the role of water supply and sanitation in the risk of helminth infections of children in West Africa. Proceedings of the National Academy of Sciences of the United States of America, 108(50), 20084-20089. doi:10.1073/pnas.1106784108

47. Karagiannis-Voules, D.-A.,Biedermann, P.,Ekpo, U. F.,Garba, A.,Langer, E.,Mathieu, E.,Midzi, N.,Mwinzi, P.,Polderman, A. M.,Raso, G. (2015). Spatial and temporal distribution of soil-transmitted helminth infection

in sub-Saharan Africa: a systematic review and geostatistical meta-analysis. Lancet Infectious Diseases, 15(1), 74-84. doi:10.1016/S1473-3099(14)71004-7

48. Booth, M.,Bundy, D. A. P. (1992). Comparative prevalences of Ascaris lumbricoides, Trichuris trichiura and hookworm infections and the prospects for combined control. Parasitology, 105, 151-157. doi:10.1017/S003118200007380

49. Brooker, S.,Clements, A. C. (2009). Spatial heterogeneity of parasite co-infection: Determinants and geostatistical prediction at regional scales. International Journal for Parasitology, 39(5), 591-597. doi:10.1016/j.ijpara.2008.10.014

50. Clements, A. C. A.,Deville, M. A.,Ndayishimiye, O.,Brooker, S.,Fenwick, A. (2010). Spatial co-distribution of neglected tropical diseases in the East African Great Lakes region: revisiting the justification for integrated control. Tropical Medicine & International Health, 15(2), 198-207. doi:10.1111/j.1365-3156.2009.02440.x

51. Dorkenoo, A. M.,Bronzan, R. N.,Ayena, K. D.,Anthony, G.,Agbo, Y. M.,Sognikin, K. S. E.,Dogbe, K. S.,Amza, A.,Sodahlon, Y.,Mathieu, E. (2012). Nationwide integrated mapping of three neglected tropical diseases in Togo: countrywide implementation of a novel approach. Tropical Medicine & International Health, 17(7), 896-903. doi:10.1111/j.1365-3156.2012.03004.x

52. Ekpo, U. F.,Huerlimann, E.,Schur, N.,Oluwole, A. S.,Abe, E. M.,Mafe, M. A.,Nebe, O. J.,Isiyaku, S.,Olamiju, F.,Kadiri, M.,Poopola, T. O. S.,Braide, E. I.,Saka, Y.,Mafiana, C. F.,Kristensen, T. K.,Utzinger, J.,Vounatsou, P. (2013). Mapping and prediction of schistosomiasis in Nigeria using compiled survey data and Bayesian geospatial modelling. Geospatial Health, 7(2), 355-366. doi:10.4081/gh.2013.92

53. Woodhall, D. M.,Wiegand, R. E.,Wellman, M.,Matey, E.,Abudho, B.,Karanja, D. M. S.,Mwinzi, P. M. N.,Montgomery, S. P.,Secor, W. E. (2013). Use of Geospatial Modeling to Predict Schistosoma mansoni Prevalence in Nyanza Province, Kenya. PLoS One, 8(8), e71635. doi:10.1371/journal.pone.0071635

54. Brooker, S.,Singhasivanon, P.,Waikagul, J.,Supavej, S.,Kojima, S.,Takeuchi, T.,Luong, T. V.,Looareesuwan, S. (2003). Mapping Soil-Transmitted Helminths in Southeast Asia and Implications for Parasite Control. Southeast Asian Journal of Tropical Medicine & Public Health, 34(1), 24-36.

55. Spear, R. C. (2012). Internal versus external determinants of Schistosoma japonicum transmission in irrigated agricultural villages. Journal of the Royal Society, Interface, 9(67), 272-282. doi:10.1098/rsif.2011.0285

56. Raso, G.,Matthys, B.,N'goran, E.,Tanner, M.,Vounatsou, P.,Utzinger, J. (2005). Spatial risk prediction and mapping of Schistosoma mansoni infections among schoolchildren living in western Côte d'Ivoire. Parasitology, 131(01), 97-108. doi:10.1017/S0031182005007432

57. Beck-Woerner, C.,Raso, G.,Vounatsou, P.,N'Goran, E. K.,Rigo, G.,Parlow, E.,Utzinger, J. (2007). Bayesian Spatial Risk Prediction of Schistosoma mansoni Infection in Western Côte d'Ivoire Using a Remotely-Sensed Digital Elevation Model. American Journal of Tropical Medicine and Hygiene, 76(5), 956-963. doi:10.4269/ajtmh.2007.76.956

58. Raso, G.,Li, Y.,Zhao, Z.,Balen, J.,Williams, G. M.,McManus, D. P. (2009). Spatial Distribution of Human Schistosoma japonicum Infections in the Dongting Lake Region, China. PLoS One, 4(9), e6947. doi:10.1371/journal.pone.0006947.

59. Pullan, R. L.,Bethony, J. M.,Geiger, S. M.,Cundill, B.,Correa-Oliveira, R.,Quinnell, R. J.,Brooker, S. (2008). Human Helminth Co-Infection: Analysis of Spatial Patterns and Risk Factors in a Brazilian Community. PLoS Neglected Tropical Diseases, 2(12), e352. doi:10.1371/journal.pntd.0000352

60. Saathoff, E.,Olsen, A.,Kvalsvig, J. D.,Appleton, C. C.,Sharp, B.,Kleinschmidt, I. (2005). Ecological Covariates of Ascaris lumbricoides Infection in Schoolchildren from Rural KwaZulu-Natal, South Africa. Tropical Medicine & International Health, 10(5), 412-422. doi:10.1111/j.1365-3156.2005.01406.x

61. Saathoff, E.,Olsen, A.,Sharp, B.,Kvalsvig, J. D.,Appleton, C. C.,Kleinschmidt, I. (2005). Ecologic Covariates of Hookworm Infection and Reinfection in Rural Kwazulu-natal/South Africa: A Geographic Information System–Based Study. American Journal of Tropical Medicine and Hygiene, 72(4), 384-391.

62. Fonseca, F.,Freitas, C.,Dutra, L.,Guimaraes, R.,Carvalho, O. (2014). Spatial modeling of the schistosomiasis mansoni in Minas Gerais State, Brazil using spatial regression. Acta Tropica, 133, 56-63. doi:10.1016/j.actatropica.2014.01.015

63. Chen, Z.,Zhou, X.-N.,Yang, K.,Wang, X.-H.,Yao, Z.-Q.,Wang, T.-P.,Yang, G.-J.,Yang, Y.-J.,Zhang, S.-Q.,Wang, J.,Jia, T.-W.,Wu, X.-H. (2007). Strategy formulation for schistosomiasis japonica control in different environmental settings supported by spatial analysis: a case study from China. Geospatial Health, 1(2), 223-231. doi:10.4081/gh.2007.270

64. Malone, J. B.,Yilma, J. M.,McCarroll, J. C.,Erko, B.,Mukaratirwa, S.,Zhou, X. Y. (2001). Satellite climatology and the environmental risk of Schistosoma mansoni in Ethiopia and east Africa. Acta Tropica, 79(1), 59-72. doi:10.1016/s0001-706x(01)00103-6

65. Abdel-Rahman, M. S.,El-Bahy, M. M.,El-Bahy, N. M.,Malone, J. B. (1997). Development and Validation of a Satellites Based Geographic Information System (GIS) Model for Epidemiology of Schistosoma Risk Assessment on Snail Level in Kafr El-Sheikh Governorate. Journal of the Egyptian Society of Parasitology, 27(2), 299-316.

66. Schuele, S. A.,Clowes, P.,Kroidl, I.,Kowuor, D. O.,Nsojo, A.,Mangu, C.,Riess, H.,Geldmacher, C.,Laubender, R. P.,Mhina, S.,Maboko, L.,Loescher, T.,Hoelscher, M.,Saathoff, E. (2014). Ascaris lumbricoides Infection and Its

Relation to Environmental Factors in the Mbeya Region of Tanzania, a Cross-Sectional, Population-Based Study. PLoS One, 9(3), e92032. doi:10.1371/journal.pone.0092032

67. Bisht, D.,Verma, A. K.,Bharadwaj, H. H. D. (2011). Intestinal parasitic infestation among children in a semi-urban Indian population. Tropical Parasitology, 1(2), 104-107. doi:10.4103/2229-5070.86946

68. Brooker, S.,Donnelly, C. A.,Guyatt, H. L. (2000). Estimating the number of helminthic infections in the Republic of Cameroon from data on infection prevalence in schoolchildren. Bulletin of the World Health Organization, 78(12), 1456-1465.

69. Chan, M.,Medley, G.,Jamison, D.,Bundy, D. (1994). The evaluation of potential global morbidity attributable to intestinal nematode infections. Parasitology, 109(03), 373-387. doi:10.1017/S0031182000078410

70. Schur, N.,Gosoniu, L.,Raso, G.,Utzinger, J.,Vounatsou, P. (2011). Modelling the geographical distribution of co-infection risk from single-disease surveys. Statistics in Medicine, 30(14), 1761-1776. doi:10.1002/sim.4243

71. Tarafder, M. R.,Balolong, E.,Carabin, H.,Belisle, P.,Tallo, V.,Joseph, L.,Alday, P.,Gonzales, R. O.,Riley, S.,Olveda, R.,McGarvey, S. T. (2006). 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, 6(61), 1-10. doi:10.1186/1471-2458-6-61

72. Abdel-Rahman, M. S.,El-Bahy, M. M.,Malone, J. B.,Thompson, R. A.,El-Bahy, N. M. (2001). Geographic information systems as a tool for control program management for schistosomiasis in Egypt. Acta Tropica, 79(1), 49-57. doi:10.1016/s0001-706x(01)00102-4

73. Brooker, S.,Hay, S. I.,Issae, W.,Hall, A.,Kihamia, C. M.,Lwambo, N. J.,Wint, W.,Rogers, D. J.,Bundy, D. A. (2001). Predicting the distribution of urinary schistosomiasis in Tanzania using satellite sensor data. Tropical Medicine & International Health, 6(12), 998-1007. doi:10.1046/j.1365-3156.2001.00798.x

74. Liu, Z.,Li, C.,Tang, L.,Zhou, X.,Ma, L.,Liu, C. (2015). Prediction of oncomelania hupensis (vector of schistosomiasis) distribution based on remote sensing data and fuzzy information theory. Paper presented at the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS).

Table 3.A1*: Saaty's pairwise comparison matrix for Land Use*

	Wet Soil	Water bodies	Agriculture land and grass	Forest and Natural Areas	Built land	Barren land
Wet Soil	1.00	3.00	5.00	7.00	9.00	9.00
Water bodies	0.33	1.00	5.00	7.00	9.00	9.00
Agriculture land and grass	0.20	0.20	1.00	5.00	7.00	9.00
Forest and Natural Areas	0.14	0.14	0.20	1.00	5.00	5.00
Built land	0.11	0.11	0.14	0.20	1.00	3.00
Barren land	0.11	0.11	0.11	0.20	0.33	1.00
sum=	1.90	4.57	11.45	20.40	31.33	36.00

Normalized relative weights						
Wet Soil	**Water bodies**	**Agriculture land and grass**	**Forest and Natural Areas**	**Built land**	**Barren land**	**Normalized pricipal eingen vector**
0.53	0.66	0.44	0.34	0.29	0.25	0.42
0.18	0.22	0.44	0.34	0.29	0.25	0.29
0.11	0.04	0.09	0.25	0.22	0.25	0.16
0.08	0.03	0.02	0.05	0.16	0.14	0.08
0.06	0.02	0.01	0.01	0.03	0.08	0.04
0.06	0.02	0.01	0.01	0.01	0.03	0.02
1.00	1.00	1.00	1.00	1.00	1.00	

Principal eingen value	7.52	
Number of factors	6.00	
Consistency index	0.30	
Random index (n factors)	1.24	
Consistency ratio	0.24	acceptable

***Table 3.A2**: Saaty's pairwise comparison matrix for Elevation*

	High Risk	Medium Risk	Low Risk	Normalized relative weights			Normalized pricipal eingen vector
High Risk	1.00	5.00	9.00	0.76	0.81	0.53	0.70
Medium Risk	0.20	1.00	7.00	0.15	0.16	0.41	0.24
Low Risk	0.11	0.14	1.00	0.08	0.02	0.06	0.06
Sum=	1.31	6.14	17	1	1	1	

Principal eingen value	3.35	
Number of factors	3.00	
Consistency index	0.18	
Random index (n factors)	0.58	
Consistency ratio	0.31	acceptable

***Table 3.A3**: Saaty's pairwise comparison matrix for Slope*

	High Risk	Medium Risk	Low Risk	Normalized relative weights			Normalized pricipal eingen vector
High Risk	1.00	5.00	7.00	0.74	0.81	0.54	0.70
Medium Risk	0.20	1.00	5.00	0.15	0.16	0.38	0.23
Low Risk	0.14	0.20	1.00	0.11	0.03	0.08	0.07
sum=	1.34	6.20	13.00	1.00	1.00	1.00	

Principal eingen value	3.31	
Number of factors	3.00	
Consistency index	0.15	
Random index (n factors)	0.58	
Consistency ratio	0.26	acceptable

Table 3.A4*: Saaty's pairwise comparison matrix for Distance to water bodies*

	High Risk	Medium Risk	Low Risk	Normalized relative weights			Normalized principal eingen vector
High Risk	1.00	7.00	9.00	0.80	0.86	0.53	0.73
Medium Risk	0.14	1.00	7.00	0.11	0.12	0.41	0.22
Low Risk	0.11	0.14	1.00	0.09	0.02	0.06	0.05
sum=	1.25	8.14	17.00	1.00	1.00	1.00	

Principal eingen value	3.61	
Number of factors	3.00	
Consistency index	0.30	
Random index (n factors)	0.58	
Consistency ratio	0.53	acceptable

Table 3.A5*: Saaty's pairwise comparison matrix for snail infection rate*

	High Risk	Medium Risk	Low Risk	Normalized relative weights			Normalized pricipal eingen vector
High Risk	1.00	3.00	7.00	0.68	0.71	0.54	0.64
Medium Risk	0.33	1.00	5.00	0.23	0.24	0.38	0.28
Low Risk	0.14	0.20	1.00	0.10	0.05	0.08	0.07
sum=	1.48	4.20	13.00	1.00	1.00	1.00	

Principal eingen value	3.10	
Number of factors	3.00	
Consistency index	0.05	
Random index (n factors)	0.58	
Consistency ratio	0.08	acceptable

Table 3.A6: *Saaty's pairwise comparison matrix for all risk factors*

	Land Use	Elevation	Slope	Distance to water bodies	Snail infection rate
Land Use	1.00	7.00	3.00	0.33	5.00
Elevation	0.14	1.00	0.20	0.11	0.33
Slope	0.33	5.00	1.00	0.20	3.00
Distance to water bodies	3.00	9.00	5.00	1.00	7.00
Snail infection rate	0.20	3.00	0.33	0.14	1.00
sum	4.68	25.00	9.53	1.79	16.33

Normalized relative weights					Normalized principal eingen vector
0.21	0.28	0.31	0.19	0.31	0.26
0.03	0.04	0.02	0.06	0.02	0.03
0.07	0.20	0.10	0.11	0.18	0.13
0.64	0.36	0.52	0.56	0.43	0.50
0.04	0.12	0.03	0.08	0.06	0.07
1.00	1.00	1.00	1.00	1.00	

Principal eingen value	5.37	
Number of factors	5.00	
Consistency index	0.09	
Random index (n factors)	1.12	
Consistency ratio	0.08	acceptable

Table 3.A7*: Total weights for all risk factors*

Risk factors	Categories	Prior probabilities per category (Weights)	Prior probabilities per risk factor (Weights)
	Wet Soil	0.42	
	Water bodies	0.29	
	Agriculture land and grass	0.16	
	Forest and Natural Areas	0.08	
	Built land	0.04	
Land Use	Barren land	0.02	0.26
	High risk	0.70	
	Medium risk	0.24	
Elevation	Low risk	0.06	0.03
	High risk	0.70	
	Medium risk	0.23	
Slope	Low risk	0.07	0.13
	High risk	0.73	
	Medium risk	0.22	
DWB	Low risk	0.05	0.50
	High risk	0.64	
	Medium risk	0.28	
SIR	Low risk	0.07	0.07

Table A4.1: *Survey data*

Barangay ID (k)	Number of positive cases (y)	Number of sampled people (N)	Barangay ID (k)	Number of positive cases (y)	Number of sampled people (N)
1	0	199	33	0	96
2	0	246	34	0	258
3	0	183	35	3	198
4	1	264	36	1	169
5	0	268	37	0	252
6	16	428	38	0	207
7	23	271	39	0	222
8	1	302	40	0	233
9	0	256	41	0	166
10	20	268	42	0	215
11	3	226	43	0	206
12	0	243	44	0	125
13	0	154	45	0	102
14	0	208	46	0	129
15	14	267	47	0	138
16	0	34	48	0	76
17	0	69	49	9	264
18	0	115	50	0	247
19	0	224	51	0	168
20	0	161	52	0	216
21	0	208	53	0	56
22	0	196	54	0	15
23	0	234	55	0	87
24	6	215	56	1	111
25	0	159	57	1	87
26	1	275	58	0	89
27	1	202	59	0	47
28	3	267	60	1	88
29	0	81	61	0	236
30	1	272	62	0	257
31	5	208	63	0	181
32	0	209	64	0	193

Barangay ID (k)	Number of positive cases (y)	Number of sampled people (N)	Barangay ID (k)	Number of positive cases (y)	Number of sampled people (N)
65	0	319	87	1	130
66	0	178	88	0	227
67	0	142	89	0	165
68	0	133	90	0	145
69	0	91	91	0	253
70	0	196	92	0	289
71	0	114	93	2	91
72	0	199	94	6	250
73	0	288	95	0	156
74	0	6	96	3	140
75	0	115	97	0	143
76	0	59	98	0	90
77	0	59	99	0	258
78	0	104	100	0	256
79	0	84	101	0	258
80	0	134	102	0	229
81	0	150	103	0	260
82	2	144	104	0	269
83	0	91	105	0	274
84	0	147	106	0	274
85	0	167	107	0	274
86	1	91	108	0	275

Code A 4.2*: BUGs code*

```
model{
  #likelihood
    for(k in 1:108){
       y[k]~dbin(p[k],n[k]);

        for(j in offset[k]:(offset[k+1]-1)){
        alfa_1[j]<-
beta0+beta1*ndvi[j]+beta2*ndwi[j]+beta3*lstd[j]+beta4*lstn[j]+beta5*e[j]
+beta6*ndwb[j]+s[k]
        alfa_2[j]<-1/(1+exp(-alfa_1[j]))
        }
        p[k]<-sum(alfa_2[offset[k]:(offset[k+1]-1)])/m[k]
    }

#PRIORS FOR FIXED EFFECTS BETAS
            beta0~dunif(-100,100)
            beta1~dnorm(0.0,prec)
            beta2~dnorm(0.0,prec)
            beta3~dnorm(0.0,prec)
            beta4~dnorm(0.0,prec)
            beta5~dnorm(0.0,prec)
            beta6~dnorm(0.0,prec)

            sigma.b~dunif(0,100)
            prec<-pow(sigma.b,-2)

#PRIORS FOR SPATIAL RANDOM EFFECTS

    s[1:108]~spatial.exp(mu[],xcoor[],ycoor[],tau,phi,kappa)

    for (i in 1:108){
        mu[i]<-0
    }

    #FIRST OPTION
    sigma ~ dnorm(0,1) I(0,) #half standard normal prior
    tau <- pow(sigma,-2)

    phi~ dunif(0.0000002,0.003)
    kappa <-1

#PREDICTIONS

            for (l in 1:108){
            y.pred_c[l]~dbin(p[l],n[l])
            }
}
```

Author’s Publication

- Araujo Navas, A. L., Hamm, N. A. S., Soares Magalhães, R. J., and Stein, A. (2016). Mapping Soil Transmitted Helminths and Schistosomiasis under Uncertainty: A Systematic Review and Critical Appraisal of Evidence. *PLoS Neglected Tropical Diseases*, *10*(12), e0005208. doi:10.1371/journal.pntd.0005208

- Araujo Navas, A. L.; Soares Magalhaes, R. J.; Osei, F.; Fornillos, R. J. C.; Leonardo, L. R.; Stein, A. (2018) Modelling local areas of exposure to *Schistosoma japonicum* in a limited survey data environment. Parasites & Vectors, 11(465). doi: 10.1186/s13071-018-3039-6.

- Araujo Navas, A. L., Osei, F., Leonardo, L. R., Soares Magalhães, R. J., and Stein, A. (2019). Modeling *Schistosoma japonicum* Infection under Pure Specification Bias: Impact of Environmental Drivers of Infection. *International Journal of Environmental Research and Public Health*, *16*(2), 176. doi:10.3390/ijerph16020176